www.sciencemag.org/content/354/6317/1305/suppl/DC1

Supplementary Materials for

The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish

Noah M. Reid, Dina A. Proestou, Bryan W. Clark, Wesley C. Warren, John K. Colbourne, Joseph R. Shaw, Sibel I. Karchner, Mark E. Hahn, Diane Nacci, Marjorie F. Oleksiak, Douglas L. Crawford, Andrew Whitehead*

*Corresponding author. Email: awhitehead@ucdavis.edu

Published 9 December 2016, *Science* **354**, 1305 (2016)
DOI: 10.1126/science.aah4993

This PDF file includes

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/354/6317/1305/suppl/DC1)

Tables S1 to S4 as Excel files

Supplemental Materials and Methods

Fish Collection and Sample Preparation

Samples in this study were collected and prepared as described in (*15*). Briefly, 60-100 adult *Fundulus heteroclitus* were collected using baited minnow traps from eight estuarine sites spanning approximately 600 km of the Atlantic Coast of the USA between 2008 and 2011 (Table S1). These specific killifish populations had previously been characterized as either tolerant or sensitive to dioxin-like compounds (DLCs), based on early life stage sensitivity to PCB126 ((*4, 5, 26*); reviewed in (*15*)) (Table S1). Each DLC-tolerant population was paired with a nearby DLC-sensitive population. Upon return to the US EPA Atlantic Ecology Division (Narragansett, RI), fish were sacrificed and stored at either -20 or -80°C prior to DNA extraction. Genomic DNA was extracted from caudal fin tissue according to the QIAGEN DNeasy protocol for animal tissue (optional RNase treatment included), quantified with the PicoGreen dsDNA assay (Invitrogen), and diluted to a standard concentration of 20 ng/µl.

Population Genomics

Sequencing and Alignment

Genomes of 384 killifish (43 to 50 fish per population) were sequenced (Illumina PE-100). Sex ratios (% female) ranged from 41% to 59% within populations. Following extraction and quantification, genomic DNA was sheared to 500bp by sonication (Covaris E220). Sheared DNA was used to construct individually-indexed sequencing libraries using the NextFlex DNA sequencing kit (Bioo Scientific). Library insert sizes were determined by TapeStation (Agilent)

using DNA high sensitivity ScreenTape, and libraries were quantified by Quant-iT PicoGreen (Life Technologies). Following quantification, libraries were normalized to a uniform concentration and 96 indexed libraries (all individuals in a T-S population pair) were pooled on an equal molar basis for sequencing, resulting in four sets of pooled libraries. Library construction, quantification, normalization, and pooling were conducted utilizing a dual-hybrid Biomek FXp automated liquid handler (Beckman Coulter).

We mapped reads to the *F. heteroclitus* reference genome (NCBI BioProject number PRJNA323589) using both bowtie2 v 2.02 (*27*) and BWA MEM v.0.7.5a-r405e (*28*). We marked duplicates and generated split and discordant read files using SAMBLASTER v 0.1.16 (*29*) then compressed, sorted, indexed and characterized depth of coverage of the resulting alignments with Samtools v 0.1.19-96b5f2294a (*30*). This generated an average of 93.2 million reads per individual in our high coverage population pair (T1 and S1) and 7.7 million reads per individual in our low coverage populations (T2, S2, T3, S3, T4, S4). Given a predicted genome size of 1.3Gb, this resulted in an expected per base coverage of 7.2x in the high coverage population pair and 0.6x coverage in the low coverage populations. Consistent with our expectations, mean per base coverage of our 0.93Gb assembly at Q30 and excluding duplicates was 5.0x and 0.5x for the high and low coverage populations respectively (Fig. S23). We excluded 7.9mb (~1%) of our reference assembly with aberrantly high coverage from population genomic analysis. Reads mapping to these regions also typically had low mapping qualities and high divergence from the reference assembly. This suggests mismapping of repetitive motifs under-represented in the reference.

Variant calling

We called variants using Freebayes v0.9.18-1-g4233a23 (*31*) discarding reads with mapping quality < 30, bases with quality < 20 and all discordantly mapped or duplicate read pairs. We retained two sets of variants. The first was unfiltered. The second was filtered to create a set of bi-allelic SNPs with between 200x and 750x coverage across all individuals, with at least 80 samples having data, minor allele frequency > 0.05 and quality scores > 30. SNP calling yielded a filtered set of 20 million biallelic variant sites.

We identified sex-linked scaffolds by looking for scaffolds with many SNPs for which individual genotypes were highly correlated with sex. We also scanned for depth of coverage differences between males and females. We identified 21 sex chromosome-derived scaffolds comprising 2.75% of our reference assembly. Killifish are thought to have homomorphic sex chromosomes, and consistent with this, we observed no substantial regions where coverage in males was half that of females. The reference genome is derived from a female, so we are missing any male-unique regions. Our approach relies on restricted recombination between the X and Y preventing alleles from crossing over, so it will fail to identify any physically sex-linked scaffolds that are inherited in pseudo-autosomal fashion.

We estimated pairwise F_{ST} values from called genotypes using Weir and Cockerham's theta (*32*), as implemented in VCFLIB (https://github.com/vcflib/vcflib). We attempted to phase our diploid genotypes using BEAGLE (*33*). In low coverage populations, this was completely ineffective. In high coverage populations we found a high indidence of "phase switching" where haplotypes seemed to be accurately inferred over short physical distances, but incorrectly broken over shorter distances, so we do not rely heavily on that analysis here. We assessed population structure through ordination using multidimensional scaling (MDS). MDS is a technique for reducing high-dimensional data, such as long vectors of individual genotypes, into low-dimension summaries. We use it here to visualize genetic relationships of individuals in 2-dimensional space. Here we

calculated MDS components based on Euclidean distances between individual genotype vectors in R, a procedure that is numerically identical to Principal Components Analysis (*34*). MDS analyses clearly identify sampling sites as distinct populations and show that paired tolerant-reference sites are most similar to one another (Fig. S25).

Estimation of population genetic summary statistics

We used the software package ANGSD (*35*) to estimate the summary statistics π, Tajiima's D and F_{ST}. We first estimated 1 and 2-dimensional allele frequency spectra using 50mb of our reference genome, filtering out sites with excessive coverage, as above, and sites with data from < 10 individuals. We set read quality filters: mapping Q >= 30, base Q >= 20, properly mapping read pairs only. We then used those frequency spectra as priors in the empirical Bayesian procedure implemented in ANGSD to estimate values per site across the genome. We combined per site estimates into sliding windows of 5kb, moved in 1kb increments, and 50kb, moved in 10kb increments. Patterns of summary statistics across the genome were not qualitatively different between 5kb and 50kb sliding window analyses. Accordingly, we report results from 5kb sliding window analyses only. We excluded from consideration any window in which the mean number of sites evaluated across all populations was <40% (907,315 out of 1,027,354 windows were retained for the 5kb set). We observe wide variation in the distributions of these summary statistics across populations (Fig. S1), but statistics are generally highly correlated among population pairs (coefficients of 0.84 to 0.95 for π and 0.71 to 0.94 for Tajima's D). Genetic diversity increases moving from from North to South. T1 and S1 are the most highly differentiated pair (Fig. S2). Consistent with overall demographic decline in tolerant populations, possibly a result of a bottleneck attending colonization of polluted habitat, we observe subtle genome-wide shifts toward reduction in genetic diversity and a slight positive shift in Tajima's D when compared to sensitive populations (Fig. S1).

Demography Estimation and Neutral Simulation

We estimated demographic models for each population and pair using the Python module dadi and folded allele frequency spectra estimated using ANGSD as input. Spectra from low coverage populations were projected down to a sample size of 12 to 24 alleles. We fit each pair to a model consisting of three epochs. Two epochs in the ancestral population with independent population sizes followed by a population split, after which both populations had constant size and independent migration rates. This model has 7 parameters (N0,N1,N2,T0,T1,M12,M21). For each population pair, we optimized the model repeatedly from different starting points, and perturbed optimal parameters and re-optimized. We used the resulting parameters and an assumed recombination rate of 10^-8 to simulate neutral distributions of π, Tajima's D and F_{ST} in 5kb windows using ms (*36*). We simulated 20,000 replicates for each population pair.

Outlier Delimitation

To identify candidate regions underlying pollution tolerance in killifish, we scanned the genome for canonical signals of selective sweeps in 5kb sliding windows: reduction in genetic diversity (measured by π), a skew in the allele frequency spectrum (measured by Tajima's D: td) and high allele frequency differentiation (F_{ST}). Because high levels of missing data can lead to stochasticity in summary statistics, and may result in higher measured F_{ST}, we first excluded windows in which fewer than 2,000 bases were evaluated by ANGSD (given criteria listed above). We looked for a correlation between Fst and 'missingness' by fitting a linear model with F_{ST} as a function of the number of bases evaluated in the window and found a significant ($p<2e-16$) but very slight correlation (slope: 9e-7, R-squared: 1e-4). We do not regard the level of missing data after filtering as having a substantial impact on our estimates. First, we examined

tolerant-sensitive pairs independently, using our simulated neutral distributions. We identified windows for which empirical statistics exceeded the 0.001,0.001 and 0.999 quantiles of 1) piT - piS, 2) tdT – tdS, and 3) F_{ST}, respectively. For F_{ST}, we used values calculated in VCFLIB. For windowed averages, these values were highly correlated with those calculated in ANGSD. On a per site basis, VCFLIB was much noisier, which is to be expected because it does not use empirical Bayesian smoothing as in ANGSD. In outlier delimitation, we used the values from VCFLIB simply because the ANGSD Fst estimation procedure took quite a long time to complete. In practice, these thresholds were close to the 0.01 or 0.99 quantiles of the empirical distribution. Windows exceeding a threshold for any statistic were retained as outliers. Outlier windows within 50kb of one another were merged into outlier regions. In order to rank outlier regions by the extent of their deviation from genome-wide expectations, we converted each statistic to a Z-score and summed up the Z-score minus the threshold value for each summary statistic for each outlier window within each region. These aggregate statistics are thus a product of the length of an outlier region and the extremity of summary statistic values within the region. We used these statistics to prioritize analysis of outlier regions. We discarded outlier regions identified by F_{ST} where values of pi and Tajima's D suggested the sensitive population was the target of selection. This approach prioritizes rapid, complete, or nearly complete selective sweeps of variants beginning at very low frequency and occurring in regions of moderate to high background genetic diversity. It is likely to miss incomplete or soft sweeps in regions with low genetic diversity. Our low coverage data and attendant inability to accurately phase genotypes made it difficult to apply methods meant to identify soft or incomplete sweeps in this system.

A weakness of the pairwise approach is that population pairs may have independent selective and demographic histories such that strong signals of selection in the tolerant population are not a result of adaptation to pollution. In practice, this appears to be the case for a number of outlier regions identified with the above procedure. Upon examination in the context of all 8 populations,

several high ranked outlier regions in all three northern population pairs appear to be inconsistent with adaptation to pollution, with identical signatures of selection present in, and linked variation shared with, one or more sensitive populations. In order to resolve this, we repeated the above procedure of identifying outlier regions using population triads with one tolerant population and the two geographically closest sensitive populations. The statistics applied were 1) max(piT - piS1,piT - piS2), 2) max(tdT - tdS1,tdT - tdS2), and 3) the population branch statistic of Yi et. al (*37*). We did not simulate 3-population models, but instead set thresholds for each statistic at the 0.01,0.01, and 0.99 quantiles, respectively. This approach either eliminated or greatly reduced the rankings of many pairwise outlier regions that close examination suggested were not associated with pollution tolerance, but otherwise produced very similar results to the pairwise approach, so we focus on this approach in the rest of the analysis under the assumption that the tails of our summary statistic distributions are more extreme than expected under a simple neutral model.

There is no chromosome-scale assembly available for *F. heteroclitus*, so in order to visualize the distribution of outlier regions across the genome, we mapped our set of scaffolds to the genome of platyfish (*Xiphophorus maculatus*), the most closely related fish species with such an assembly published (*38*). To do this, we used BLAST to map all annotated exon sequences in *F. heteroclitus* to *X. maculatus.* After discarding exons with more than five BLAST hits, we assigned *F. heteroclitus* scaffolds to *X. maculatus* chromosomes if more than 75% of BLAST hits originating from the scaffold mapped to the chromosome. Otherwise, the scaffold was left unmapped. This portion of the analysis was conducted in R. We did not break *F. heteroclitus* scaffolds during this process, but simply ordered them by their mean position on the *X. maculatus* chromosome. A visual inspection of the results indicated that synteny was coarsely preserved, with the ordering of exons on most large scaffolds highly similar between the species, although various rearrangements were evident. With the exception of one scaffold, all scaffolds discussed in the main text were successfully mapped.

Phylogeny Estimation

We calculated allele frequencies for bi-allelic SNPs and used the CONTML module of the package Phylip (through Rphylip (*39*)) to estimate population trees for 1) a subset of SNPs from across the genome, 2) all 50kb windows in the genome and 3) delimited outlier regions. The genome-wide population tree reiterates population structure observed in ordination analysis (Fig. S25) and clusters tolerant-sensitive pairs (Fig. 1). In addition, by far the most common bipartitions across all 50kb windows match the genome-wide population tree. We scanned the set of population trees for trees that conflicted with the dominant pattern by clustering sets of tolerant populations.

Copy number variation

We searched for large structural changes in the genome relevant to pollution adaptation by scanning for changes in depth of coverage among population pairs. Large changes in coverage might indicate duplications or deletions with strong frequency differences among population pairs. We calculated coverage per individual in several ways: 1) read coverage per base per individual using Samtools depth, 2) calculated fragments per 5kb window per individual using bedtools and 3) calculated fragments per annotated gene per individual, also using bedtools. We did not do statistical analysis on the per base coverage, and used edgeR to model the counts per genomic region. While many regions of the genome show significant differences in copy number among population pairs, the vast majority involve strong deviations from the expected coverage in both members of a population pair and are often associated with gaps in scaffolds of our assembly. This suggests read mis-mapping and/or assembly problems and makes interpretation difficult. However, we consistently identified two genomic regions with large changes in

coverage between tolerant and sensitive pairs, where the coverage changes affect regions with high quality read mapping and which are also within high ranking outlier regions. In the first of these regions (Fig. 3A,B) three tolerant populations (T1,T3, and T4) show signatures of deletion (Fig. S7 A-C) that spans genes AHR1a and AHR2a. In the second of these regions (Fig. 3C,D) the three northern tolerant populations (T1,T2, and T3) have increased coverage relative to expected (Fig. S7 D,E, and Fig. S11) which suggests an increase in copy number; this duplication spans gene CYP1A.

We confirmed the deletion in T4 with PCR. PCR primers were designed flanking the left and right junctions of the putative deleted region (LF1 and RR2), and within the deletion (RF2) (Fig. S26). Genomic DNA (10 ng) from 8 fish from each of T4 and S4 populations were amplified with the LF1/RR2 and RF2/RR2 primer pairs using Advantage DNA polymerase (Clontech) with the following cycling conditions: [94^{o}C, 1 min]; [94^{o}C, 5 sec; 68^{o}C, 2 min] 25 X; [68^{o}C, 5 min]. The amplification products were resolved in 1% agarose gels and stained with ethidium bromide. The 1.3 kb LF1/RR2 PCR products from fish #13 and #14 were ligated into pGEM-T Easy (Promega) and sequenced from both ends. Primer Sequences: LF1: 5'-AGTATGCATTTACGCAACAGAGCG-3'; RF2: 5'-GAGTGACGCAGCATCACAATAAGC-3'; RR2: 5'-ACAACAAACGTAGAACCACACAGC-3'.

Pathway Analysis

Genes (human orthologs) that were differentially expressed upon PCB challenge between tolerant and sensitive populations (see RNA-seq analysis below) were used for pathway and network analysis in Ingenuity Pathway Analysis (Ingenuity.com). Similarly, genes that were in popualtion genetic outlier regions for each tolerant-sensitive popualtion pair were used for network analysis in IPA. IPA uses a Z-score algorithm to predict upstream regulators (see description at

http://ingenuity.force.com/ipa/articles/Feature_Description/Upstream-Regulator-Analysis). Canonical pathway enrichment analysis was also performed in IPA for genes that were differentially expressed and for genes that were within population genetic outlier windows, again using a Z-score algorithm as described (http://ingenuity.force.com/ipa/articles/Feature_Description/Canonical-Pathways-for-a-Dataset).

RNA-seq analysis

For each of our eight populations, we exposed developing embryos (two generations removed from field-collected) from 1 day post fertilization to post-organogenesis (stage 35, ~10 days post fertilization) to model toxicant PCB126 and vehicle (DMSO) control as described in (*9*). We included 3-5 biological replicates per treatment. RNA was extracted as described in (*9*) and indexed RNA-seq libraries prepared using NEB Next Ultra RNA library prep kits for Illumina according to the manufacturer's protocol. Indexed samples were pooled and sequenced (Illumina PE-100). We quality trimmed reads using Trimmomatic (*40*) according to recommendations in (*41*). We aligned reads to the *Fundulus heteroclitus* reference genome using TopHat (*42*) and counted reads falling in annotated gene regions using featureCounts (*43*) and tested for differential expression using the quasi-likelihood method (*44*) implemented in edgeR (*45*) and retained as differentially expressed genes with p-values that put their false discovery rate below 5%. Critical contrasts tested were: 1) dose responses (PCB *versus* DMSO control, 2) dose by evolved tolerance responses, and 3) dose by evolved tolerance by population pair responses.

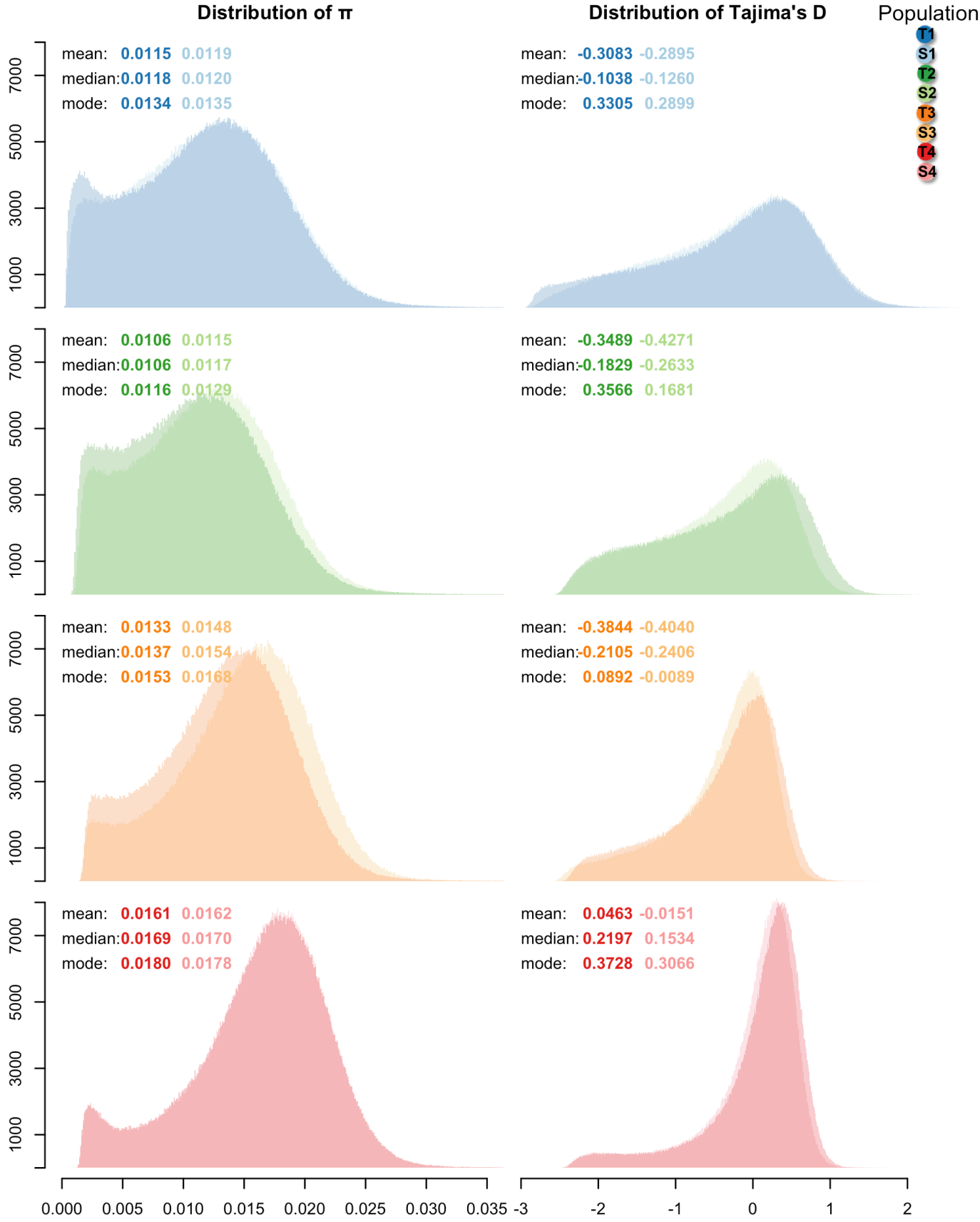

Fig. S1. Distributions of pi (left panel) and Tajima's D (right panel) in 5 kb windows for each population. Pi is reduced genome-wide, and Tajima's D shifted positive, in tolerant (T) populations compared to their sensitive (S) population counterparts, consistent with reduced effective population size in T populations.

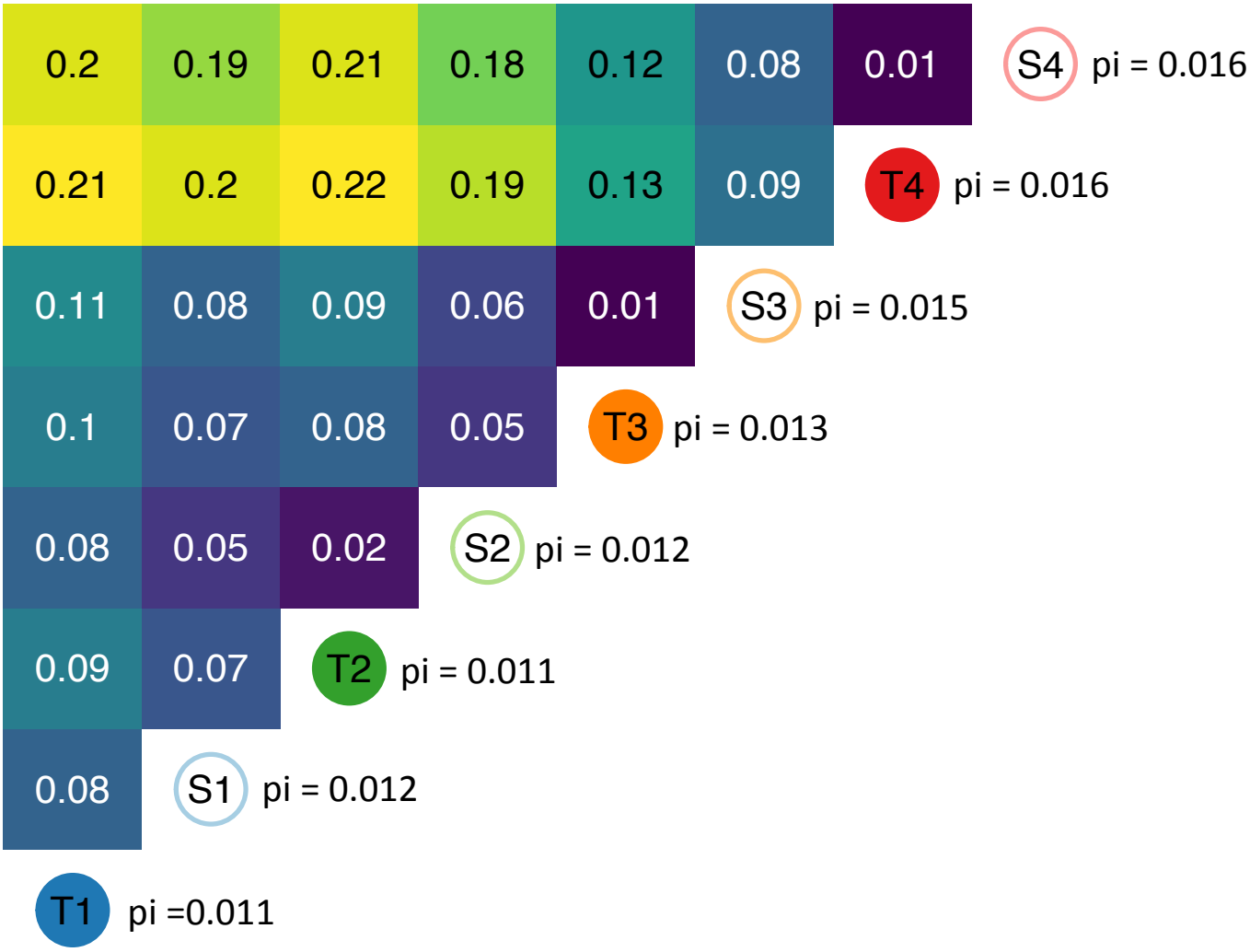

Fig. S2. Fst between pairs of populations, calculated from genome-wide SNP variation. Boxes are colored, from cool to warm, with increasing Fst. Geographic pairs have very low Fst (~0.1 or below), where the largest genetic differentiation is between northern (T1, S1, T2, S2, T3, S3) and southern (T4, S4) populations. Genome-wide average nucleotide diversity (pi) is reported for each population on the diagonal. Nucleotide diversity within *F. heteroclitus* populations is extremely high, ranking them as the most genetically diverse among vertebrates compared to other species reported in (*25*).

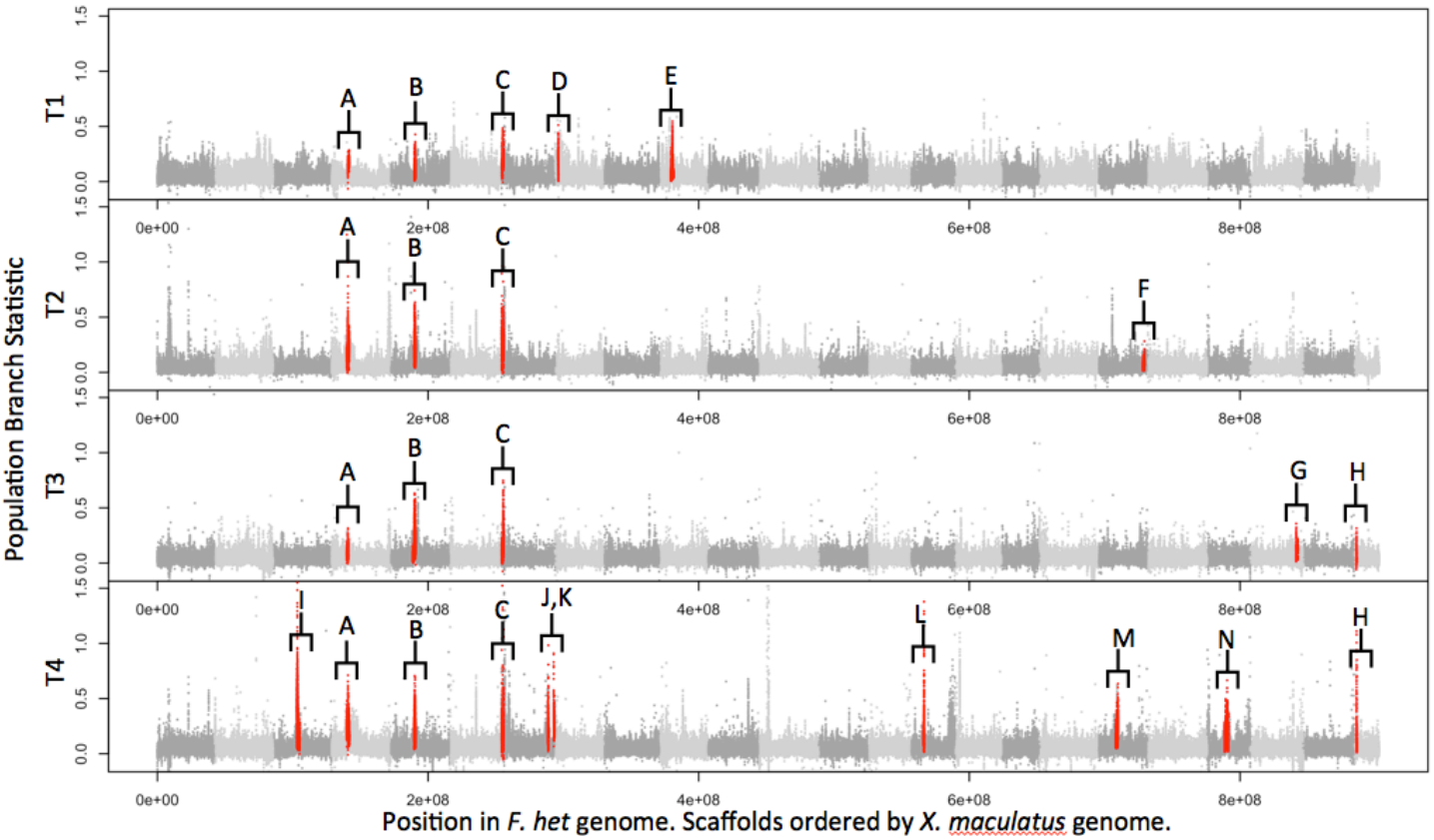

Fig. S3. A plot of the population-branch statistic (PBS) genome wide for each tolerant population. Each point represents a 5kb genomic window. Points are ordered by position in the genome assembly of *Xiphophorus maculatus,* as described in the Supplemental Methods. Points with alternating shades of gray indicate different chromosomes in *X. maculatus*. Points are colored red to indicate that the 5kb window falls within an identified outlier region discussed or depicted in the main text. Not all outlier regions are colored. Outlier scaffolds referred to in the main text are identified in red, and letters indicate candidate genes: A): CYP1A, B) AIP, C) AHR 1a/2a, D) T1 outlier 2, E) T1 outlier 1, F) ER2b, G) EPAS1, H) AHR 1b/2b, I) ARNT, J) GST-theta, K) Interleukin and Cytokine receptors, L) HSP90, M) CYP1C/GFRP, N) KCNB2. *Scaffold522 containing KCNC3 did not map to the* X. maculatus *assembly.*

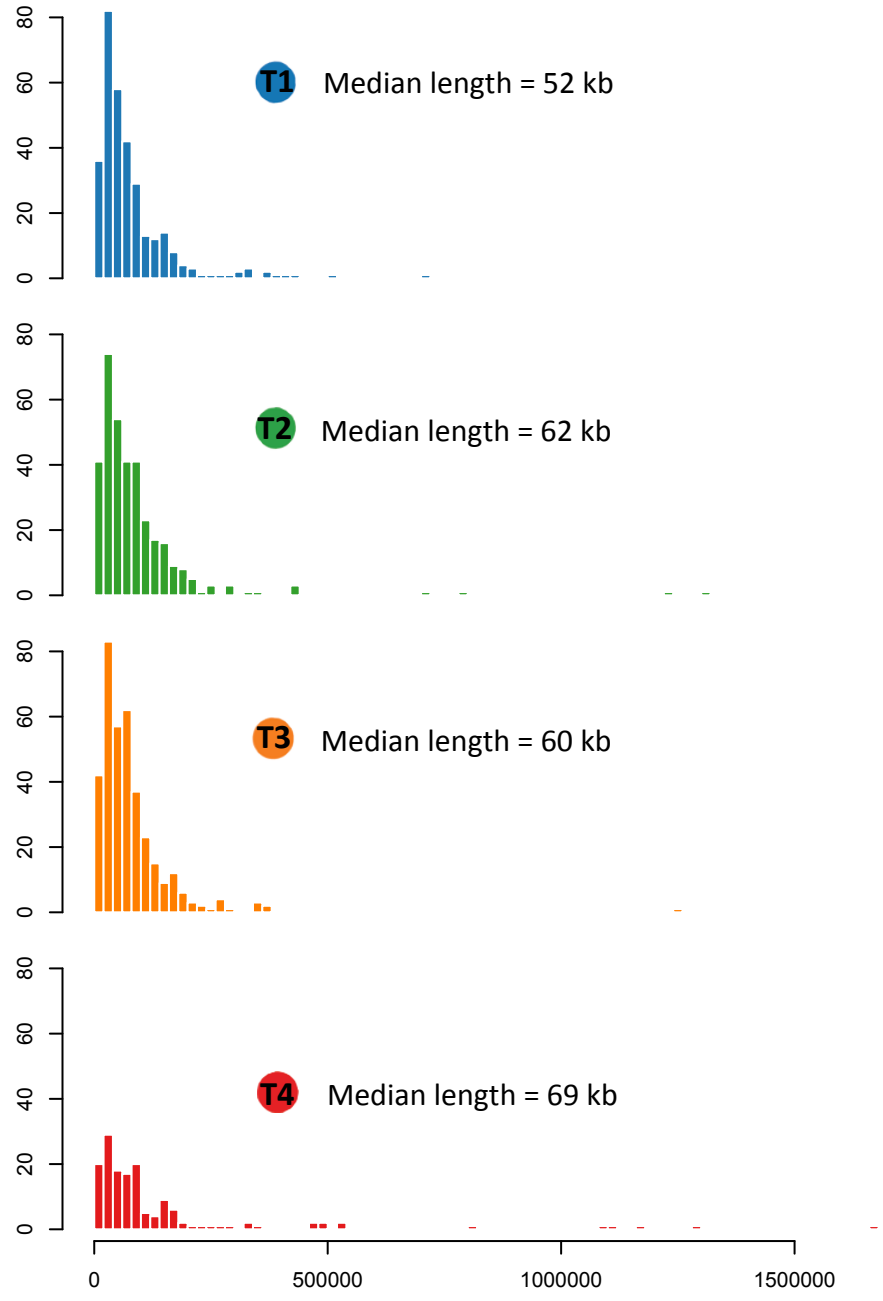

Fig. S4. Histograms of the lengths of outlier windows for tolerant populations. Most outliers tend to be small (median lengths indicated) with a small number that are large.

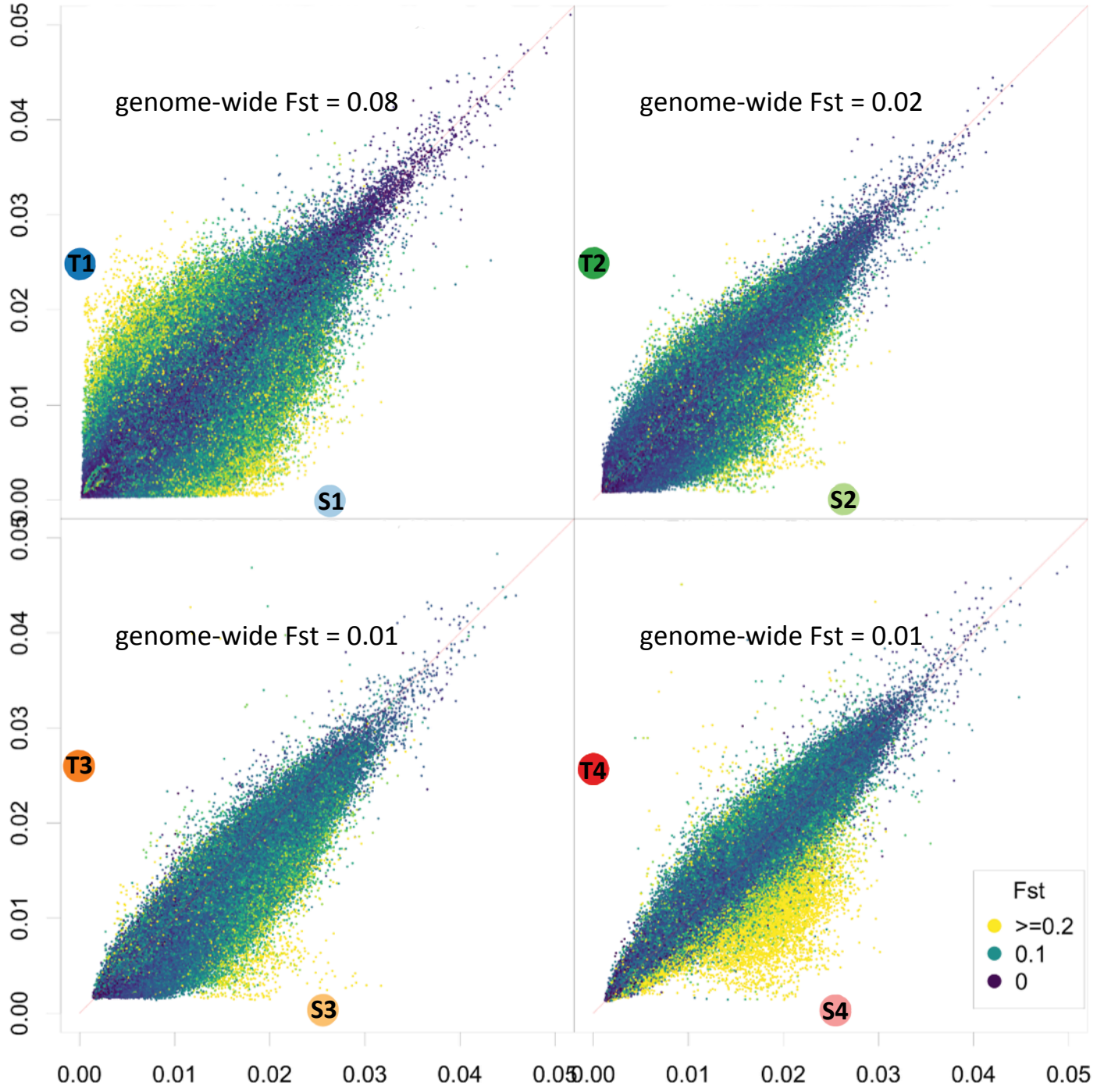

Fig. S5. Correlation in nucleotide diversity (pi) between members of tolerant-sensitive population pairs. Each dot represents a single 5-kb sliding window. All dots represent all 5-kb sliding windows genome wide. Each 5-kb window is colored by Fst between the population pair, where warmer colors indicate higher Fst. For population pairs 2-4, windows with high Fst (yellow dots) and low genetic diversity in only one member of the pair (suggesting divergent selection pressure) tend to indicate selection in the tolerant member of the population pair. In pair 1 both populations appear to have been targeted by diverging selection pressures.

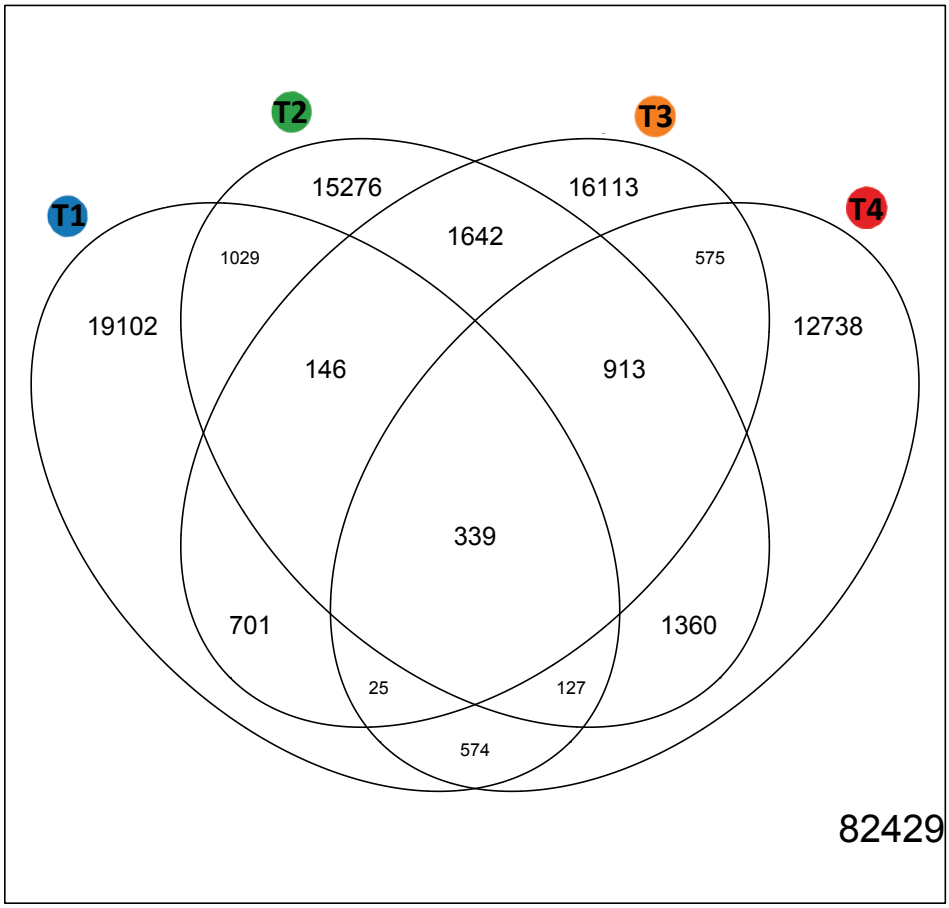

Fig. S6. Venn diagram showing the overlap in outlier windows between tolerant populations. Most outlier windows tend to be specific to particular tolerant populations. But a few are shared between populations. Those that are shared tend to be the most highly ranked outliers for each population pair – those with the strongest signals of recent selection.

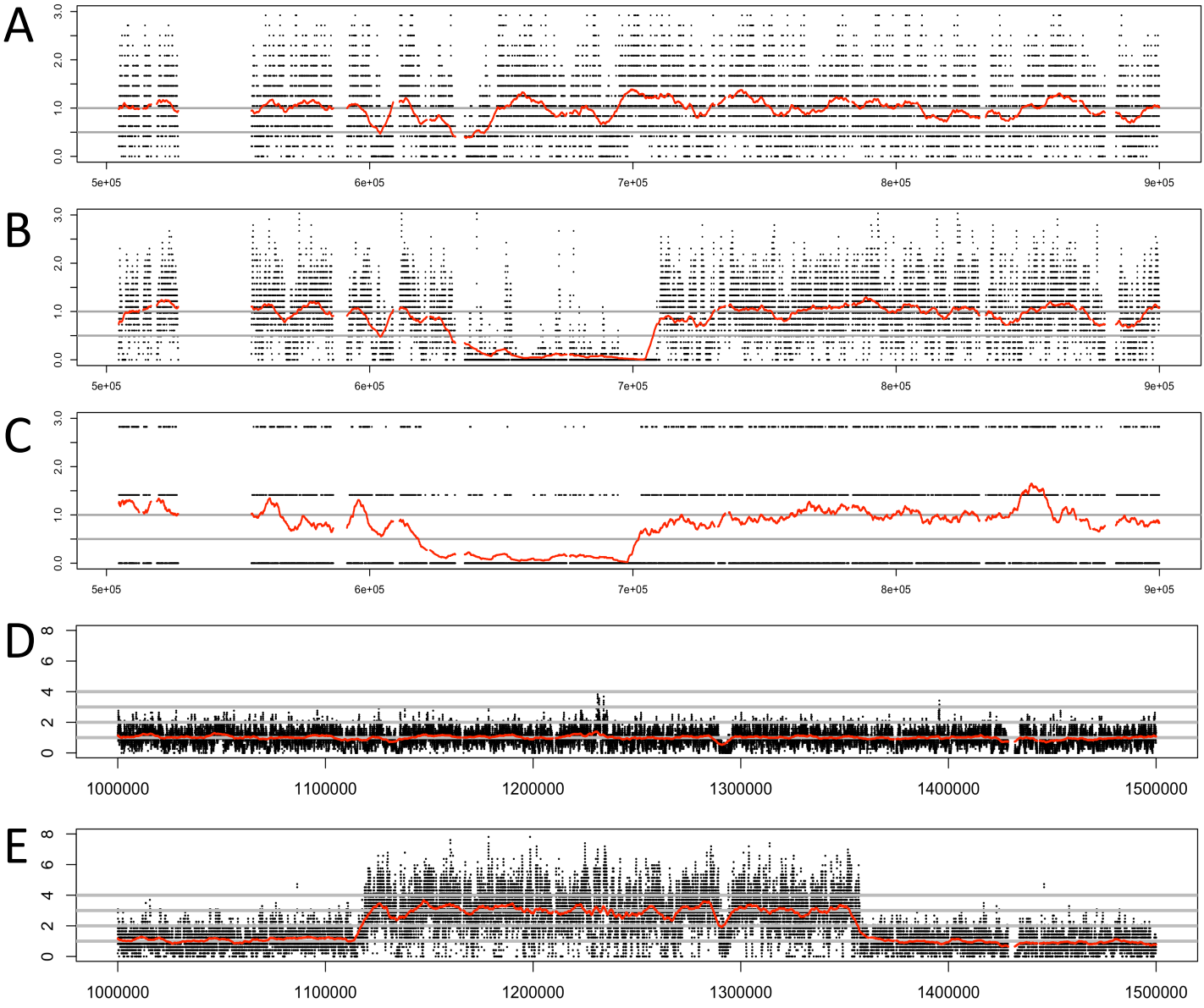

Fig. S7. Per base read mapping coverage showing evidence of copy number variation. Panels A-C are for three representative individuals for the deletion region spanning genes AHR2a and AHR1a. A) Individual from S1 showing no evidence of deletion, where black dots represent coverage per base, and red line represents the sliding window average per base coverage. Y axis is coverage divided by expected coverage given no deletion. The expectation is that the red line should hover near 1 for no deletion, should drop to 0.5 for a deletion heterozygote, and drop to zero for a homozygote deletion. B) Individual from T1 that appears homozygous for a deletion. C) Individual from T4 that appears homozygous for a deletion. Panels D-E are for two representative individuals for the duplication spanning gene CYP1A. D) Individual from S1 showing no evidence of copy number increase. E) Individual from T1 showing evidence of four extra copies.

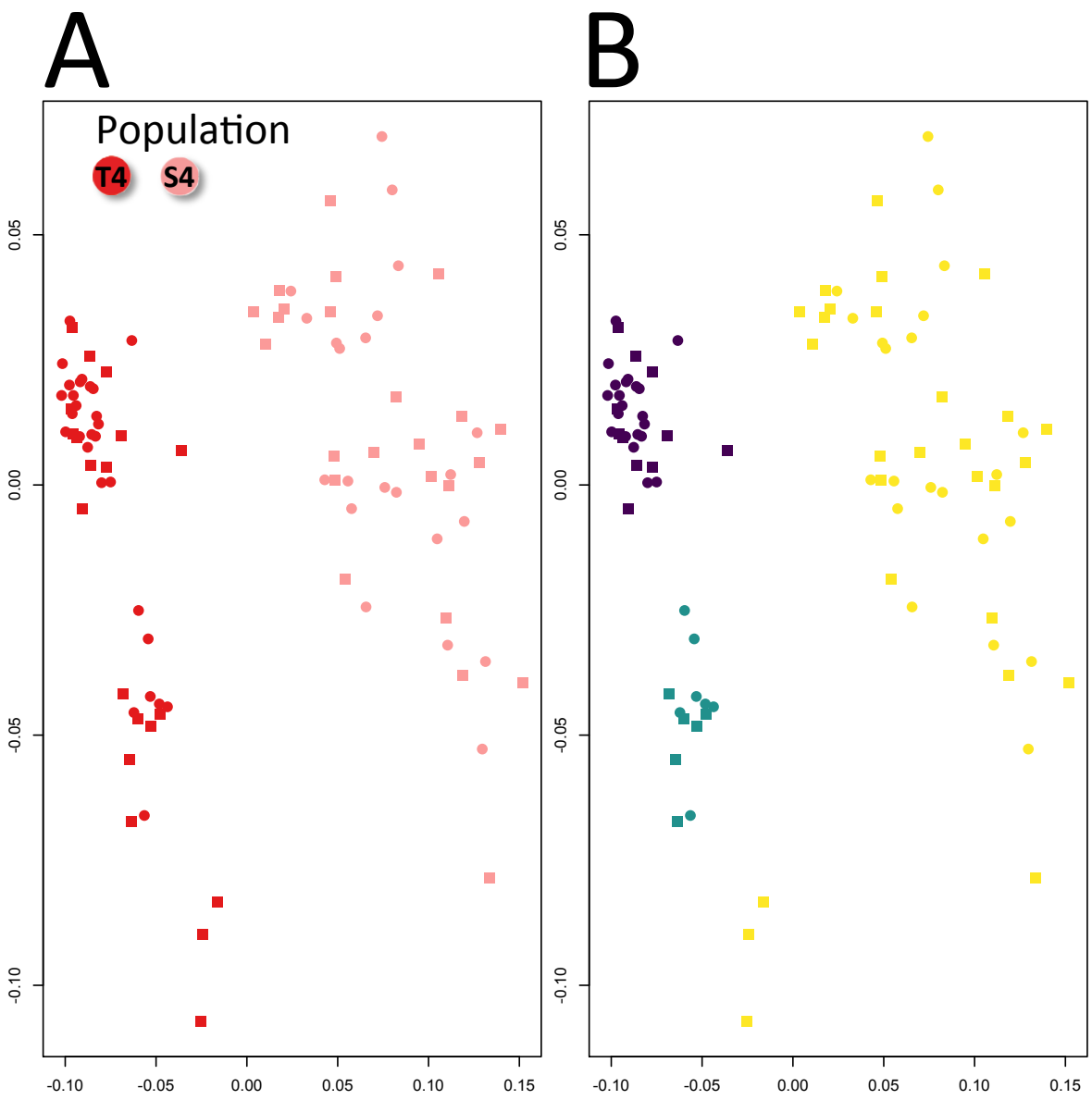

Fig. S8. MDS plots of genotypic similarity for the scaffold containing AHR2a and AHR1a genes for all individuals from the T4 and S4 populations. A) individuals colored by population of origin. B) individuals colored by homozygous for the deletion (purple), heterozygous for the deletion (teal), or no deletion (yellow).

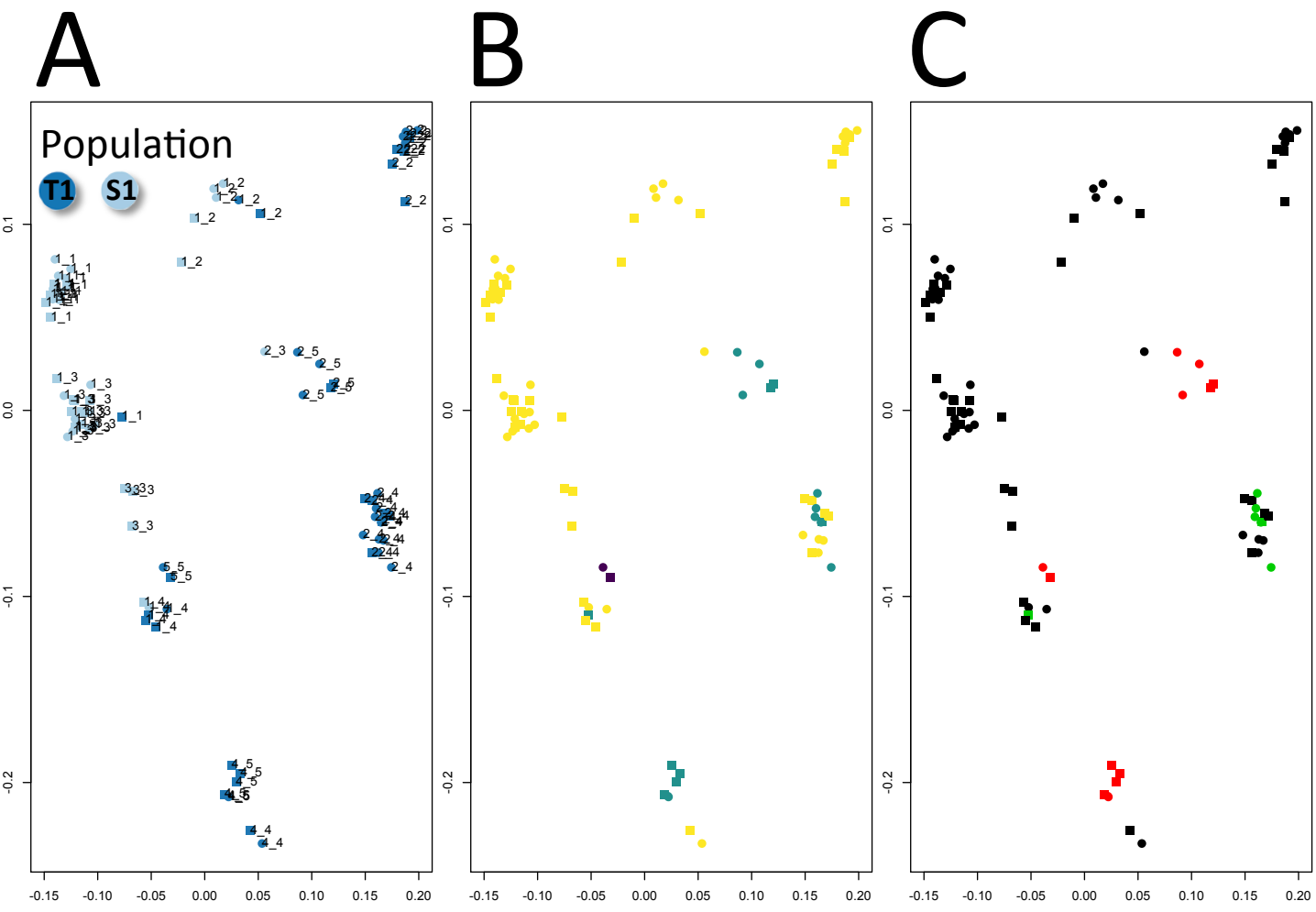

Fig. S9. MDS plots of genotypic similarity for the scaffold containing AHR2a and AHR1a genes for all individuals from the T1 and S1 populations. A) individuals colored by population of origin. Numbers indicate diploid haplotype identity. We detect five haplotypes. B) individuals colored by homozygous for a deletion (purple), heterozygous for a deletion (teal), or no deletion (yellow). C) Individuals colored by which deletion they bear: red is for the deletion that spans the same region in T1 and T3 (see figure 3A), green is for the deletion found only in T1 (see figure 3A), black is no deletion.

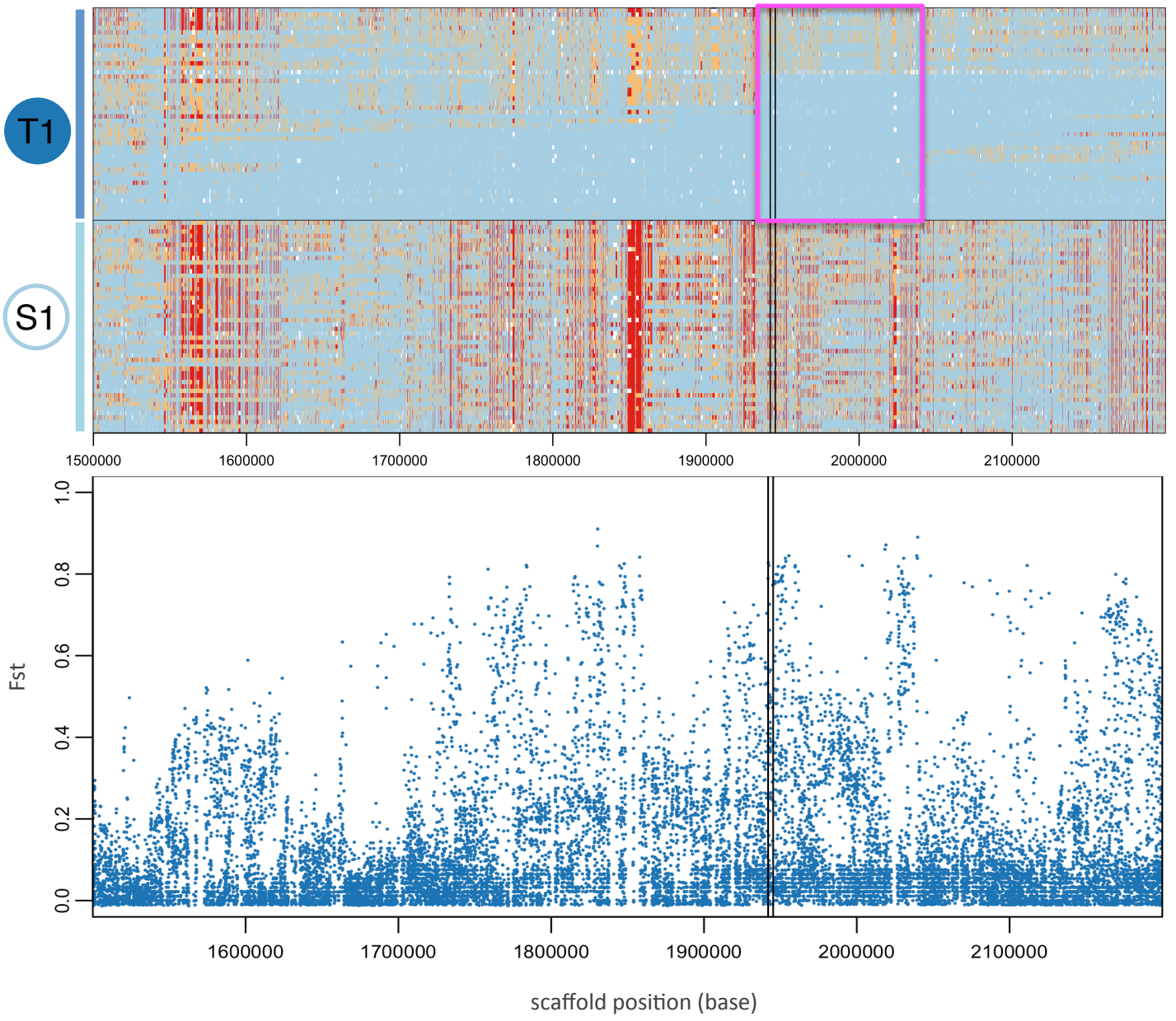

Fig. S10. Haplotypic variation at the AIP locus in T1 and S1 individuals, where each row is an individual, each column is a variable site on the genomic scaffold, blue is homozygous for the allele that matches the sweeping haplotype, red is homozygous for the alternate allele, and orange represents a heterozygote. Vertical gray line indicates AIP locus. A single core haplotype of ~100kb has swept to high frequency in T1 (pink box), and to fixation in T2 and T3 (see MDS plots in Figure 3C). A different haplotype has swept to fixation in T4 (see MDS plots in Figure 3C). Bottom panel is Fst between populations T1 and S1. The core haplotype coincides with peak divergence.

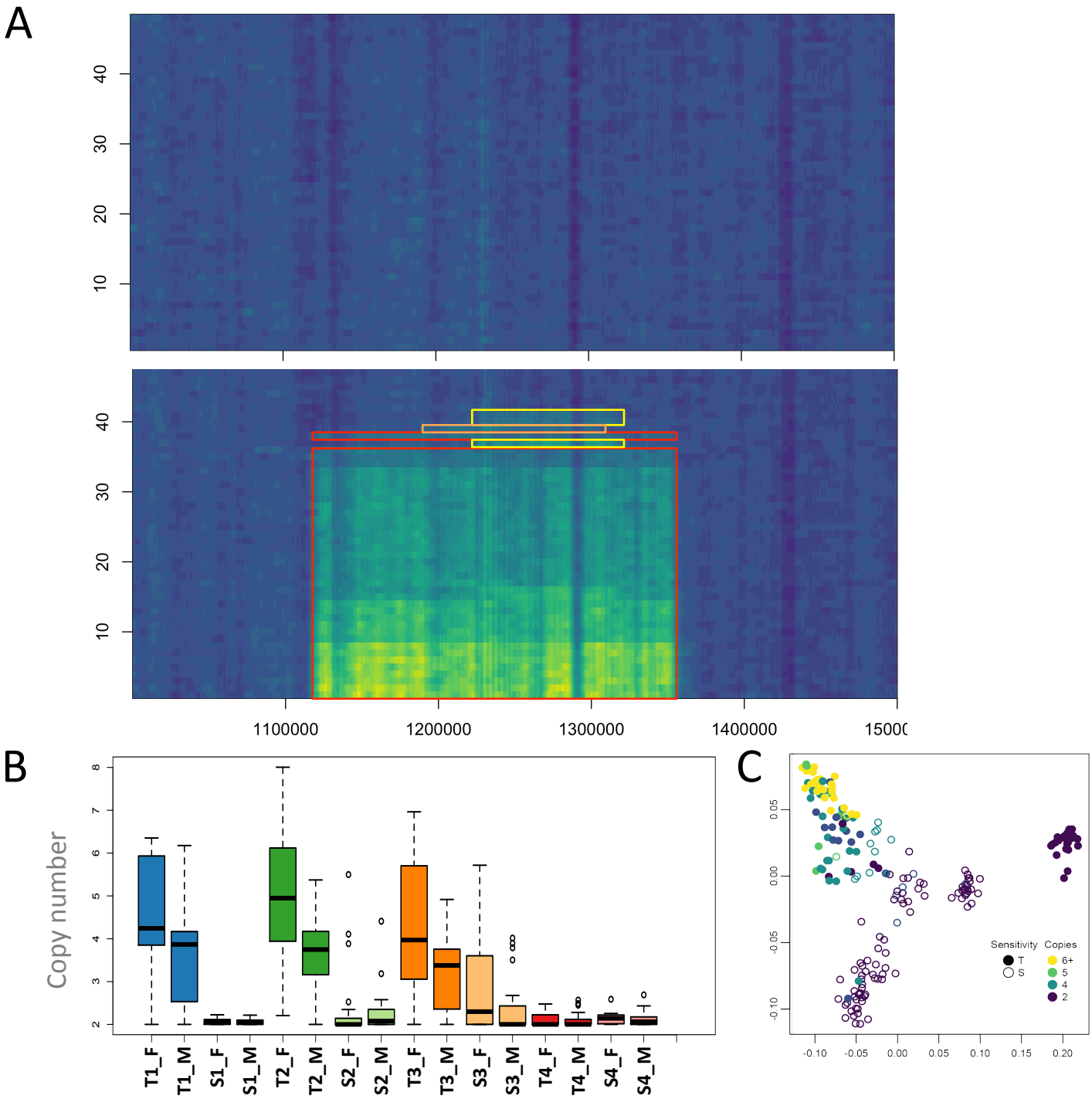

Fig. S11. Mapping depth evidence for three copy number alleles that have swept to high frequency in T populations. A) Top panel are 48 individuals from S1 population, and second panel are 48 individuals from T1 population, where each row is an individual and each column is a SNP position on the scaffold. Color is scaled by copy number from blue (2 copies) to bright green (8 copies). We detect 3 independently duplicated regions with different genomic spans in T1. They are C1 (100kb: yellow box), C2 (120kb: orange box), and C3 (250kb: red box). All three variants are supported by increased coverage, and C3 is supported by discordantly mapping paired end reads, which suggest at least one tandem duplication. When we estimate individual copy number based on ratios of coverage inside to outside putative duplicated regions, this ranges from 2 (1 per chromosome, no extra copies, colored blue) to 8 (six extra copies, colored bright green). All three variants completely encompass gene CYP1A, the most strongly up-regulated transcriptional target of the ligand-activated AHR pathway. Intriguingly, the scaffold on which CYP1A is found is sex-linked. Our analysis suggests at least one extra copy of the duplicated region exists on the X chromosome, as females have more copies on average than males (B). Population T4 shows no signs of increased copy number in this region, though this remains a significant outlier region in T4. C) MDS plot of genotypic variation on the scaffold containing the CYP1A gene (as in Fig. 3D), but where individual genotypes are colored by copy number. Clustering of genotypes with high copy number of the duplications around CYP1A suggests that extra copies arose from a single haplotypic background. Though this region is also a top-ranked outlier in T4, differentiation is not associated with a change in copy number.

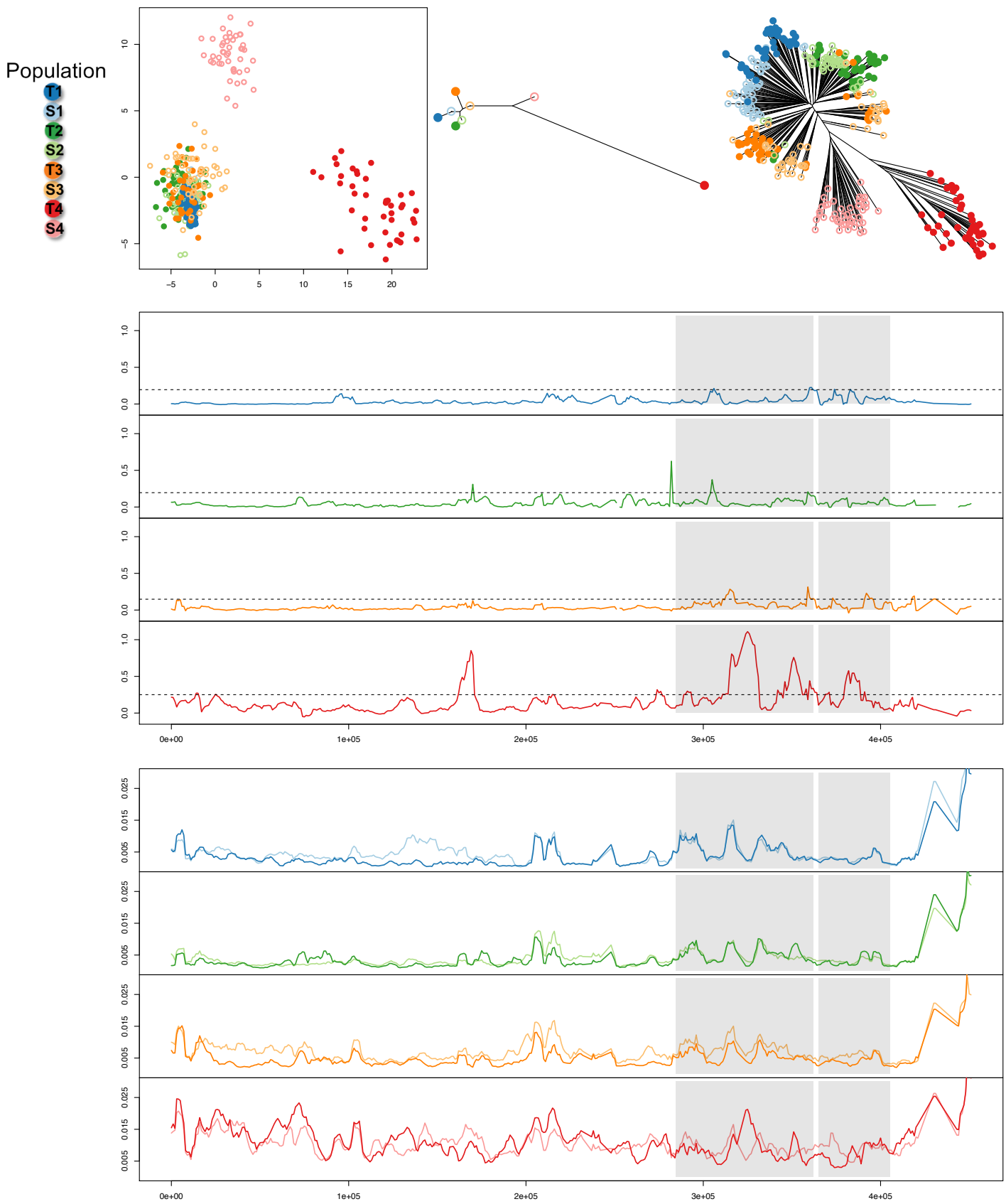

Fig. S12. Signatures of selection in the outlier region containing genes AHR2b and AHR2b (scaffold 217 in). Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene models for AHR2b (left) and AHR1b (right).

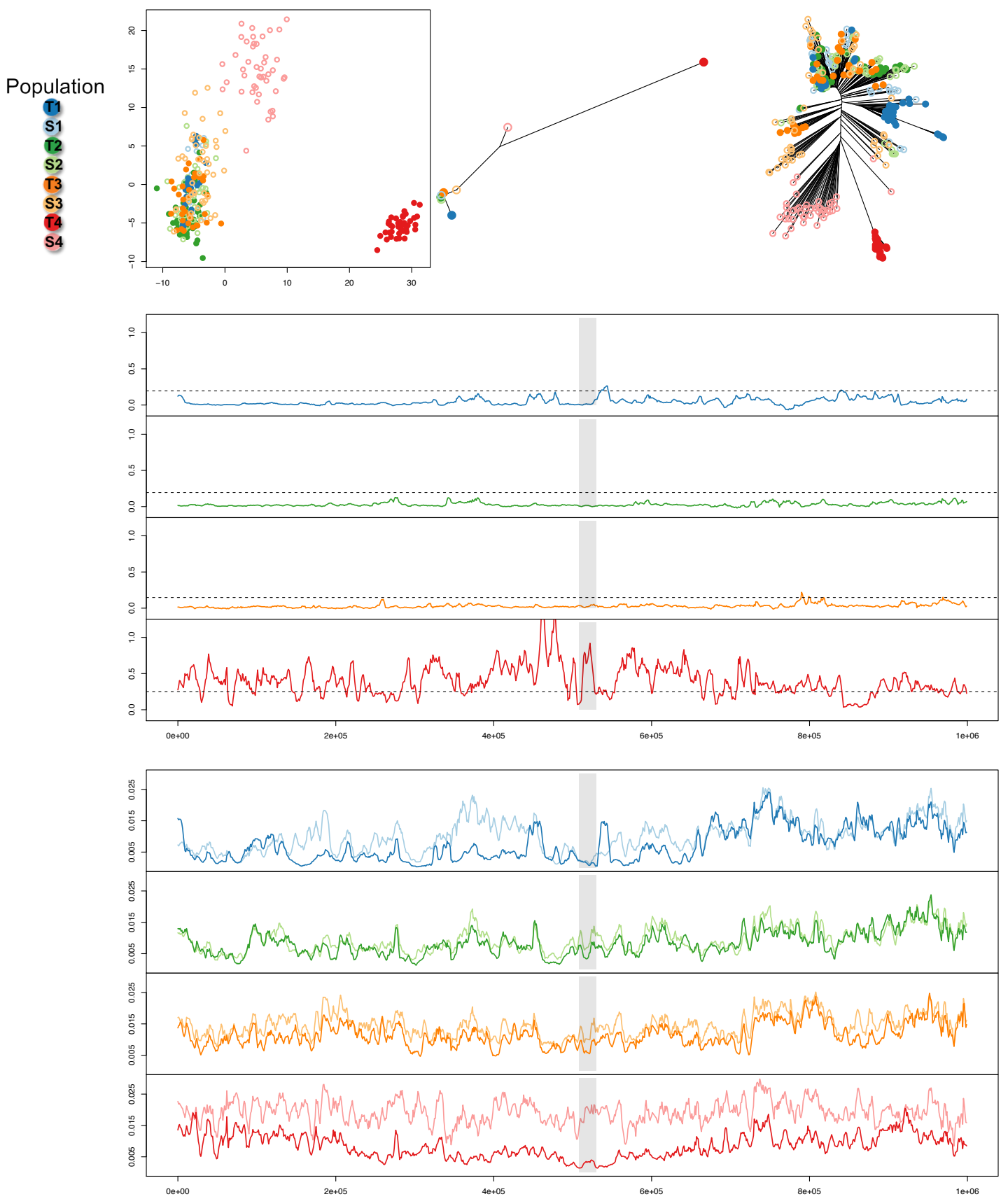

Fig. S13. Signatures of selection in the outlier region containing gene ARNT1c. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for ARNT1c.

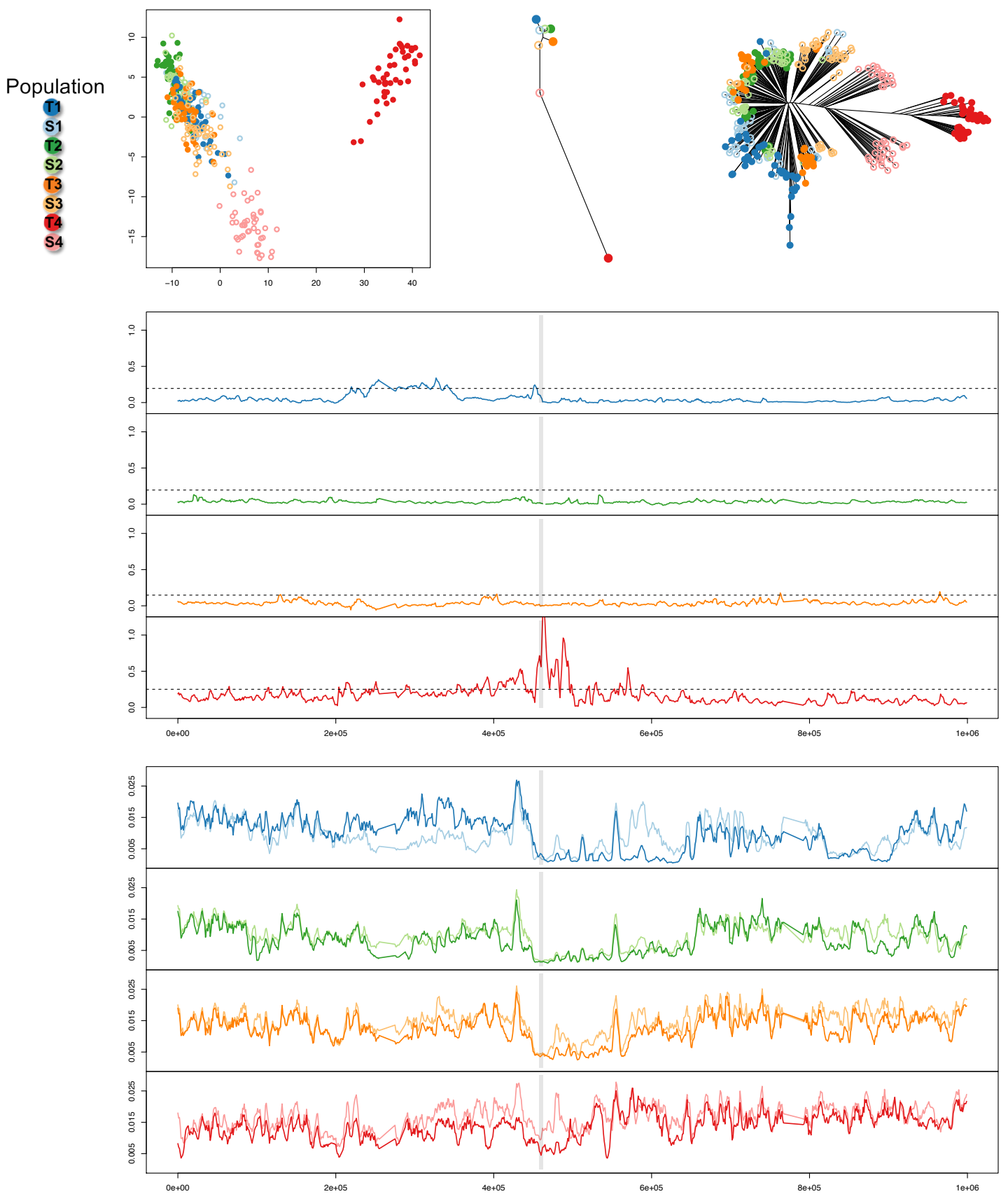

Fig. S14. Signatures of selection in the outlier region containing gene HSP90. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for HSP90.

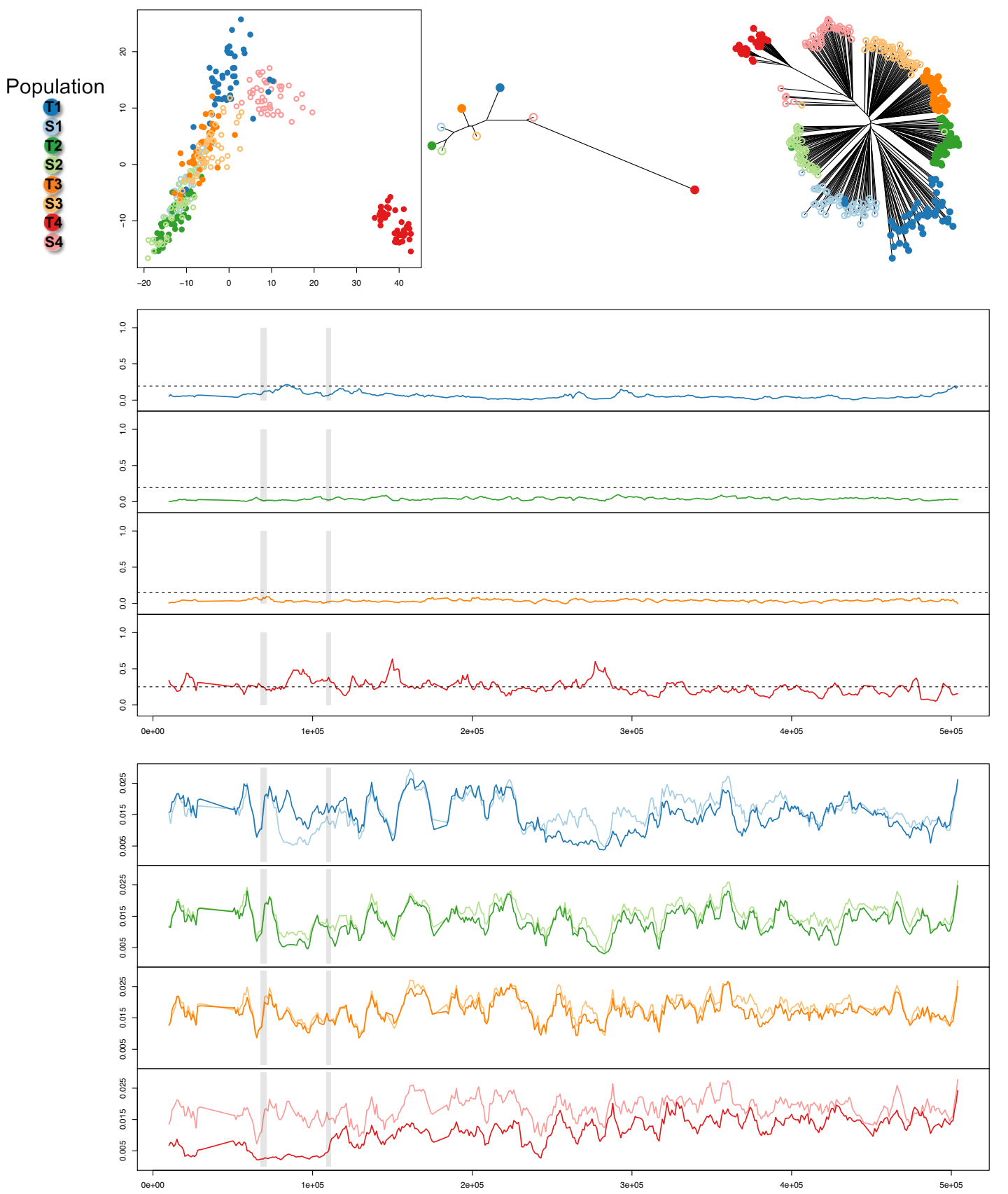

Fig. S15. Signatures of selection in the outlier region containing genes CYP1C and GFRP. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene models for CYP1C1 and 1C2 (tandem) (left) and GFRP (right).

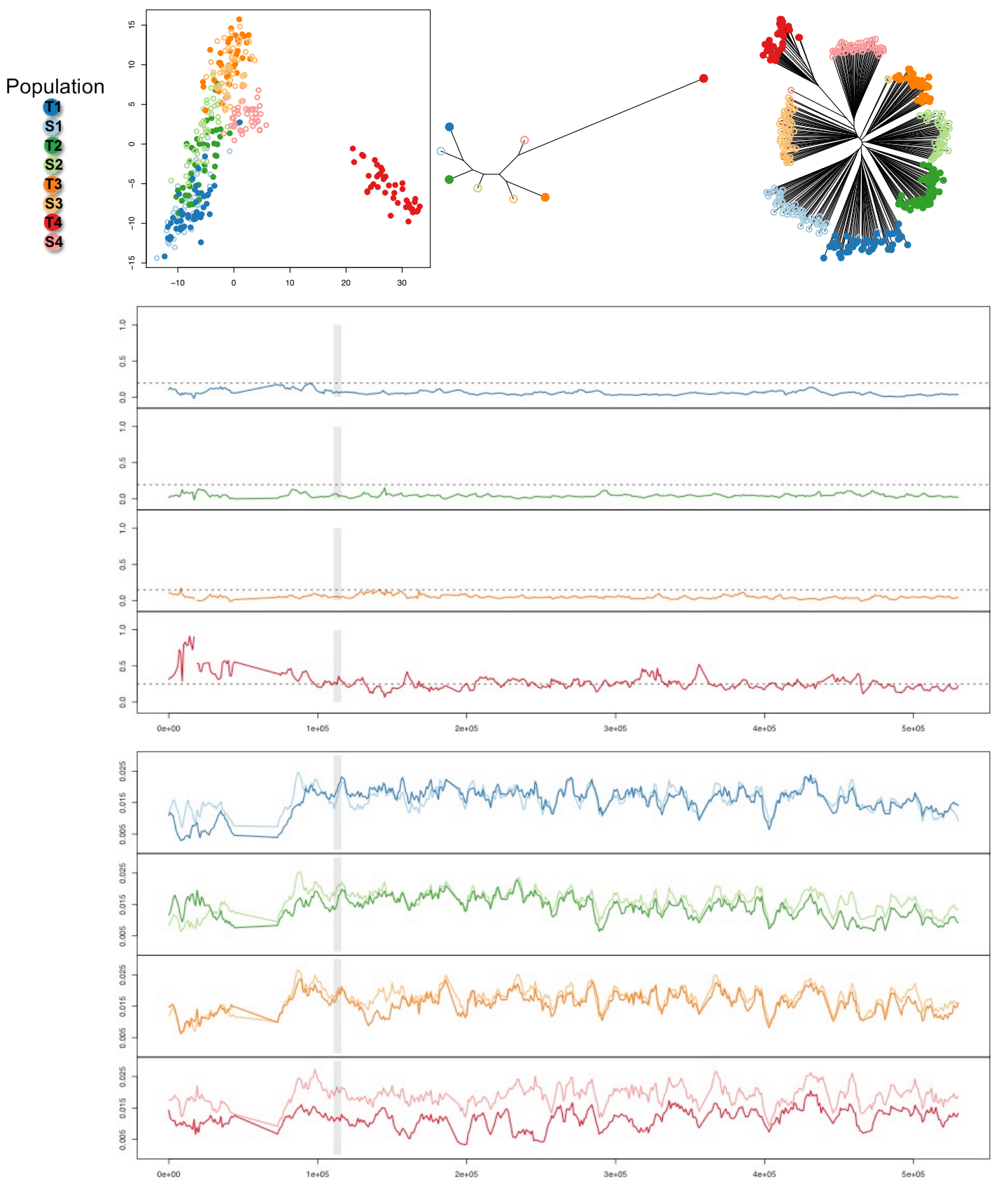

Fig. S16. Signatures of selection in the outlier region containing gene GST-theta. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panel indicates position of gene models for GST-theta.

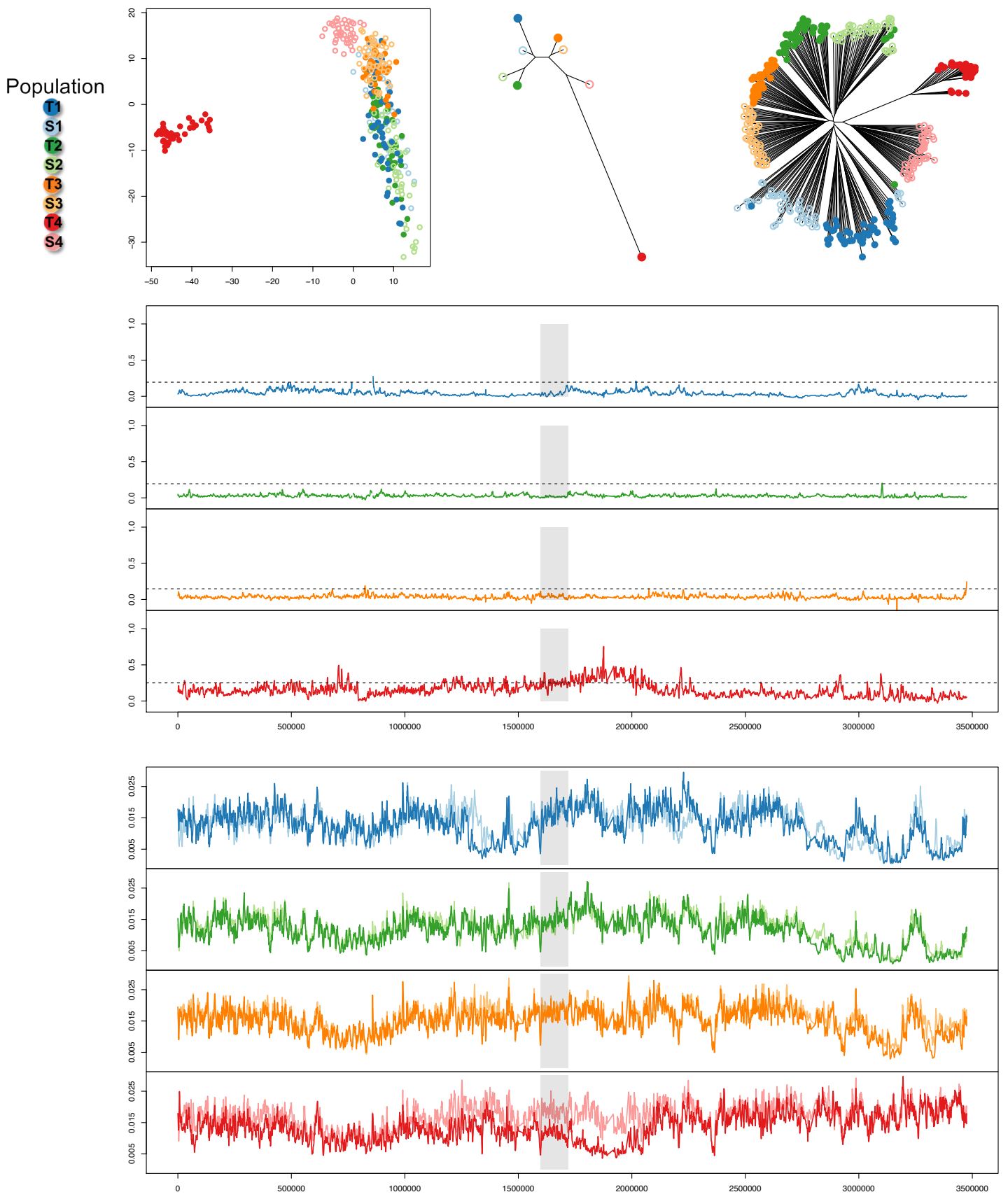

Fig. S17. Signatures of selection in the outlier region containing gene KCNB2. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for KCNB2.

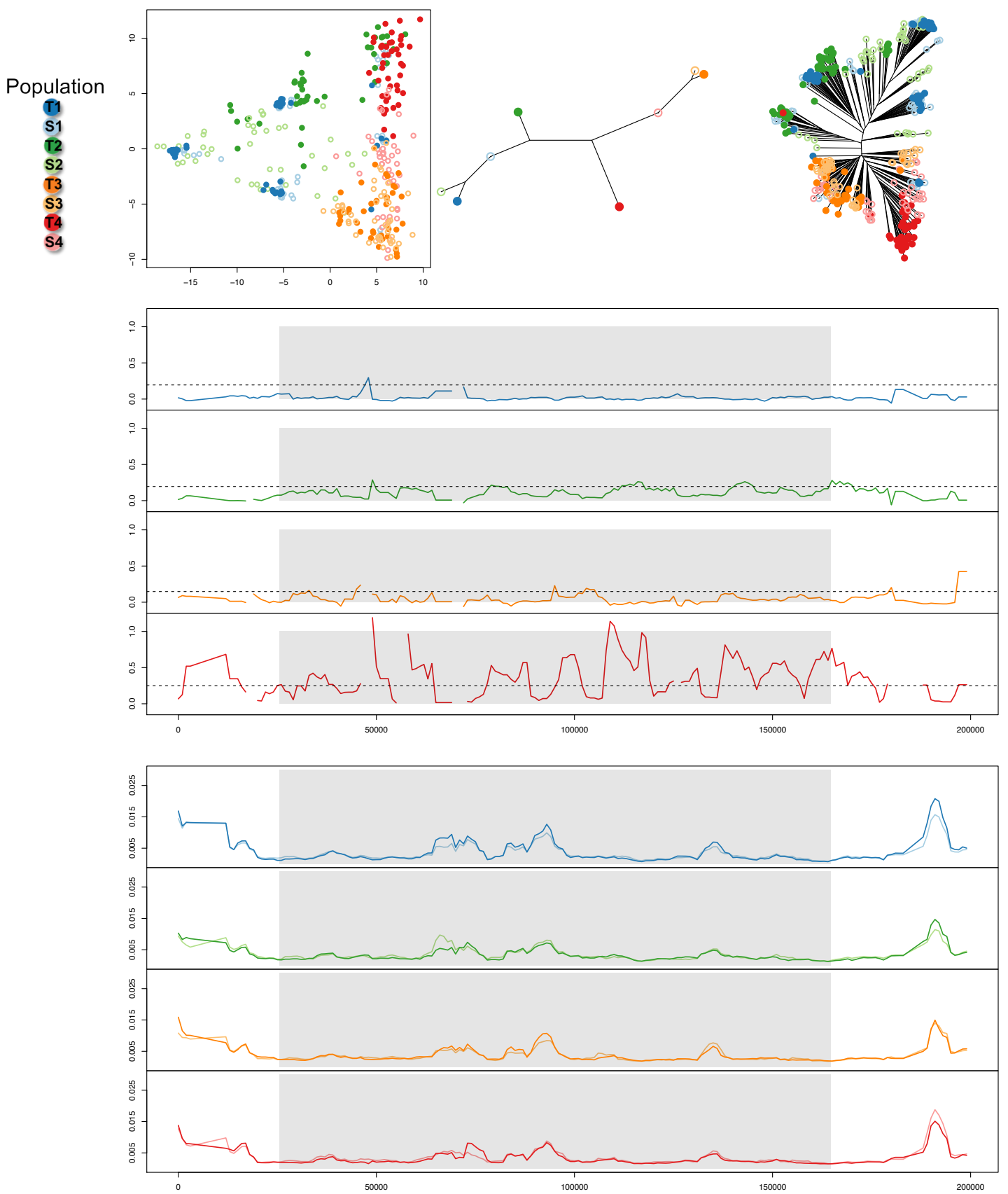

Fig. S18. Signatures of selection in the outlier region containing gene KCNC3. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for KCNC3.

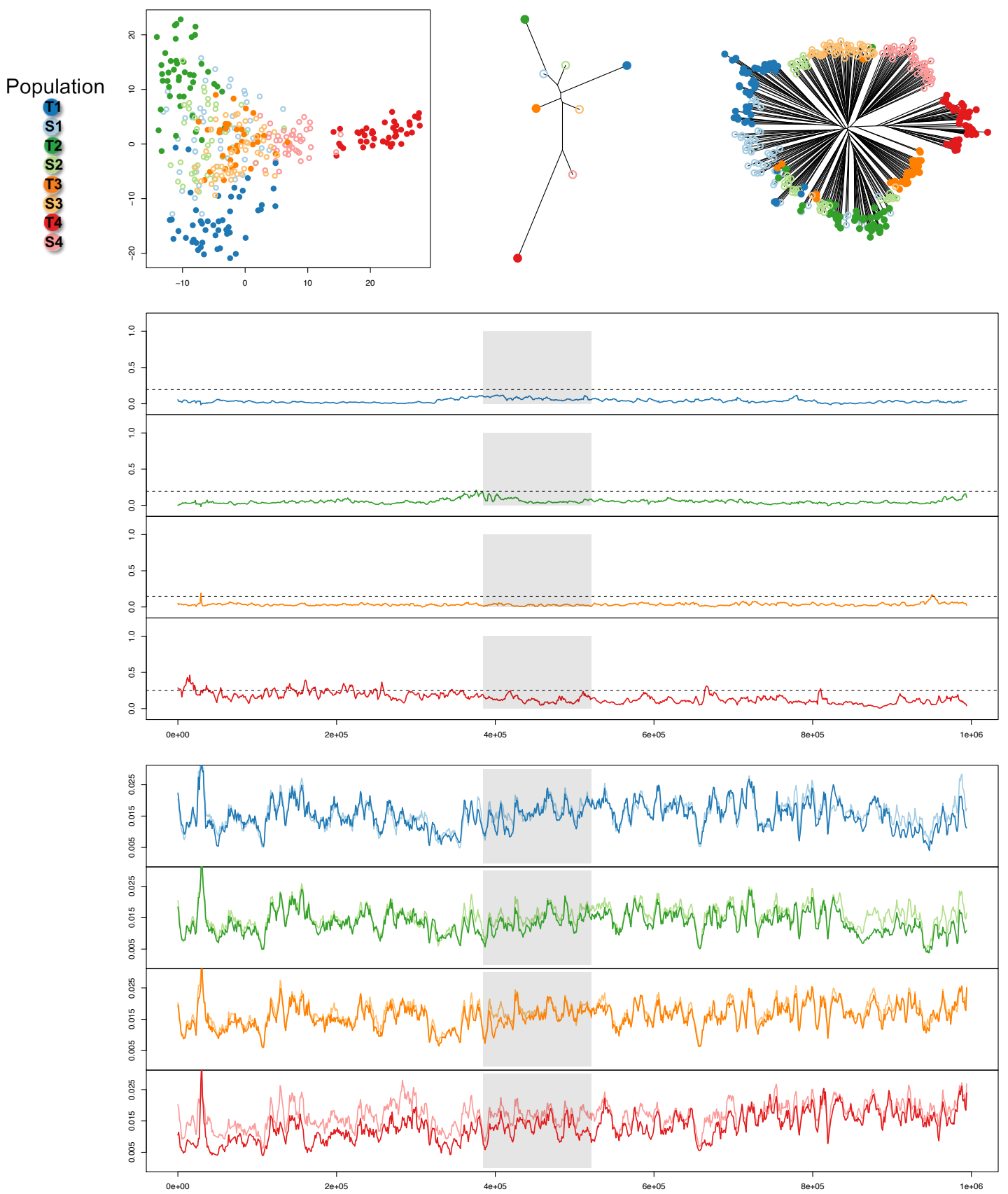

Fig. S19. Signatures of selection in the outlier region containing gene RYR3. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for RYR3.

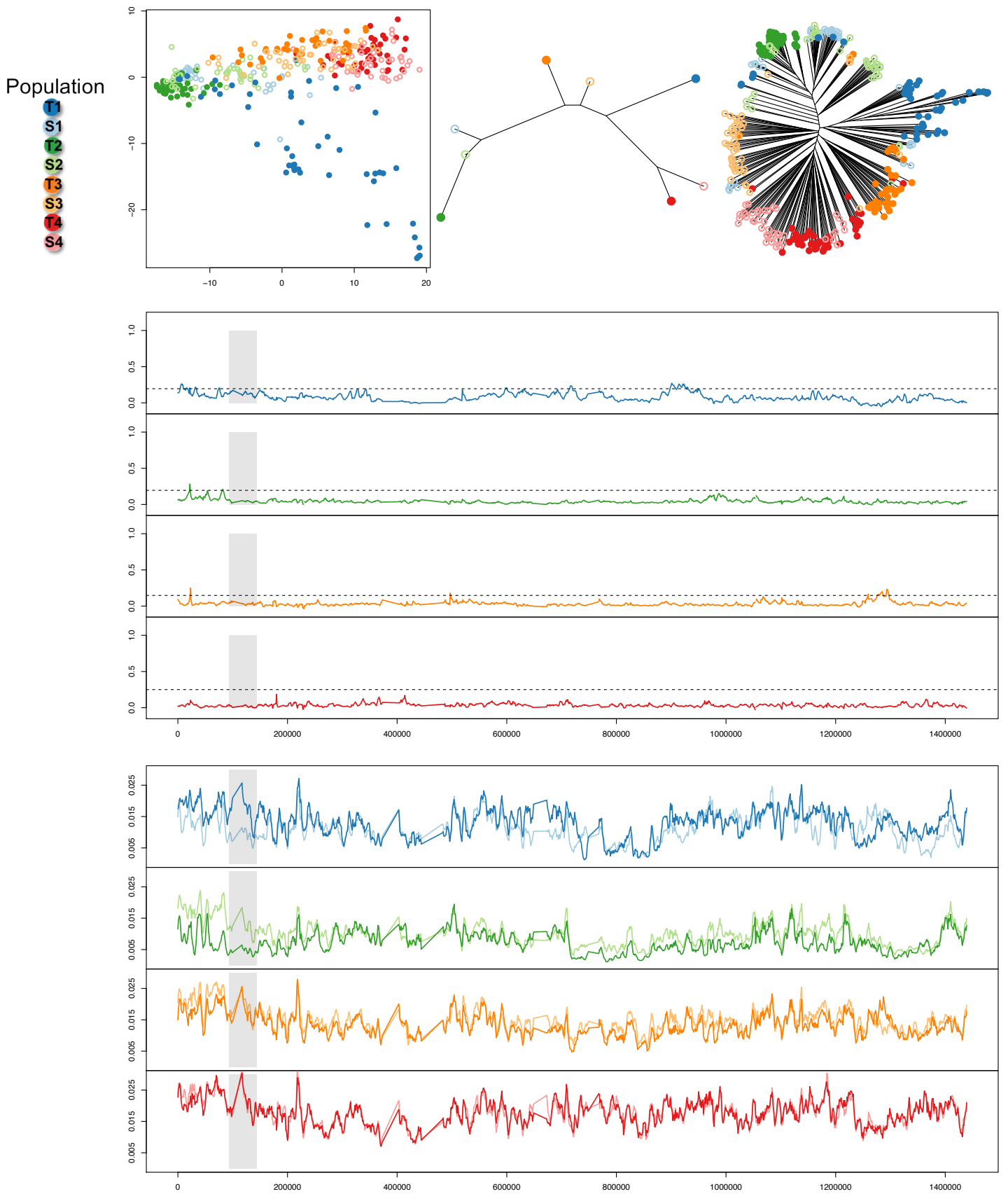

Fig. S20. Signatures of selection in the outlier region containing gene ESR2b. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for ESR2b.

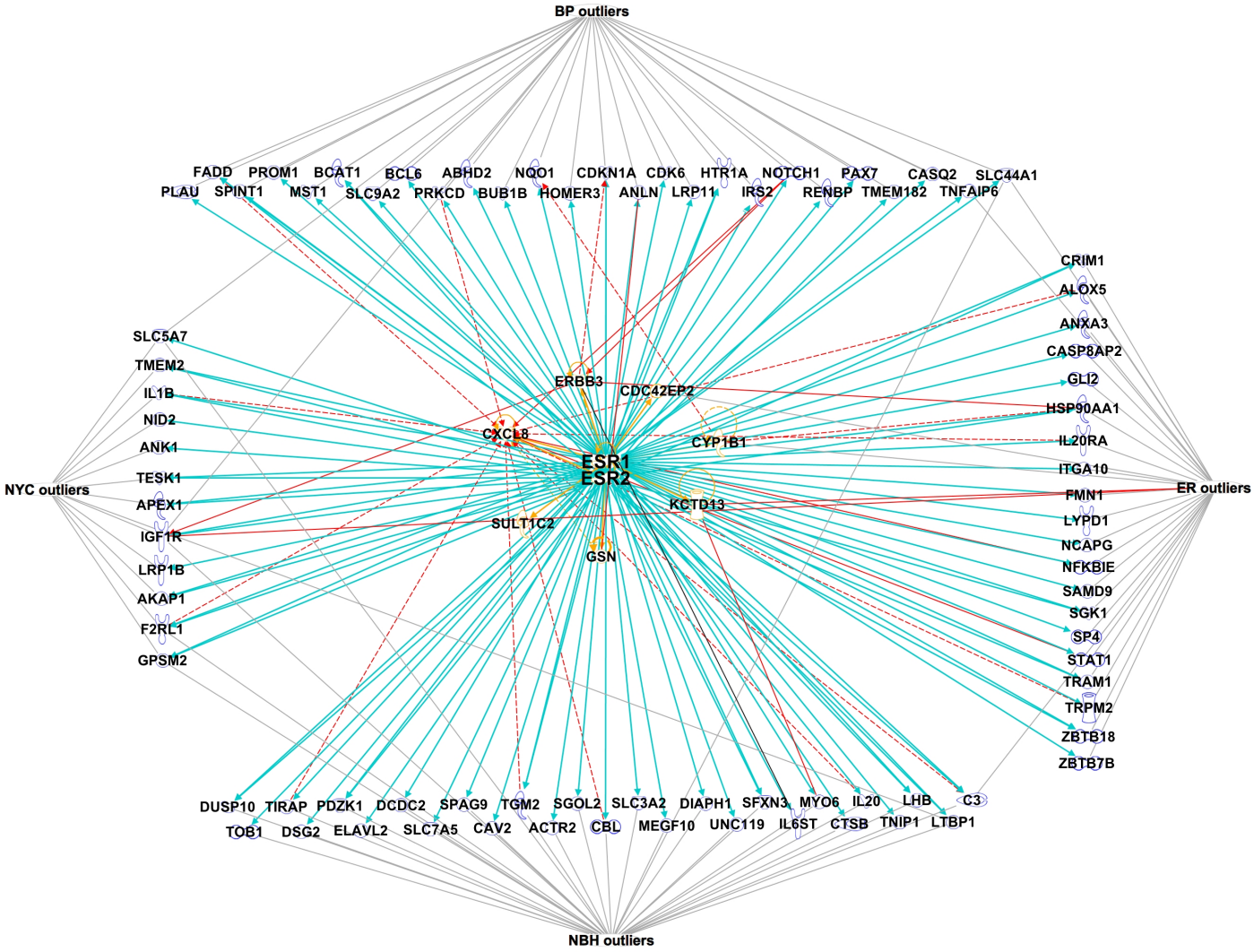

Fig. S21. Estrogen receptors (ESR) are in center. Genes that show differences in expression between tolerant and sensitive populations form the inner circle around ESRs (genes from Fig 2C). Genes that form the outer box are popgen outliers. Yellow lines indicate functional connection between ESR and genes with population-variable expression. Blue lines indicate functional connection between ESR and genes that are within population genomic outlier windows. Gray lines connect genes that are population genomic outliers to the population(s) within which they are outliers.

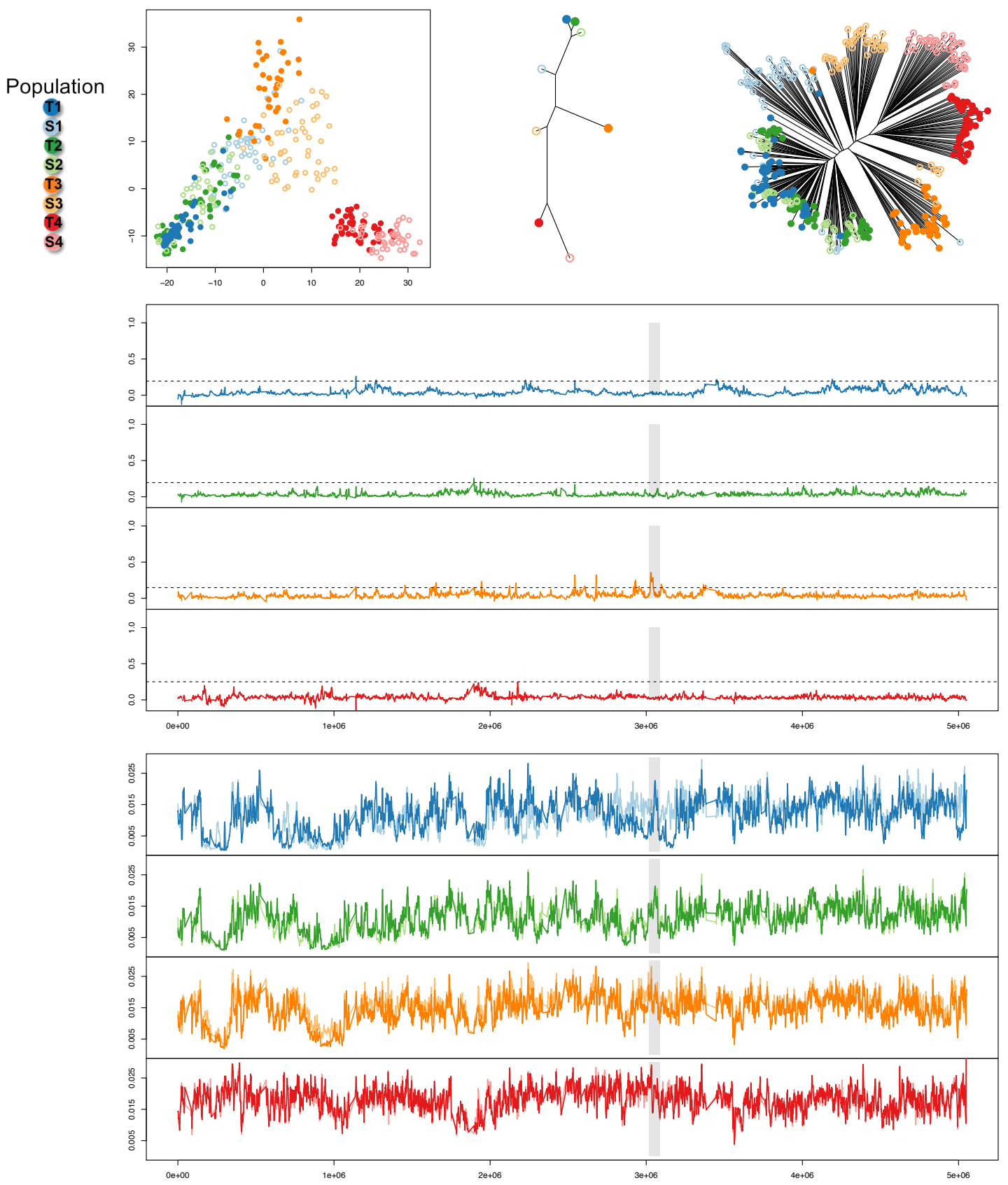

Fig. S22. Signatures of selection in the outlier region containing gene HIF2α. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for HIF2α.

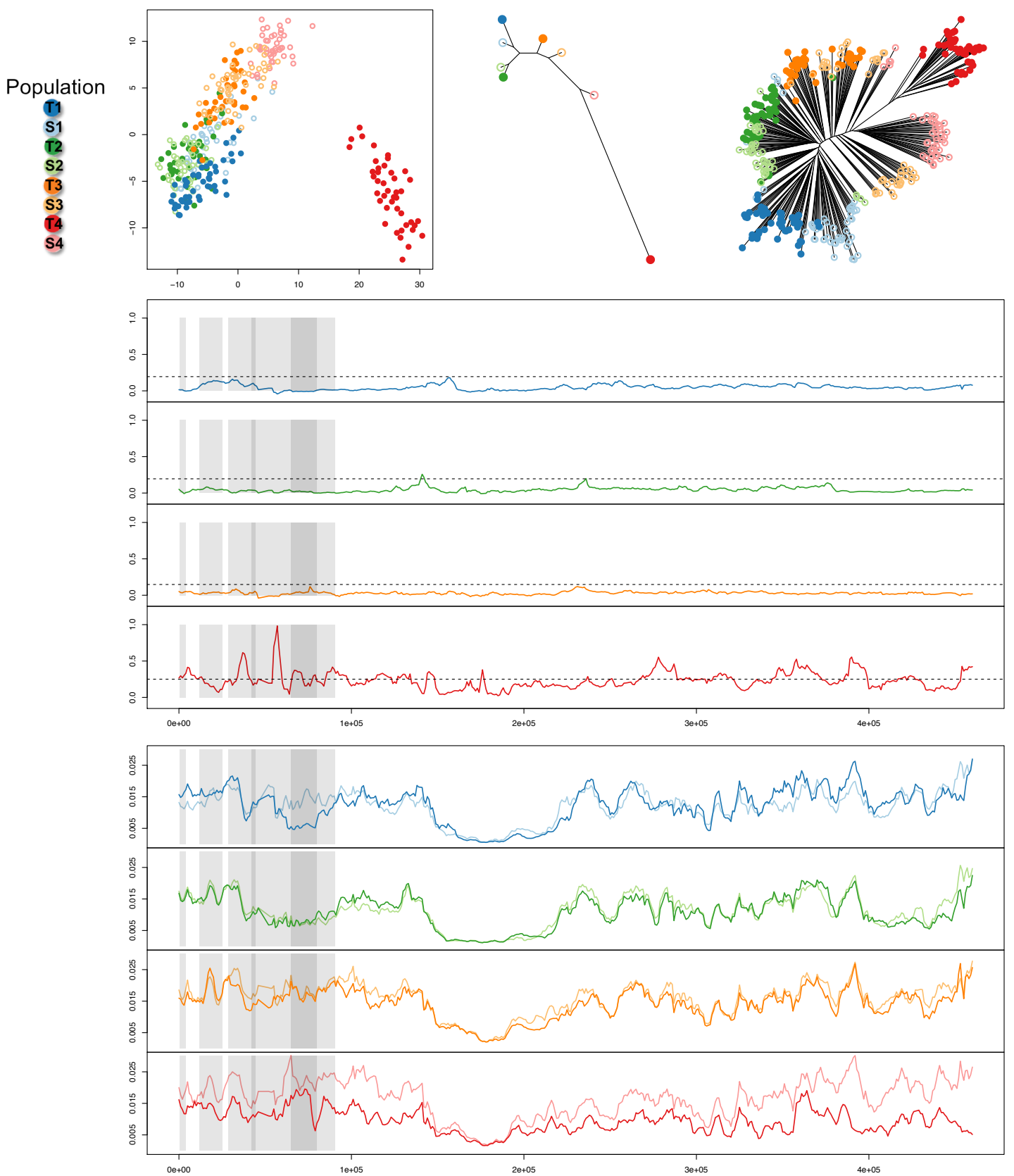

Fig. S23. Signatures of selection in the outlier region containing a cluster of immune system genes. Top panel (A) includes plots of genetic differentiation including individual MDS plots (left), population phylogenetic tree (middle), and individual phylogenetic tree (right). Middle panel is Fst between S-T population pairs, where the horizontal dotted line is the outlier threshold. Bottom panel is nucleotide diversity (pi) difference (Sensitive pi – Tolerant pi) for S-T population pairs. Gray panels indicate position of gene model for several cytokine receptors.

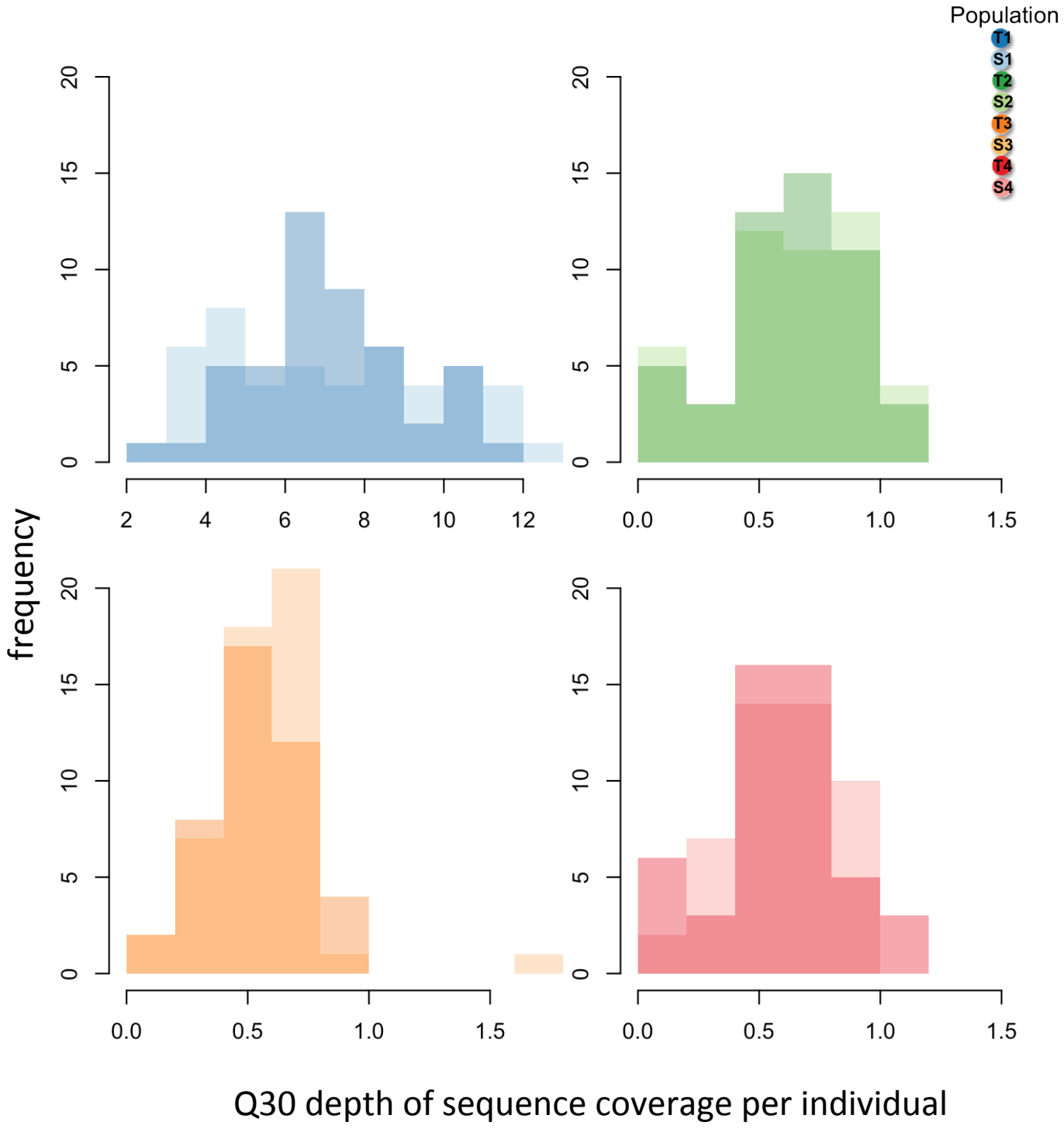

Fig. S24. Histogram of depth of coverage for individual samples for all eight populations.

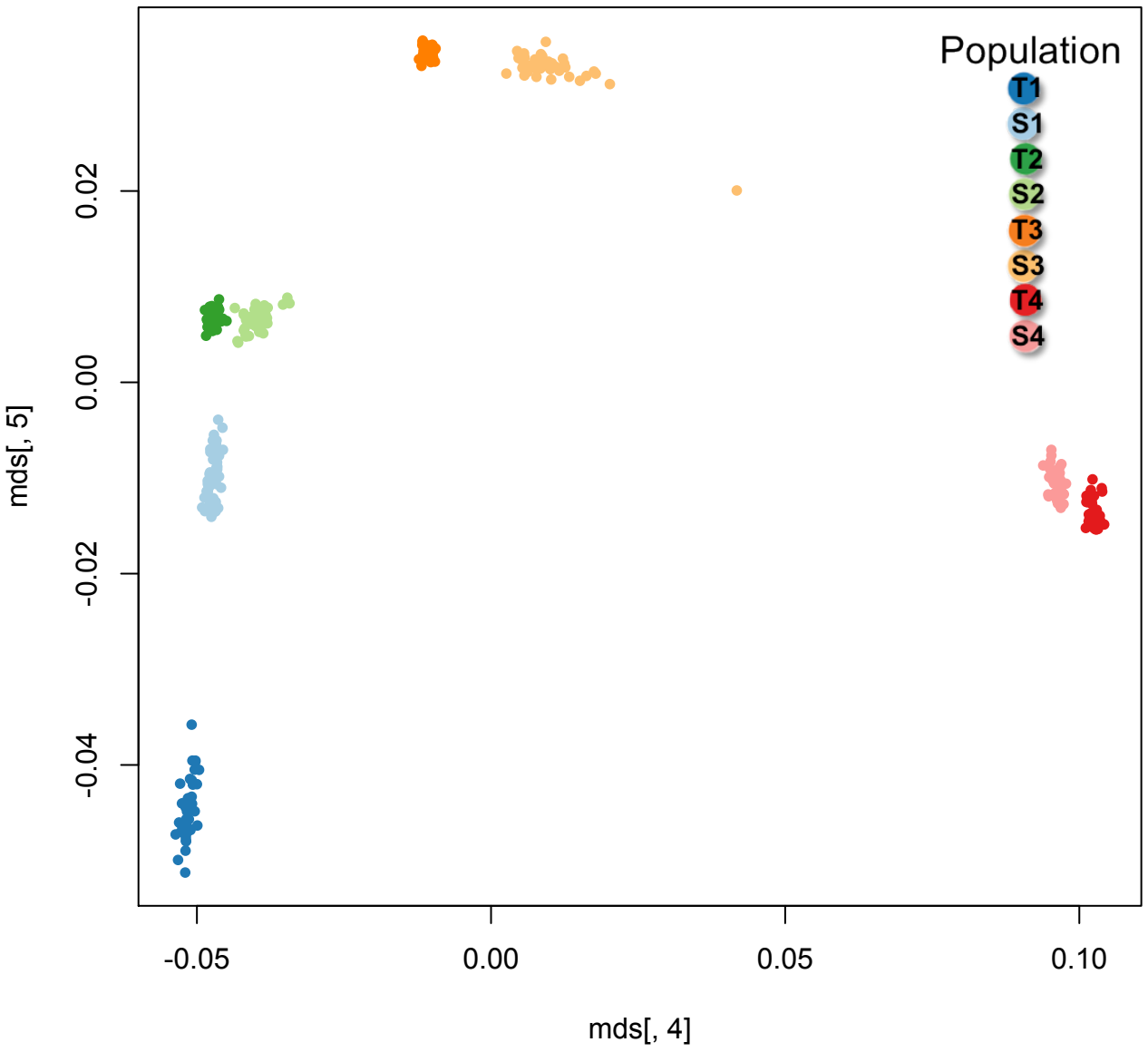

Fig. S25. Multi-dimensional scaling (MDS) plot of genome-wide genotypic variation for all individuals. Sampling sites are distinct populations and paired tolerant-reference sites are most similar to one another.

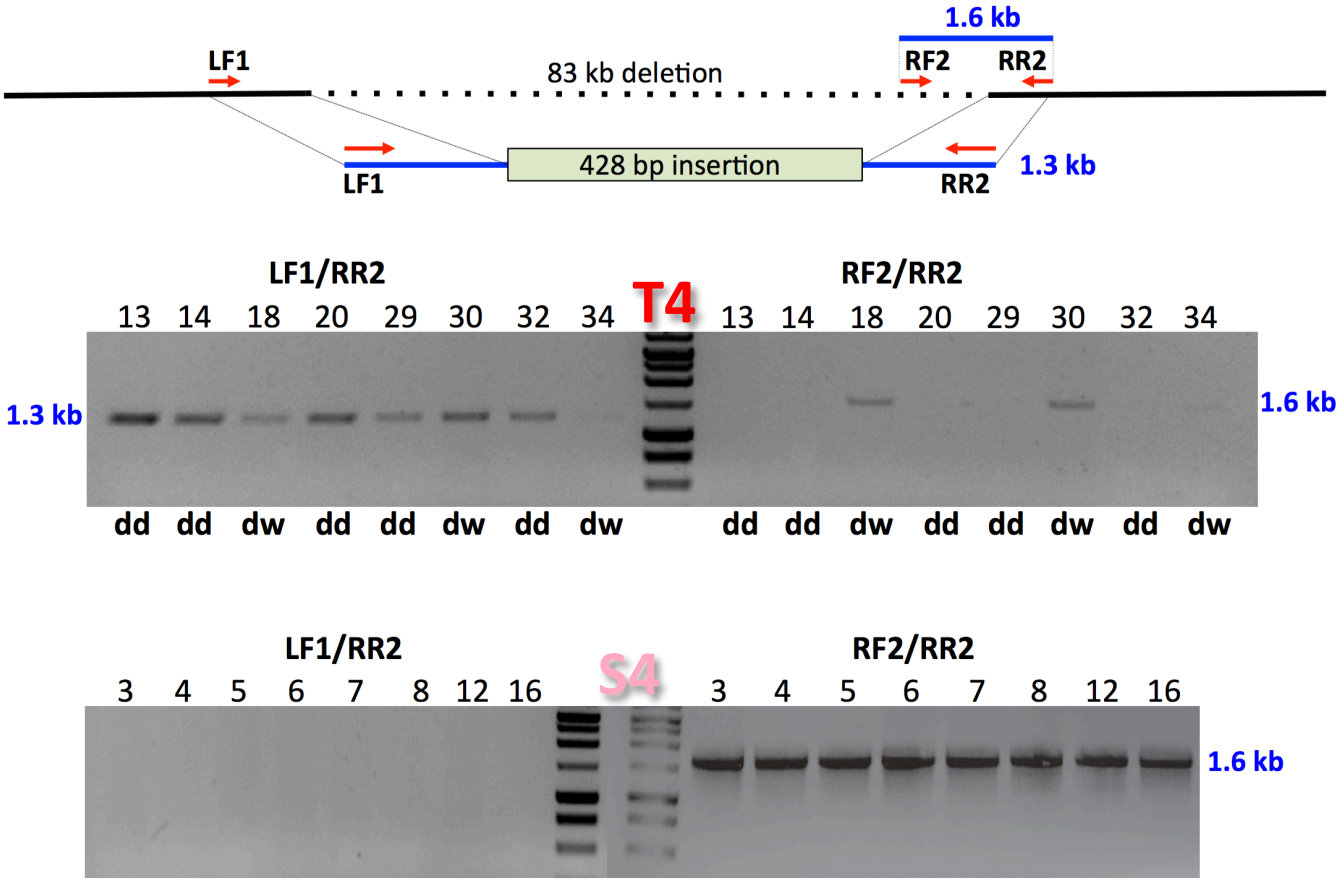

Fig. S26. Confirmation of the deletion spanning AHR2a and AHR1a (Fig. 3A) by PCR. Eight individual fish from each of T4 and S4 populations were assayed. Genomic DNA samples from these fish were amplified with primers flanking the left and right junctions of the deleted region (LF1/RR2), as well as within the deletion (RF2/RR2). Numbers above the lanes indicate fish ID numbers. Primers straddling the deletion (LF1/RR2) resulted in a 1.3 kb fragment in all T4 fish (lanes to the left of the ladder in the T4 gel image), whereas no amplification product was observed in any of the S4 fish (lanes to the left of the ladder in the S4 gel image). The 1.3 kb products from fish #13 and 14 were sequenced and found to match the genomic sequence flanking the deleted region, except for a 428 bp insertion. The insertion aligned perfectly to a different scaffold in the reference genome, in addition to multiple other scaffolds with high % identity. The RF2/RR2 primer pair produced the expected 1.6 kb product from all S4 fish, and only from the ER fish #18, 30, and 34. Deletion heterozygotes were annotated as “dw”, and deletion homozygotes as “dd”.

Supplemental tables are separate excel files that are hosted online.

Table S1. Populations sampled, including site and population characteristics.

Table S2. All gene models in the *Fundulus heteroclitus* genome, including scaffold number and position, annotation, and ranking of outlier region if the gene model was found within an outlier region for each population pair (right 4 columns).

Table S3. Expression levels for genes showing a different transcriptional response to PCB exposure between tolerant and sensitive populations (significant dose-by-population interaction).

Table S4. Expression levels for genes known to be targets of ligand-activated AHR.

References and Notes

1. A. P. Hendry, T. J. Farrugia, M. T. Kinnison, Human influences on rates of phenotypic change in wild animal populations. *Mol. Ecol.* **17**, 20–29 (2008). doi:10.1111/j.1365-294X.2007.03428.x Medline

2. G. Bell, Evolutionary rescue and the limits of adaptation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120080 (2012). doi:10.1098/rstb.2012.0080 Medline

3. I. Valiela, J. E. Wright, J. M. Teal, S. B. Volkmann, Growth, production and energy transformations in the salt-marsh killifish *Fundulus heteroclitus*. *Mar. Biol.* **40**, 135–144 (1977). doi:10.1007/BF00396259

4. D. Nacci, L. Coiro, D. Champlin, S. Jayaraman, R. McKinney, T. R. Gleason, W. R. Munns Jr., J. L. Specker, K. R. Cooper, Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Mar. Biol.* **134**, 9–17 (1999). doi:10.1007/s002270050520

5. D. Nacci, D. Champlin, S. Jayaraman, Adaptation of the estuarine fish *Fundulus heteroclitus* (Atlantic killifish) to polychlorinated biphenyls (PCBs). *Estuaries Coasts* **33**, 853–864 (2010). doi:10.1007/s12237-009-9257-6

6. Materials and methods are available as supplementary materials on *Science* Online.

7. D. D. Duvernell, J. B. Lindmeier, K. E. Faust, A. Whitehead, Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus*. *Mol. Ecol.* **17**, 1344–1360 (2008). doi:10.1111/j.1365-294X.2007.03648.x Medline

8. R. Pohjanvirta, *The AH Receptor in Biology and Toxicology* (Wiley, Hoboken, NJ, 2012).

9. A. Whitehead, W. Pilcher, D. Champlin, D. Nacci, Common mechanism underlies repeated evolution of extreme pollution tolerance. *Proc. Biol. Sci.* **279**, 427–433 (2012). doi:10.1098/rspb.2011.0847 Medline

10. A. M. Reitzel, S. I. Karchner, D. G. Franks, B. R. Evans, D. Nacci, D. Champlin, V. M. Vieira, M. E. Hahn, Genetic variation at aryl hydrocarbon receptor (AHR) loci in populations of Atlantic killifish (*Fundulus heteroclitus*) inhabiting polluted and reference habitats. *BMC Evol. Biol.* **14**, 6 (2014). doi:10.1186/1471-2148-14-6 Medline

11. B. W. Clark, C. W. Matson, D. Jung, R. T. Di Giulio, AHR2 mediates cardiac teratogenesis of polycyclic aromatic hydrocarbons and PCB-126 in Atlantic killifish (*Fundulus heteroclitus*). *Aquat. Toxicol.* **99**, 232–240 (2010). doi:10.1016/j.aquatox.2010.05.004 Medline

12. D. Nacci, D. Proestou, D. Champlin, J. Martinson, E. R. Waits, Genetic basis for rapidly evolved tolerance in the wild: Adaptation to toxic pollutants by an estuarine fish species. *Mol. Ecol.* **25**, 5467–5482 (2016). doi:10.1111/mec.13848 Medline

13. I. Wirgin, N. K. Roy, M. Loftus, R. C. Chambers, D. G. Franks, M. E. Hahn, Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. *Science* **331**, 1322–1325 (2011). doi:10.1126/science.1197296 Medline

14. M. Nukaya, B. C. Lin, E. Glover, S. M. Moran, G. D. Kennedy, C. A. Bradfield, The aryl hydrocarbon receptor-interacting protein (AIP) is required for dioxin-induced hepatotoxicity but not for the induction of the *Cyp1a1* and *Cyp1a2* genes. *J. Biol. Chem.* **285**, 35599–35605 (2010). doi:10.1074/jbc.M110.132043 Medline

15. D. A. Proestou, P. Flight, D. Champlin, D. Nacci, Targeted approach to identify genetic loci associated with evolved dioxin tolerance in Atlantic killifish (*Fundulus heteroclitus*). *BMC Evol. Biol.* **14**, 7 (2014). doi:10.1186/1471-2148-14-7 Medline

16. J. V. Schmidt, G. H. T. Su, J. K. Reddy, M. C. Simon, C. A. Bradfield, Characterization of a murine *Ahr* null allele: Involvement of the Ah receptor in hepatic growth and development. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6731–6736 (1996). doi:10.1073/pnas.93.13.6731 Medline

17. M. S. Denison, A. A. Soshilov, G. He, D. E. DeGroot, B. Zhao, Exactly the same but different: Promiscuity and diversity in the molecular mechanisms of action of the aryl hydrocarbon (dioxin) receptor. *Toxicol. Sci.* **124**, 1–22 (2011). doi:10.1093/toxsci/kfr218 Medline

18. J. P. Incardona, M. G. Carls, H. Teraoka, C. A. Sloan, T. K. Collier, N. L. Scholz, Aryl hydrocarbon receptor-independent toxicity of weathered crude oil during fish development. *Environ. Health Perspect.* **113**, 1755–1762 (2005). doi:10.1289/ehp.8230 Medline

19. F. Brette, B. Machado, C. Cros, J. P. Incardona, N. L. Scholz, B. A. Block, Crude oil impairs cardiac excitation-contraction coupling in fish. *Science* **343**, 772–776 (2014). doi:10.1126/science.1242747 Medline

20. T. V. Beischlag, J. Luis Morales, B. D. Hollingshead, G. H. Perdew, The aryl hydrocarbon receptor complex and the control of gene expression. *Crit. Rev. Eukaryot. Gene Expr.* **18**, 207–250 (2008). doi:10.1615/CritRevEukarGeneExpr.v18.i3.20 Medline

21. P. M. Fernandez-Salguero, D. M. Hilbert, S. Rudikoff, J. M. Ward, F. J. Gonzalez, Aryl-hydrocarbon receptor-deficient mice are resistant to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-induced toxicity. *Toxicol. Appl. Pharmacol.* **140**, 173–179 (1996). doi:10.1006/taap.1996.0210 Medline

22. S. R. Greytak, A. M. Tarrant, D. Nacci, M. E. Hahn, G. V. Callard, Estrogen responses in killifish (*Fundulus heteroclitus*) from polluted and unpolluted environments are site- and gene-specific. *Aquat. Toxicol.* **99**, 291–299 (2010). doi:10.1016/j.aquatox.2010.05.009 Medline

23. J. J. Berg, G. Coop, A Coalescent Model for a Sweep of a Unique Standing Variant. *Genetics* **201**, 707–725 (2015). doi:10.1534/genetics.115.178962 Medline

24. B. Wilson, P. Pennings, D. Petrov, *bioRxiv* (2016). 10.1101/052993

25. E. M. Leffler, K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, M. Przeworski, Revisiting an old riddle: What determines genetic

diversity levels within species? *PLOS Biol.* **10**, e1001388 (2012). doi:10.1371/journal.pbio.1001388 Medline

26. D. E. Nacci, D. Champlin, L. Coiro, R. McKinney, S. Jayaraman, Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish *Fundulus heteroclitus*. *Environ. Toxicol. Chem.* **21**, 1525–1532 (2002). doi:10.1002/etc.5620210726 Medline

27. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

28. H. Li, *arXiv* preprint, arXiv:1303.3997 (2013).

29. G. G. Faust, I. M. Hall, SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014). doi:10.1093/bioinformatics/btu314 Medline

30. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

31. E. Garrison, G. Marth, arXiv preprint, arXiv: 1207.3907 (2012).

32. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358 (1984). doi:10.2307/2408641

33. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007). doi:10.1086/521987 Medline

34. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi:10.1086/519795 Medline

35. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014). doi:10.1186/s12859-014-0356-4 Medline

36. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002). doi:10.1093/bioinformatics/18.2.337 Medline

37. X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, P. Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, J. Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang,

R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, J. Wang, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010). doi:10.1126/science.1190371 Medline

38. A. Amores, J. Catchen, I. Nanda, W. Warren, R. Walter, M. Schartl, J. H. Postlethwait, A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. *Genetics* **197**, 625–641 (2014). doi:10.1534/genetics.114.164293 Medline

39. L. J. Revell, S. A. Chamberlain, Rphylip: An R interface for PHYLIP. *Methods Ecol. Evol.* **5**, 976–981 (2014). doi:10.1111/2041-210X.12233

40. A. M. Bolger, M. Lohse, B. Usadel, *Bioinformatics* 10.1093/bioinformatics/btu170 (2014).

41. M. D. MacManes, *bioRxiv* 10.1101/000422 (2014).

42. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). doi:10.1093/bioinformatics/btp120 Medline

43. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). Medline

44. S. P. Lund, D. Nettleton, D. J. McCarthy, G. K. Smyth, Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* **11**, 8 (2012). doi:10.1515/1544-6115.1826 Medline

45. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). doi:10.1093/bioinformatics/btp616 Medline