Corrected 22 January 2010; see below

www.sciencemag.org/cgi/content/full/science.1183863/DC1

Supporting Online Material for

A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection

Sharon R. Grossman,* Ilya Shylakhter,* Elinor K. Karlsson, Elizabeth H. Byrne, Shannon Morales, Gabriel Frieden, Elizabeth Hostetter, Elaine Angelino, Manuel Garber, Or Zuk, Eric S. Lander, Stephen F. Schaffner, Pardis C. Sabeti*

*To whom correspondence should be addressed. E-mail: psabeti@oeb.harvard.edu (P.C.S.), shari.grossman@post.harvard.edu (S.R.G), ilya_shl@alum.mit.edu (I.S.)

Published 7 January 2010 on *Science* Express
DOI: 10.1126/science.1183863

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1183863/DC1)

Table S6 (Excel format)

Correction: The complete table S6 is now provided as an Excel file. The original version of the supporting online material contained only the first 10 pages of the table as part of the PDF file (pp. 64 to 73). In the corrected PDF file, these pages have been omitted.

Supporting Online Material:
Methods, Figure and Table Legends

METHODS

Simulations

We simulated population genetics datasets designed to mimic the 1000 genomes project dataset (S*1*) as well as, after thinning, the HapMapII dataset (S*2*). We used cosi, a coalescent simulator with a previously published demographic model that has been shown to match real data on several metrics (S*3*), with one change, earlier termination of migration, to facilitate simulation of selective sweeps.

Each simulation replica represented a 1MB genomic region of a human autosome (about 10,000 SNPs), genotyped in three populations (West African, East Asian and European). 120 chromosomes were sampled from each population, which had present-day effective sizes of 7700 for the non-African populations and of 2400 for the West African population. The population history was as follows. An ancestral population split into an African and a Eurasian population 3500 generations ago; the latter then split into Asian and European populations 2000 generations ago. The ancestral population had effective size of 12,500, which expanded to 24,000 at time 17,000 generations ago. A bottleneck was modeled for the Eurasian population shortly after its split from the ancestral population (inbreeding coefficient = .085). Bottlenecks were modeled for the three terminal populations shortly after the Asian/European split, with inbreeding coefficients of 0.008, 0.67 and 0.02 for the African, Asian and European populations respectively. Migration was modeled between Africa and the two non-African populations, in the 500 generations following the Asian/European split, at the probability per chromosome per generation of 32e-6 with Europe and 8e-6 with Africa. In the

published model, migration continues for the full 2000 generations following the Asian/European split. Recombination was modeled as varying along the region, with a hierarchical model that included both regional variation in recombination rates (estimated from deCode data) and local hotspots of recombination (S*2*).

Selective sweeps were modeled for a single population within cosi, using an established method (S*4*-S*6*). In this approach, the simulation outside the period of the selective sweep, and for populations not undergoing the sweep, is carried out with the usual neutral coalescent machinery. During the sweep, the population under selection is treated as two pools, one with and one without the selected allele. Coalescence occurs only within the same pool, while recombination (which is handled by the normal cosi machinery) can move a sequence segment onto either a selected or unselected background, with the probability determined by the frequency of the beneficial allele. The frequency of the beneficial allele is modeled by the deterministic approximation of Stephan *et al.* (S*7*). Our implementation differs from that of Kim and Stephan (S*5*) only in that we choose the initial and final frequencies of the beneficial allele to be $1-1/2N_e$ and $1/2N_e$, respectively. We have also added the option of a partial sweep; in this case, the final frequency is supplied by the user, with chromosomes randomly assigned to the two pools with probabilities give by that frequency.

Implementation of the sweep algorithm was validated by direct inspection of the beneficial allele frequency trajectory and the associated coalescence and recombination rates. We tested the code's large-N_e behavior by comparing the predicted heterozygosity within a selective sweep with the approximate model in Durrett and Schweinsberg (2004) ((S*8*) Proposition 1, in that paper*)* and found excellent agreement. We have previously

used the sweep code for power calculation simulations and other analyses on iHS, and we found very good agreement between our results and those of Voight et al. (2006) (S*9*), which did their calculations using sweeps simulated by the program SelSim (S*10*).

The sweep machinery is in the publicly available version of cosi, and is invoked by the addition of a sweep parameter to the parameter file, using the following syntax

pop_event sweep "string" <pop index> <sweep time> <sel_coeff> <pos> <end_freq>

where "string" is a user-supplied name for the sweep, <pop index> is the numerical identifier of the population in which the sweep takes place, <sweep time> is the end time (in generations) of the sweep, <pos> is the location of the sweep expressed as a decimal fraction of the sequence being simulated, and <end freq> is the frequency of the selected allele at the end of the sweep (i.e. at <sweep time>). The causal variant is included in the simulated output sequence. Historical events (e.g. changes in population size) during a sweep are disallowed within the population under selection, and only for that population; they are processed normally otherwise. Any migration during the sweep is also disallowed.

Our simulations included sweeps in each population; the end time of the sweeps ranged over 5ky, 10ky 15ky, 20ky, 25ky and 30ky; and the final frequency of the selected allele ranged over 0.2, 0.4, 0.6, 0.8 and 1.0. For each combination of these parameters, we created 100 simulation replicas, with a single selected SNP in the middle of the 1MB region. We classified simulations as low- or high- frequency sweeps, based on the actual present-day frequency of the selected allele (<50% or >50% respectively).

Thinned simulations modeling SNP ascertainment were created from full-sequence simulations by randomly removing SNPs, with each SNP's probability of

removal based on its minor-allele frequency. The per-frequency removal probabilities were chosen to match HapMapII densities of SNPs for each minor-allele frequency.

iHS test

Following Voight et al. (S*9*), we define the iHS test with respect to a given core SNP and a given population. We perform the test only for bi-allelic SNPs whose minor allele frequency is above 5%. We partition the chromosomes in the population according the core SNP allele they carry. Let A denote the ancestral allele and D, the derived allele. For all the chromosomes carrying A, we calculate EHH scores between the core SNP and every bi-allelic SNP within 2.5 Mb. We integrate EHH with respect to genetic distance (cM), linearly interpolating between successive bi-allelic SNPs, until the point at which EHH drops to 0.05. If EHH does not drop below 0.05 within 2.5Mb, we skip the iHS test for that SNP. We denote the integral by iHH_A. We similarly calculate iHH_D for the chromosomes carrying D. The unstandardized integrated haplotype score, iHS, is defined as $\ln(iHH_A / iHH_D)$.

In analysis of simulations, we calculated unstandardized iHS scores for every bi-allelic SNP in each simulated population. As iHS scores are approximately normal for SNPs with comparable derived allele frequencies, we split the unstandardized scores into 20 equally sized bins according to their derived allele frequencies. We calibrated each of the bins in the neutral simulations to have mean zero and variance one, and then used the same parameters to normalize scores in all other simulated scenarios.

For analysis of the HapMapII dataset, we used information on the ancestral state of each SNP provided by the International Haplotype Map Consortium, based on the chimpanzee and macaque bases (S*2*). The ancestral allele was taken to be the chimpanzee

bases, where available, otherwise the macaque base. If neither base was available, no ancestral state was inferred. For analysis of the 1000 Genomes dataset, we used information on the ancestral state of each SNP provided by the 1000 Genomes Consortium. These were based on 4-way EPO alignment, of human, chimpanzee, orangutan, and rhesus macaque (S*1*). For SNPs whose ancestral state is not available (~7%), iHH and iHS scores were not calculated. We calculated iHS scores for every SNP in the genome in each population, and normalized the scores in each bin.

XP-EHH test

Following Sabeti et al. (S*11*), we define the XP-EHH test for a given core SNP with respect to two populations, A and B, and a given direction, centromere distal or proximal. We take the set of all SNPs for which we have data in both A and B within 1MB of the core SNP in the given direction. We pick the SNP X in this region that has an EHH score with respect to all chromosomes in both populations that is closest to 0.04. If there is no SNP with such an EHH between 0.03 and 0.05, we skip the XP-EHH test for the core SNP. Next, we split the chromosomes from the two populations, and calculate EHH in each population at all the SNPs between the core SNP and X. Similarly to iHS, we integrate the EHH within these bounds with respect to genetic distance, and call the resulting integrals I_A and I_B. We define the XP-EHH logratio as $\ln(I_A/I_B)$.

For all SNPs in the simulated regions, we calculated XP-EHH logratio scores in both directions for each population pair. We normalized the set off all logratio scores in neutral simulations to have mean zero and variance one, and used the same parameters to normalize all other scenarios. For the SNPs in HapMapII, we calculate XP-EHH

logratios for each population pair and in each direction, and normalized all the scores similarly.

For each candidate SNP in the selected regions, we have a putative selected population. We took the maximum of the XP-EHH scores for the selected population with respect to the two other populations in both directions.

ΔiHH test

Beginning with iHH defined by Voight et al. (S*9*), we partition the chromosomes in the population according to the core SNP allele they carry, and calculate iHH_A and iHH_D for the core SNP. We define the unstandardized ΔiHH as $|iHH_A - iHH_D|$. We calculate ΔiHH for each bi-allelic SNP in the putative selected population. In analysis of the simulated regions and the HaMap Phase II and the selected 1000 Genomes regions, we sort the unstandardized scores in each region into 20 equally sized frequency bins, and normalize the scores within each bin to have mean zero and variance one.

We observed that iHS is very sensitive to fluctuations in the length of the ancestral haplotype (iHH_A), since it is a ratio. Neutral SNPs linked to the causal variant that happen to have particularly short ancestral haplotypes thus can have higher iHS scores than the true causal variant. More generally, for a given derived haplotype, the maximum iHS score a SNP can achieve is determined by the length of the ancestral haplotypes at each position. This effect is especially pronounced in low-frequency sweeps, where there is more variability in the length of the ancestral haplotypes. To address this problem, we created the ΔiHH statistic, which is more robust to fluctuations in iHH_A. ΔiHH captures the magnitude of the haplotype length, as opposed to iHS, which captures the relative sizes of the ancestral and derived haplotypes.

F_{ST} test

We calculated individual marker F_{ST}s between the hypothesized selected population and each of the other populations using the unbiased estimator of Weir and Cockerham (S*12*). For each SNP, we calculated the mean result for the two population comparisons.

What is considered a non-selected population??? The cross population + the target population? or two other populations?

ΔDAF test

For each SNP, we find the mean derived allele frequency in the two non-selected populations, denoted $\overline{D_{NS}}$, and the derived allele frequency in the putative selected population, denoted D_S. The ΔDAF score, developed in this study, is defined as $D_S - \overline{D_{NS}}$. ΔDAF scores range between -1 and 1, with positive scores indicating SNPs with high derived allele frequencies in the selected population. This test is designed to detect variants with a high derived allele frequency only in one population. The ancestral and derived state of the allele for HapMapII and 1000 Genomes datasets were determined as described in the iHS Methods section.

Calculation of likelihood tables and CMS test

We computed the empirical distributions of each of the tests described above for causative variants, neutral variants within regions under selection, and neutral variants in regions with no selection (Fig. 1). The distributions were built from 1000 neutral regions and 7500 regions under selected pressure, simulated using the model calibrated to the three HapMapII populations (West African, European, East Asian). The simulations were each 1 MB, and included in the distributions represented a range of selective strengths and selected allele frequencies (Table S1).

For each test, the distribution of selected (causal) variants approximates the probability a SNP will have a given score *s* if it is selected, and likewise the distribution of neutral variants approximates the probability a SNP will have a score of *s* if it is neutral. Treating the tests as independent, the probability that a selected SNP will have a set of scores $(s_1,\ldots s_n)$ is $\prod_{i=1}^{n} P(s_i \mid selected)$, and analogously the probability that a neutral SNP will have the scores is $\prod_{i=1}^{n} P(s_i \mid neutral)$. From these we calculate the Bayes factor for each SNP, where the null hypothesis is that the SNP is neutral, and the alternative hypothesis is that the SNP is selected (causal).

$$BF = \prod_{i=1}^{n} \frac{P(s_i \mid selected)}{P(s_i \mid neutral)} \qquad \textbf{Equation 2}$$

The CMS score is the posterior probability that the SNP is selected:

$$CMS = \prod_{i=1}^{n} \frac{P(s_i \mid selected) \times \pi}{P(s_i \mid selected) \times \pi + P(s_i \mid neutral) \times (1-\pi)} \qquad \textbf{Equation 3}$$

We assume a uniform prior probability $\pi = \frac{1}{N_{SNP}}$ that each SNP is the selected variant. However, one could also incorporate information into the prior about biological function, such as proximity to genes, conservation, or copy number variation.

Genome-wide scores are computed using the distribution of scores of variants in regions without selection in the null hypothesis. For analysis of regions with evidence of selection, we find that using the distribution of scores for neutral variants within selected regions (i.e., within 500 kb of a selected variant) as the null model results in greater sensitivity to detect the causal variant.

Power and FPR Calculations

To compare the effectiveness of CMS and the individual tests to detect the causal variant within regions with selection, we estimated the specificity of the tests at equal levels of power. Power can be estimated by observing the fraction of simulations in which the true causal SNPs is detected. We adjusted the test thresholds for significant results so that all tests had equal power (90% and 50%) to detect the causal SNP. The specificity can be estimated by observing the average number of neutral SNPs above the significance threshold in each selected region.

We assign significance using the distribution of CMS scores in simulated neutral regions (for simulated regions) and the empirical genome-wide distribution (for real data). The empirical *P*-value provides a measure of how likely a SNP is to be selected relative to the rest of the genome, as opposed to a formal significance test.

To assess the false positive rate in neutral regions, we defined genome-wide CMS thresholds that captured 89% of causal SNPs at frequencies above 50% (and 40% of causal SNPs at frequencies less than 50%). We calculated the per-SNP false posive rate in all neutral regions, and in neutral regions scoring in the top 1.3% of regions by each of the long haplotype tests (yielding an overall FPR of 5% by the three tests combined). For neutral regions with SNP(s) above the threshold, we also calculated the number of SNPs significant scores within the 1MB region.

Construction of confidence regions

Posterior probability curves for the location of the causal SNPs are computed by interpolating a spline between CMS scores across the region. To smooth the curve, we average scores within 0.01 cM windows. We normalize the curve so that the entire 1 MB

region integrates to one, and define 90% credible interval by choosing the narrowest region for which the integral equals 0.9. Regions defined in this way include the selected position in 92% of the simulations, and had an average size of 89 kb.

An alternative approach to constructing confidence regions, which has been used in previous scans for selection (S*9, 13*) is to look for windows with multiple extreme scoring SNPs. This approach defines approximately equivalent regions, and is less sensitive to variations in ascertainment and SNP density. In this windowed approach, we divide the region into 0.02 cM regions, each overlapping the next one by 0.01 cM, and include all windows that contain at least 3 SNPs (for thinned data) or 7 SNPs (for full sequence data) with normalized CMS scores above 0.5. This method defined regions that include the selected position in 85% of thinned simulations and 91% of full sequence simulations.

Table of Significant SNPs in HapMap II

We include all SNPs in the candidate regions above the CMS threshold that captures 90% of the causal SNPs in the simulated regions. To capture potential selected ancestral alleles, we recalculated CMS scores without the ΔDAF score and using the absolute value of the iHS score, modification that made all the input statistics agnostic to the derived state. We include all additional SNPs that score as highly as the initial set of significant SNPs using the modified CMS score, marked 'ancestral'.

SNP annotation information

Annotation information was obtained from dbSNP build 129 for all SNPs, including the classification of SNP functional states: non-synonymous, synonymous,

within an intron or mRNA UTR, or within 2 kb of a gene. We developed a database for our candidate regions to annotate all potentially functional DNA changes, including non-synonymous variants, variants disrupting predicted transcription factor and miRNA binding sites, variants within lincRNAs, variants within regions of conservations among vertebrates, and variants previously associated with human phenotypic differences. We evaluated the potential effect of the non-synonymous SNPs using the PolyPhen predictions (S*14*).

Conservation

Conservation scores were computed for each base using SiPhy (S*15*) on 29 placental mammals from the 44-way alignment (S*16*). For each alignment column SiPhy estimates the stationary distribution of the evolutionary model by maximum likelihood and reports the log-odds ratio of the fitted and neutral stationary distributions.

Structural Model of PCDH15's Cadherin Domain 4

We generated a homology model of the PCDH15 4th cadherin domain (Cad4) using six solved cadherin type II structures: MN-cadherin EC1, Cadherin-11 EC1-2, Cadherin-8 EC1-3, and Protocadherin-9 (S*17,* S*18*). We aligned the corresponding protein sequences using SALIGN (S*19*). We then added the amino acid sequence of PCDH15's Cad4 (residues 396-509) to this structural alignment using Modeller 9v6 (S*20*). The resulting alignment was used as the input to Modeller 9v6 to build ten PCDH15 Cad4 structure models, and the best model was selected based on the Objective Function Score. We performed a loop refinement on regions with high DOPE score using Modeler9v1, significantly reducing the energy of the regions. We further evaluated the model by examining the distribution of conserved residues using ConSurf (S*21*) with an

alignment of PCDH15 Cad4 sequences from 22 species. We observed a bias of conserved residues in the calcium-binding pocket, which supports our PCDH15 Cad4 model. To identify potential Ca^{2+} binding regions of PCDH15 Cad4, we superimposed the model to the Cadherin-11-Ca^{2+} complex structure. Asp-435 and Asp-90 of the Cad4 correspond to Cad-11 residues interacting with Ca^{2+} at the EC1-EC2 interface. The figure was generated with PyMOL (S*22*).

Expression analysis

We obtained normalized expression intensities of 44,000 probes representing the majority of the human gene complement for the HapMapII individuals from the Wellcome Trust Sanger Institute's website (S*23,* S*24*). We used intensities from each gene in the CMS regions as quantitative traits in a standard association test to the SNPs within the larger 1 MB regions. We identified SNPs that are significant at $p < 10^{-4}$ for association with expression levels, consistent with Kudaravalli et. al. (S*25*).

Gene Ontology Analysis

In order to identify gene ontology terms that are over- or under-represented among our localized regions, we made use of the PANTHER Gene Ontology database (S*26*). Since PANTHER consists of three ontologies, entitled 'Biological Processes,' 'Pathways,' and 'Molecular Functions,' we performed a separate assay for each. To begin, we identified a list of genes whose genomic positions overlap with our localized regions for each population (European American: CEU, West African: YRI, and East Asian: JPT+CHB). For each of the three lists generated, we then determined the number of genes in our list and the total number of genes in the genome related to each PANTHER ontology term. We obtained one-tailed p-values for over- or under-expression of each term using the

binomial test as described by Cho and Campbell (S*27*). Finally, we corrected for multiple testing with a modified Bonferroni correction, in which any two tests were considered independent so long as the PANTHER terms related to these tests were not directly linked or nested in the ontology. Despite this conservative correction, several PANTHER terms were significantly enriched among our lists of genes, as is shown in Table S7.

Supplemental Figures

Fig. S1. Comparison of ΔDAF and DAF's ability to localize signal and distinguish causal variant.

Fig. S2. Correlation between the individual tests for selection.

Fig. S3. Each test's ability to distinguish the selected variant for high (50-100%) and low (0-50%) frequency variants.

Fig. S4. Distribution of CMS scores of variants in regions without selection (neutral) simulated under different demographic models.

Fig. S5. Application of CMS to large selected regions in the Human Haplotype Map

Fig. S6. Applying CMS to selected regions in preliminary 1000 genomes data corroborates findings in HapMap data.

Fig. S7. High scoring coding change in *PCDH15* is validated in 1000 genomes data and highly conserved position across species.

Fig. S8. Global distribution of *PCDH15* D435A.

Fig. S9. SNPs in *USF1* and *PAWR*, under selection in West Africa, are associated with changes in gene expression.

Fig. S10. Near-perfect proxies for selected and neutral SNPs.

Supplemental Tables

Table S1. Simulation parameters for demographic models analyzed

Table S2. CMS performance to localize region and distinguish causal variant.

Table S3. Number of significant non-causal SNPs per region at a given 90% or 50% power to detect the causal SNP for individual tests and CMS.

Table S4. False discovery rate for all neutral regions and neutral loci that have haplotype scores in the most tail 5% of the null distribution

Table S5. Selected regions identified by CMS test in HapMapII.

Table S6. All HapMapII SNPs in selected regions with significant CMS scores, at a threshold with 90% power to capture the causal allele, with genome annotations.

Table S7. *p*-Values for enrichment of GO categories among genes in regions identified by CMS.

Supplemental References

S1. www.1000genomes.org.
S2. K. A. Frazer *et al.*, *Nature* **449**, 851 (Oct 18, 2007).
S3. S. F. Schaffner *et al.*, *Genome Res* **15**, 1576 (Nov, 2005).
S4. J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, W. Stephan, *Genetics* **140**, 783 (Jun, 1995).
S5. Y. Kim, W. Stephan, *Genetics* **160**, 765 (Feb, 2002).
S6. M. Przeworski, *Genetics* **160**, 1179 (Mar, 2002).
S7. W. Stephan, T. H. E. Wiehe, M. Lenz, *Theor Popul Biol* **41**, 237 (1992).
S8. R. Durrett, J. Schweinsberg, *Theor Popul Biol* **66**, 129 (Sep, 2004).
S9. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol* **4**, e72 (Mar 7, 2006).
S10. C. C. Spencer, G. Coop, *Bioinformatics* **20**, 3673 (Dec 12, 2004).
S11. P. C. Sabeti *et al.*, *Nature* **449**, 913 (Oct 18, 2007).
S12. B. S. Weir, C. C. Cockerham, *Evolution* **38**, 1358 (1984).
S13. P. C. Sabeti *et al.*, *Science* **312**, 1614 (Jun 16, 2006).
S14. V. Ramensky, P. Bork, S. Sunyaev, *Nucleic Acids Res* **30**, 3894 (Sep 1, 2002).
S15. M. Garber *et al.*, *Bioinformatics* **25**, i54 (Jun 15, 2009).
S16. http://genome.ucsc.edu/.
S17. S. D. Patel *et al.*, *Cell* **124**, 1255 (Mar 24, 2006).
S18. M. Sato *et al.*, *www.pdb.org*.
S19. M. A. Marti-Renom, M. S. Madhusudhan, A. Sali, *Protein Sci* **13**, 1071 (Apr, 2004).
S20. M. A. Marti-Renom *et al.*, *Annual review of biophysics and biomolecular structure* **29**, 291 (2000).
S21. M. Landau *et al.*, *Nucleic Acids Res* **33**, W299 (Jul 1, 2005).
S22. W. L. DeLano. (DeLano Scientific LLC, Palo Alto, CA, USA, 2007).
S23. B. E. Stranger *et al.*, *Science* **315**, 848 (Feb 9, 2007).
S24. www.sanger.ac.uk/humgen/genevar/.
S25. S. Kudaravalli, J. B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, J. K. Pritchard, *Mol Biol Evol* **26**, 649 (Mar, 2009).
S26. P. D. Thomas *et al.*, *Nucleic acids research* **31**, 334 (Jan 1, 2003).
S27. R. J. Cho, M. J. Campbell, *Trends Genet* **16**, 409 (Sep, 2000).

Fig. S1. Comparison of ΔDAF and DAF's ability to localize signal and distinguish causal variant. Left: top (red) and bottom (black) 5% of scores and mean score (black, dashed) of DAF (A) and ΔDAF (C) in 1MB surrounding causal mutation (located at red dashed line). Right: distribution of scores for the causal variant (red), nearby unselected variants (blue) and variants in regions without selection (grey, below axis) for DAF (B) and ΔDAF (D).

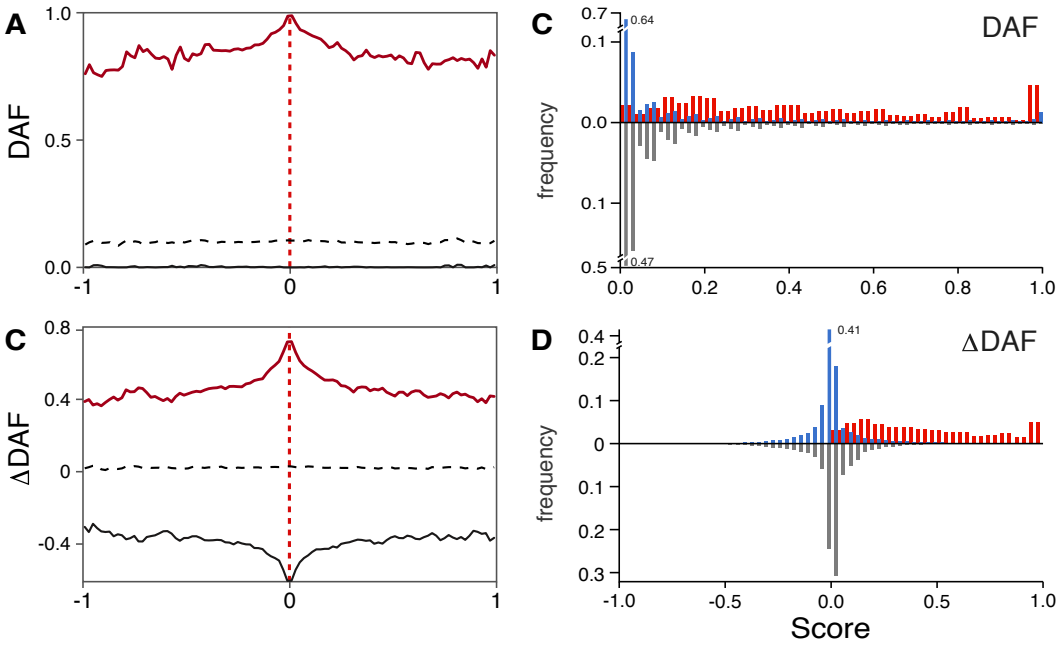


Fig. S2. Correlation between the individual tests for selection. The overall strength of correlation between scores for each of the 5 tests examined for neutral regions (A), and for neutral variants in regions under positive selection (B). The correlations of the individual tests are strongest around the selected position, and fall off as a function of distance from the causal allele (C). Correlation coefficient for all tests with overall correlation greater than 0.2, calculated in sliding windows of 10kb across the 1MB selected region.

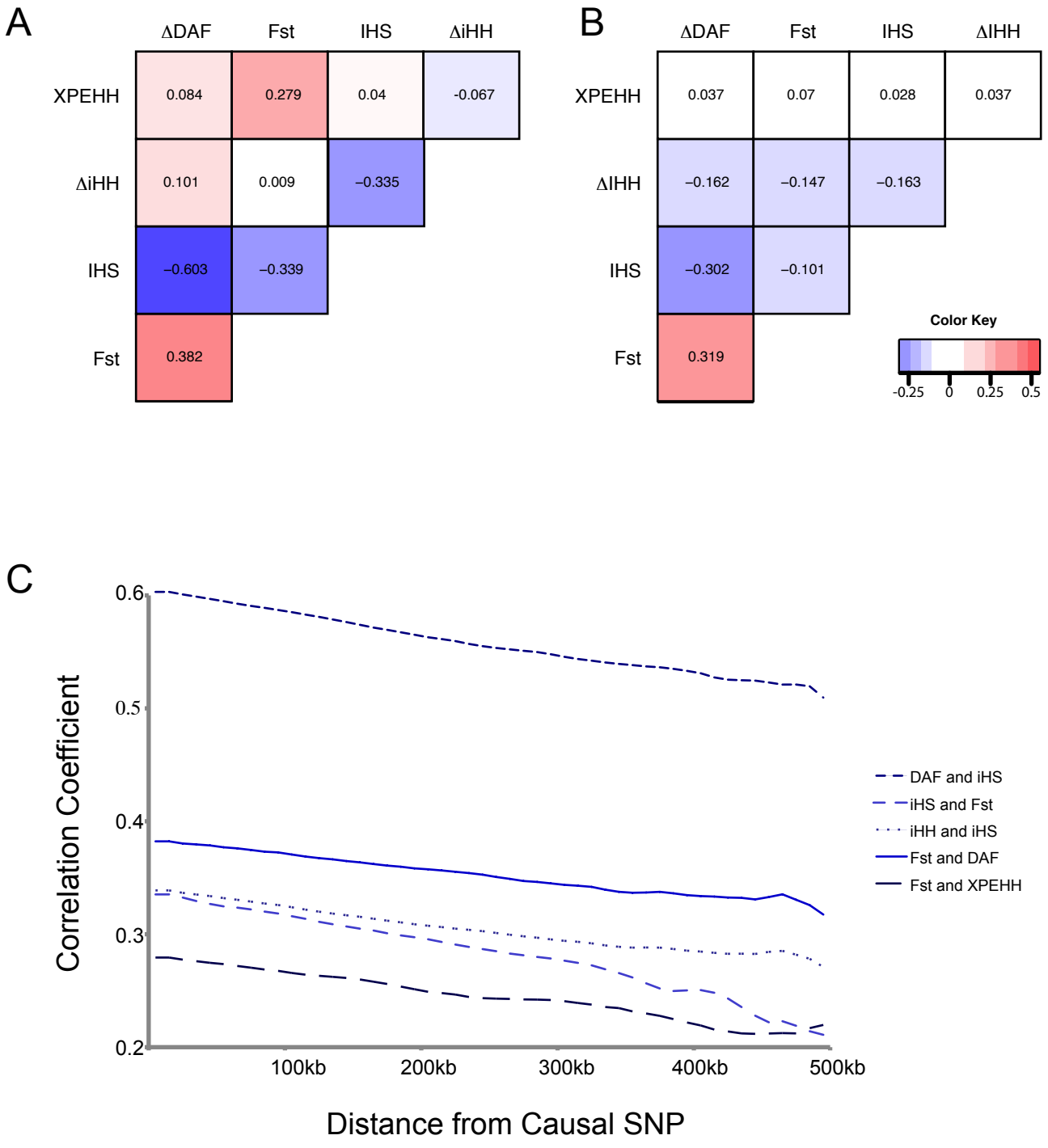
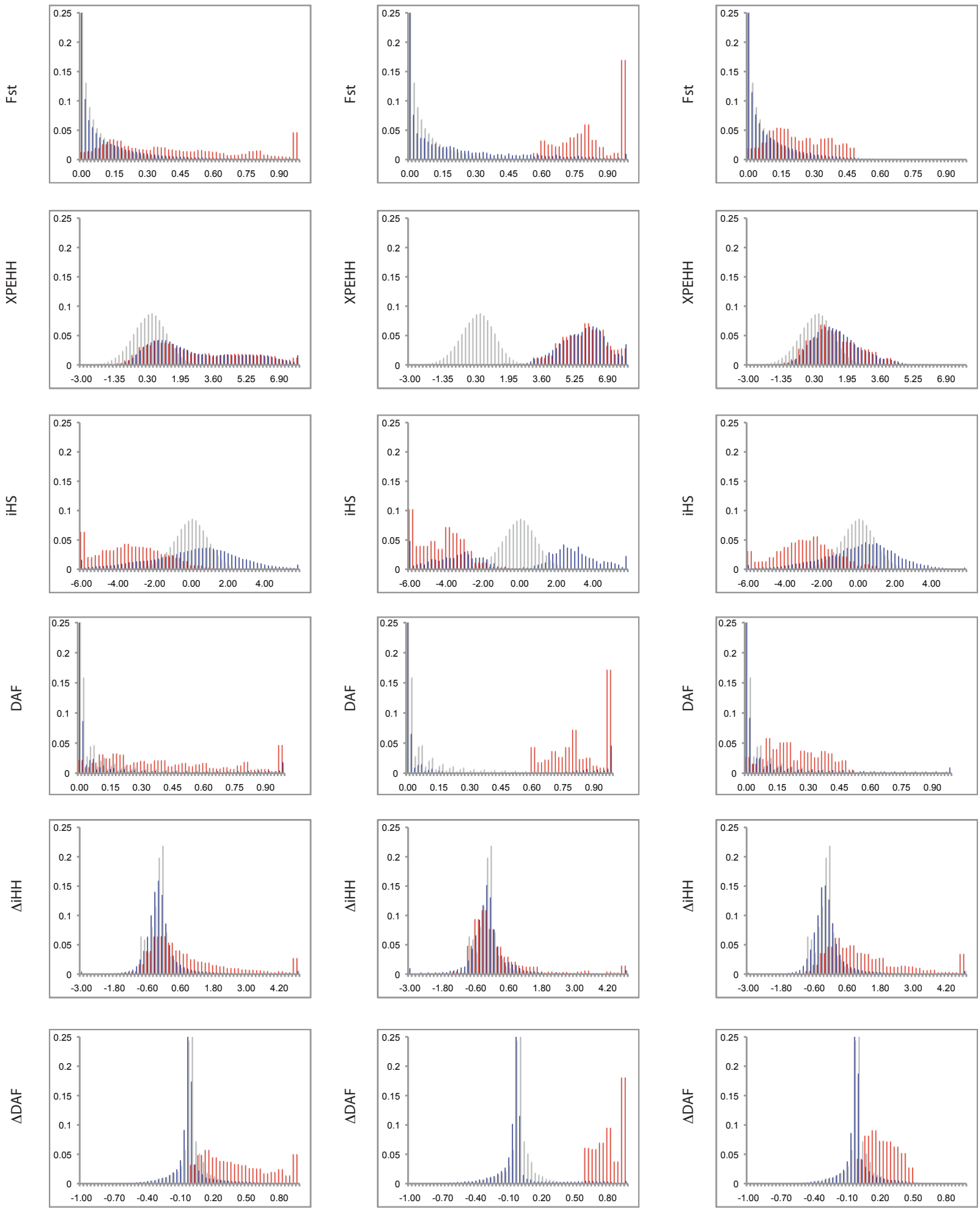


Fig. S3. Each test's ability to distinguish the selected variant for high (50-100%) and low (0-50%) frequency variants. We show the distribution of scores for the causal variant, compared to unselected variants in the same simulated region (blue), and variants in regions without selection (grey, below axis). Blue bars in A show the distribution of scores for neutral variants within 1MB of selected variant, and in B for neutral variants within 10kb of selected variant. While XP-EHH and FST are the most effective for high frequency selected variants, ΔiHH, and iHS are better at distinguishing the low frequency causal variants.


B
All Frequency
High Frequency
Low Frequency
Fst
XPEHH
iHS
DAF
ΔiHH
ΔDAF

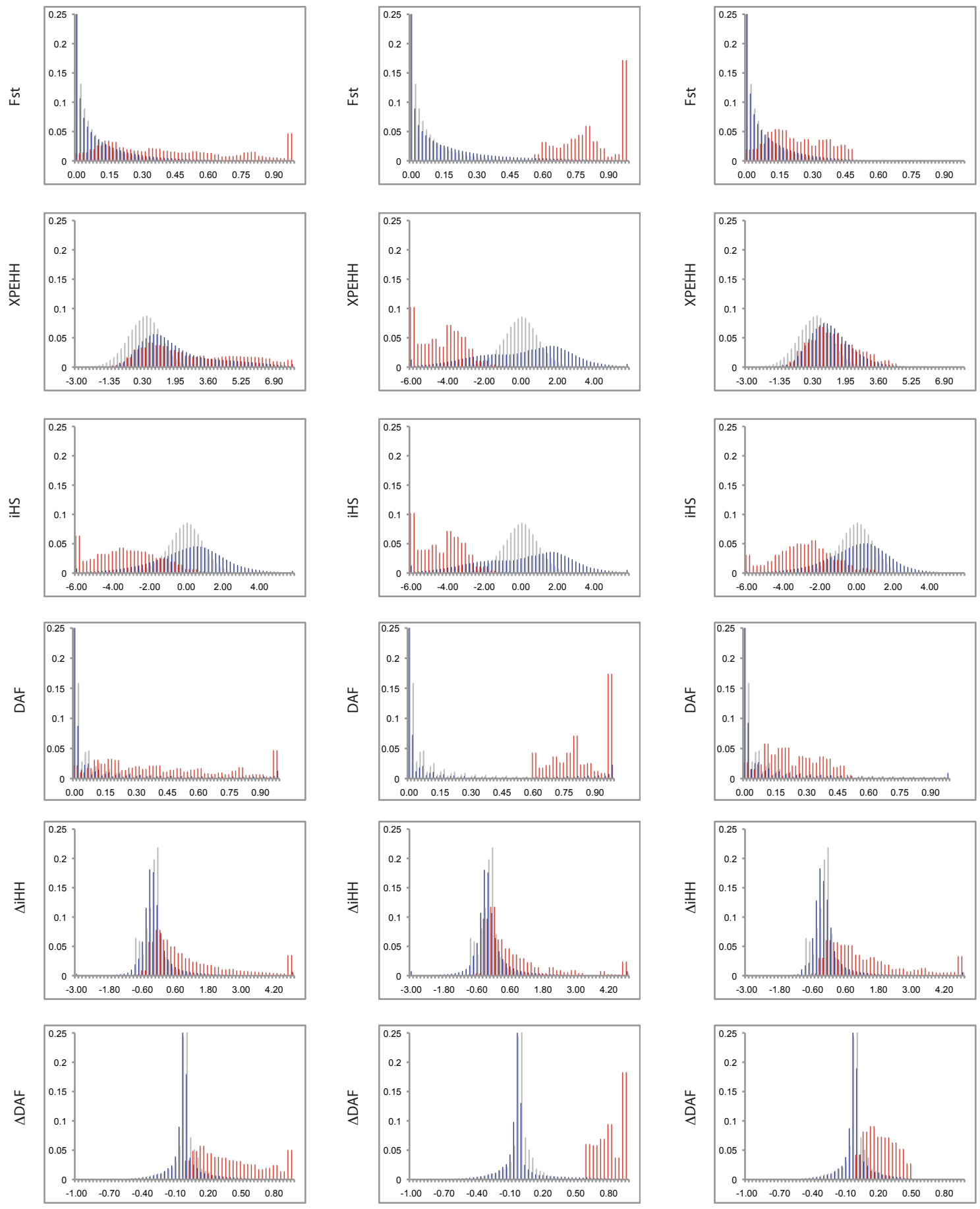

A
All Frequency
High Frequency
Low Frequency
Fst
XPEHH
iHS
DAF
ΔiHH
ΔDAF

Figure S4. Distribution of CMS scores of variants in regions without selection (neutral) simulated under different demographic models. The null distribution under the a constant sized population model (magenta circles), the calibrated European model (blue crosses), East Asian model (green triangles), and West African model (red squares), as well as a more extreme bottleneck (cyan triangles).

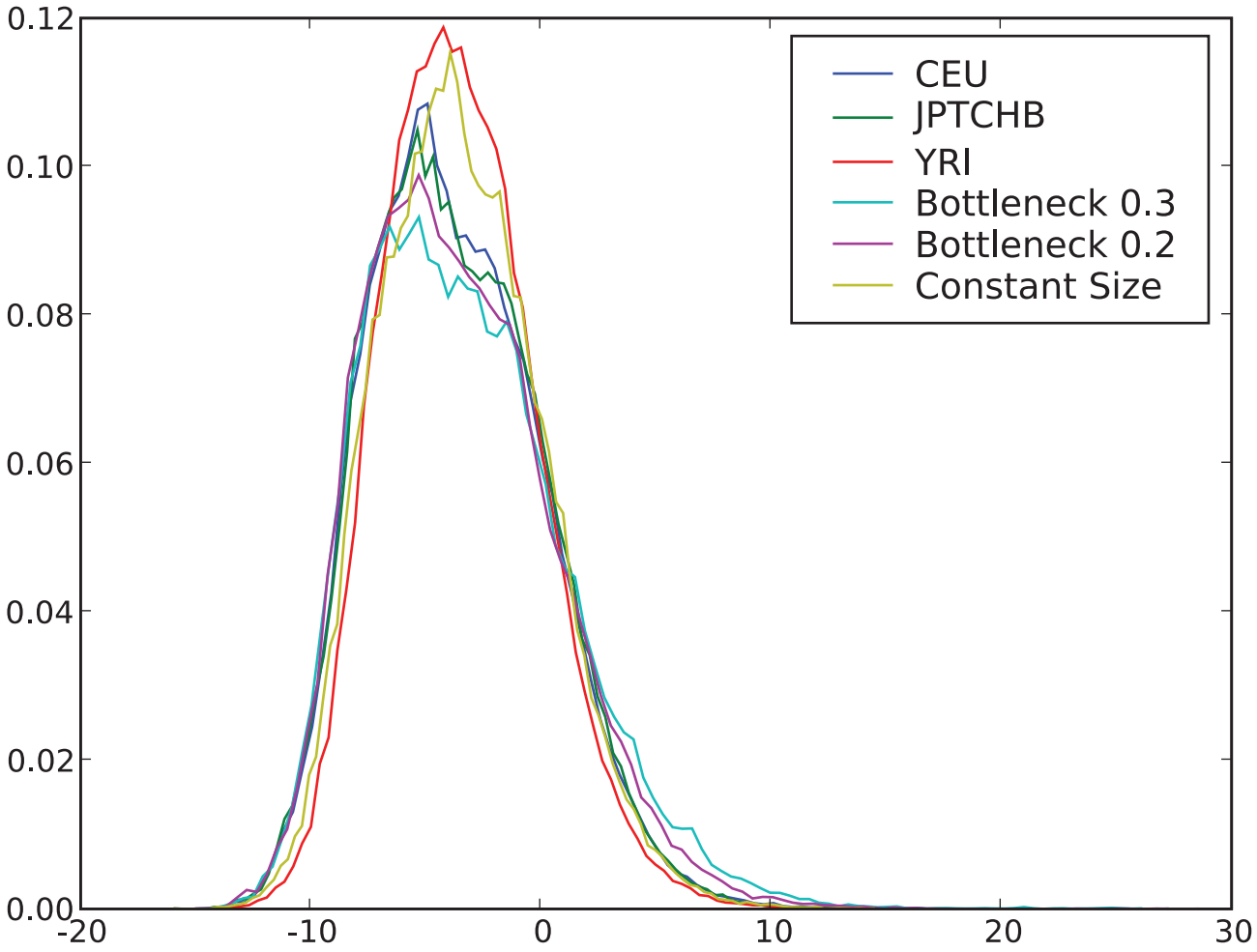
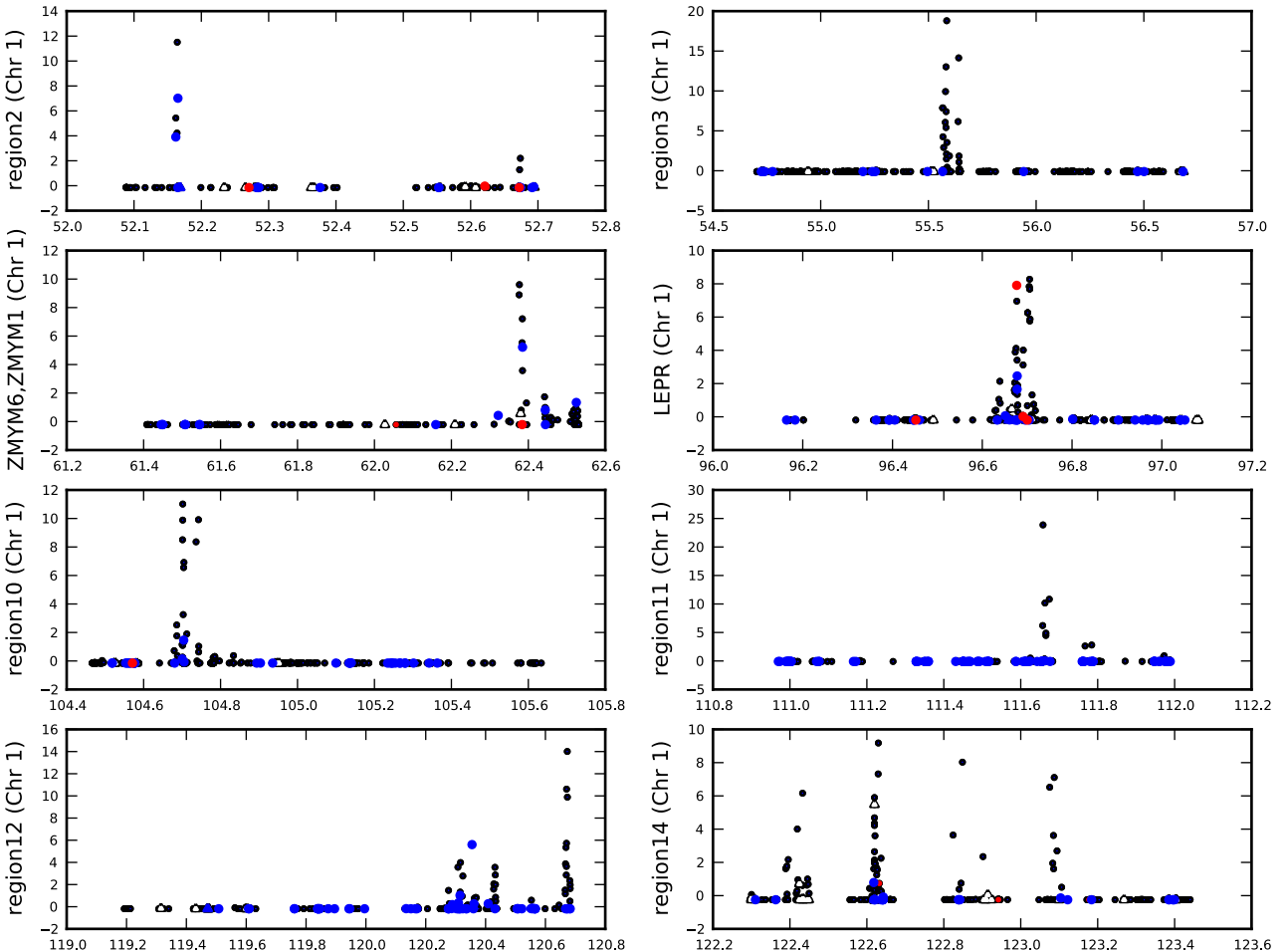
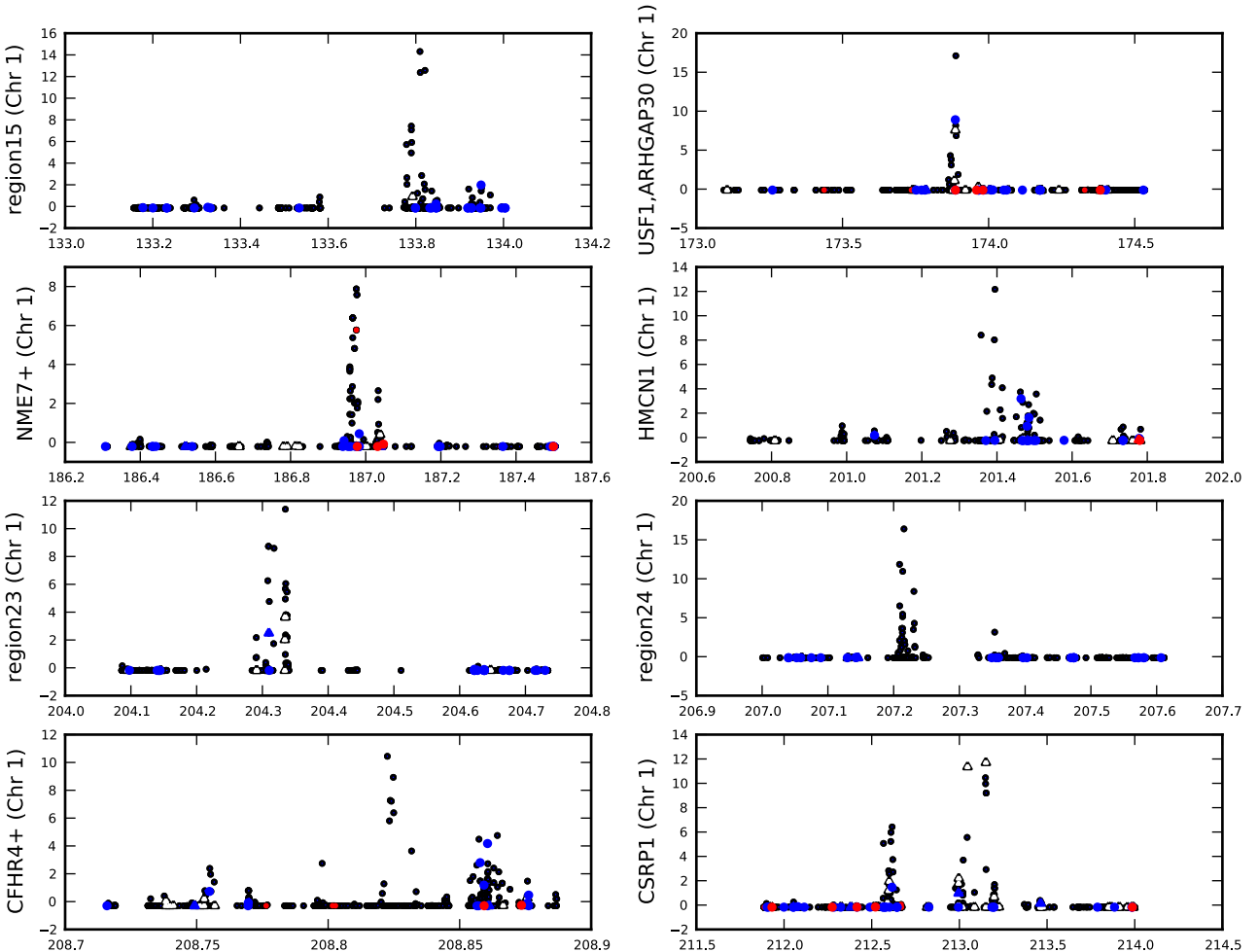
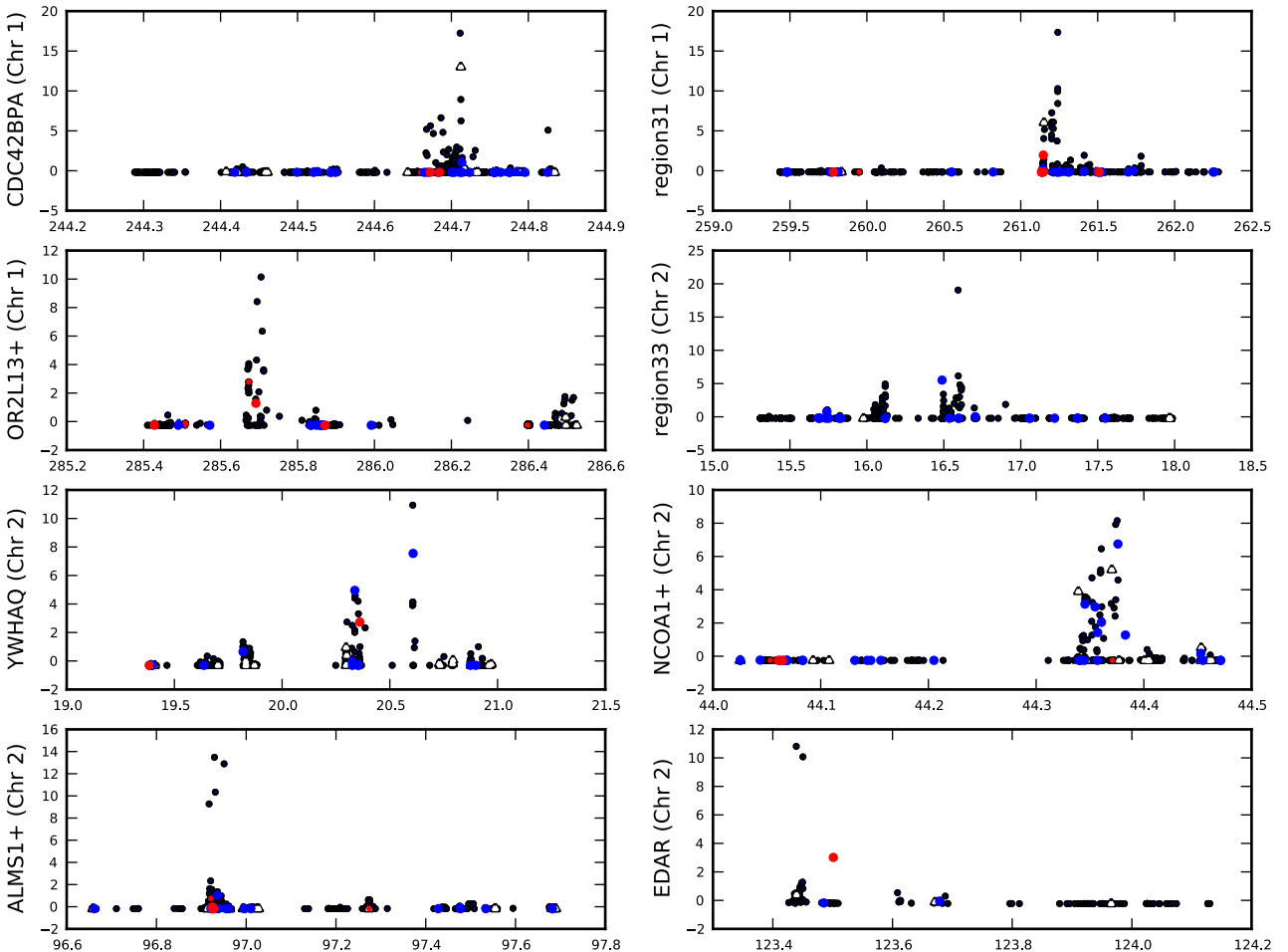
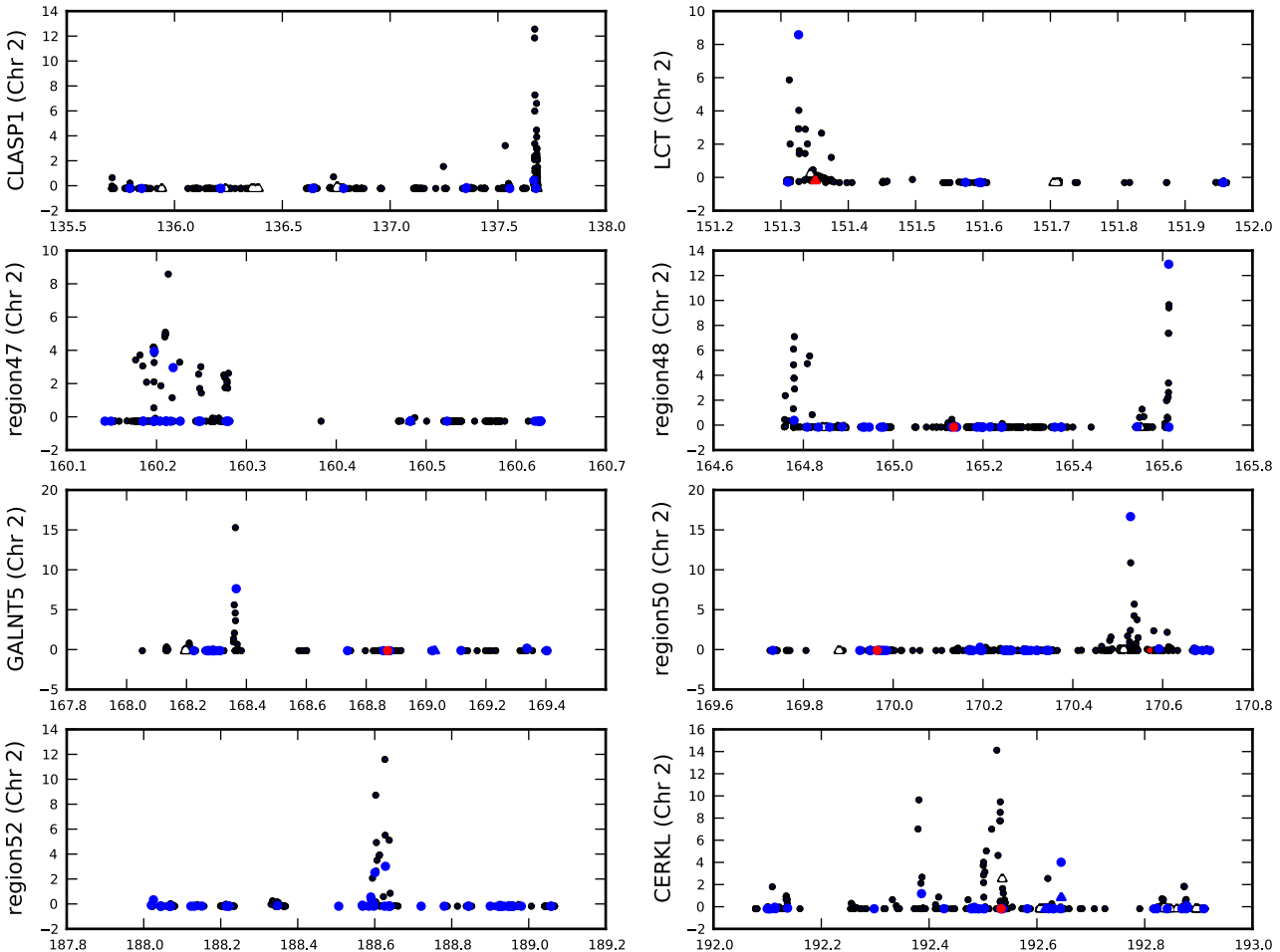


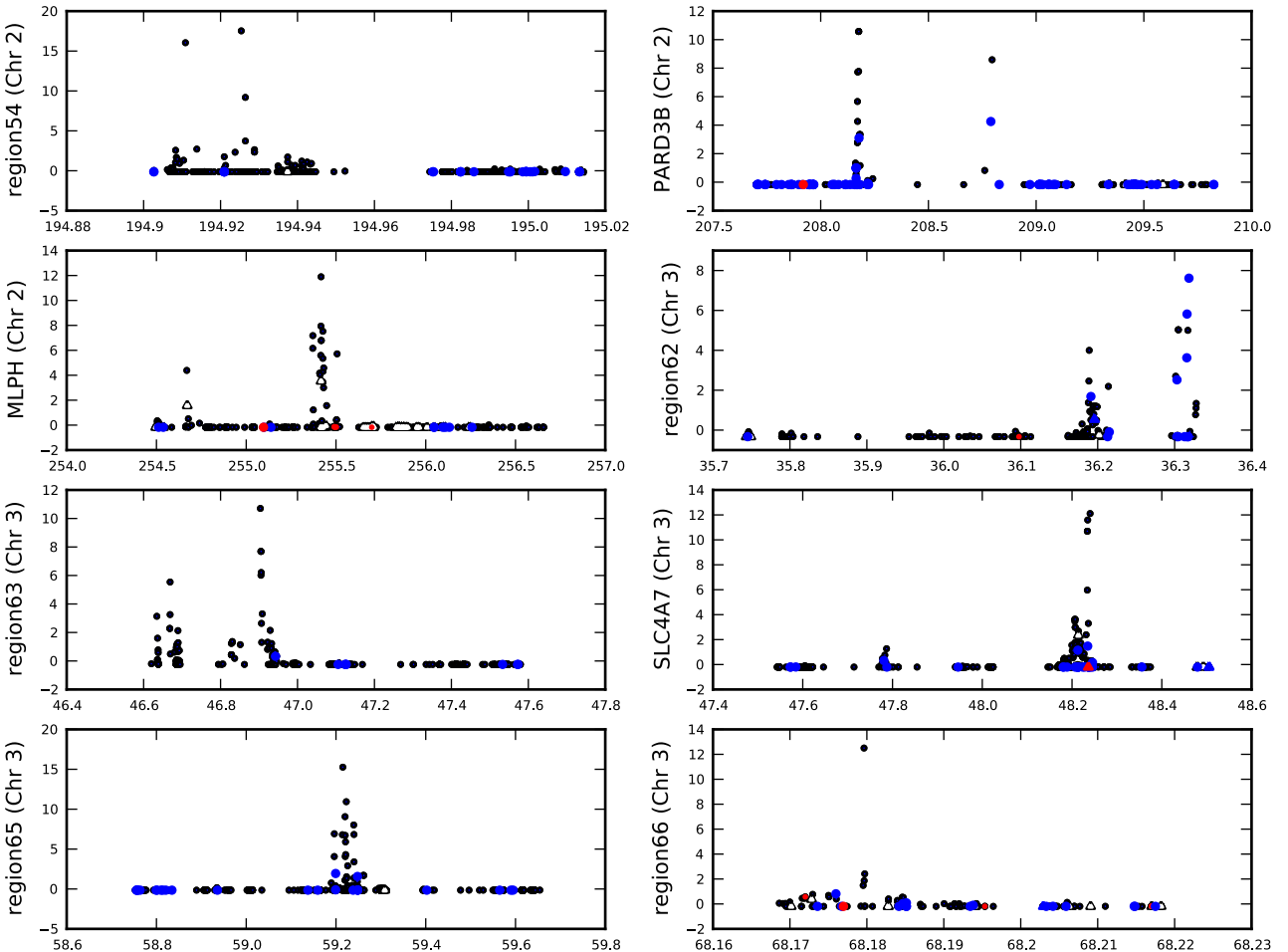
Fig. S5. Application of CMS to large selected regions in the Human Haplotype Map For each of the (1 MB) candidate selected regions, we plot the scores for CMS for all variants in the region, distinguishing non-synonymous SNPs (red), variants in conserved regions (blue), and variants in experimentally determined transcription factor binding sites (white triangle). The left axis is labeled with the genes (if any) that lie within the CMS peak. If more than two genes lie within the peak, the axis is labeled with the name of the first genes with a "+" to indicate the presence of the remaining genes, listed in Table S3.

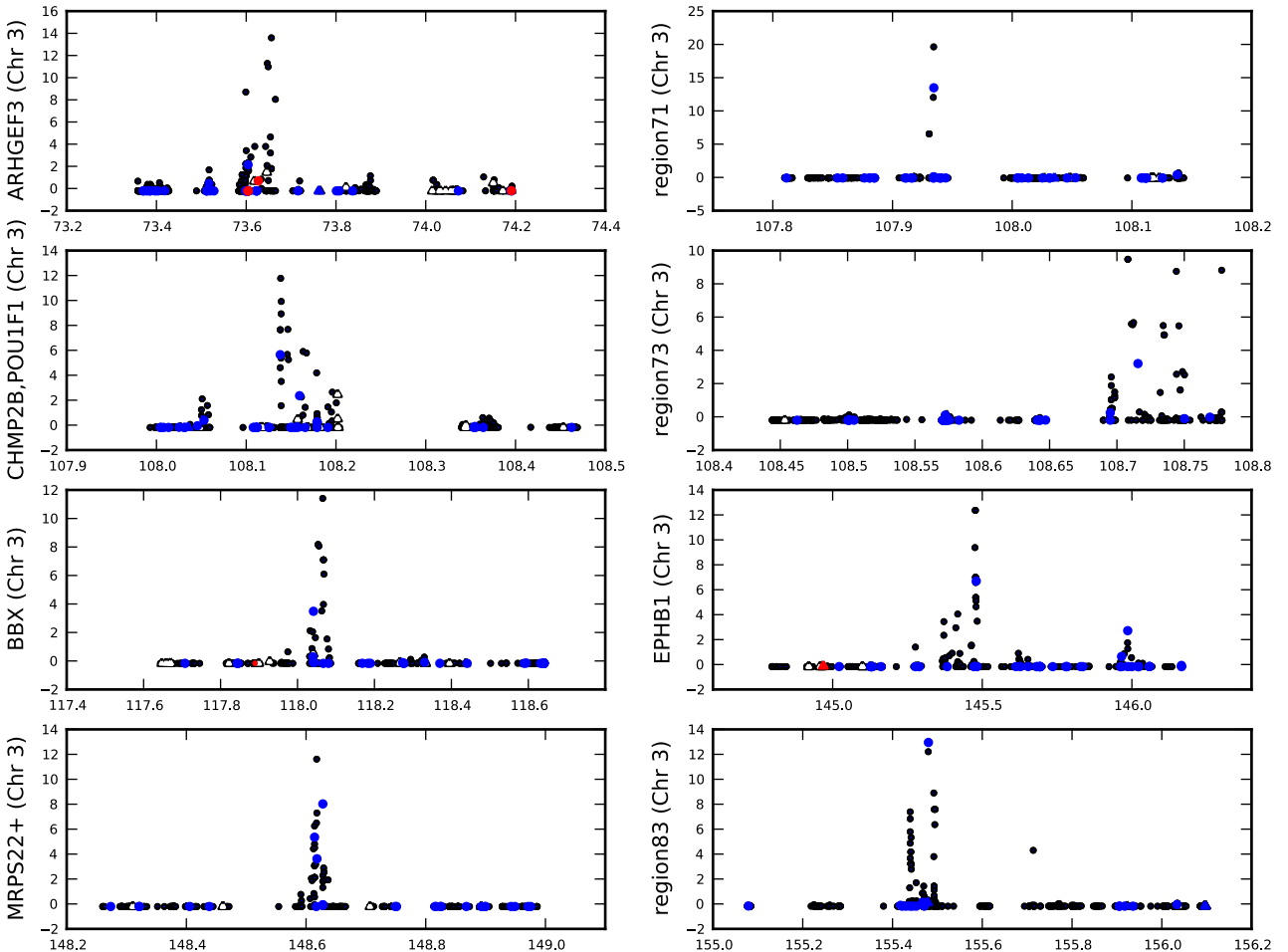

region2 (Chr 1)
region3 (Chr 1)
ZMYM6,ZMYM1 (Chr 1)
LEPR (Chr 1)
region10 (Chr 1)
region11 (Chr 1)
region12 (Chr 1)
region14 (Chr 1)

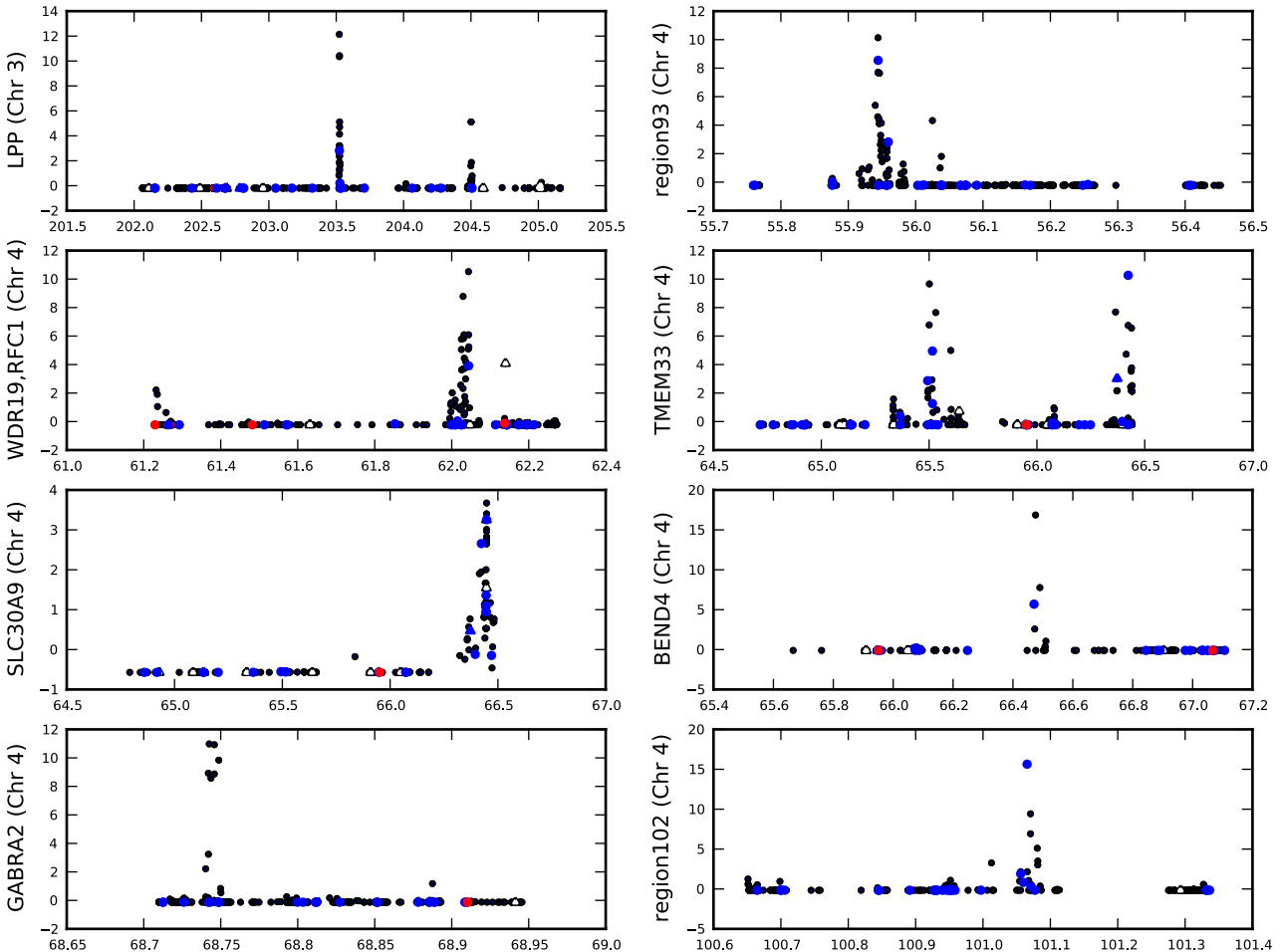

region15 (Chr 1)
USF1,ARHGAP30 (Chr 1)
NME7+ (Chr 1)
HMCN1 (Chr 1)
region23 (Chr 1)
region24 (Chr 1)
CFHR4+ (Chr 1)
CSRP1 (Chr 1)

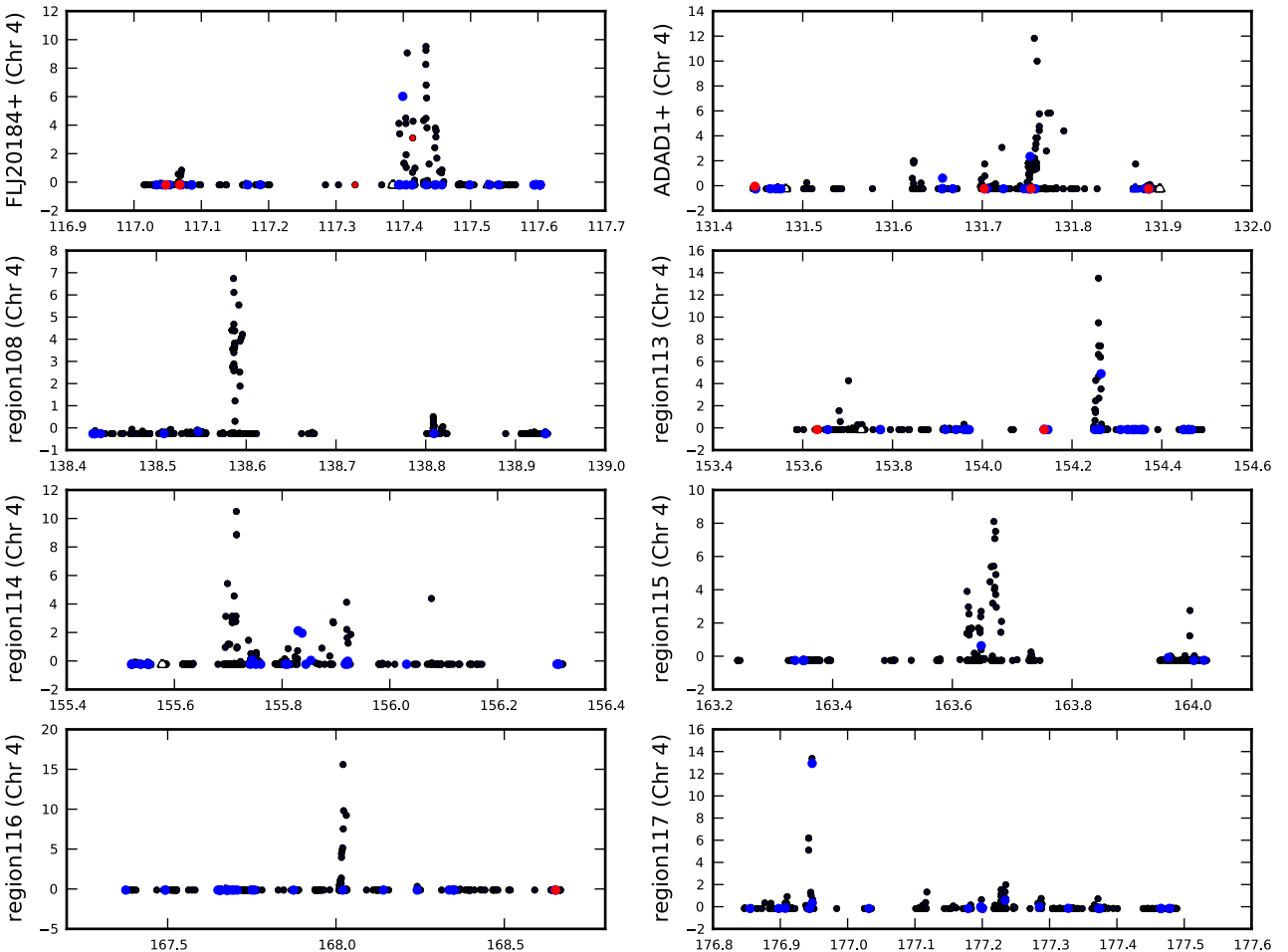

CDC42BPA (Chr 1)
region31 (Chr 1)
OR2L13+ (Chr 1)
region33 (Chr 2)
YWHAQ (Chr 2)
NCOA1+ (Chr 2)
ALMS1+ (Chr 2)
EDAR (Chr 2)

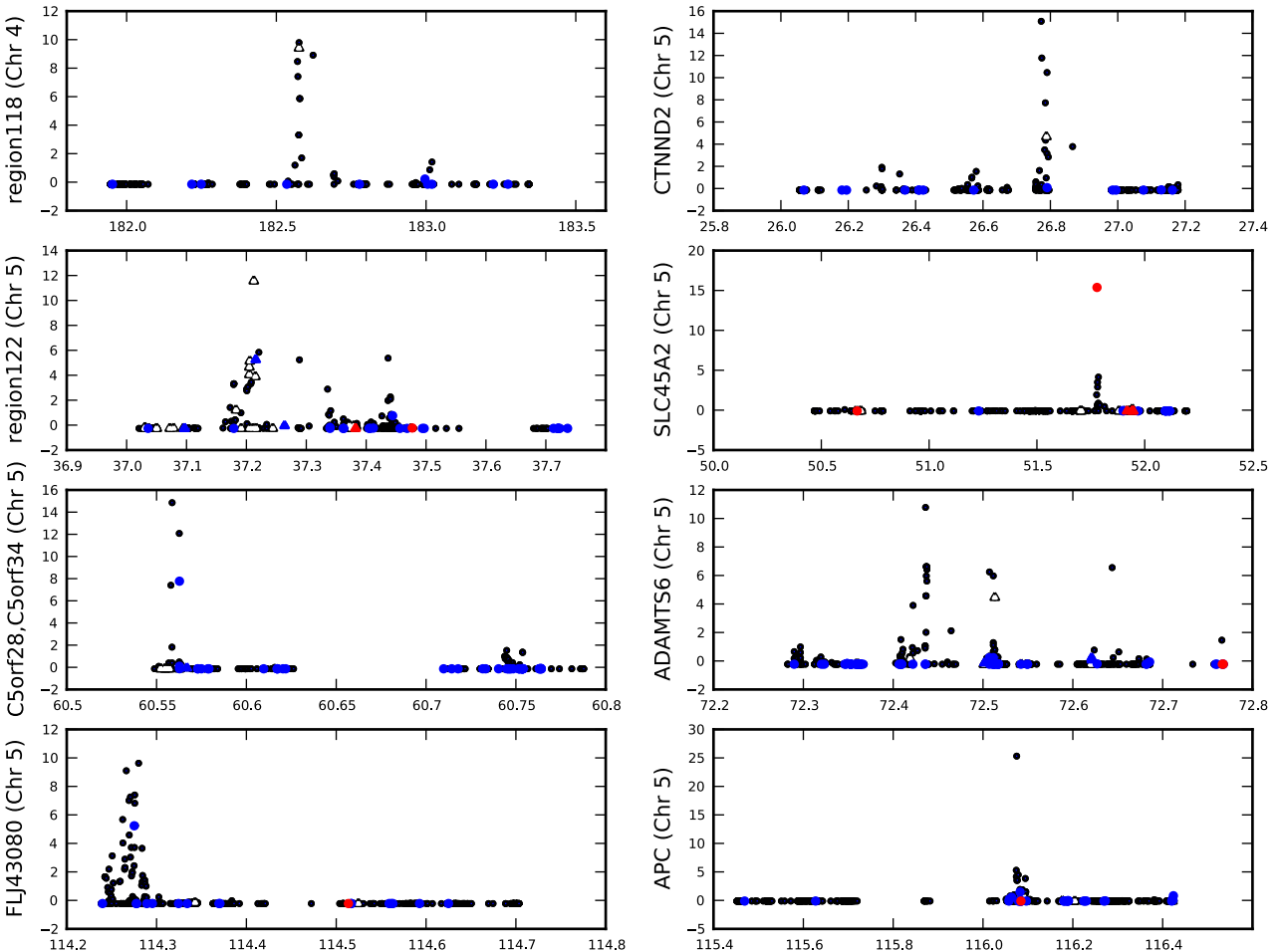

CLASP1 (Chr 2)
LCT (Chr 2)
region47 (Chr 2)
region48 (Chr 2)
GALNT5 (Chr 2)
region50 (Chr 2)
region52 (Chr 2)
CERKL (Chr 2)

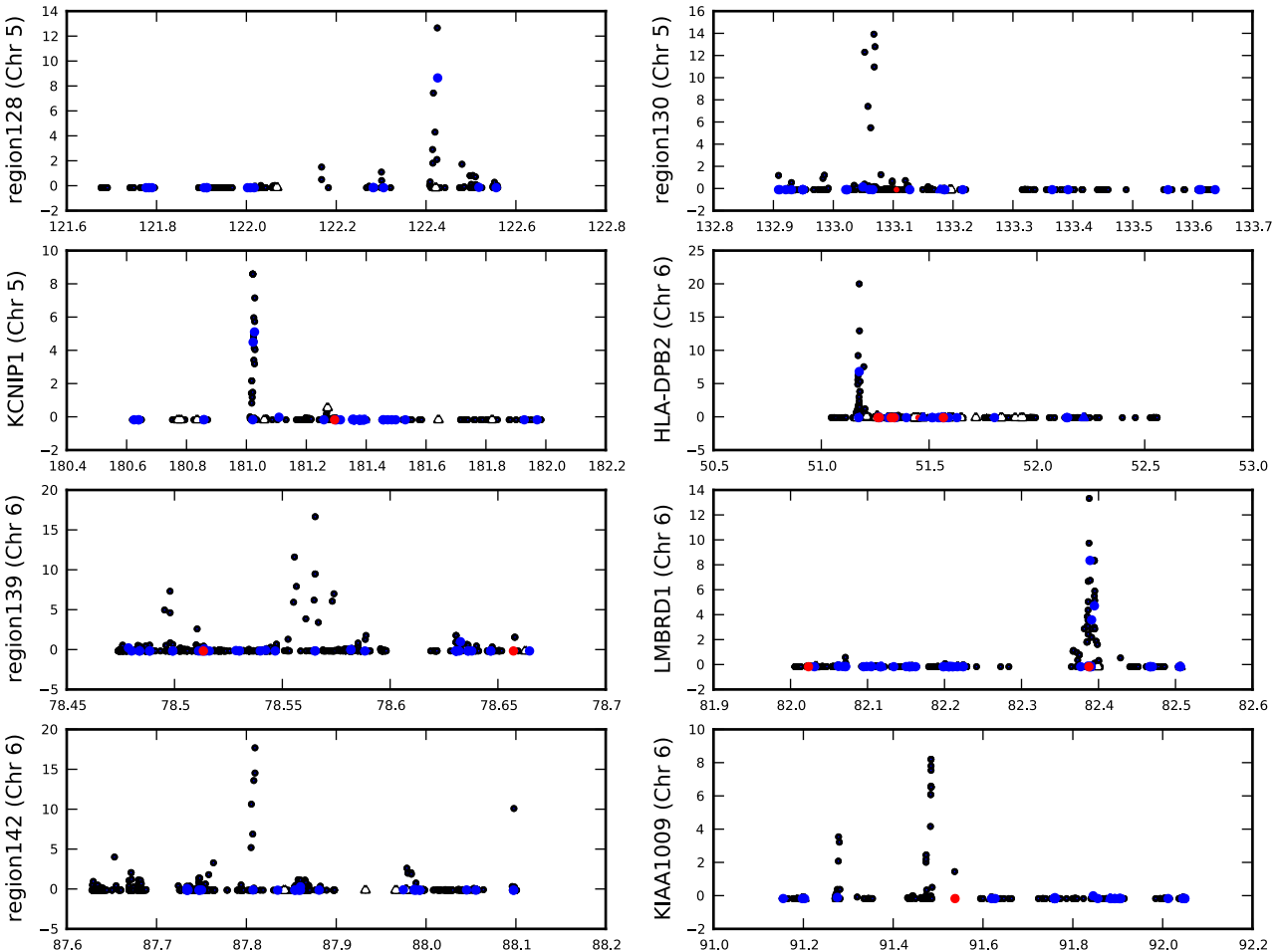

region54 (Chr 2)
PARD3B (Chr 2)
MLPH (Chr 2)
region62 (Chr 3)
region63 (Chr 3)
SLC4A7 (Chr 3)
region65 (Chr 3)
region66 (Chr 3)

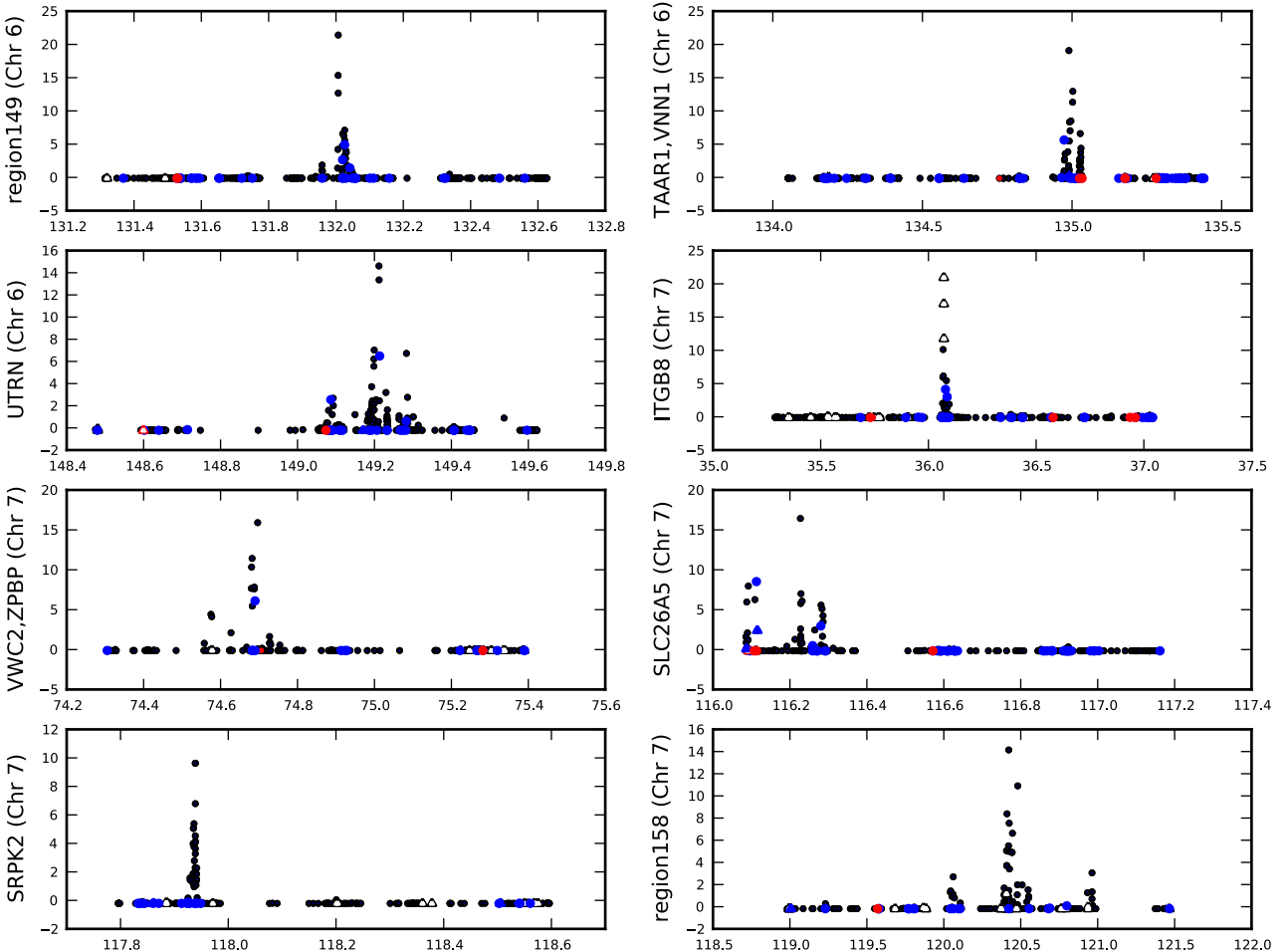

ARHGEF3 (Chr 3)
region71 (Chr 3)
CHMP2B,POU1F1 (Chr 3)
region73 (Chr 3)
BBX (Chr 3)
EPHB1 (Chr 3)
MRPS22+ (Chr 3)
region83 (Chr 3)

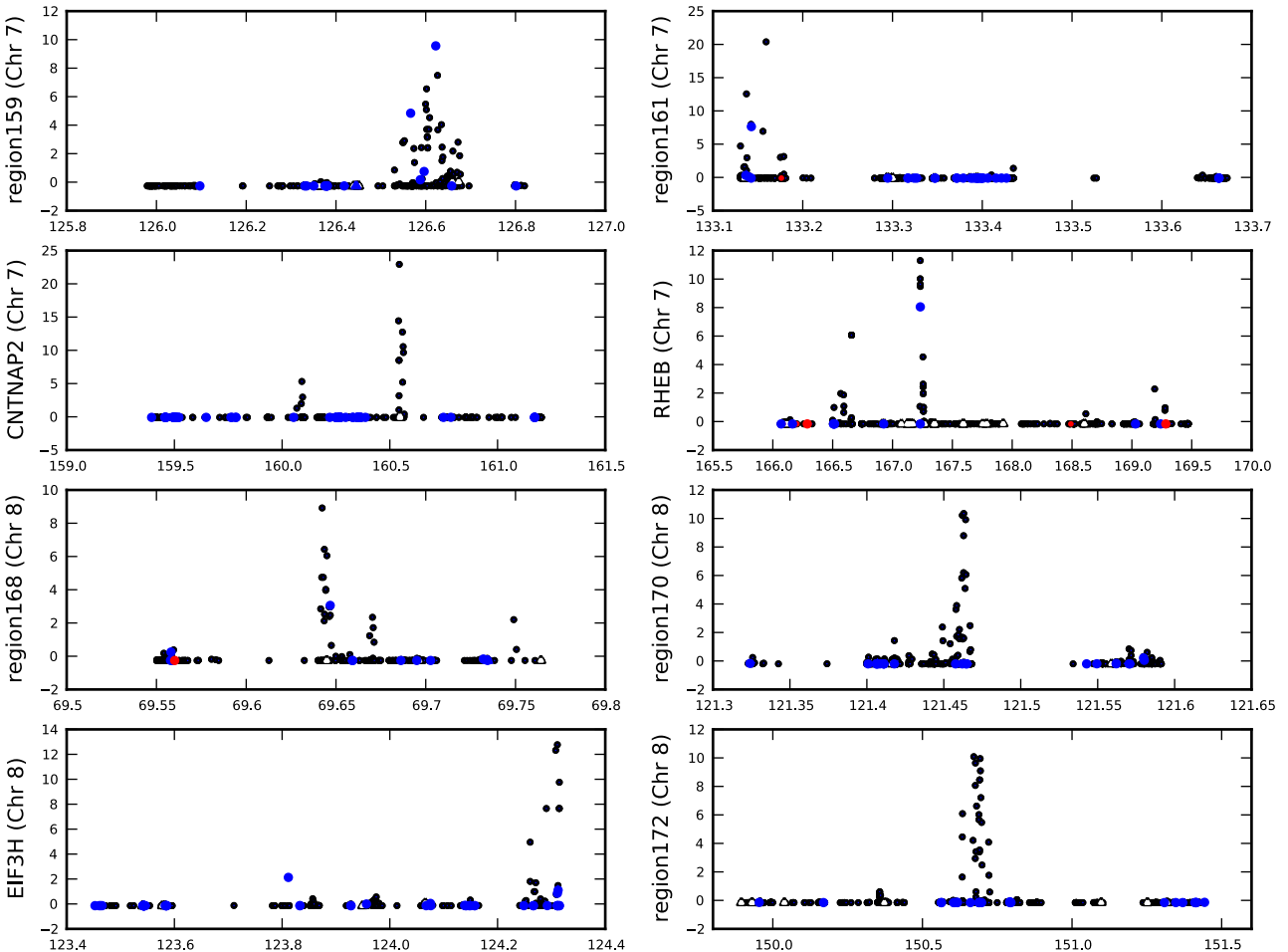

LPP (Chr 3)
region93 (Chr 4)
WDR19,RFC1 (Chr 4)
TMEM33 (Chr 4)
SLC30A9 (Chr 4)
BEND4 (Chr 4)
GABRA2 (Chr 4)
region102 (Chr 4)

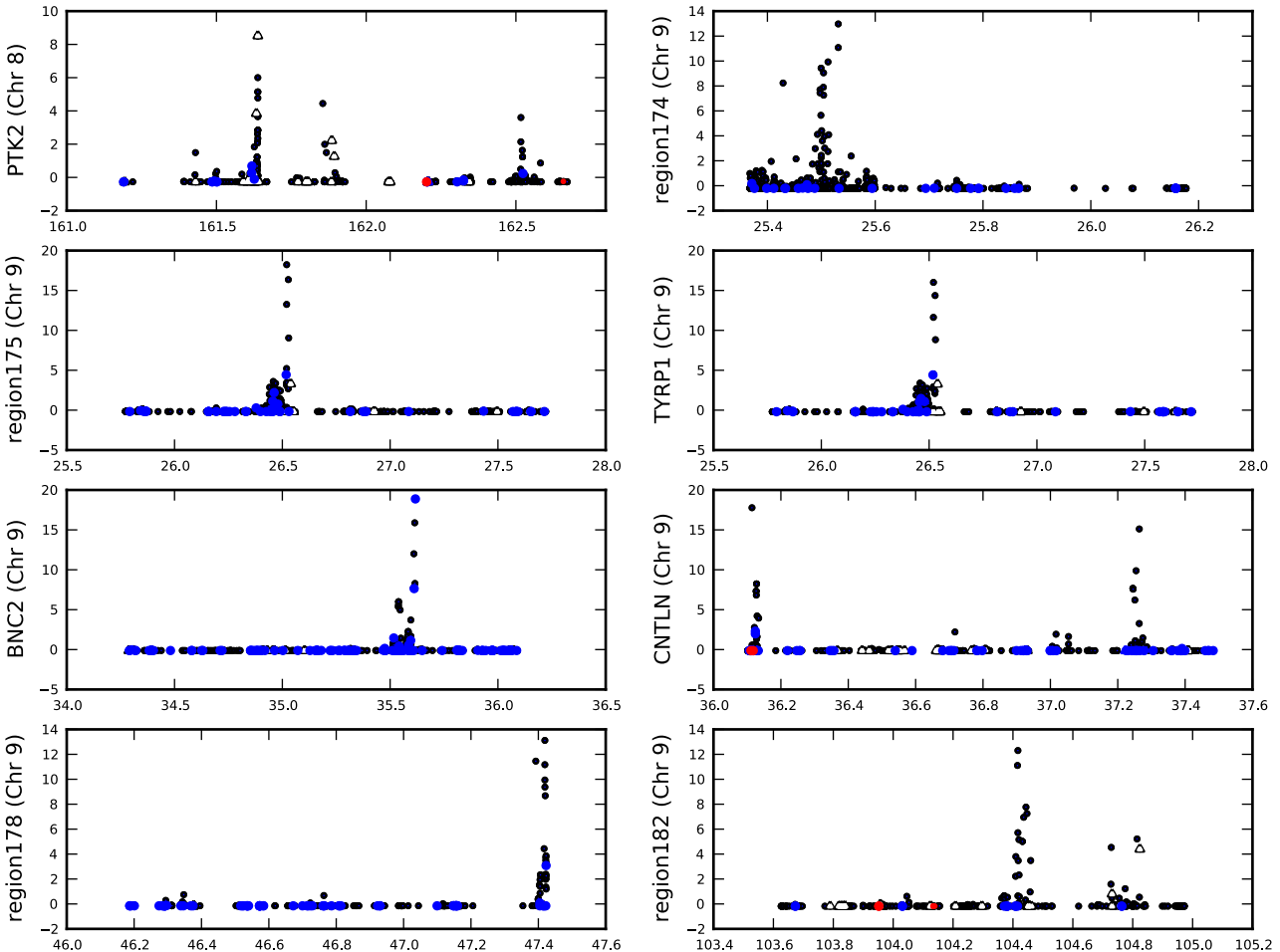

FLJ20184+ (Chr 4)
ADAD1+ (Chr 4)
region108 (Chr 4)
region113 (Chr 4)
region114 (Chr 4)
region115 (Chr 4)
region116 (Chr 4)
region117 (Chr 4)

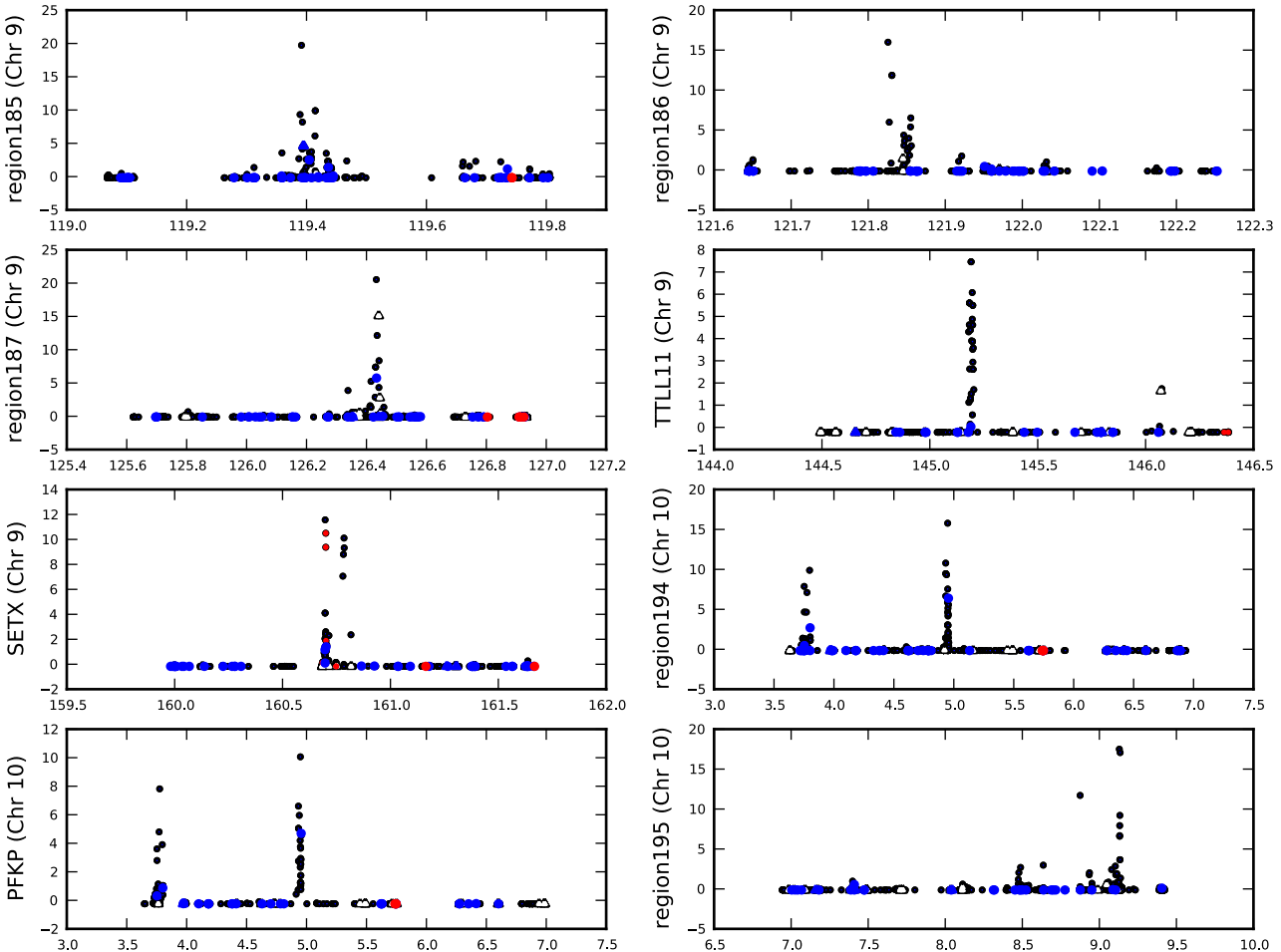

region118 (Chr 4)
CTNND2 (Chr 5)
region122 (Chr 5)
SLC45A2 (Chr 5)
C5orf28,C5orf34 (Chr 5)
ADAMTS6 (Chr 5)
FLJ43080 (Chr 5)
APC (Chr 5)

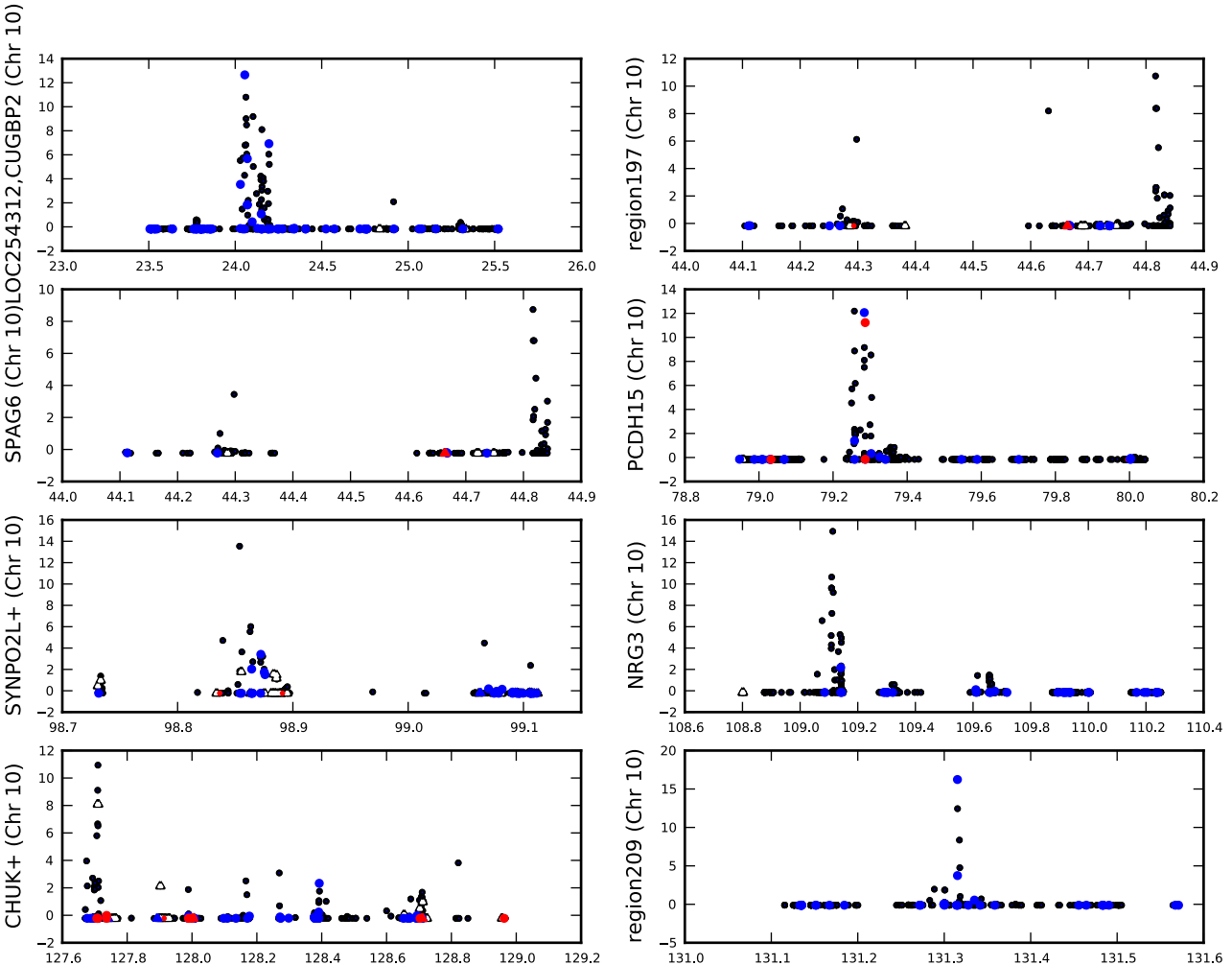

region128 (Chr 5)
region130 (Chr 5)
KCNIP1 (Chr 5)
HLA-DPB2 (Chr 6)
region139 (Chr 6)
LMBRD1 (Chr 6)
region142 (Chr 6)
KIAA1009 (Chr 6)

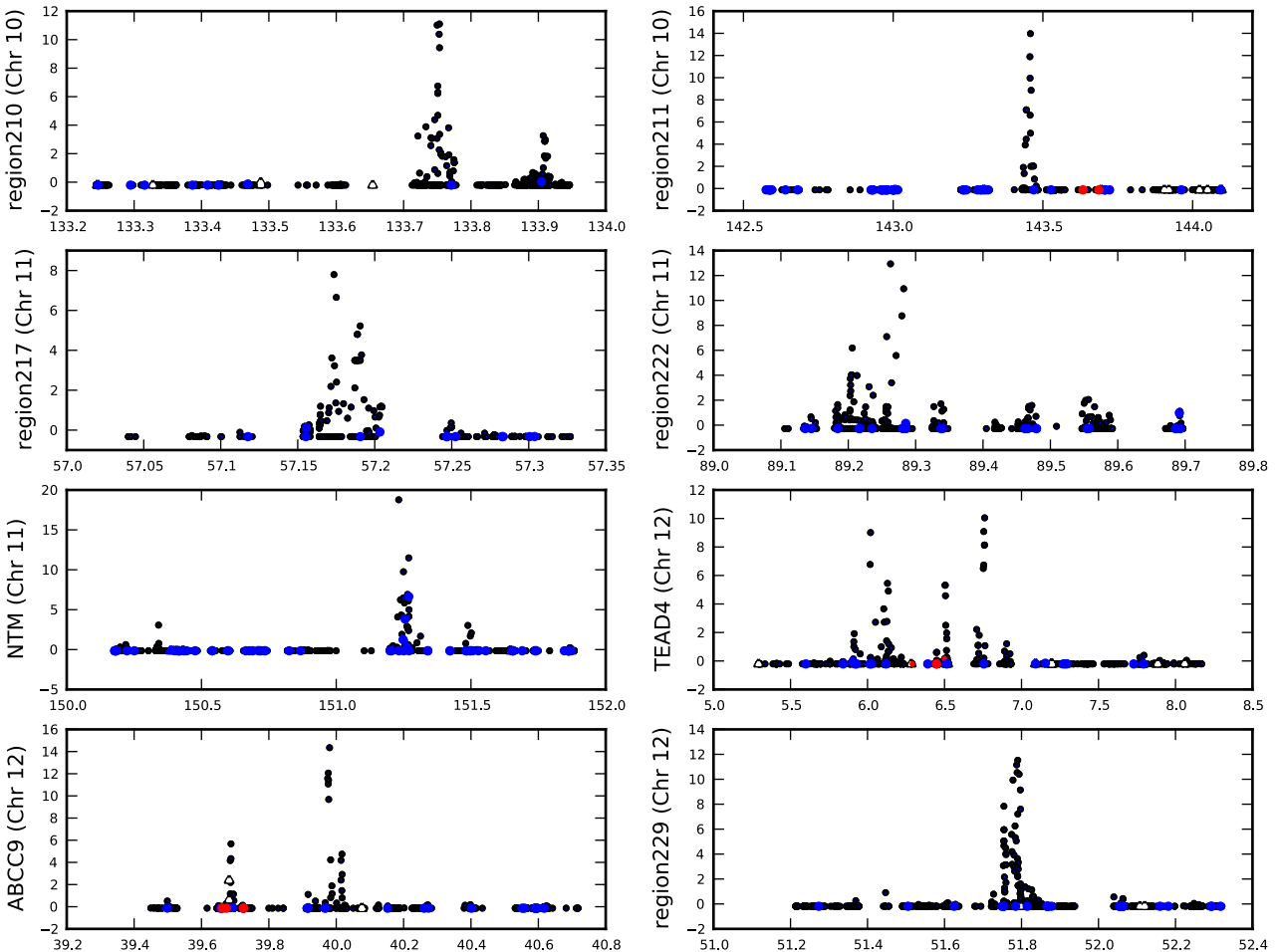

region149 (Chr 6)
TAAR1,VNN1 (Chr 6)
UTRN (Chr 6)
ITGB8 (Chr 7)
VWC2,ZPBP (Chr 7)
SLC26A5 (Chr 7)
SRPK2 (Chr 7)
region158 (Chr 7)

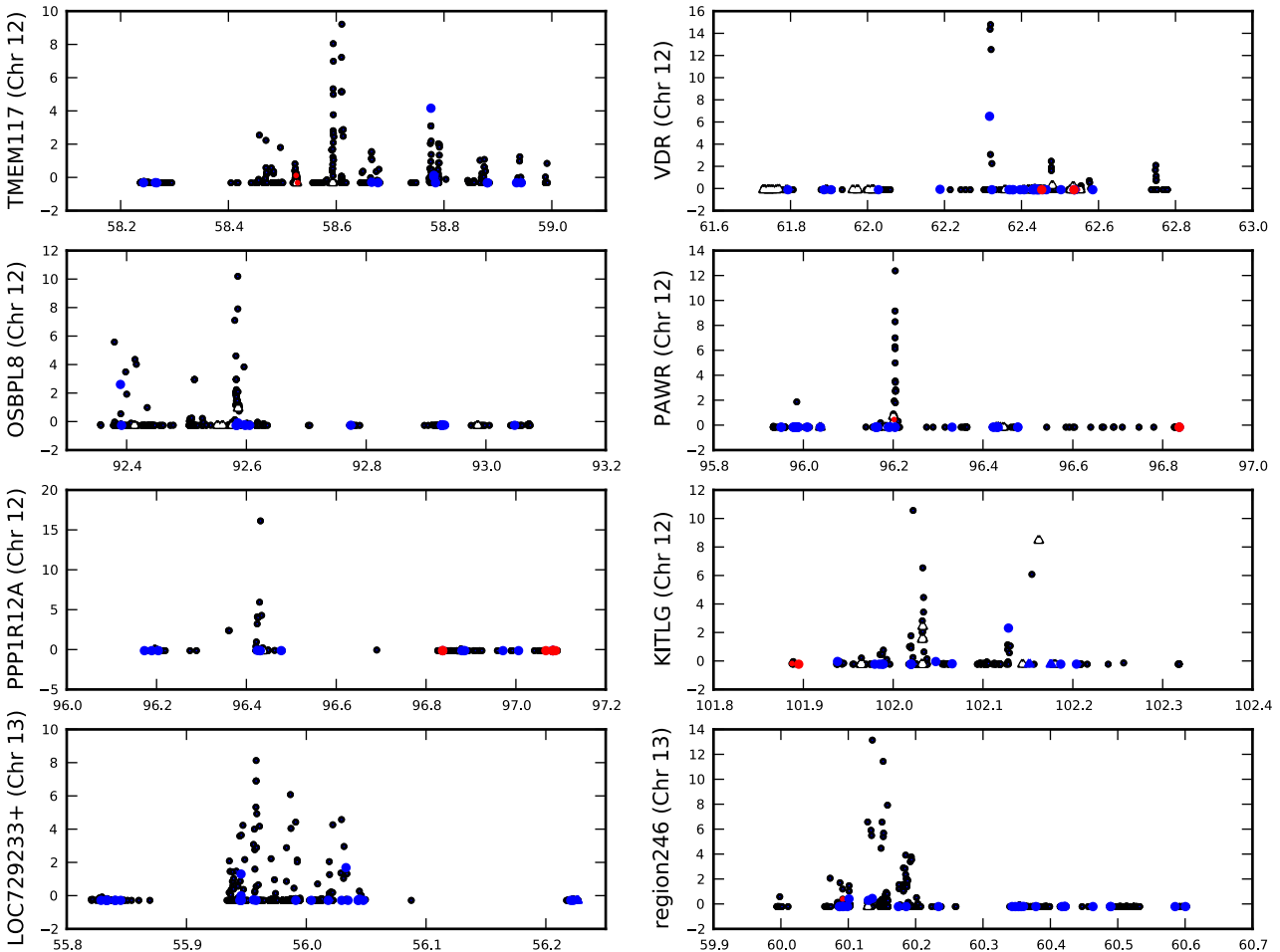

region159 (Chr 7)
region161 (Chr 7)
CNTNAP2 (Chr 7)
RHEB (Chr 7)
region168 (Chr 8)
region170 (Chr 8)
EIF3H (Chr 8)
region172 (Chr 8)

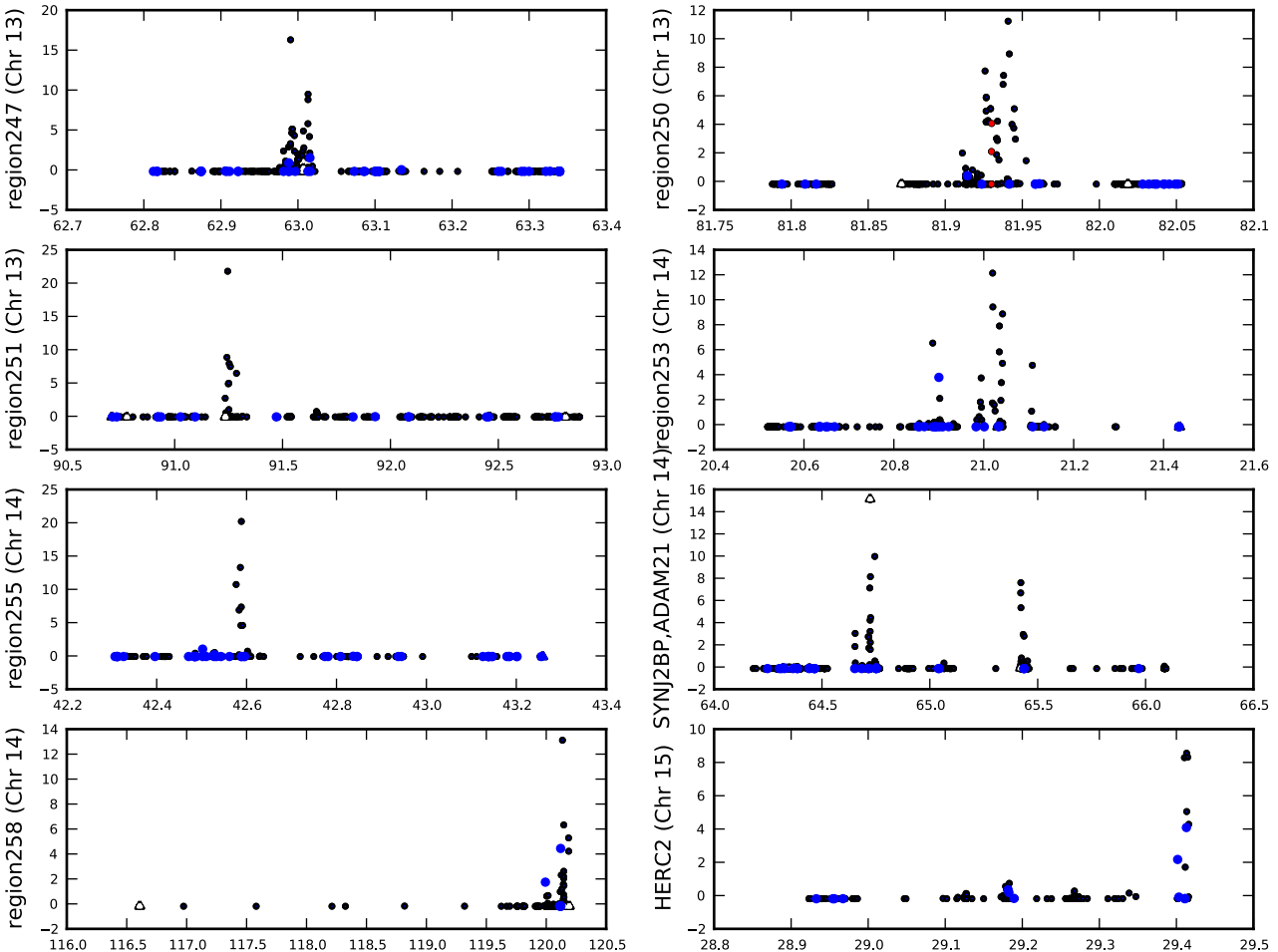

PTK2 (Chr 8)
region174 (Chr 9)
region175 (Chr 9)
TYRP1 (Chr 9)
BNC2 (Chr 9)
CNTLN (Chr 9)
region178 (Chr 9)
region182 (Chr 9)

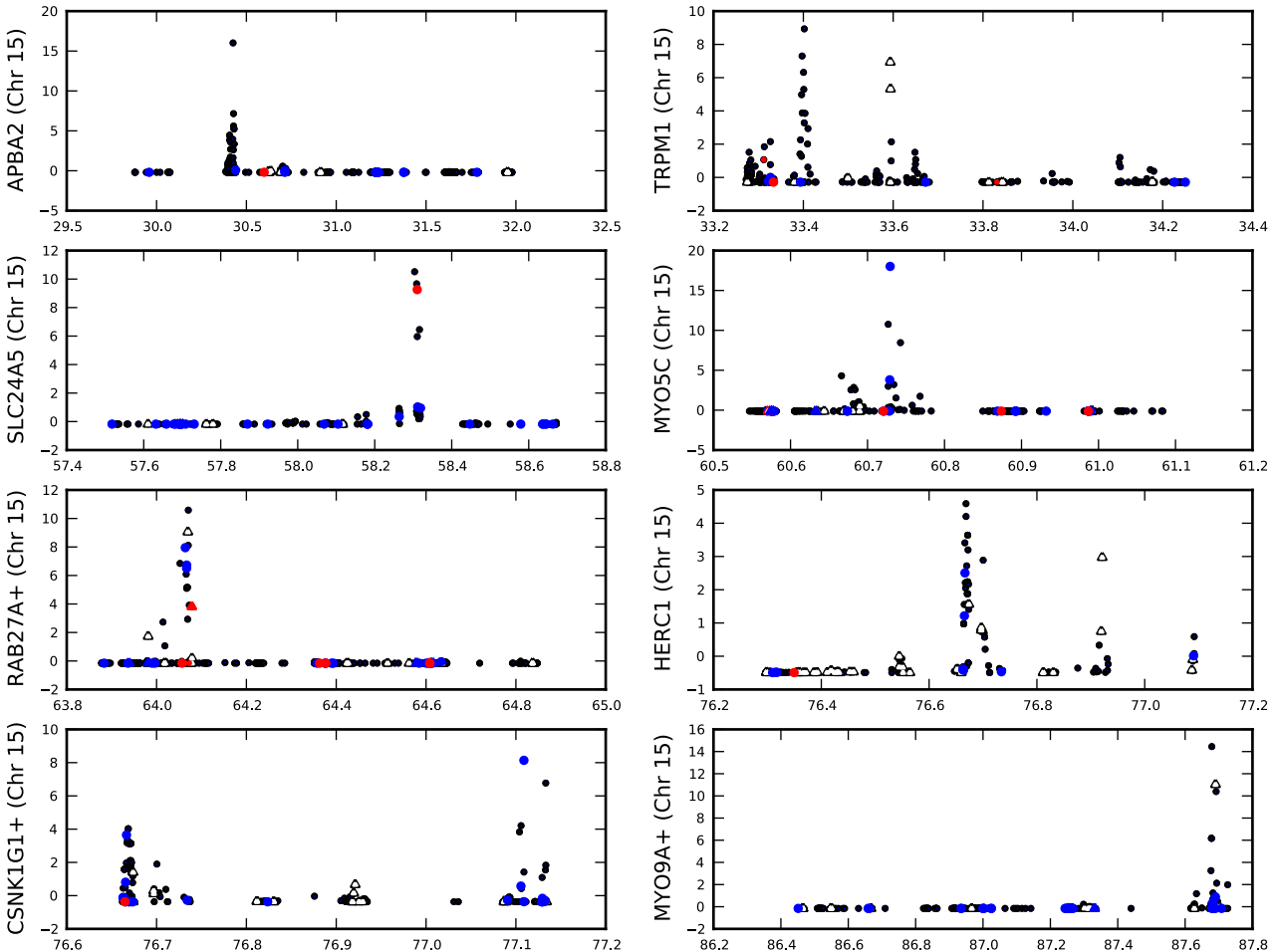

region185 (Chr 9)
region186 (Chr 9)
region187 (Chr 9)
TTLL11 (Chr 9)
SETX (Chr 9)
region194 (Chr 10)
PFKP (Chr 10)
region195 (Chr 10)

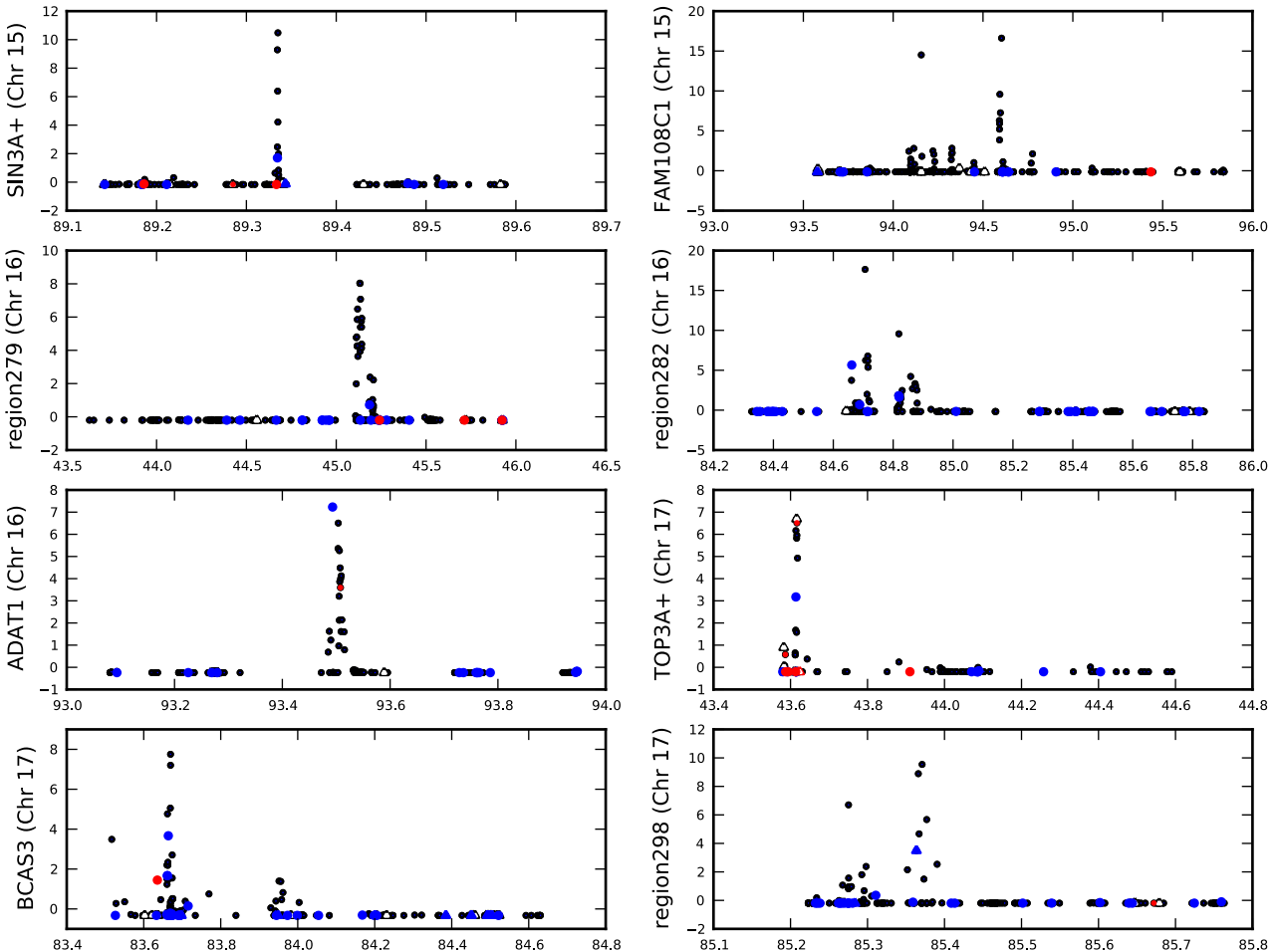

SPAG6 (Chr 10)LOC254312,CUGBP2 (Chr 10)
region197 (Chr 10)
PCDH15 (Chr 10)
SYNPO2L+ (Chr 10)
NRG3 (Chr 10)
CHUK+ (Chr 10)
region209 (Chr 10)

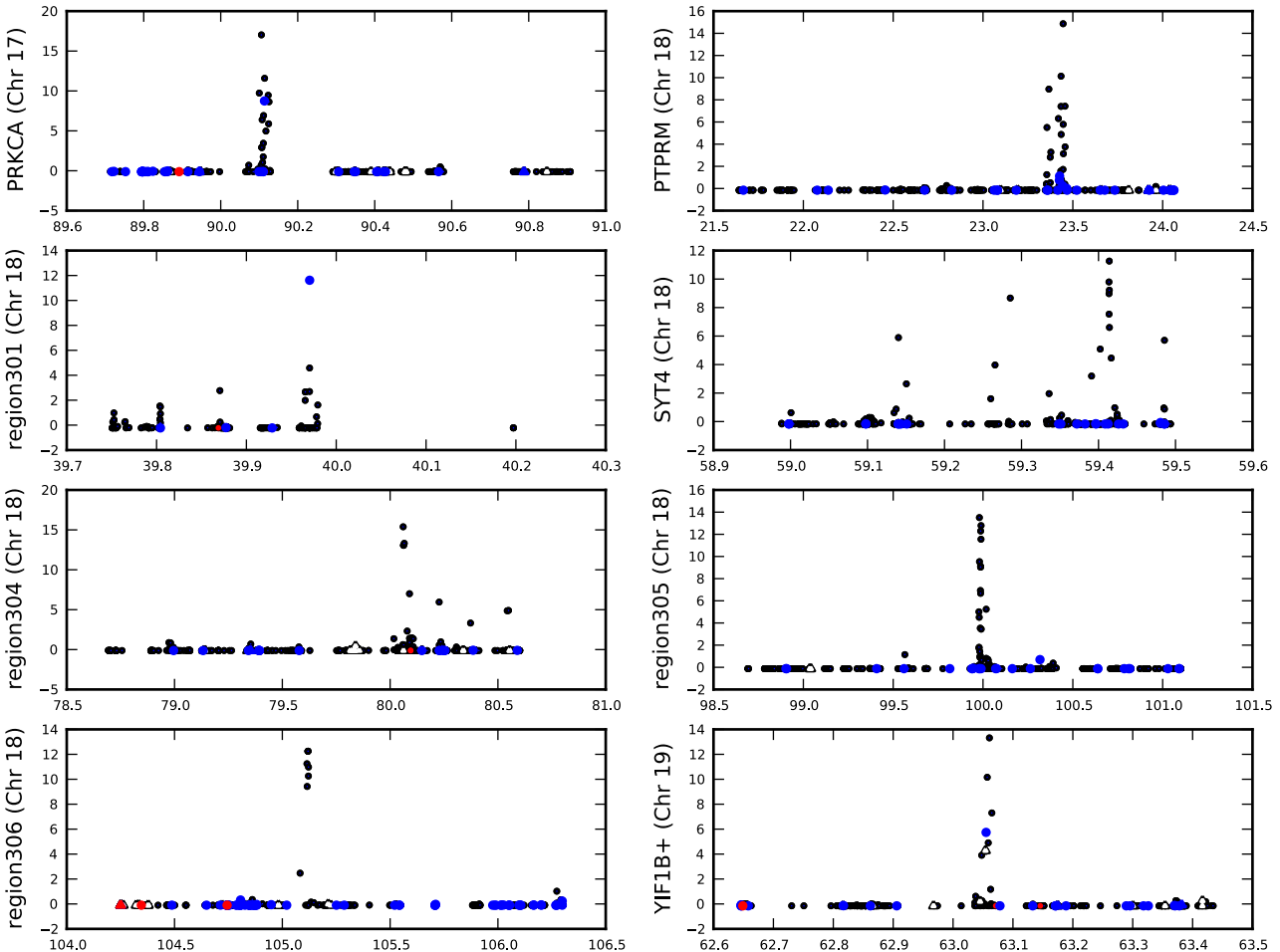

region210 (Chr 10)
region211 (Chr 10)
region217 (Chr 11)
region222 (Chr 11)
NTM (Chr 11)
TEAD4 (Chr 12)
ABCC9 (Chr 12)
region229 (Chr 12)


TMEM117 (Chr 12)
VDR (Chr 12)
OSBPL8 (Chr 12)
PAWR (Chr 12)
PPP1R12A (Chr 12)
KITLG (Chr 12)
LOC729233+ (Chr 13)
region246 (Chr 13)


region247 (Chr 13)
region250 (Chr 13)
region251 (Chr 13)
region253 (Chr 14)
region255 (Chr 14)
SYNJ2BP,ADAM21 (Chr 14)
region258 (Chr 14)
HERC2 (Chr 15)


APBA2 (Chr 15)
TRPM1 (Chr 15)
SLC24A5 (Chr 15)
MYO5C (Chr 15)
RAB27A+ (Chr 15)
HERC1 (Chr 15)
CSNK1G1+ (Chr 15)
MYO9A+ (Chr 15)


SIN3A+ (Chr 15)
FAM108C1 (Chr 15)
region279 (Chr 16)
region282 (Chr 16)
ADAT1 (Chr 16)
TOP3A+ (Chr 17)
BCAS3 (Chr 17)
region298 (Chr 17)


PRKCA (Chr 17)
PTPRM (Chr 18)
region301 (Chr 18)
SYT4 (Chr 18)
region304 (Chr 18)
region305 (Chr 18)
region306 (Chr 18)
YIF1B+ (Chr 19)

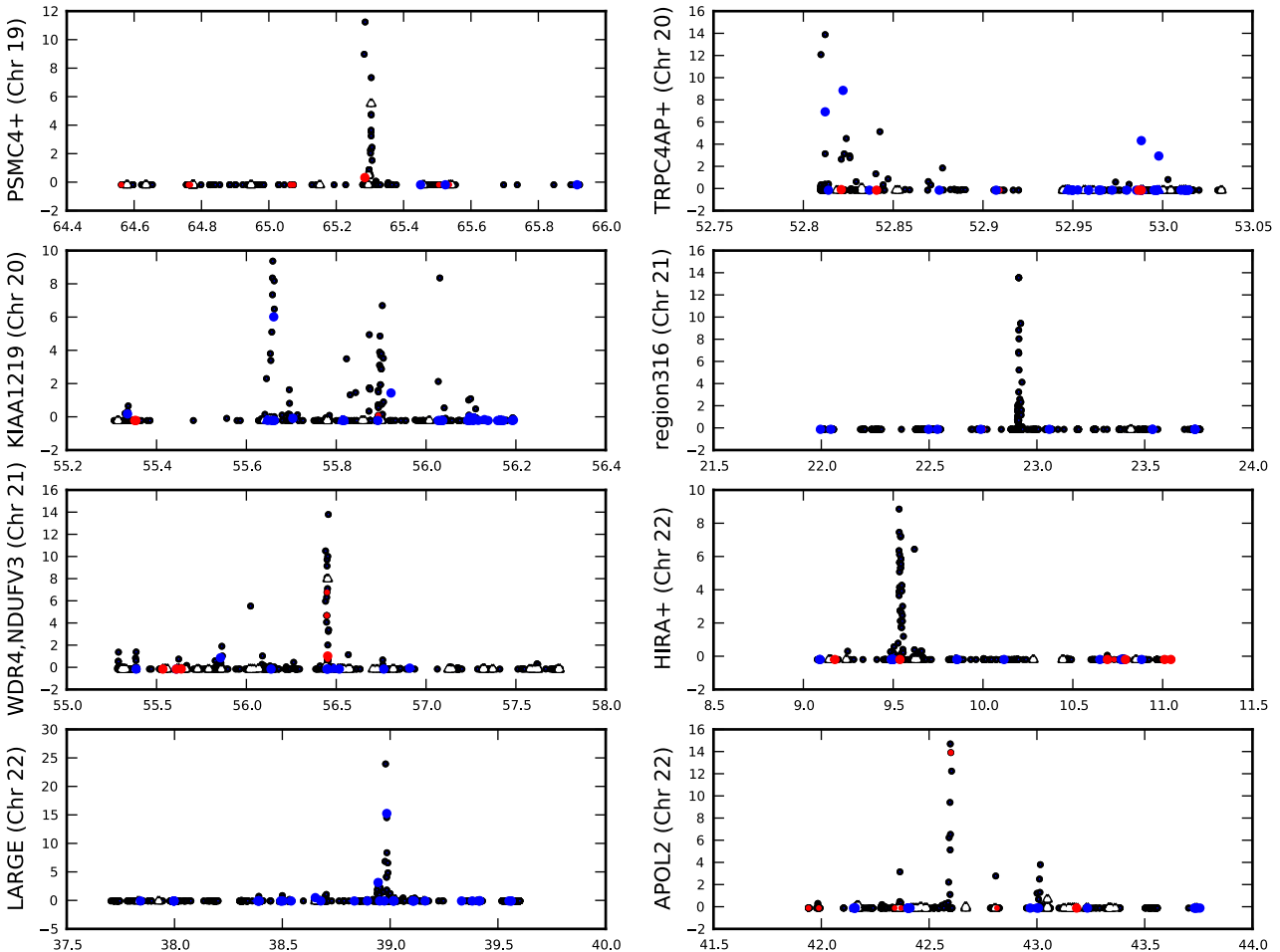
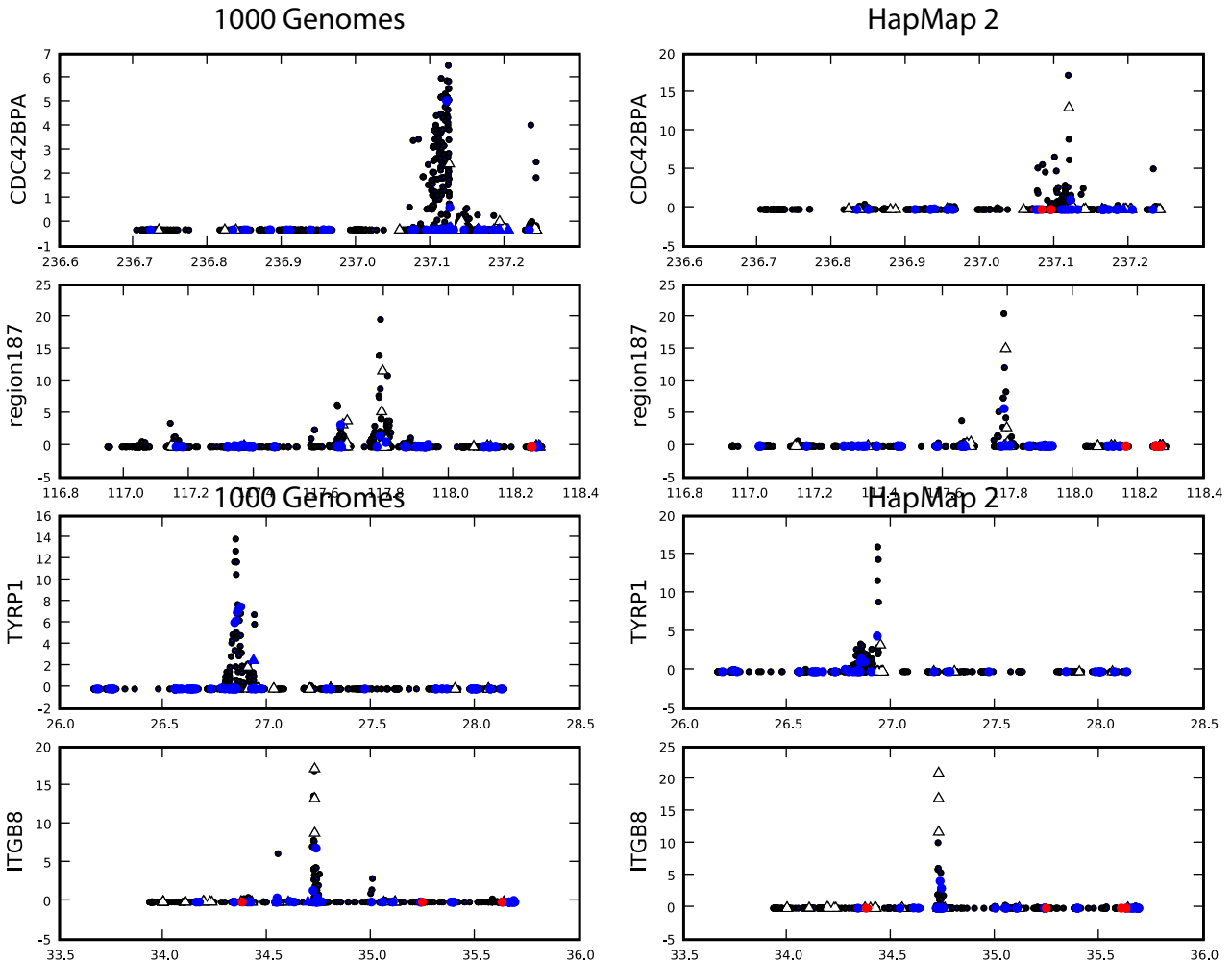
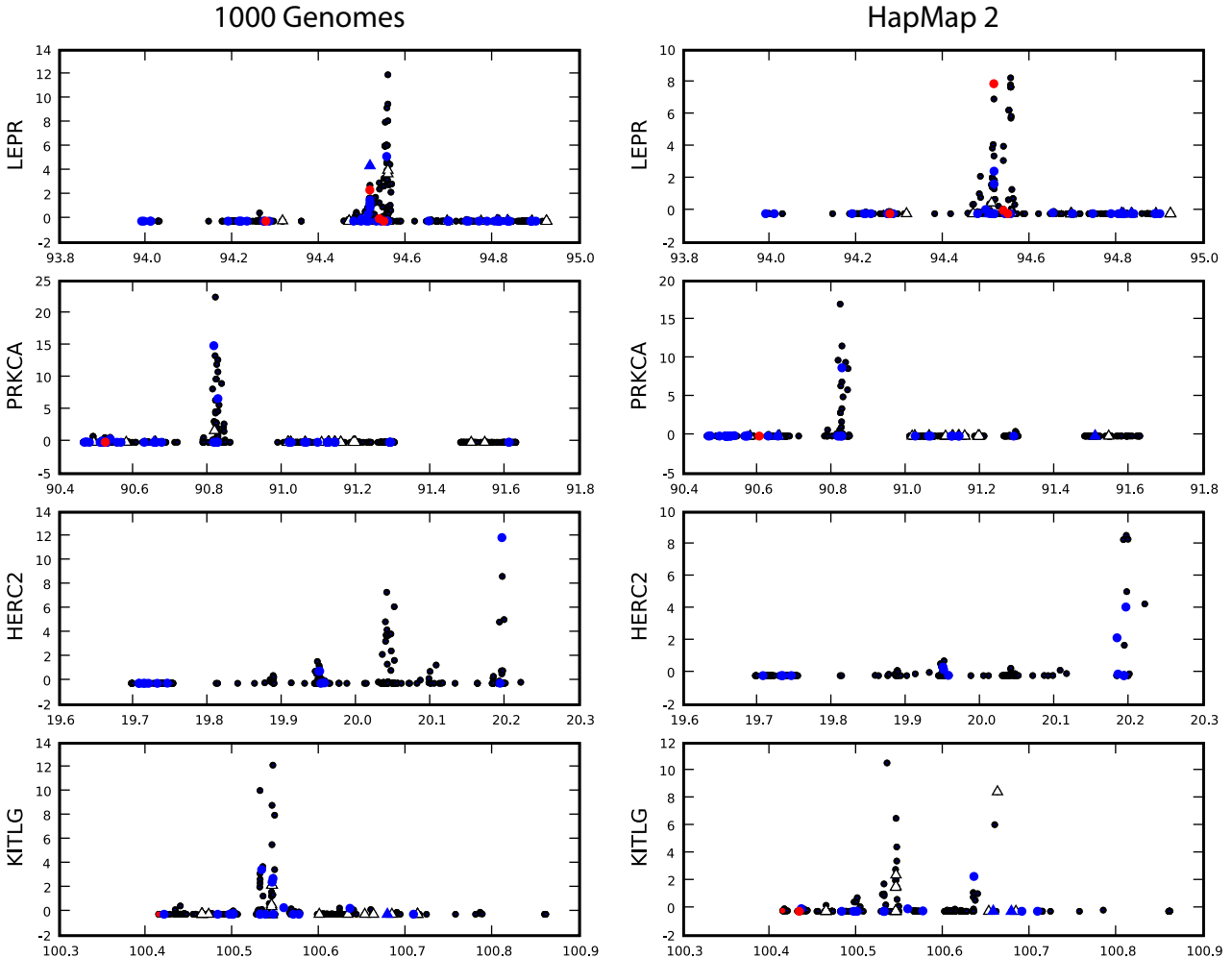
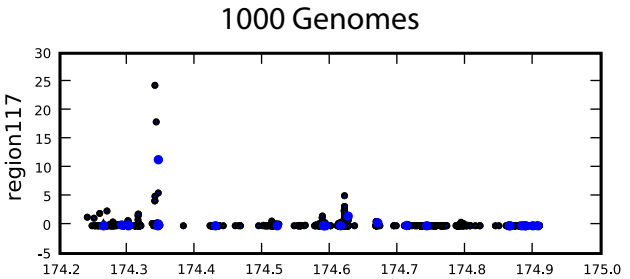

PSMC4+ (Chr 19)
TRPC4AP+ (Chr 20)
KIAA1219 (Chr 20)
region316 (Chr 21)
WDR4,NDUFV3 (Chr 21)
HIRA+ (Chr 22)
LARGE (Chr 22)
APOL2 (Chr 22)

Fig. S6. Applying CMS to selected regions in preliminary 1000 genomes data corroborates findings in HapMap data. For 17 regions shown, we show CMS scores of all variants in the region in 1000 Genomes data (left) and HapMap 2 data (right), distinguishing non-synonymous SNPs (red), variants in conserved regions (blue), and variants in experimentally determined transcription factor binding sites (white triangle).


1000 Genomes
HapMap 2
CDC42BPA
region187
TYRP1
ITGB8


1000 Genomes
HapMap 2
LEPR
PRKCA
HERC2
KITLG


1000 Genomes
region117
30
25
20
15
10
5
0
-5
174.2
174.3
174.4
174.5
174.6
174.7
174.8
174.9
175.0

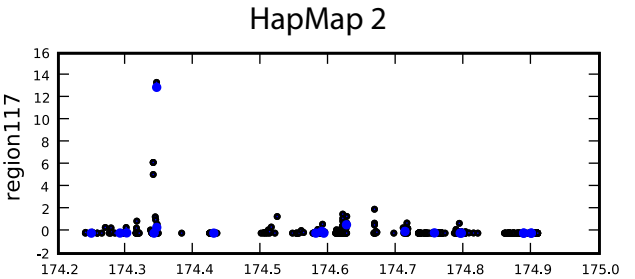

HapMap 2
region117
16
14
12
10
8
6
4
2
0
-2
174.2
174.3
174.4
174.5
174.6
174.7
174.8
174.9
175.0

Fig. S7. High scoring coding change in PCDH15 is validated in 1000 genomes data and highly conserved position across species. (A) CMS analysis of HapMap data for a selective signal around the gene PCDH15 scores an aspartic acid to alanine mutation (D435A) among the top variants; the same analysis carried out using full sequence data from the 1000 Genomes Project scores the D435A mutation highest. (B) The residue lies within a series of cadherin domains in the protein (red) and changes the amino acid from a highly acidic aspartic acid to neutral, non-polar alanine in roughly 90% of East Asians. It is completely conserved as an aspartic acid among all 34 sequenced mammalsm and conserved as an acidic residue throughout the vertebrate lineage. Darker shading indicates an amino acid change. Single base changes are shown as asterixes and insertions/deletions as red numbers, indicating indel length.

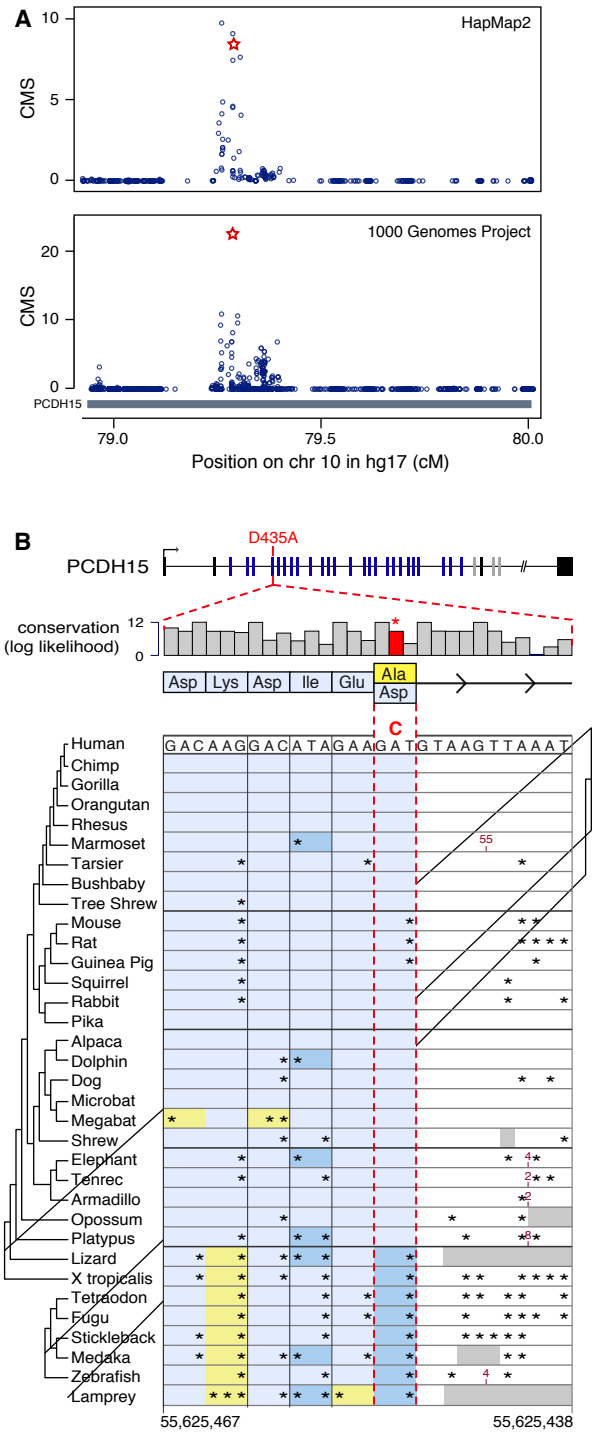


Fig. S8. Global distribution of PCDH15 D435A. Genotyping in the CEPH (Centre d'Etudie du Polymorphisme Human) global diversity panel (25) shows D435A is common in East Asia and present in Western Asia and the Americas, but rare in Europe and Africa.

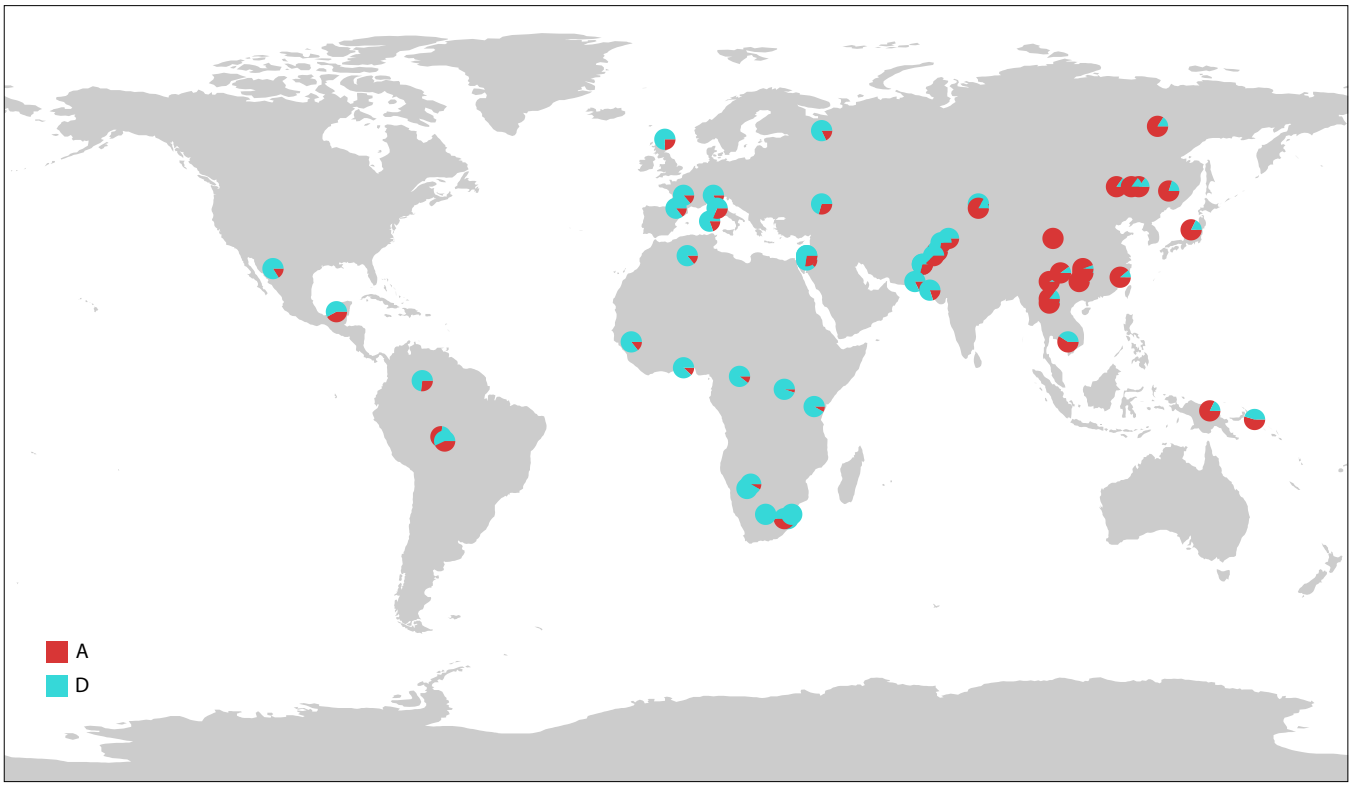


Fig. S9. SNPs in USF1 and PAWR, under selection in West Africa, are associated with changes in gene expression. A. At top, CMS scoring of SNP genotypes for the HapMapII populations (blue circles) localizes a 1Mb selected region on chromosome 12, containing three genes, to a 57kb region (grey, dashed line) in the gene PAWR, a transcriptional repressor. The SNPs in this 57kb region are significantly associated with PAWR gene expression intensity, also measured in the HapMapII individuals (41, 42) (middle, red circles) but not to Syt1 expression, a neighboring gene which encodes a synaptic vesicle protein (bottom, pink circles) - suggesting that variants altering the expression level of PAWR are positively selected in the Yoruba of West Africa. SNPs were considered significant at $p < 10\text{-}4$, consistent with Kudaravalli et. al. (24). (B) CMS scoring localizes a 766kb region on chromosome 12, containing 18 genes, to a 22kb region (grey, dashed line) containing 2 genes, the transcription factor and hyperlipidemia gene USF1, and the GTPase activator ARHGAP30. SNPs in this region are strongly associated with USF1 gene expression intensity (middle, red circles). Although probes for ARHGAP30 were not included in the expression data-set, a nearby gene, PVRL1, related to the poliovirus receptor gene, shows no pattern of association (bottom, pink circles), suggesting that variants in or near USF1 that alter its expression level are positively selected in the Yoruba.

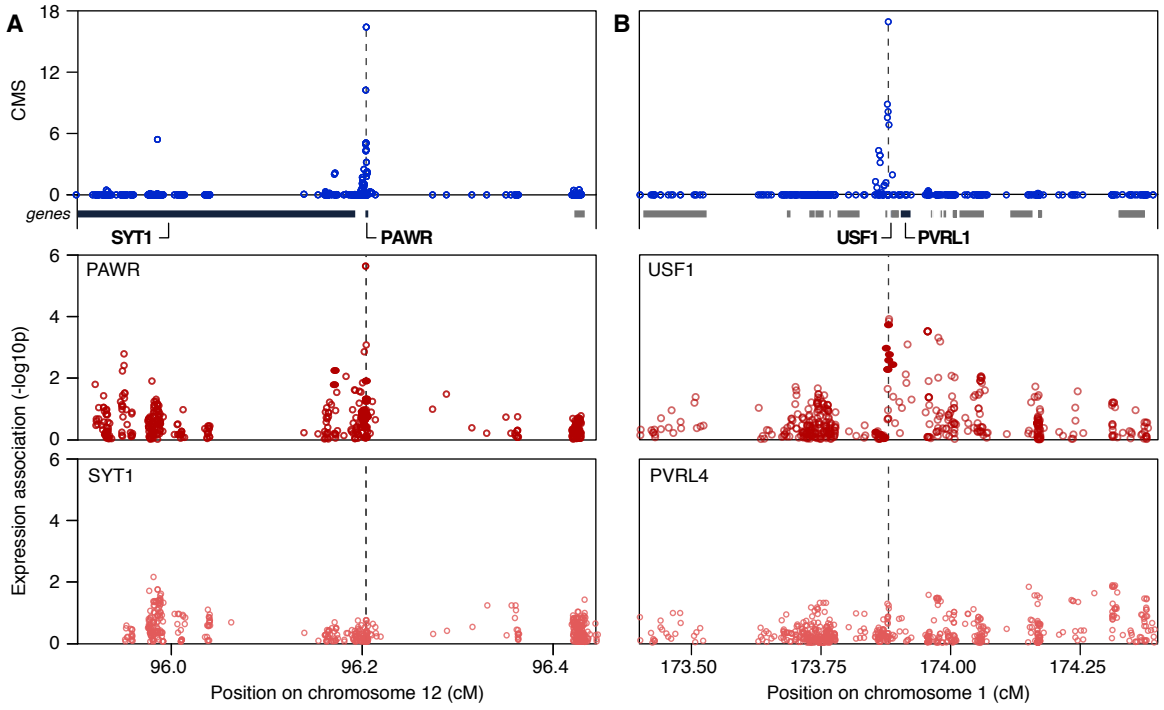
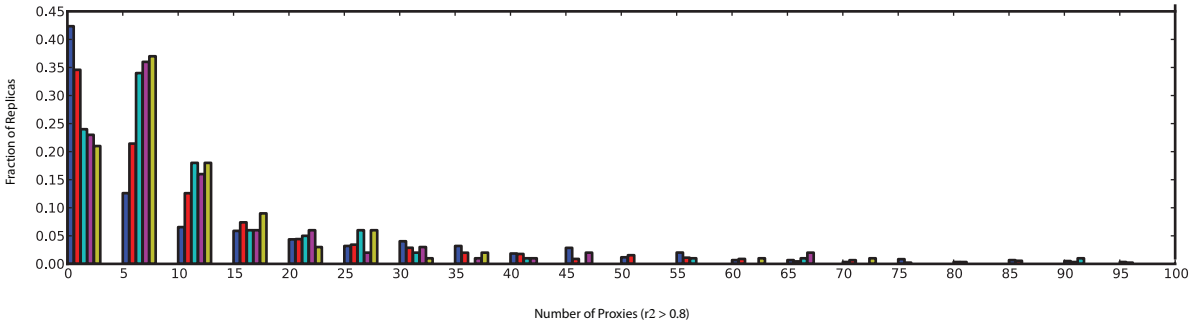
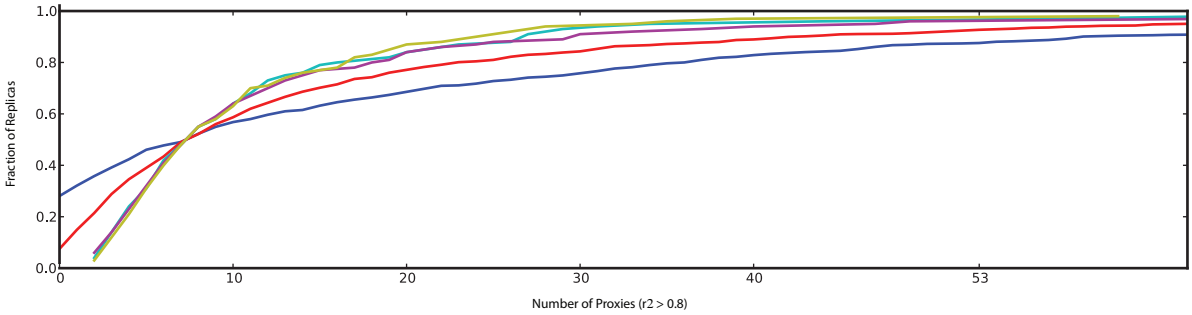


Fig. S10. Near-perfect proxies for selected and neutral SNPs. Normalized (A) and cumulative (B) histograms of the number of SNPs with r2 > 0.8 for variants show that variants in both selected and neutral region have ~20 perfect proxies in our simulations. (C) Histogram of the distance of the proxies from each other for selected variants and variants in regions without selection. In neutral regions, the perfect proxies are all within 0.02 cM of each other, whereas in selected regions, perfect proxies extend much farther, up to 1 cM. High-frequency (50-100%) selected SNPs are shown in red, low-frequency (0-50%) selected variants in blue, SNPs in neutral regions in the European model in cyan, in the East Asian model in magenta, and in the West African in yellow.

A



B



C

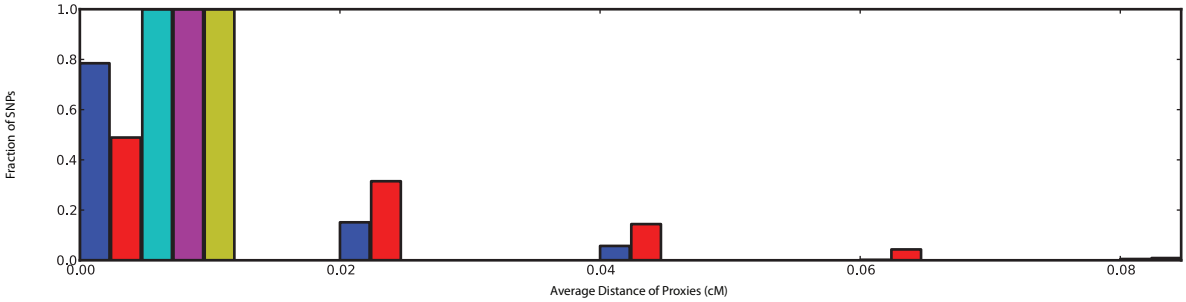


Table S1. Simulation parameters for demographic models analyzed

Scenarios for Calibrated Model fo European, East Asia, and West African Populations

Demographic model: (time t measured in generations ago)

gene_conversion_rate= 0.0000000045
mutRate = 0.000000015
recombRate, see SFS et al. 2005
pop_size = European: 7700, African: 24000, Asian: 770
sample_size=European:120, African: 120, Asian: 120

migration_rate (probability per chromosome per generation):
European<->African: 0.000032 at t=1505
African<->Asian: 0.000008 at t=1505

Population split (out-of-Africa): t=3500
Bottleneck: out of Africa, in African, t=3500, inbreed_coeff=.0
Pop split: European and Asian split, t=2000
Pop expansion: in African from 12500 to 24000 at t=17000:

Replicas per scenario: 100

In addition to Neutral simulations for the calibrated model we did the following selected scenarios:

mutAge(ky)	mutPop	mutFreq	selCoeff
5	European	0.2	0.033
5	European	0.4	0.0369
5	European	0.6	0.0402
5	European	0.8	0.0441
5	European	1	0.0771
5	EastAsian	0.2	0.033
5	EastAsian	0.4	0.0369
5	EastAsian	0.6	0.0402
5	EastAsian	0.8	0.0441
5	EastAsian	1	0.0771
5	WestAfrican	0.2	0.0376
5	WestAfrican	0.4	0.0415
5	WestAfrican	0.6	0.0447
5	WestAfrican	0.8	0.0487
5	WestAfrican	1	0.0862
10	European	0.2	0.0165
10	European	0.4	0.0185
10	European	0.6	0.0201
10	European	0.8	0.0221
10	European	1	0.0386
10	EastAsian	0.2	0.0165
10	EastAsian	0.4	0.0185
10	EastAsian	0.6	0.0201

Table S1. Simulation parameters for demographic models analyzed

10	EastAsian	0.8	0.0221
10	EastAsian	1	0.0386
10	WestAfrican	0.2	0.0188
10	WestAfrican	0.4	0.0207
10	WestAfrican	0.6	0.0224
10	WestAfrican	0.8	0.0243
10	WestAfrican	1	0.0431
15	European	0.2	0.011
15	European	0.4	0.0123
15	European	0.6	0.0134
15	European	0.8	0.0147
15	European	1	0.0257
15	EastAsian	0.2	0.011
15	EastAsian	0.4	0.0123
15	EastAsian	0.6	0.0134
15	EastAsian	0.8	0.0147
15	EastAsian	1	0.0257
15	WestAfrican	0.2	0.0125
15	WestAfrican	0.4	0.0138
15	WestAfrican	0.6	0.0149
15	WestAfrican	0.8	0.0162
15	WestAfrican	1	0.0287
20	European	0.2	0.0083
20	European	0.4	0.0092
20	European	0.6	0.01
20	European	0.8	0.011
20	European	1	0.0193
20	EastAsian	0.2	0.0083
20	EastAsian	0.4	0.0092
20	EastAsian	0.6	0.01
20	EastAsian	0.8	0.011
20	EastAsian	1	0.0193
20	WestAfrican	0.2	0.0094
20	WestAfrican	0.4	0.0104
20	WestAfrican	0.6	0.0112
20	WestAfrican	0.8	0.0122
20	WestAfrican	1	0.0216
25	European	0.2	0.0066
25	European	0.4	0.0074
25	European	0.6	0.008
25	European	0.8	0.0088
25	European	1	0.0154
25	EastAsian	0.2	0.0066
25	EastAsian	0.4	0.0074
25	EastAsian	0.6	0.008
25	EastAsian	0.8	0.0088
25	EastAsian	1	0.0154
25	WestAfrican	0.2	0.0075
25	WestAfrican	0.4	0.0083

Table S1. Simulation parameters for demographic models analyzed

25	WestAfrican	0.6	0.0089
25	WestAfrican	0.8	0.0097
25	WestAfrican	1	0.0172
29	European	0.2	0.0057
29	European	0.4	0.0064
29	European	0.6	0.0069
29	European	0.8	0.0076
29	European	1	0.0133
29	EastAsian	0.2	0.0057
29	EastAsian	0.4	0.0064
29	EastAsian	0.6	0.0069
29	EastAsian	0.8	0.0076
29	EastAsian	1	0.0133
29	WestAfrican	0.2	0.0065
29	WestAfrican	0.4	0.0072
29	WestAfrican	0.6	0.0077
29	WestAfrican	0.8	0.0084
29	WestAfrican	1	0.0149

Scenarios for Different Demographic Models

Demographic model: (time t measured in generations ago)

gene_conversion_rate= 0.0000000045
mutRate = 0.000000015
recombRate, see SFS et al. 2005
pop_size = European: 10000, African: 10000, Asian: 10000
sample_size=European:120, African: 120, Asian: 120

Population split (out-of-Africa): t=3500
Bottleneck: out of Africa, in African, t=3500, inbreed_coeff=.085
Pop split: European and Asian split, t=2000
Pop expansion: in African from 12500 to 24000 at t=17000:

Replicas per scenario: 100

In addition to Neutral simulations for the different demographic models we did the following selected scenarios:

mutAge(ky)	mutPop	mutFreq	selCoeff	inbreedCoeff for European bottleneck
10	European	0.2	0.017	0
10	European	0.4	0.019	0
10	European	0.6	0.0206	0
10	European	0.8	0.0226	0
10	European	1	0.0396	0
10	EastAsian	0.2	0.017	0

Table S1. Simulation parameters for demographic models analyzed

10	EastAsian	0.4	0.019	0
10	EastAsian	0.6	0.0206	0
10	EastAsian	0.8	0.0226	0
10	EastAsian	1	0.0396	0
10	WestAfrican	0.2	0.017	0
10	WestAfrican	0.4	0.019	0
10	WestAfrican	0.6	0.0206	0
10	WestAfrican	0.8	0.0226	0
10	WestAfrican	1	0.0396	0
10	European	0.2	0.017	0.2
10	European	0.4	0.019	0.2
10	European	0.6	0.0206	0.2
10	European	0.8	0.0226	0.2
10	European	1	0.0396	0.2
10	EastAsian	0.2	0.017	0.2
10	EastAsian	0.4	0.019	0.2
10	EastAsian	0.6	0.0206	0.2
10	EastAsian	0.8	0.0226	0.2
10	EastAsian	1	0.0396	0.2
10	WestAfrican	0.2	0.017	0.2
10	WestAfrican	0.4	0.019	0.2
10	WestAfrican	0.6	0.0206	0.2
10	WestAfrican	0.8	0.0226	0.2
10	WestAfrican	1	0.0396	0.2
10	European	0.2	0.017	0.3
10	European	0.4	0.019	0.3
10	European	0.6	0.0206	0.3
10	European	0.8	0.0226	0.3
10	European	1	0.0396	0.3
10	EastAsian	0.2	0.017	0.3
10	EastAsian	0.4	0.019	0.3
10	EastAsian	0.6	0.0206	0.3
10	EastAsian	0.8	0.0226	0.3
10	EastAsian	1	0.0396	0.3
10	WestAfrican	0.2	0.017	0.3
10	WestAfrican	0.4	0.019	0.3
10	WestAfrican	0.6	0.0206	0.3
10	WestAfrican	0.8	0.0226	0.3
10	WestAfrican	1	0.0396	0.3

Table S2. Performance of CMS to localize region and distinguish causal variant

	stat	*all*	*low frequency 20%-50%*	*high frequency 50%-100%*
CMS	Size of localized region	89.6kb	118.4kb	67.1kb
	Number of SNPs in 90% confidence set	109.98	146.31	90.77
	Number of SNPs in 50% confidence set	18.41	23.77	15.55
	% of causal SNPs that are the top SNP out of all SNPs	20.6%	2.6%	30.1%
	% of causal SNPs that are in the top 10 SNPs out of all SNPs	40.0%	20.8%	50.2%
	% of causal SNPs that are in the top 100 SNPs out of all SNPs	82.0%	87.5%	71.5%

Table S3. Number of significant non-causal SNPs per region at a given 90% or 50% power to detect the causal SNP for individual tests and CMS

All Frequency Sweeps

	Top SNP	In Top 50
XPEHH	1460.34	386.98
F_{ST}	447.05	119.84
iHS	407.18	140.28
iHH	1241.83	242.78
DAF	407.56	127.27
CMS	**0.24**	**18.41**

High Frequency Sweeps (>0.5)

	90%	50%
XPEHH	1463.94	664.18
F_{ST}	487.99	174.22
iHS	431.22	170.33
iHH	1100.51	233.22
DAF	431.14	182.65
CMS	**90.77**	**15.55**

Low Frequency Sweeps (<.05)

	90%	50%
XPEHH	1456.16	65.47
F_{ST}	399.55	56.77
iHS	371.28	106.39
iHH	1339.07	245.97
DAF	380.66	63.28
CMS	**146.31**	**23.77**

Table S4. False discovery rate for all neutral regions and neutral loci that have haplotype scores in the most tail 5% of the null distribution

		Fraction of SNPs above Threshold	1 Significant SNP	2-6 significant SNPs	>6 significant SNPs
Low Threshold	All Neutral Regions	1.10E-04	3.33%	3.44%	0.89%
	Outlier Neutral Regions	2.42E-04	3.03%	5.30%	0.76%
High Threshold	All Neutral Regions	3.64E-07	0.11%	<0.1%	<0.1%
	Outlier Neutral Regions	3.64E-07	0.11%	<0.1%	<0.1%

Low threshold: 90% power for high frequency causal alleles (>50%)
40% power for low frequency causal alleles (<50%)

High threshold: 65% power for high frequency causal alleles (>50%)
17% power for low frequency causal alleles (<50%)

Table S5. Selected regions identified by CMS test in HapMapII.

Chr	Start (hg18)	End (hg18)	Size	Peak SNP	Peak Score	p-value	Old Start	Old End	Pop	Genes in Region
1	26789601	26802164	12563	26798430	6.24	1.50E-03	27139629	27652238	YRI	
1	30459963	30489581	29618	30489581	18.25	4.02E-05	30154160	30762854	CEU	
1	35246775	35364963	118188	35267397	19.67	2.54E-05	34824768	35781247	CEU	ZMYM6, ZMYM1
1	65775602	65912101	136499	65873972	12.56	7.84E-04	65499104	66321085	JPT+CHB	LEPR
1	76271901	76302897	30996	76277500	19.76	4.71E-05	75813598	76717927	JPT+CHB	
1	83031668	83060499	28831	83035522	11.58	1.18E-03	82913557	83057121	JPT+CHB	
1	91016931	91025451	8520	91024714	6.12	1.07E-02	90820011	90829408	CEU	
1	93991061	94104739	113678	94104739	9.34	3.22E-03	93663879	94310148	JPT+CHB	
1	106579382	106671147	91765	106640062	5.78	1.99E-03	106005990	106943274	YRI	
1	159281070	159303096	22026	159295759	11.16	8.28E-05	159063726	159911579	YRI	USF1, ARHGAP30
1	167443021	167633980	190959	167614169	18.61	7.31E-05	167209453	167966779	JPT+CHB	NME7, BLZF1, C1orf114
1	183952167	184133467	181300	184031442	4.98	1.84E-02	183628536	184619838	CEU	HMCN1
1	188052112	188096282	44170	188089763	11.82	6.81E-04	188030378	188257912	CEU	
1	192278226	192348839	70613	192305203	13.77	2.56E-04	192123622	192849020	CEU	
1	195099548	195412005	312457	195099548	9.15	2.73E-04	194624369	195513591	YRI	CFHR4, CFHR2, CFHR5, F13B, ASPM, ZBTB41
1	199722533	199726875	4342	199725145	11.22	9.10E-04	199209008	199826786	CEU	CSRP1
1	225242425	225488493	246068	225471703	7.56	6.82E-03	224887929	225663992	JPT+CHB	CDC42BPA
1	234861564	234904018	42454	234904018	12.21	5.58E-04	234813780	235227559	CEU	
1	246235220	246384632	149412	246346107	13.29	3.29E-04	246126046	247130153	CEU	OR2L13, OR2L2, OR2L3, OR2M1P, OR2M5
2	7957879	7973932	16053	7957879	11.34	8.66E-04	7479337	8404110	CEU	
2	9635559	9660585	25026	9651729	16.98	1.40E-04	9066229	9753029	JPT+CHB	YWHAQ
2	24566434	24899859	333425	24896129	11.58	6.61E-05	24199211	24983596	YRI	NCOA1, C2orf79, CENPO, ADCY3
2	73492331	73760106	267775	73593933	10.92	1.05E-03	73464789	74175977	CEU	ALMS1, NAT8, ALMS1P
2	108448484	108880033	431549	108448484	25.68	2.88E-05	108382462	109014473	JPT+CHB	GCC2, LIMS1, RANBP2, CCDC138, EDAR
2	121860035	121958191	98156	121891975	12.54	4.78E-04	121116746	122093327	CEU	CLASP1
2	135808531	136137160	328629	136123949	36.39	0.00E+00	135640937	136633825	CEU	ZRANB3, R3HDM1
2	136039146	136325116	285970	136123949	36.39	8.38E-06	135826599	136809279	CEU	R3HDM1, UBXN4, LCT, MCM6
2	146436279	146743324	307045	146660748	9.18	3.47E-03	146436279	147347656	JPT+CHB	
2	153784770	153797429	12659	153792538	15.65	2.18E-04	152861979	153797429	JPT+CHB	
2	157858690	157870279	11589	157868082	14.71	1.72E-04	157409064	158362537	CEU	GALNT5
2	159330454	159347591	17137	159330454	12.10	9.55E-04	158582334	159435451	JPT+CHB	
2	177341050	177382985	41935	177355234	22.28	2.18E-05	177092418	177387738	JPT+CHB	
2	182203716	182244638	40922	182212528	6.76	1.08E-03	181823305	182777089	YRI	CERKL
2	186238080	186463172	225092	186428579	2.16	1.87E-02	186142658	187124844	YRI	
2	205913806	205934985	21179	205934985	16.46	8.09E-05	205679048	206161103	CEU	PARD3B
2	238018515	238066523	48008	238034021	9.96	2.48E-03	238011106	238174962	JPT+CHB	MLPH
3	17883552	17960264	76712	17960264	15.63	2.21E-04	17297436	17960264	JPT+CHB	
3	26005011	26048418	43407	26005011	16.76	1.50E-04	25769776	26681386	JPT+CHB	
3	27408401	27487899	79498	27487899	11.79	1.08E-03	26743389	27524192	JPT+CHB	SLC4A7
3	36169171	36274027	104856	36197653	10.18	1.47E-03	36113231	36622465	CEU	
3	49186993	49532861	345868	49336795	17.32	1.21E-04	49063116	50014478	JPT+CHB	
3	56770059	56801805	31746	56781243	7.71	6.13E-04	56204345	57006701	YRI	ARHGEF3
3	86711778	86743626	31848	86738189	14.78	9.76E-06	86598463	87304060	YRI	
3	87256871	87451172	194301	87266302	7.79	5.89E-04	86865272	87804944	YRI	CHMP2B, POU1F1
3	88509607	88672957	163350	88509607	9.35	2.18E-03	87971626	88679694	CEU	
3	108881301	108962444	81143	108930547	34.29	2.79E-07	108418493	109388029	JPT+CHB	BBX
3	136053970	136073444	19474	136055178	7.50	5.42E-03	135982109	136369580	CEU	EPHB1
3	140550580	140697160	146580	140662599	18.87	6.64E-05	140523205	141092976	JPT+CHB	MRPS22, COPB2, RBP2
3	146766081	146805632	39551	146768691	8.95	3.08E-04	146451580	147096721	YRI	
3	189732111	189751267	19156	189732111	15.01	2.88E-04	189728951	190165670	JPT+CHB	LPP
4	34057924	34094101	36177	34066869	8.98	3.00E-04	33891511	34442958	YRI,JPT+CHB	
4	38885011	39028263	143252	39010782	8.88	3.93E-03	38479458	39224328	JPT+CHB	WDR19, RFC1
4	41621468	41676123	54655	41651170	10.90	1.06E-03	40705728	41689402	CEU,JPT+C	TMEM33
4	41621468	41790757	169289	41769390	13.70	4.73E-04	40863907	41815266	JPT+CHB	TMEM33, WDR21B, SLC30A9
4	41807491	41849931	42440	41815266	19.62	5.02E-05	41344700	42251414	JPT+CHB	BEND4
4	45927013	46007334	80321	45941857	5.37	2.53E-03	45811397	46688413	YRI	GABRA2
4	85579348	85620193	40845	85579348	10.34	1.38E-03	85547286	85910744	CEU	

Table S5. Selected regions identified by CMS test in HapMapII.

4	106768545	106882395	113850	106848458	9.40	2.12E-03	106516945	107041027	CEU	FLJ20184, INTS12, GSTCD
4	123544376	123777780	233404	123544376	7.19	8.32E-04	122963243	123920063	YRI	ADAD1, IL2, IL21
4	132755974	132859693	103719	132766493	7.13	6.48E-03	132755974	132983931	CEU	
4	149692178	149747508	55330	149702941	11.58	1.18E-03	149243863	150087213	JPT+CHB	
4	153011996	153059485	47489	153054770	12.71	7.53E-04	152787360	153600251	JPT+CHB	
4	158862019	158921890	59871	158904521	19.40	5.52E-05	158322304	159236204	JPT+CHB	
4	163874331	163934336	60005	163893227	14.80	3.12E-04	163375223	164137266	JPT+CHB	
4	171411665	171441966	30301	171441539	10.46	1.97E-03	171271317	172184617	JPT+CHB	
4	176513915	176543673	29758	176531263	15.02	1.49E-04	176030860	176985929	CEU	
5	11738818	11778721	39903	11738818	14.69	1.75E-04	11475555	12391990	CEU	CTNND2
5	21669440	21733182	63742	21723727	13.76	2.57E-04	21641250	22059926	CEU	
5	33986873	33999967	13094	33987450	14.43	3.41E-04	33715552	34077096	CEU	MATP
5	43470168	43523631	53463	43488009	7.59	6.53E-04	43465721	44404007	YRI	C5orf28, C5orf34
5	64728681	64789915	61234	64751417	10.36	2.06E-03	64410884	65304426	JPT+CHB	ADAMTS6
5	109840968	110015127	174159	110015127	15.67	1.13E-04	109736860	110709762	CEU	FLJ43080
5	112134957	112172343	37386	112146237	12.15	9.30E-04	111900911	112645236	CEU,JPT+C	APC
5	117781195	117824923	43728	117823223	29.89	1.14E-05	117123231	117938519	JPT+CHB	
5	128157339	128227028	69689	128215792	16.29	1.73E-04	127870188	128466278	CEU,JPT+CHB	
5	170069398	170099591	30193	170069398	11.23	1.38E-03	169925098	170423604	JPT+CHB	KCNIP1
6	33164766	33191175	26409	33182496	15.07	8.09E-06	33130040	34120889	YRI	HLA-DPB2
6	63643099	63716242	73143	63668752	6.78	1.06E-03	63102720	64035369	YRI	
6	70456980	70539581	82601	70474847	9.14	2.74E-04	69828303	70684123	YRI	LMBRD1
6	77971029	77990756	19727	77990380	3.51	8.13E-03	77966036	78470746	YRI	
6	84865763	84918371	52608	84904563	13.05	3.68E-04	84492929	85257403	CEU	KIAA1009
6	130590670	130649133	58463	130591278	11.16	8.31E-05	130126363	131103166	YRI	
6	132988064	133059399	71335	132988064	8.86	3.18E-04	132584498	133542601	YRI	TAAR1, VNN1
6	145131358	145186074	54716	145174593	16.28	8.81E-05	144874527	145613986	CEU	UTRN
7	20350404	20359312	8908	20358755	10.65	1.14E-04	19926545	20751063	YRI	ITGB8
7	49899465	50091815	192350	50091815	12.96	6.80E-04	49681331	50574974	JPT+CHB	VWC2,ZPBP
7	102858219	102872149	13930	102858219	12.81	4.17E-04	102411428	103308119	CEU	SLC26A5
7	104621529	104709860	88331	104684165	10.16	1.49E-03	104454570	105109259	CEU	SRPK2
7	105878575	105985048	106473	105915816	12.50	8.06E-04	105741221	106043520	JPT+CHB	
7	111949105	112076246	127141	112052143	19.50	5.22E-05	111538084	112222399	JPT+CHB	
7	123832302	123865845	33543	123865845	7.34	5.86E-03	123815097	124660474	CEU	
7	147390394	147441955	51561	147410581	10.07	1.55E-03	146782453	147589814	CEU	CNTNAP2
7	150821867	150831070	9203	150821867	4.41	4.61E-03	150550640	151052551	YRI	RHEB
8	49218318	49265300	46982	49224721	7.18	8.36E-04	48848652	49513940	YRI	
8	111936968	112004972	68004	111985213	9.84	1.72E-03	111661201	112414608	CEU	
8	117711526	117804773	93247	117738383	15.06	1.44E-04	117224170	117804773	CEU	EIF3H
8	135150044	135223753	73709	135150044	12.38	8.49E-04	134758391	135422815	JPT+CHB	
8	142006723	142046179	39456	142023392	6.96	7.05E-03	141733154	142402607	CEU	PTK2
9	11738986	11834287	95301	11834287	9.77	1.89E-04	11314415	12292928	YRI	
9	12739596	12753450	13854	12740812	13.70	2.68E-04	12189215	13161437	CEU	
9	12739596	12753450	13854	12740812	13.70	4.73E-04	12189215	13161437	CEU	
9	16782200	16792118	9918	16792118	18.41	3.85E-05	16200330	17177853	CEU	BNC2
9	17367629	17425929	58300	17367629	7.62	5.10E-03	17286630	18236545	CEU	CNTLN
9	24858290	24904235	45945	24902363	8.17	4.63E-04	24041640	24925087	YRI	
9	90005303	90042056	36753	90017614	13.58	2.84E-04	89654521	90126027	CEU	
9	105777116	105894414	117298	105789611	11.13	1.46E-03	105404938	106389461	JPT+CHB	
9	107672621	107770610	97989	107672621	5.73	2.05E-03	107541256	108241335	YRI	
9	110323834	110349122	25288	110328982	16.07	1.87E-04	109884763	110859876	JPT+CHB	
9	123796085	123894371	98286	123839737	12.41	8.33E-04	123796085	124275969	JPT+CHB	TTLL11
9	134186267	134193052	6785	134186267	15.85	1.05E-04	134125804	134768688	CEU	SETX
10	3021204	3048375	27171	3038498	15.86	1.04E-04	2531943	3478513	CEU	
10	3021204	3048375	27171	3038498	15.86	1.84E-04	2539157	3198583	CEU	
10	4391589	4408953	17364	4391589	10.87	1.64E-03	3646199	4409830	JPT+CHB	
10	11018493	11109544	91051	11042809	10.23	1.45E-04	10747188	11442053	YRI	LOC254312, CUGBP2
10	23133288	23151926	18638	23133288	6.94	7.12E-03	22817881	23176100	CEU	
10	23133288	23204056	70768	23133288	11.58	1.03E-04	22462638	22966999	YRI	
10	55558565	55641433	82868	55561007	21.56	2.76E-05	55258501	56224535	JPT+CHB	PCDH15
10	75080058	75269133	189075	75149131	11.36	8.57E-04	75079883	75738823	CEU	SYNPO2L, AGAP5, BMS1P4, SEC24C, FUT11, CHCHD1, KIAA0913, NDST2, CAMK2G
10	83622183	83678139	55956	83675782	17.21	5.89E-05	83591444	84483347	CEU	NRG3
10	101965380	102017437	52057	102004791	6.27	1.47E-03	101854614	102770628	YRI	CHUK, CWF19L1, SNORA12
10	107309552	107333364	23812	107309552	14.06	4.21E-04	106871024	107379174	JPT+CHB	
10	109730823	109754205	23382	109753450	13.58	5.33E-04	109455143	110149085	JPT+CHB	
10	118138624	118163461	24837	118155869	18.69	3.54E-05	117766918	118635824	CEU	
11	38421494	38504753	83259	38427247	11.02	1.00E-03	38375883	38774275	CEU,YRI	
11	81157604	81220349	62745	81162715	11.81	1.07E-03	80901887	81879165	JPT+CHB	

Table S5. Selected regions identified by CMS test in HapMapII.

11	131429751	131489208	59457	131429751	16.95	1.41E-04	130969126	131616127	JPT+CHB	NTM
12	2971194	2981996	10802	2981996	11.40	8.43E-04	2483590	3310245	CEU	TEAD4
12	21885378	21895465	10087	21895465	6.78	1.07E-03	21478581	22349372	YRI	ABCC9
12	30282738	30369091	86353	30349886	8.66	3.55E-04	29989195	30942713	YRI	
12	42569052	42611328	42276	42611328	8.29	3.68E-03	42158820	43112740	CEU	TMEM117
12	46584059	46601157	17098	46599002	7.87	5.61E-04	46413304	47266706	YRI	VDR
12	75305346	75511542	206196	75449556	6.65	8.23E-03	75259508	75684469	CEU	OSBPL8
12	78508830	78565670	56840	78565670	16.92	3.91E-06	78032902	79011091	YRI	PAWR
12	78712972	78878339	165367	78840229	23.47	1.03E-05	78333387	79328607	CEU	PPP1R12A
12	87442874	87486142	43268	87442874	9.61	3.39E-03	87362094	87859455	CEU	KITLG
13	56505490	56860534	355044	56599239	13.66	1.76E-05	56242093	57189327	YRI	LOC729233, PRR20, LOC729240, LOC729246, LOC729250
13	62449023	62704071	255048	62593706	23.03	1.92E-05	62304229	63244537	JPT+CHB	
13	66994702	67084816	90114	66994702	6.30	1.45E-03	66781935	67750861	YRI	
13	87975718	88030309	54591	88021698	9.26	2.52E-04	87695036	88313165	YRI	
13	97336394	97364000	27606	97345312	14.36	2.00E-04	96993001	97706507	CEU	
14	27906468	28016233	109765	27906468	14.89	3.02E-04	27161173	28063623	JPT+CHB	
14	47974651	48027781	53130	48027781	10.47	1.26E-04	47627747	48384795	YRI	
14	69944448	70055629	111181	69953696	9.68	1.85E-03	69614648	70441657	CEU	SYNJ2BP, ADAM21
14	106210628	106257669	47041	106241455	3.29	9.29E-03	105468820	106247090	YRI	
15	26101581	26213429	111848	26186959	19.23	5.18E-05	25855959	26203777	CEU	HERC2
15	27037285	27097430	60145	27083162	19.97	2.32E-05	#VALUE!	#VALUE!	CEU	APBA2
15	29126965	29140680	13715	29135904	14.40	3.63E-04	29128753	29493205	JPT+CHB	TRPM1
15	46179457	46273218	93761	46179457	33.56	1.97E-06	45943472	46713294	CEU	SLC24A5, MYEF2, CTXN2
15	50315458	50340745	25287	50321636	20.19	2.23E-05	49805388	50731106	CEU	MYO5C
15	53357705	53500149	142444	53500149	9.54	2.15E-04	53168177	54084472	YRI	RAB27A, PIGB, CCPG1, DYX1C1
15	61782476	61948404	165928	61818807	16.42	3.20E-04	61520316	62223508	JPT+CHB	HERC1
15	62267941	62683351	415410	62424144	24.71	1.59E-05	61746382	62700937	JPT+CHB	CSNK1G1, KIAA0101, TRIP4, ZNF609
15	69932782	70253366	320584	69966232	12.33	5.27E-04	69466523	70425766	CEU	MYO9A, SENP8, GRAMD2
15	73457132	73505722	48590	73505722	8.12	4.82E-04	73107486	73831377	YRI	SIN3A
15	78781318	78815609	34291	78815609	8.78	4.09E-03	78252355	79130226	JPT+CHB	FAM108C1
16	22852656	22907132	54476	22873559	21.43	0.00E+00	22663833	23411542	YRI	
16	64293212	64324782	31570	64293212	27.18	1.34E-05	64177220	64852680	JPT+CHB	
16	64293212	64324782	31570	64293212	5.78	1.99E-03	64118269	64788109	YRI	
16	74131531	74218308	86777	74131531	17.35	2.29E-04	74107386	74605127	JPT+CHB	FLJ22167, FLJ22167, FLJ22167, GABARAPL2, ADAT1
17	18151009	18188005	36996	18165207	3.70	7.17E-03	18114705	18761904	YRI	TOP3A, SMCR8, SHMT1
17	56337706	56516428	178722	56460501	15.91	1.98E-04	56023823	56643678	JPT+CHB	BCAS3
17	57602509	57626706	24197	57620375	10.20	1.46E-03	57290505	57764547	CEU	
17	61777799	61854712	76913	61806679	13.57	2.87E-04	61531072	62253791	CEU	PRKCA
18	7528274	7617199	88925	7599137	19.11	3.07E-05	7205275	7912931	CEU	PTPRM
18	15072789	15098540	25751	15098540	7.18	6.33E-03	14571819	15308075	CEU,JPT+CHB	
18	39026284	39150615	124331	39145395	10.10	1.53E-03	38747498	39257766	CEU	SYT4
18	55805023	55813095	8072	55805023	7.59	5.18E-03	55737016	56267340	CEU	
18	68926250	68974704	48454	68926250	13.39	3.13E-04	68534238	69388250	CEU	
18	70932991	70944224	11233	70939361	8.84	2.82E-03	70361170	71103592	CEU	
19	43486735	43542599	55864	43532736	13.13	2.32E-05	42997553	43960418	YRI	YIF1B, C19orf33, KCNK6, C19orf15
19	45139097	45271682	132585	45141888	11.84	6.72E-04	44792292	45610005	CEU	PSMC4, ZNF546, ZNF780B, ZNF780A
20	33063405	33241273	177868	33090765	6.00	1.73E-03	33063405	34048166	YRI,JPT+CH	TRPC4AP, EDEM2,PROCR
20	36560765	36624040	63275	36589701	8.57	3.75E-04	36399507	37356861	YRI	KIAA1219
21	24579072	24633417	54345	24594331	12.96	2.68E-05	24345834	25073839	YRI	
21	43105249	43210983	105734	43210983	9.93	1.71E-04	42636795	43527220	YRI	WDR4, NDUFV3
22	17700844	17876423	175579	17718412	9.34	2.19E-03	17538116	18177745	CEU	HIRA, MRPL40, C22orf39, UFD1L, CDC45L
22	32521318	32541867	20549	32521318	10.97	9.48E-05	31946088	32919581	YRI	LARGE
22	34941252	34960895	19643	34948899	14.99	8.37E-06	34432424	35345275	YRI	APOL2

Table S7. p-Values for enrichment of GO categories among genes in regions identified by CMS.

Panther Category	CEU	ASN	YRI	Panther Subcategory
Biological Processes				
Signal transduction	6.56E−05	0.010	0.63	Calcium mediated signaling
	0.81	0.81	8.50E−05	Other signal transduction
	0.88	0.13	4.47E−07	Steroid hormone−mediated signaling
Sensory perception	0.01	0.59	0.57	Chemosensory perception
	0.01	0.61	0.59	Olfaction
	0.94	3.99E−05	0.93	Other sensory perception
Muscle contraction	6.76E−04	0.48	0.54	Muscle contraction
Protein metabolism and modification	8.52E−04	0.49	0.01	Protein metabolism and modification
	0.001	0.56	0.39	Protein modification
	0.002	0.41	0.29	Protein phosphorylation
	0.06	0.32	0.04	Proteolysis
Homeostasis	0.54	0.004	0.14	Homeostasis
	0.81	0.001	0.80	Other homeostasis activities
	0.87	0.009	0.86	Antioxidation and free radical removal
Immunity and defense	0.75	0.26	0.04	Blood clotting
	0.82	0.82	0.001	Complement−mediated immunity
	0.80	0.02	0.78	Detoxification
Nucleoside, nucleotide and nucleic acid metabolism	0.04	0.27	0.29	DNA replication
	0.56	0.003	0.23	mRNA transcription
	0.61	0.01	0.30	mRNA transcription regulation
	0.10	0.90	0.006	Other nucleoside, nucleotide and nucleic acid metabolism
	0.89	0.007	0.87	Pyrimidine metabolism
Amino acid metabolism	0.93	0.92	0.003	Other amino acid metabolism
Lipid, fatty acid and steroid metabolism	0.02	0.81	0.79	Cholesterol metabolism
	0.96	0.05	0.95	Lipid and fatty acid binding
	0.92	0.91	0.004	Regulation of lipid, fatty acid and steroid metabolism
Cell structure and motility	0.13	0.15	0.02	Cell structure and motility
Electron transport	0.83	0.82	0.02	Oxidative phosphorylation
Cell proliferation and differentiation	0.63	0.17	0.03	Cell proliferation and differentiation
Molecular Functions				
Kinase	0.90	0.90	6.65E−06	Carbohydrate kinase
	0.03	0.09	0.11	Kinase
	5.95E−04	0.09	0.68	Non−receptor serine/threonine protein kinase
	0.86	0.01	0.84	Nucleotide kinase
	0.87	0.86	1.99E−05	Other kinase
	0.008	0.28	0.42	Protein kinase
Defense / immunity protein	0.87	0.86	2.13E−05	Complement component
	0.31	0.30	0.04	Defense/immunity protein
Select calcium binding protein	0.85	6.24E−04	0.84	Other select calcium binding proteins
Receptor	0.79	0.78	0.03	Nuclear hormone receptor
	0.54	0.001	0.58	Other receptor
Transferase	0.24	0.75	0.004	Acetyltransferase
Oxidoreductase	0.90	0.005	0.89	Peroxidase
Membrane traffic protein	0.008	0.13	0.86	Vesicle coat protein
Nucleic acid binding	0.02	0.81	0.80	DNA helicase
	0.25	0.01	0.59	Nucleic acid binding
	0.05	0.95	0.95	Replication origin binding protein
Transcription factor	0.32	0.02	0.17	Transcription factor
Lyase	0.97	0.97	0.03	Adenylate cyclase
Select regulatory molecule	0.31	0.03	0.37	G−protein modulator
	0.52	0.03	0.17	Other G−protein modulator
Molecular function unclassified	0.51	0.04	0.13	Molecular function unclassified
Pathways				
	2.08E−06	0.81	0.80	Ionotropic glutamate receptor pathway
	0.98	0.98	1.78E−04	Serine glycine biosynthesis
	0.94	0.002	0.93	De novo pyrmidine ribonucleotides biosythesis
	0.94	0.002	0.93	De novo pyrimidine deoxyribonucleotide biosyn
	0.93	0.92	0.003	Vitamin D metabolism and pathway
	0.004	0.70	0.67	Inflammation mediated by chemokine and cyto
	0.89	0.006	0.88	De novo purine biosynthesis
	0.70	0.31	0.008	Cytoskeletal regulation by Rho GTPase
	0.85	0.16	0.02	Blood coagulation
	0.57	0.56	0.02	Interleukin signaling pathway
	0.03	0.74	0.73	Ubiquitin proteasome pathway