Science Advances
AAAS

# Supplementary Materials for

## Chromosomal inversions from an initial ecotypic divergence drive a gradual repeated radiation of Galápagos beetles

Carl Vangestel *et al.*

Corresponding author: Carl Vangestel, cvangestel@naturalsciences.be

*Sci. Adv.* **10**, eadk7906 (2024)
DOI: 10.1126/sciadv.adk7906

**The PDF file includes:**



**Other Supplementary Material for this manuscript includes the following:**

**Supplementary Text**

Structure Analysis

We used the individual-based Bayesian analysis implemented in STRUCTURE 2.3.4. (69) to assess the number of distinct genetic clusters (K) and level of admixture in 121 *Calosoma* beetles. After discarding RADtags located within the chromosomal inversions and those identified as outlier RADtags (i.e. containing outlier SNPs in at least two within-island ecotype comparisons), a single SNP was randomly selected from each of the remaining RADtags (n=900). On this SNP dataset we applied an admixture model with 10 independent replicate runs for each K = 2–8 using 100 000 Markov chain Monte Carlo repetitions with a burn-in period of 30 000, correlated allele frequencies and no prior information on the population of origin. All other default setting were retained.

We identified K=5 as the optimal cluster size as the log probability of the data reached an asymptote at this K value (Fig. S1). Results of the cluster analysis confirmed the gradual divergence of highland species with island age and confirmed results of the PCoA. At K=2 the highland species of the oldest island San Cristobal (*C. linelli*) formed a distinct genetic cluster from the remaining species/populations. Increasing the number of genetic clusters from K=3 to K=5 revealed respectively the lowland species at the oldest island San Cristobal, the highland species at Santa Cruz (*C. leleuporum*) and at Santiago (*C. galapageium*) as diverged genetic clusters (Fig. S2).

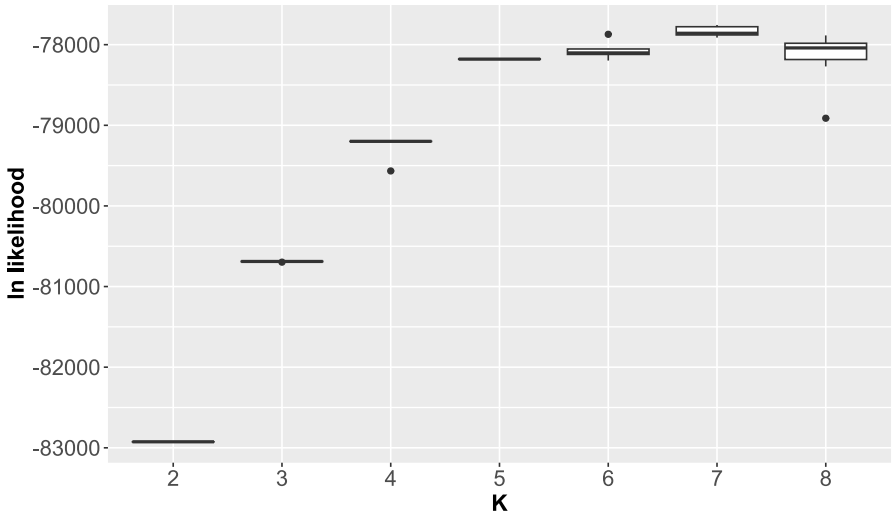



**Fig. S1 | Likelihood profile of Structure models with increasing number of genetic clusters (K).**

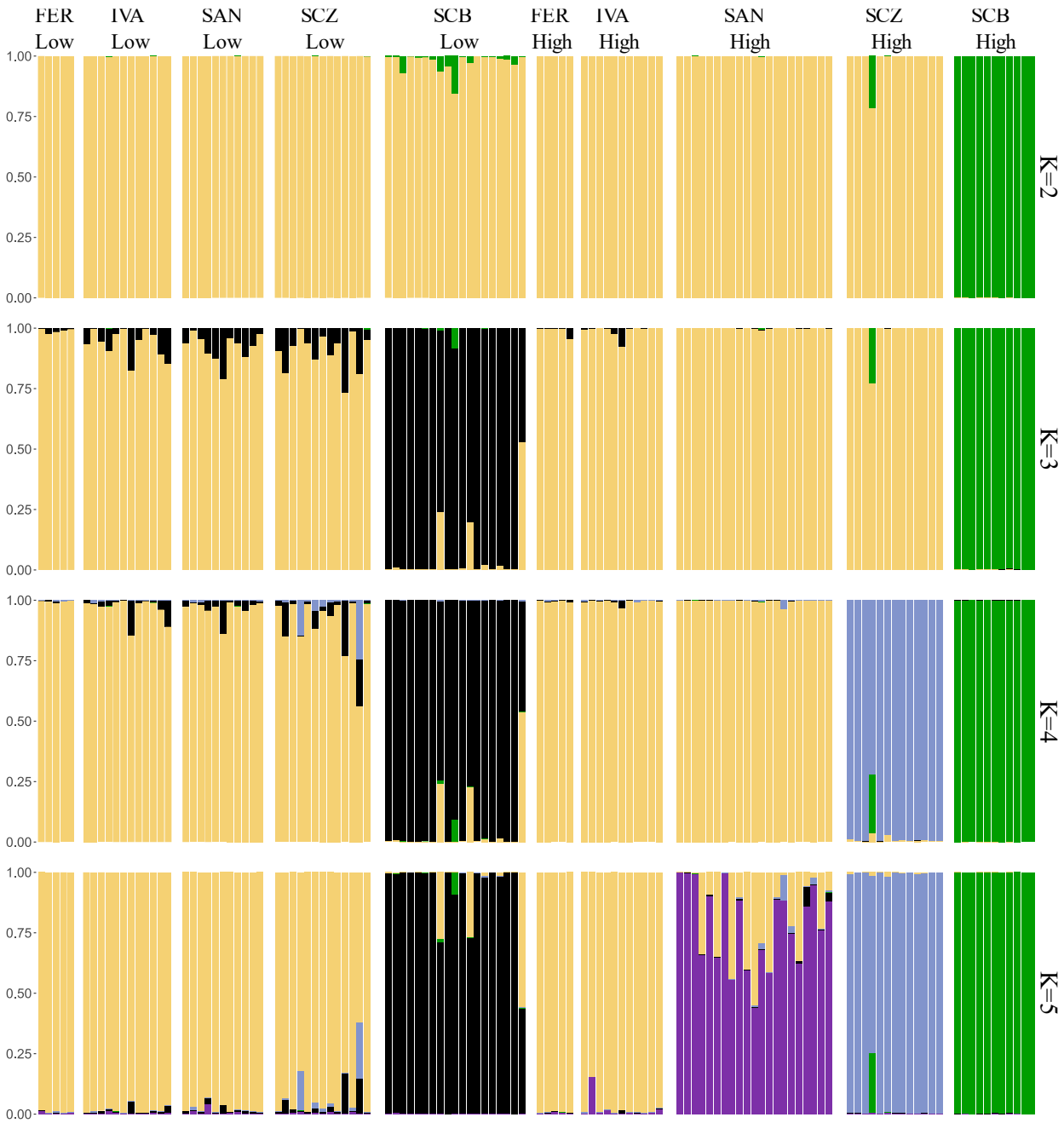



**Fig. S2 | Ancestry proportions per genetic cluster.** Each vertical bar represents one individual, and colors represent the proportional contributions of each genetic lineage K. As the outcome of different runs for each K resulted in consistent patterns, only the run with the highest likelihood for each K was plotted.

Admixture Analysis

We used TreeMix v1.13 (33) to estimate genetic relationships and admixture between 32 resequenced Calosoma specimens sampled across the archipelago and included *C. Sayi* as an outgroup. After inferring an initial maximum likelihood tree we allowed up to 10 migration events to improve the data fit. For each setting we ran the model for 10 iterations. As adding complexity to the model (i.e. migration events) will increase the likelihood, we attempted to delineate the optimal number of migration events as the one for which the mean likelihood reached a plateau, i.e. when the mean likelihood did not significantly increase when comparing it with that of a model containing 1 additional migration event. Pairwise t-tests indicated that at 6 migration events such a plateau was reached as the mean log likelihood of a model with 7 migration events did not significantly differ from one with 6 migration events ($t_{18}$=1.497, p=0.15) (Fig. S3).

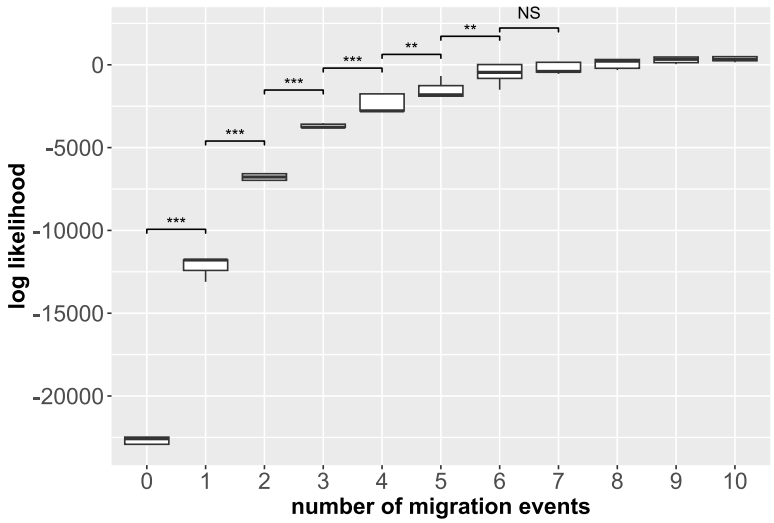



**Fig. S3 | Likelihood profile of TreeMix models with increasing number of migration events.** Plot depicts an increasing median log likelihood as more migration events are added to the model. Symbol above bracket connecting two adjacent number of migration events indicates the p-value of a t-test between their mean log likelihood (* p<0.05, ** p<0.01, *** p<0.001, NS p>0.05).

The corresponding tree suggested interspecific admixture or retention of ancestral variation between the highland species *C. linelli* (SCB High) inhabiting the oldest island and respectively i) the other highland species *C. galapageium* (SAN High) and *C. leleuporum* (SCZ High), ii) the lowland species *C. granatense* residing on the oldest island (SCB Low) and iii) an ancestor of *C. granatense* that gave rise to the lineages found on the youngest islands Isabela and Fernandina. Additional evidence for interspecific admixture was found between the highland species *C. galapageium* (SAN High) and the high- and lowland populations of *C. granatense* inhabiting the youngest islands (FER High, IVA High, IVA Low). Finally, TreeMix results indicated intraspecific admixture between the *C. granatense* populations residing on the youngest islands Fernandina and Isabela (Table S5).

We further explored admixture events across the archipelago by formally testing interspecific migration using $F_4$ statistics (34). When no admixture takes place, the $F_4(a,b;c,d)$ statistic should result in differences in allele frequencies between population *a* and *b* that are uncorrelated to those observed between population *c* and *d*, i.e. $F_4(a,b;c,d)=0$. We assigned the outgroup *C. sayi*

to population *a*, assuming that no admixture has taken place between this outgroup and any of the Calosoma species or populations of the Galápagos archipelago, such that a significant positive $f_4$ statistic would indicate admixture between populations *b* and *d*, and a significant negative $f_4$ statistic would point towards admixture between populations *b* and *c*. When testing the interspecific admixture events previously identified by TreeMix, we pooled samples according to the tree topology and migration events, i.e. specimens belonging to respectively i) *C. leleuporum* and *C. galapageium*, ii) the highland populations of Isabela and Fernandina, iii) the lowland populations of Isabela and Fernandina, and iv) the lowland populations Santa Cruz and Santiago. $F_4$ statistics were estimated using the *fourpop* module of TreeMix v1.13 (33) and associated standard errors were calculated in blocks of 500 SNPs.

Results from the $f_4$ analysis corroborated those indicated by the TreeMix graph (Table S5). Ancient variation of *C. linelli* (SCB High) could be traced back into the genomes of the other two highland species *C. leleuporum* and *C. galapageium*, the lowland species *C. granatense* of the oldest island and the highland populations of *C. granatense* of the youngest islands. The analysis further highlighted the absence of such admixture signatures between *C. linelli* and the lowland populations of *C. granatense* found on the youngest islands. In congruence with TreeMix analysis, we found statistical support for admixture between *C. galapageium* and the highland populations of *C. granatense* of the youngest islands.

**Fig. S4 | Structural variations (SV) underlie distinct alleles associated with high-lowland divergence. Each column shows the result of a single scaffold on which the SV was located. First row: $F_{ST}$ distribution (20kb windows) based on a comparison between all high- versus lowland individuals. Blue bottom line shows the location of the chromosomal inversion detected by *BreakDancer* (78). Red dots show the location of RADtags identified as outlier loci in at least one within-island highland-lowland comparison. Second row: Location of SNPs in perfect linkage disequilibrium ($r^2 = 1$). Grey dots above diagonal show $r^2 = 1$ values for all 32 resequenced individuals. Grey dots below diagonal show $r^2 = 1$ values for homozygous LL individuals only. Third row: PCoA based on SNPs located at the inversion (blue line in upper row panels). HH, LL and HL refer to the cluster of individuals genotyped at the inversion as homozygous for the highland allele (HH), lowland allele (LL) and heterozygous (HL). Forth row: Differences in nucleotide diversity at the inversion between individuals genotyped as HH, HL and LL in the third row panels. Fifth row: Maximum likelihood tree of the nucleotide sequence at the inversion excluding individuals genotyped as heterozygotes (HL). Node values represent bootstrap values based on 1000 replicates. The tree was rooted with the mainland species *C. sayi*. Sixth row: Relationship between individual nucleotide diversity at the inversion and the progression of the islands. Only individuals genotyped as homozygous for the lowland allele (LL, yellow) and highland allele (HH, remaining colors) are included. Color codes are as in Fig. 2.**

**Fig. S4 (continued).**

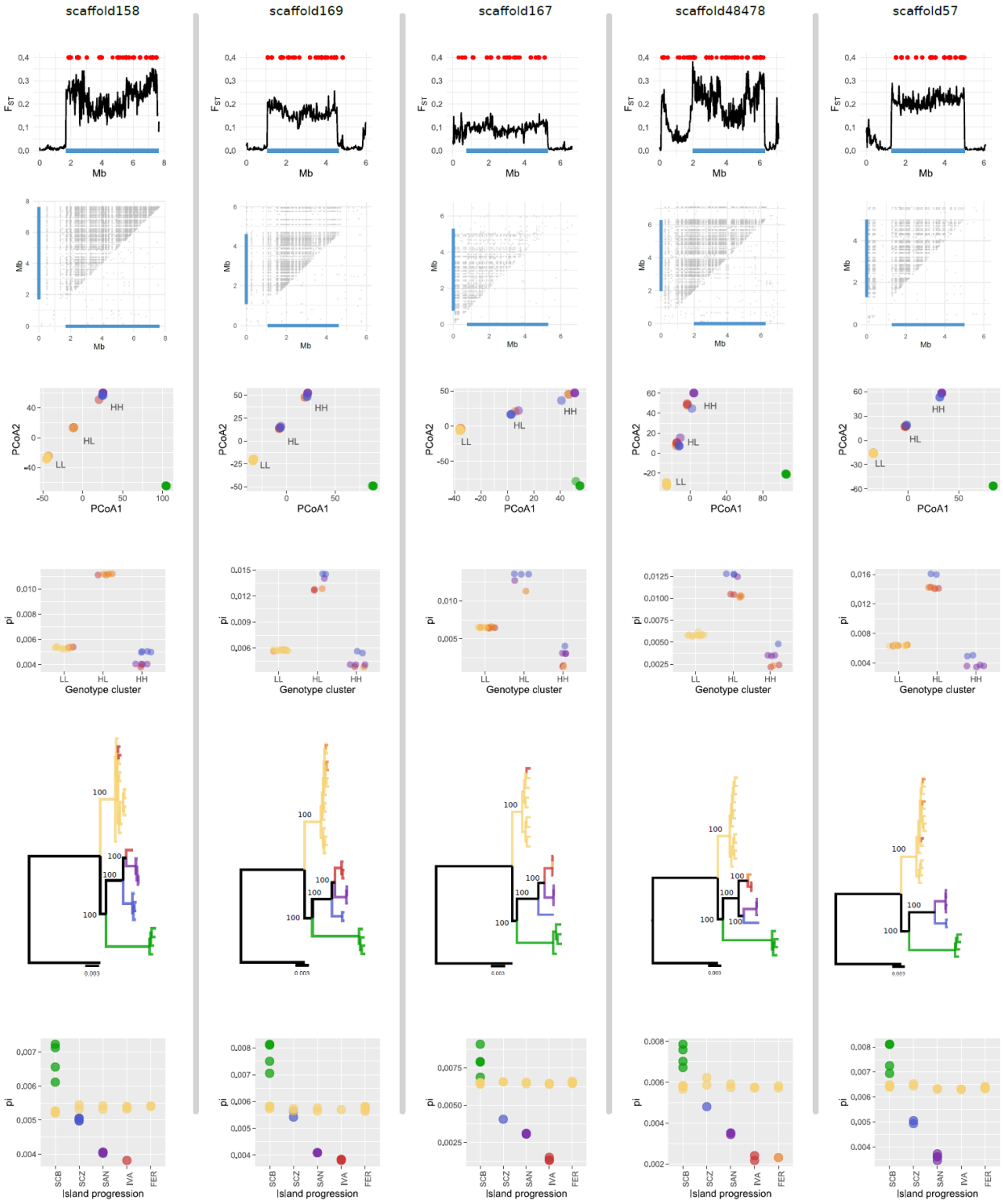


**Fig. S4 (continued).**

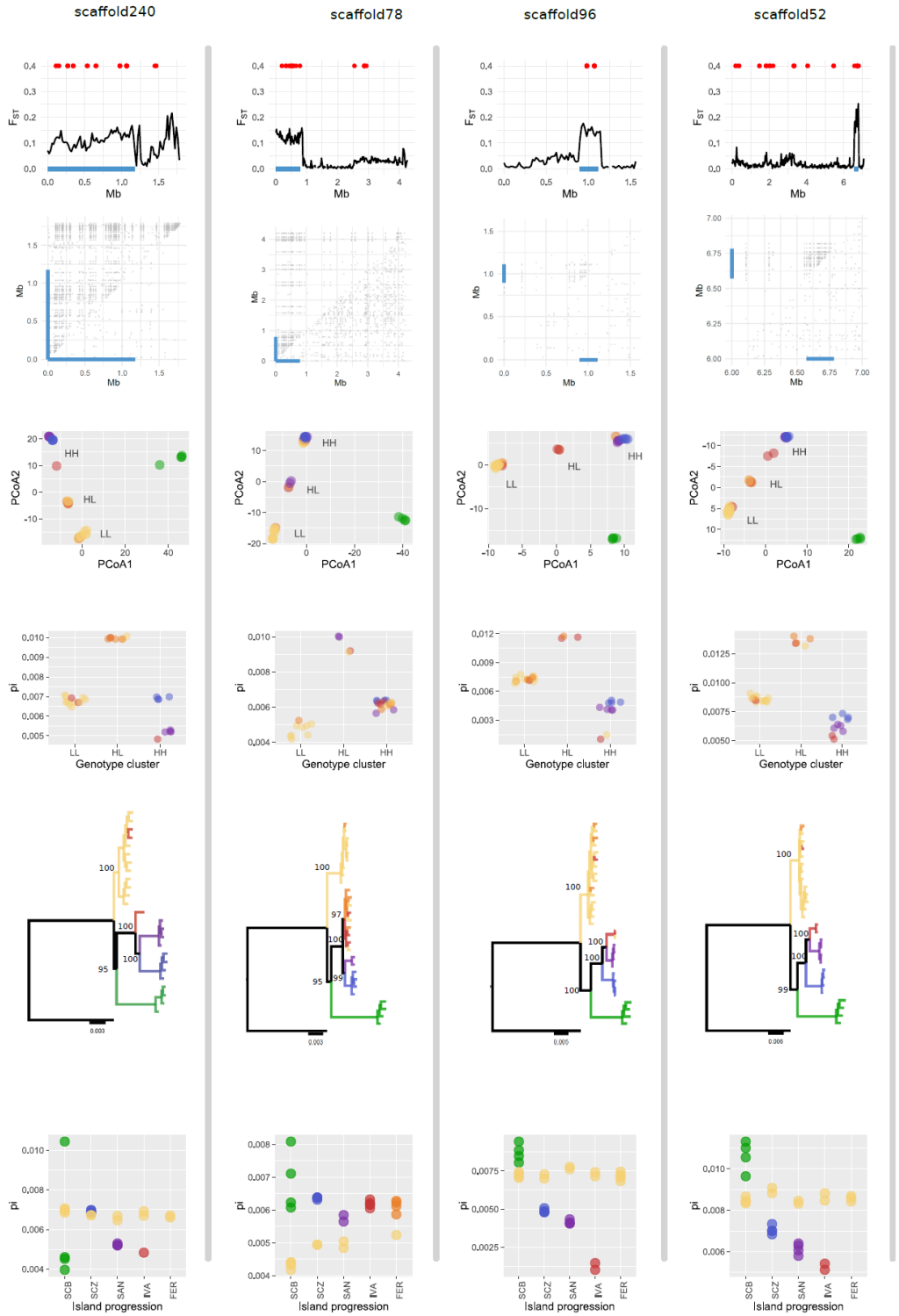


**Fig. S5 | Patterns of genetic differentiation and nucleotide diversity at scaffolds with structural variations associated with highland-lowland divergence. Upper panel denotes $F_{ST}$ distribution (20kb windows) based on a comparison between all high- versus lowland individuals. Blue bottom line shows the location of the chromosomal inversion detected by *BreakDancer* (78). Red dots show the location of RADtags identified as outlier loci in at least one within-island highland-lowland comparison. Lower panel illustrates increased nucleotide diversity at chromosomal inversions for heterozygous individuals compared to individuals homozygous for either lowland or highland allele.**

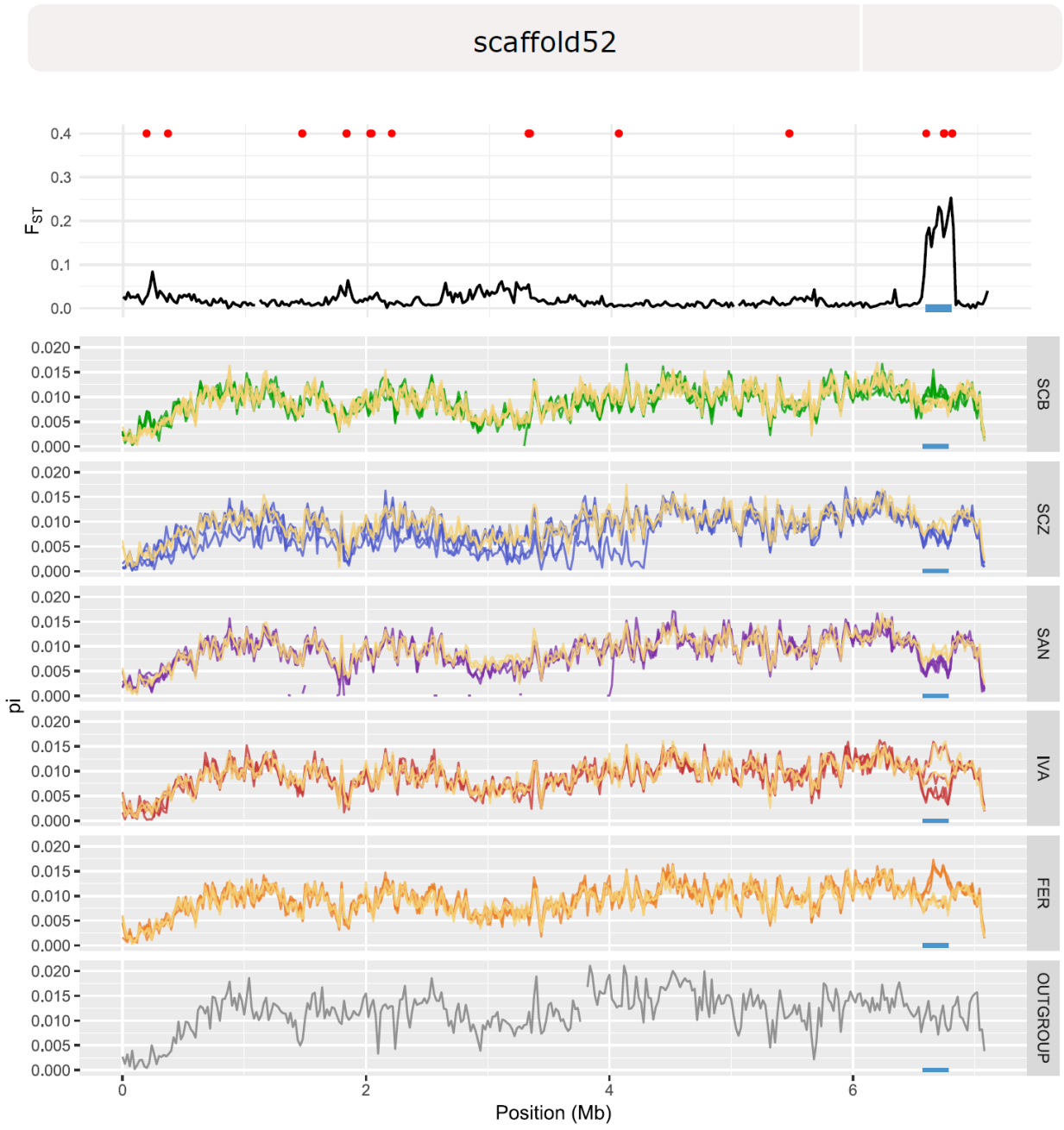


**Fig. S5 (continued).**

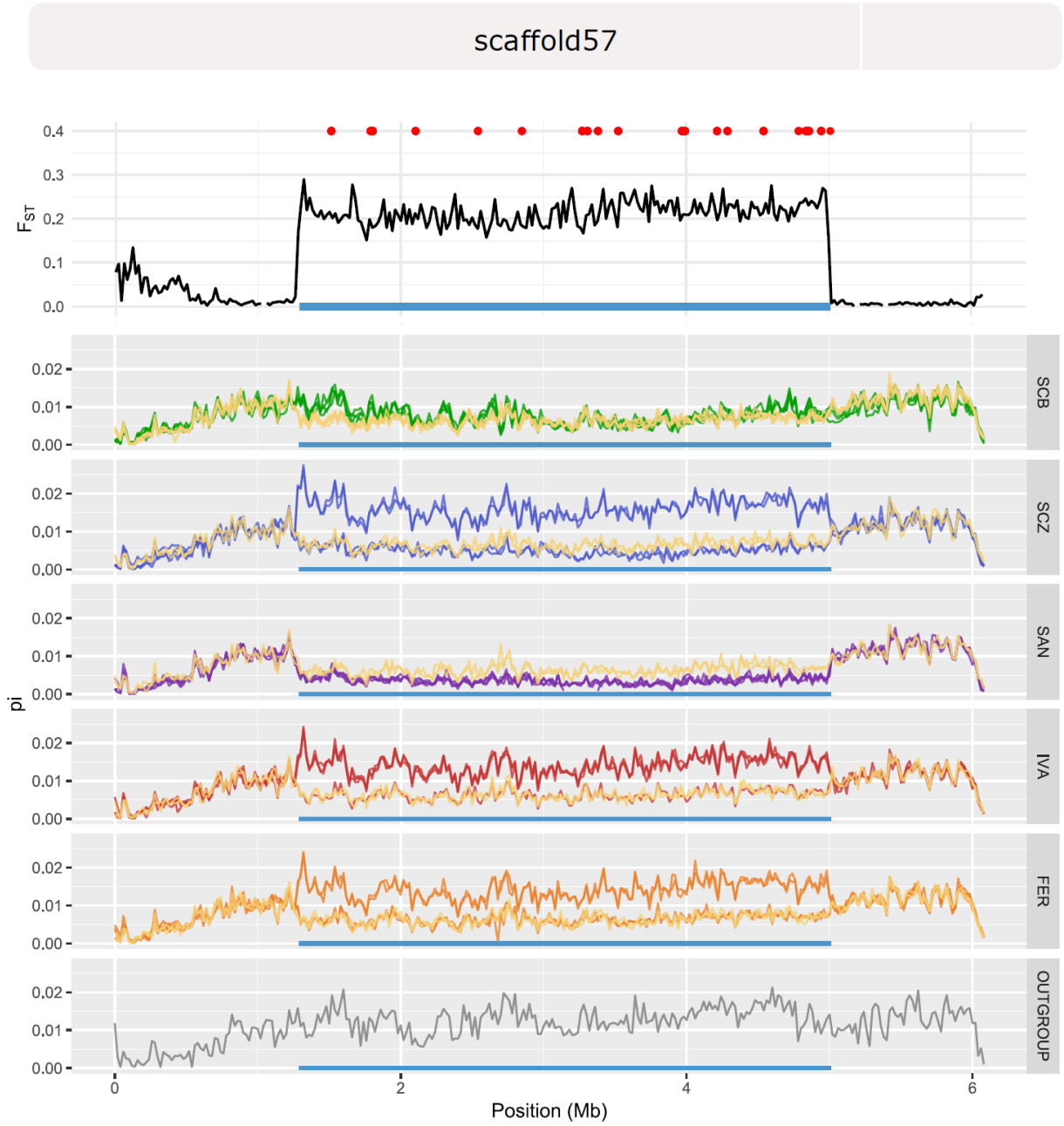


**Fig. S5 (continued).**

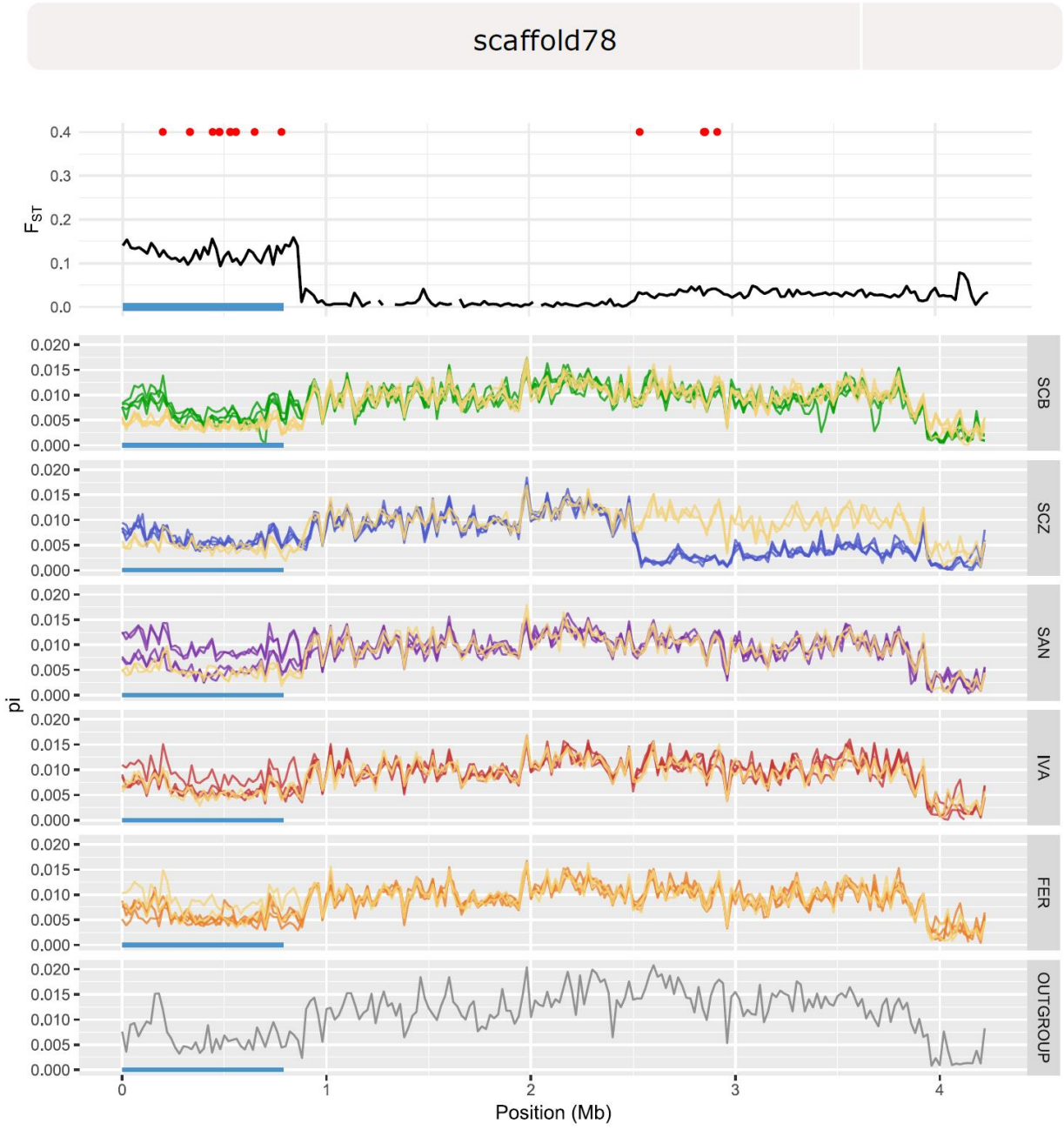


**Fig. S5 (continued).**

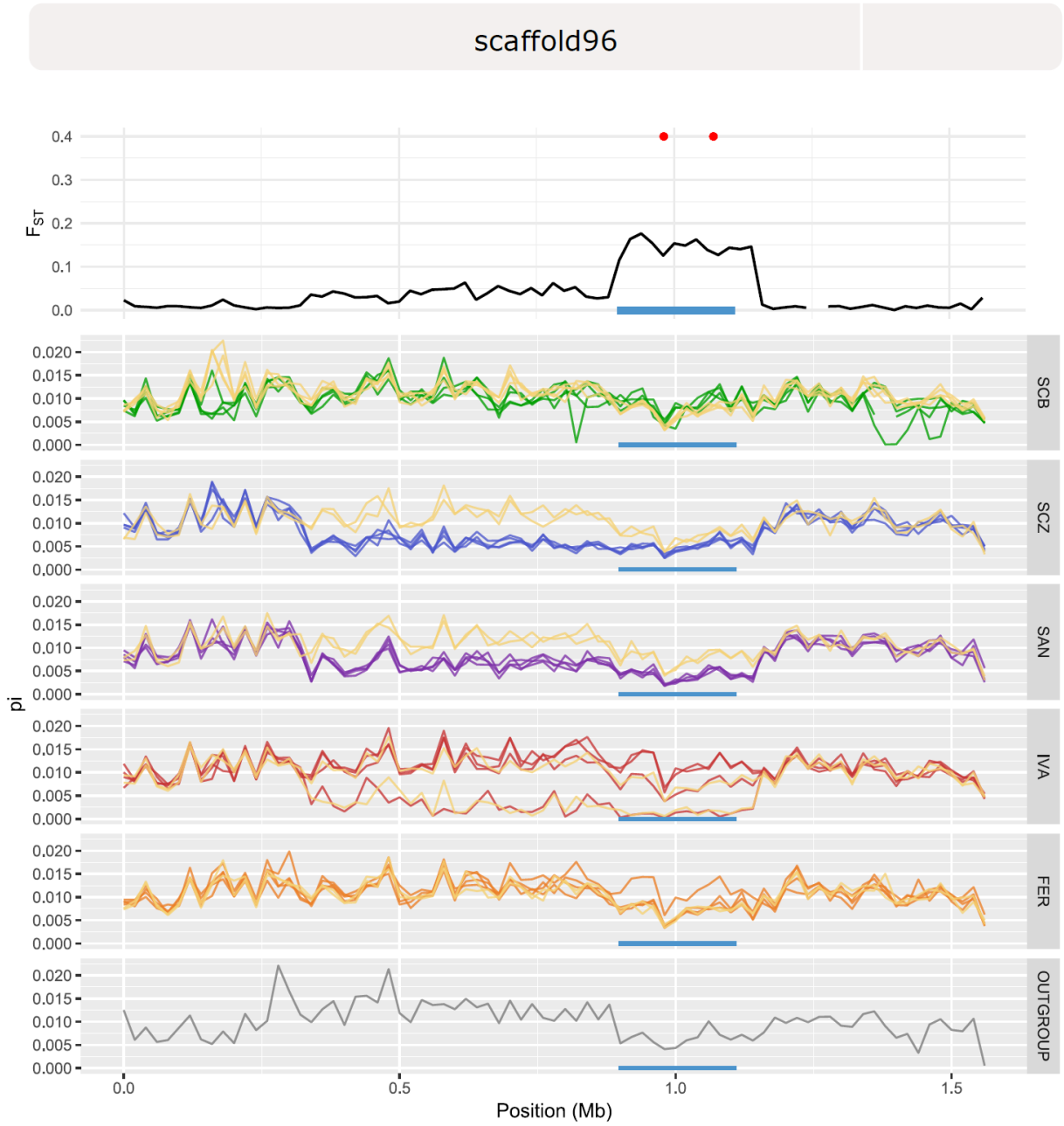


**Fig. S5 (continued).**

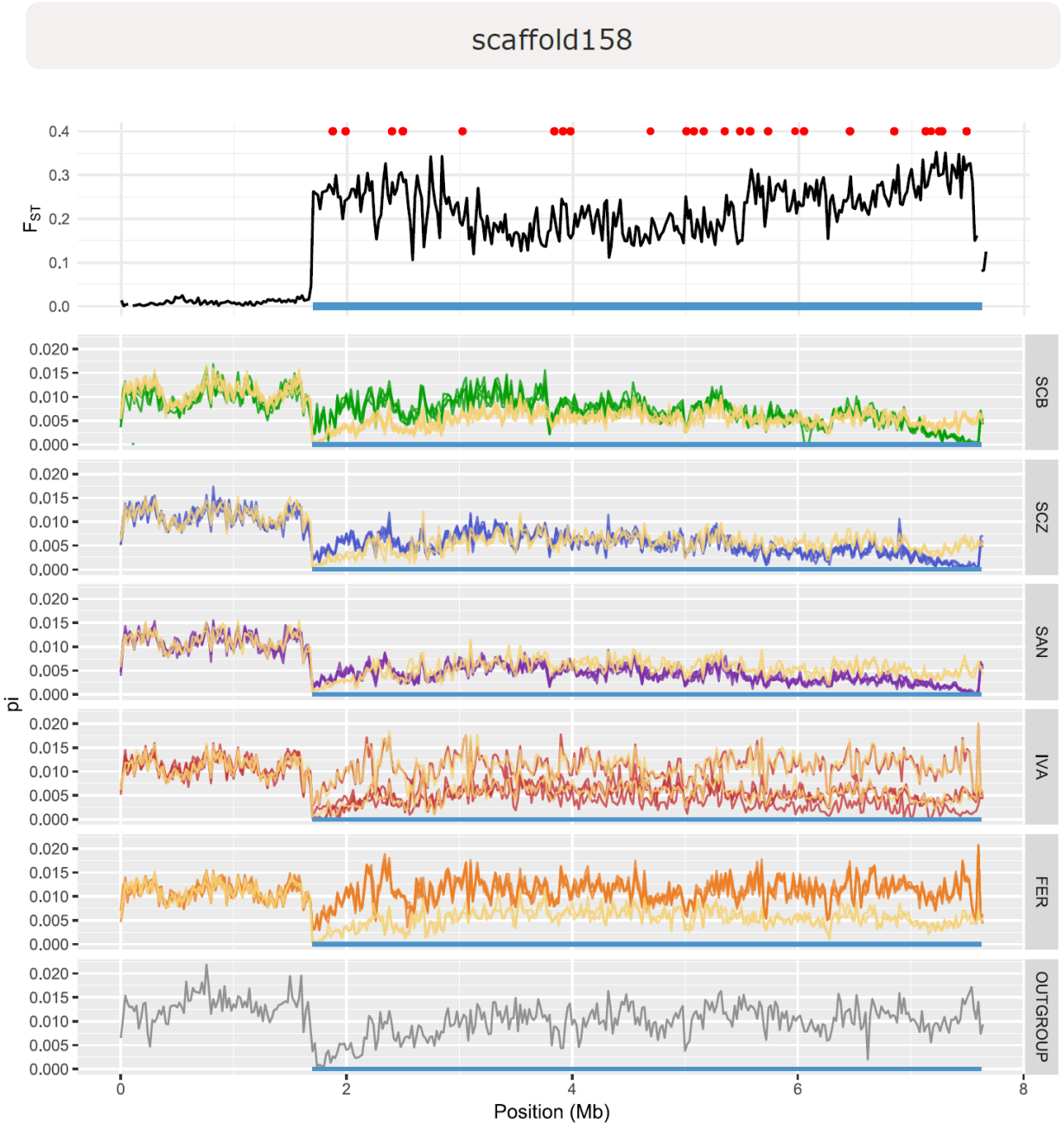


**Fig. S5 (continued).**

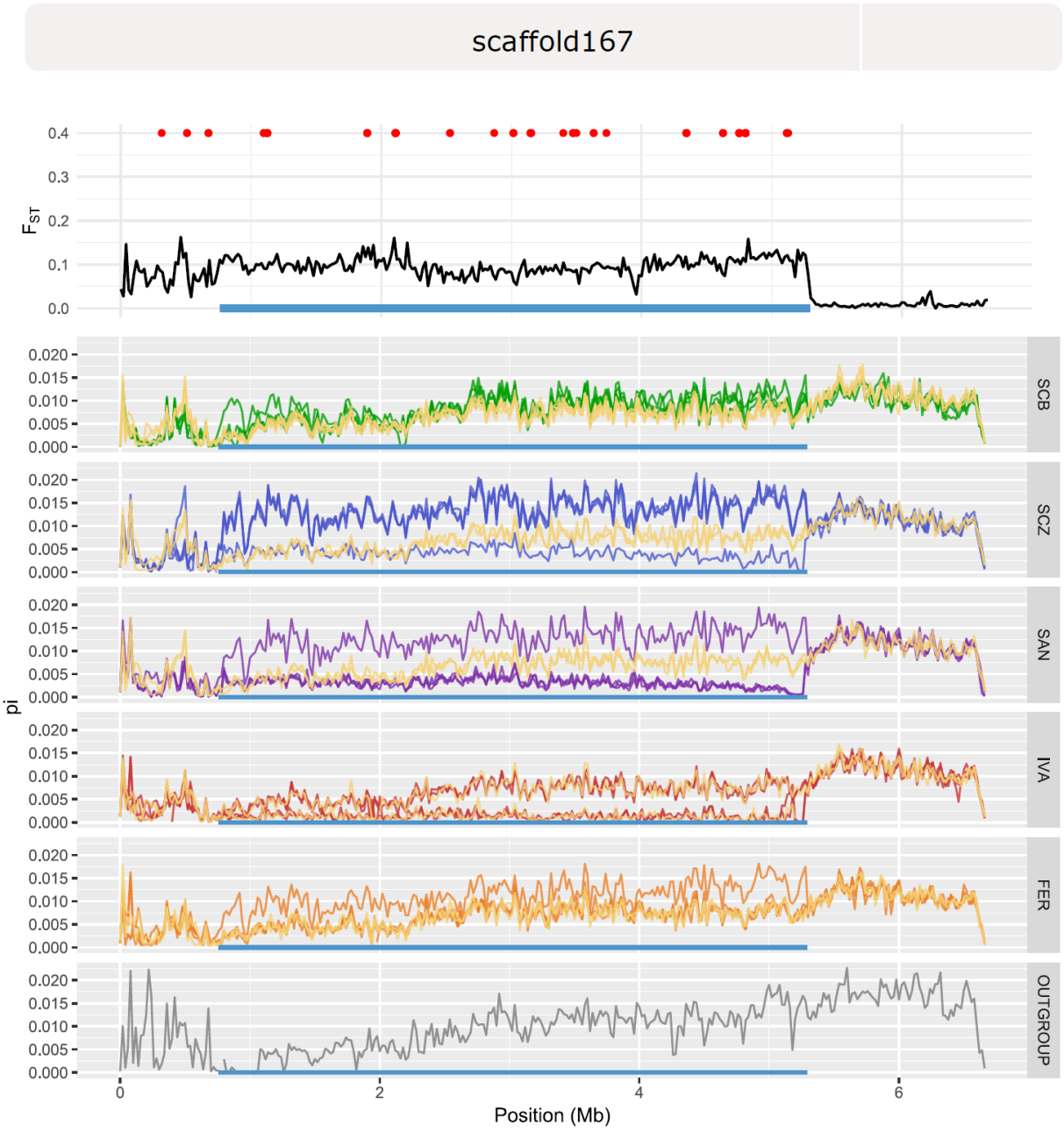


**Fig. S5 (continued).**

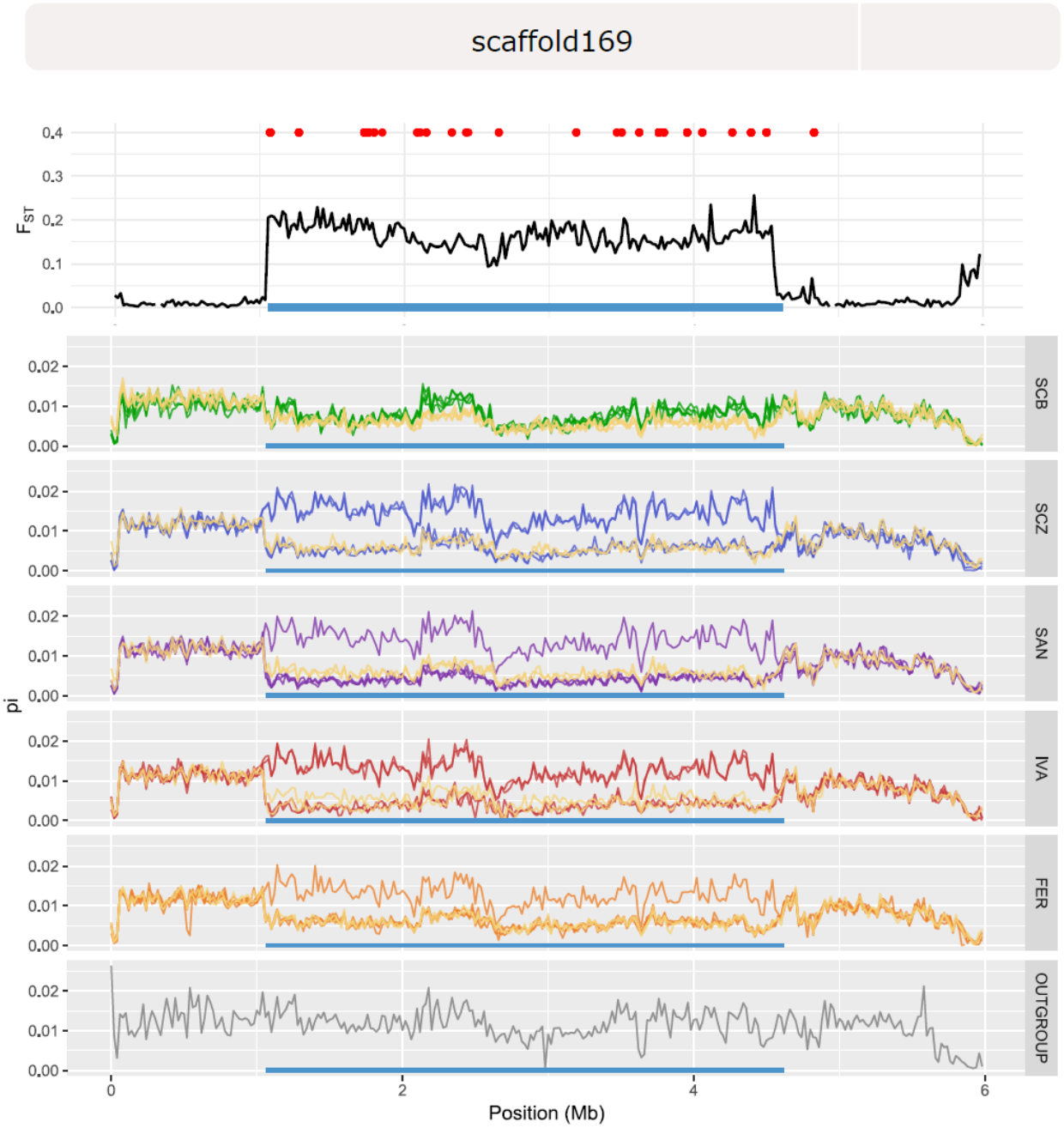


**Fig. S5 (continued).**

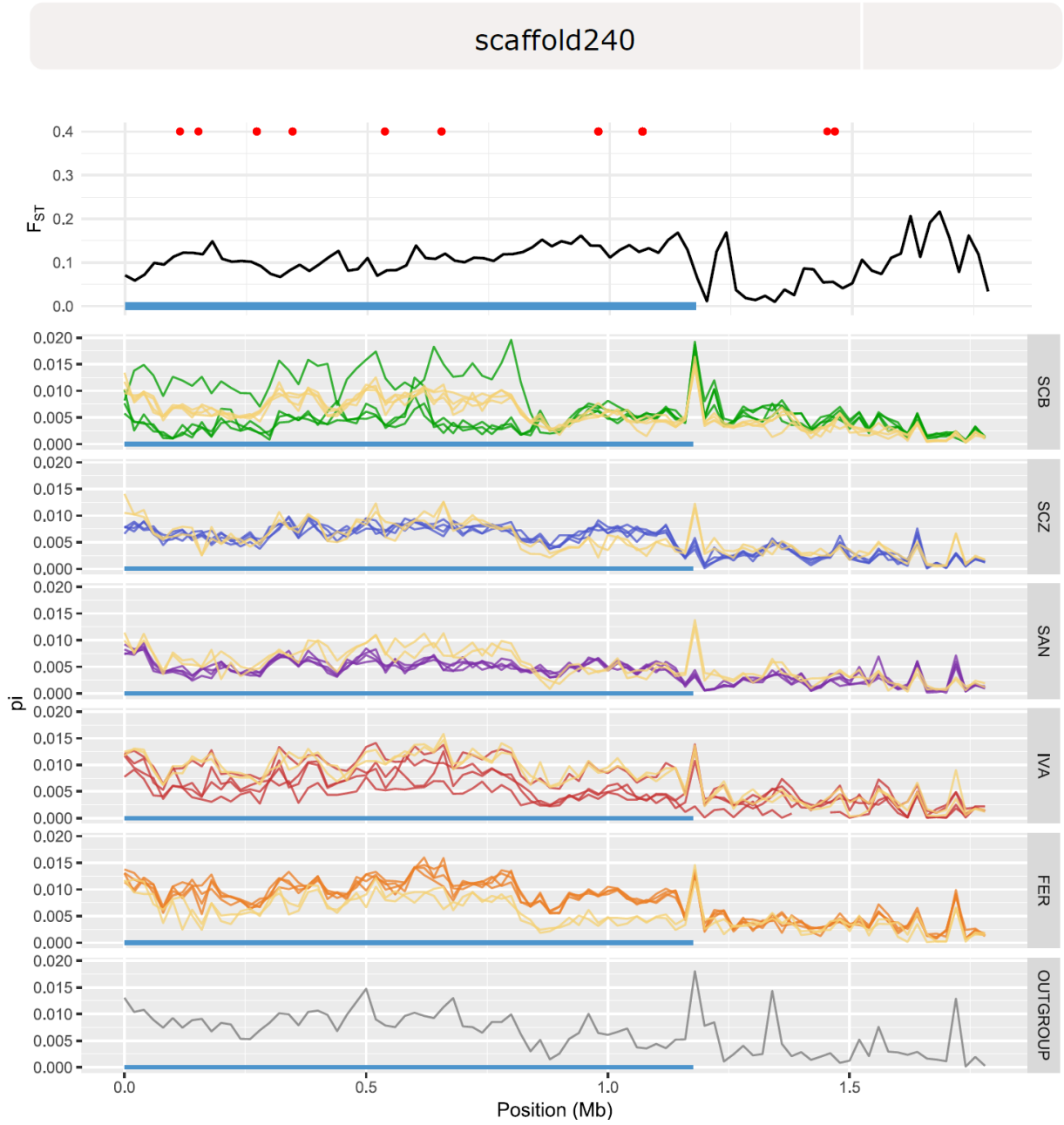


**Fig. S5 (continued).**

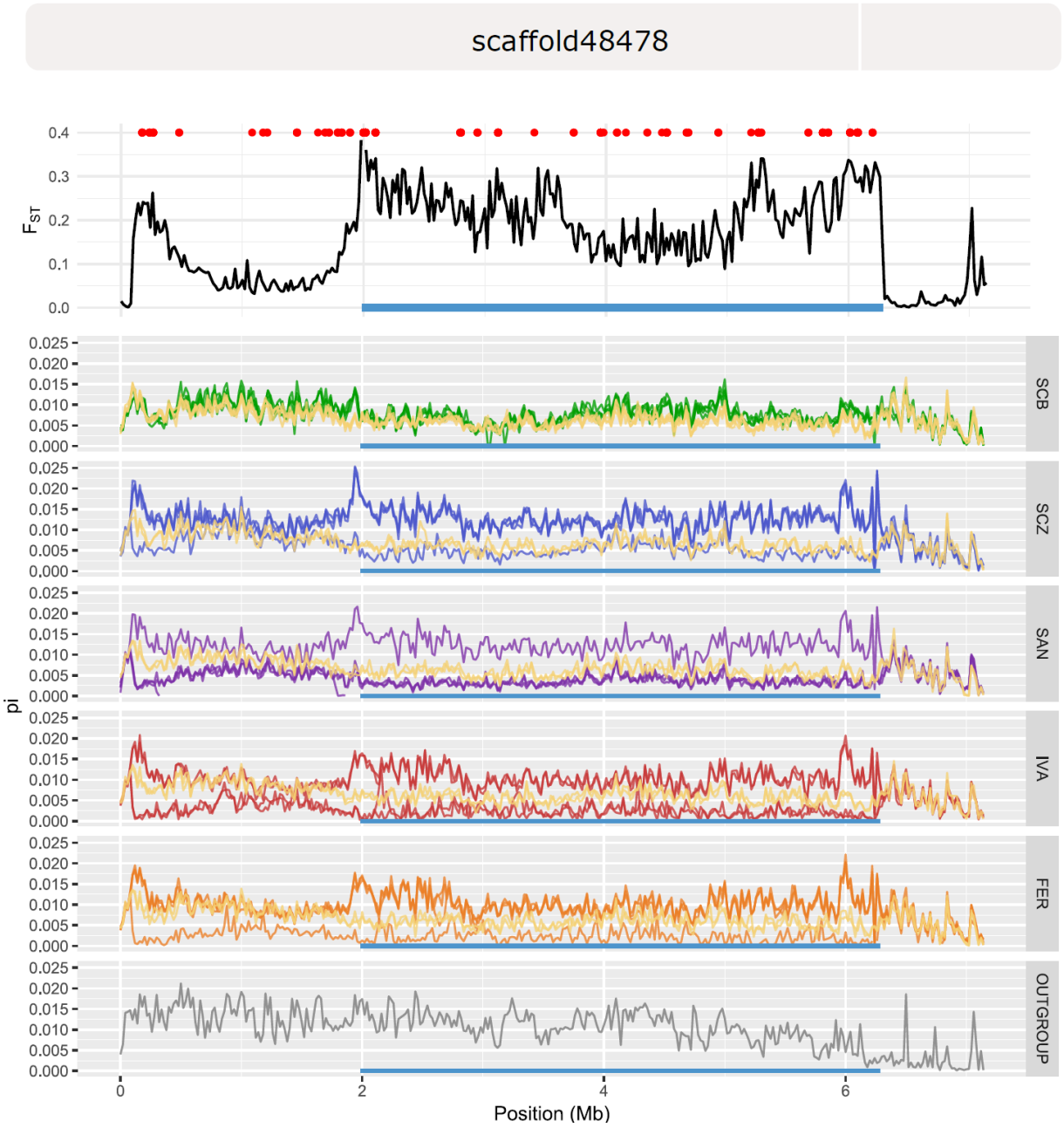
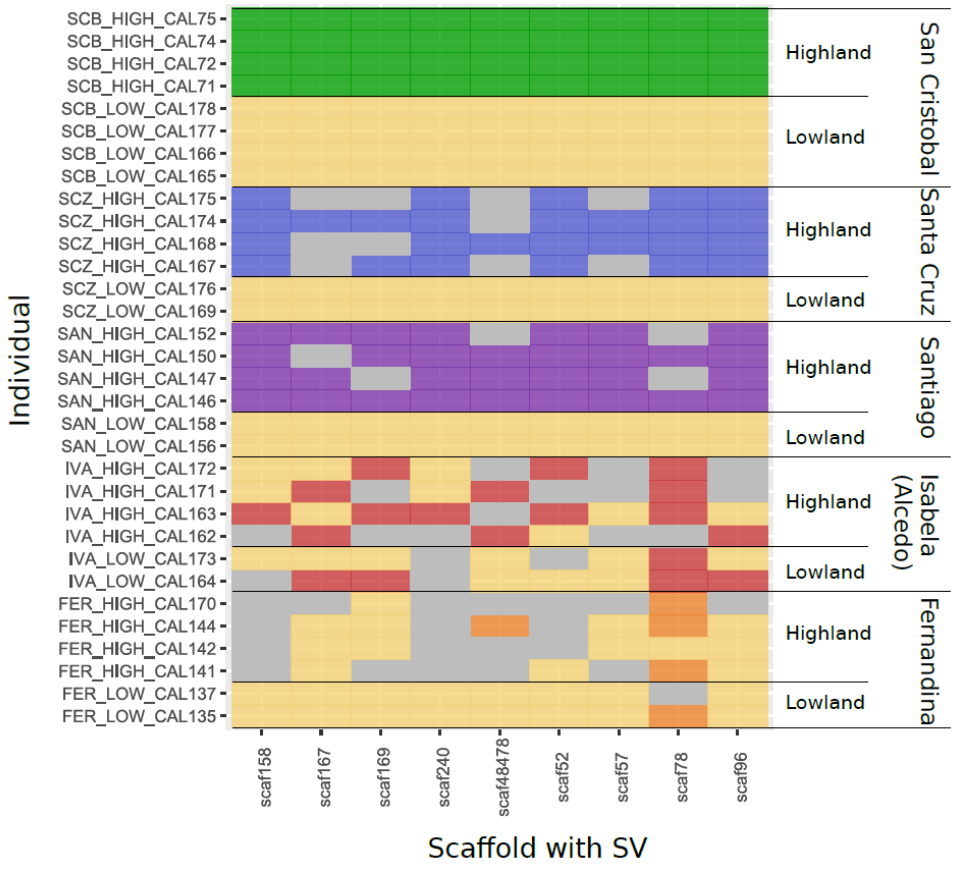


**Fig. S6 | Segregation patterns of high- and lowland alleles at the nine investigated SV associated with high-lowland divergence. Yellow squares represent individuals homozygote for the lowland allele; grey squares are individuals heterozgote for the high and lowland allele; remaining colored squares are individuals homozygote for the highland allele, with the colors corresponding to the highland population and species colors shown in Fig. 2.**



**Data S1 | Overview of the specimens used for genome assembly, restriction-site associated DNA (RAD) sequencing and whole genome resequencing.**

**Data S2 | Summary of the number of individuals of each population or species that were used for RAD sequencing and whole genome resequencing.**

**Data S3 | Assembly statistics and repeat content of the Calosoma granatense genome assembly.**

**Data S4 | Overview of pairwise Weir and Cockerham's FST estimates among all population and species pairs.**

**Data S5 | Detailed representation of admixture events identified by TreeMix v1.13. Type: intra- or interspecific admixture. Weight and migration edge: weights of inferred gene flow events (migration edge) in modified Newick format. The following columns depict results from the f4 analysis. f4 test: the four population tree. f4 value, SE and Z: the f4 estimate, its standard error and associated Z score (f4 estimates with an associated absolute value of a z-score larger than 2.58 were regarded as significant deviations from their null value (no admixture) and are indicated in bold). Last column provides a brief description of interspecific admixture event that is being tested.**

**Data S6 | Summary of the SV detected by BreakDancer v1.3.6 that overlap with the largest genomic regions with elevated divergence. Scaffold: scaffold with the inversion. SV type, start, end and length: type, start and end of the SV reported by BreakDancer. Q: quality-score of the SV reported by BreakDancer, with Q = 99 being the maximal score. Phigh-low: P-value (Welch-test) of the difference in the number reads that support the inversion between high- and lowland individuals. n RADtags: number of RADtags with an outlier SNP that is shared in at least two within-island ecotype comparisons. Pgeno, PHH-HL, PHH-LL and PHL-LL: P-value (one-way ANOVA) of the difference in nucleotide diversity at the SV between individuals identified as homozygous highland (HH), homozygous lowland (LL) and heterozygous (HL) and the respective pairwise comparisons between these genotypes by mean of a Tukey HSD test. rS(HH) and rS(LL): Spearman rank order correlation between island age and nucleotide diversity of individuals homozygous for the highland allele (HH) or lowland (LL) allele. Next columns show corresponding P -values of the correlation.**