Supplementary Material for the manuscript:

Genomic tests of the species-pump hypothesis: recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands

by Anna Papadopoulou and L. Lacey Knowles

Supplementary Text

Processing of Illumina data from RAD libraries

Sequencing of two RAD libraries generated approximately 300 million reads in total, prior to any quality filtering. At the demultiplexing step, 5% of the reads were removed due to ambiguous barcode or cut site. After the quality-filtering of data (see Fig. S1) for the 100bp and 50bp datasets (see methods for details), the average number of reads per individual was 965,345 (range from 3,000 to 2.8 million) and 1,608,058 (range from 7,000 to 3.7 million) respectively. 27 samples containing less than 1 million reads in the 50bp dataset (Fig. S1) were removed from downstream analyses. After removing barcode and cut site, the length of the retained sequences was 85 and 35bp respectively. When the 148 samples that passed the quality control were processed in USTACKS (Table S1) and repetitive loci were removed, the average of retained reads per individual were 490,000 and 1.4 million respectively for the 100bp and 50bp datasets. The mean coverage depth per locus ranged from 5 to 24 across samples, with an average of 6 for the longer reads or 8 for the shorter reads. The average numbers of unique stacks, polymorphic loci and SNPs found per population increased greatly (almost doubled in many cases) when the short reads were used (Table S1; Fig. S2). The catalog constructed by CSTACKS contained 849,944 loci, when based on the long reads or 996,789 loci when based on the short reads, with approximately 40% of the loci being polymorphic in both cases. However, when looking at the subset of loci that were shared by at least half of the individuals (>80 samples), there were three times as many such loci in the short-read catalog (72,987 vs. 24,278) with 95% of them being polymorphic. The above comparisons suggested that using the full-length reads resulted in an important loss of informative loci, as 40% of the reads were removed due to adapter contamination or low quality.

In the short-read catalog, a 30% of the polymorphic loci contained a single SNP, 21% contained 2 SNPs, 14% contained 3 SNPs and there was an 18% with more than 5 SNPs. However, as the short reads were 35bp we decided to consider only single-SNP loci for the majority of the analyses (or loci with 1-3 SNPs for divergence time estimation) because we would not expect higher numbers of segregating sites to be common in such short fragments and they might be partly due to merging of non-homologous loci. 95% of the single-SNP loci were bi-allelic.

Table S1: Processing of illumina reads using the full (100bp) and trimmed (50bp) reads. Average numbers of filtered reads (after quality control), utilized reads (after removing repetitive sequences), unique stacks, polymorphic loci and SNPs per individual are presented for each of the sampled populations. Averages were calculated after removing the individuals with less than 1 million high quality reads (see Figure S1), which were not included in the analyses.

Pop	*N*	*Filtered Reads*		*Utilized Reads*		*Unique stacks*		*Polymorphic Loci*		*SNPs*	
		100bp	**50bp**	**100bp**	**50bp**	**100bp**	**50bp**	**100bp**	**50bp**	**100bp**	**50bp**
ANEG	6	1262164	1661597	392699	1284596	55655	128280	7624	15615	14767	26758
BI	17	939518	1359312	378122	967608	58990	94780	9195	13764	17159	25958
JVD	14	988529	1360479	427597	941439	64631	108120	9692	14655	18209	27411
PI	14	942815	1873125	420992	1556640	58983	164619	9796	24100	18420	41613
PR	14	830735	1715644	581531	1455867	79996	151738	13816	21069	25835	37243
STC	14	895274	1821748	383273	1512198	54585	155792	9258	23360	17548	40536
STJ	15	891568	1788633	386452	1451996	55234	143679	9661	21437	18399	37280
STT	17	959904	1339387	392478	960418	61233	127466	9751	18916	18179	34577
TOR	16	1769173	2372901	865293	1847122	106910	173993	20732	24831	40057	45464
VGB	15	997099	1434081	417438	1028196	67113	118131	10032	15121	18780	26766
VGP	17	1286476	2633368	669064	2223376	84174	202974	14986	31141	28002	53146

Table S2. Population genetic statistics for ten *A. sanctaecrucis* populations, based on the 5,558 loci dataset and calculated only for polymorphic positions. The number of loci sequenced in each population (and shared by at least half of the individuals) is indicated in the second column. Average values across loci are presented for major allele frequency (P), nucleotide diversity (π) and observed heterozygosity (H_{obs}) as calculated by the *populations* program in *Stacks*. Two values are shown for expected heterozygosity (H_{exp}) and the Wright's inbreeding coefficient (F_{IS}), as calculated by STACKS and GENODIVE respectively. For bi-allelic SNPs π is a measure of expected heterozygosity.

Pop	loci	P	π	H_{obs}	H_{exp}	F_{IS}
ANEG	3079	0.977	0.036	0.022	0.030 / 0.038	0.327 / 0.444
BI	4032	0.956	0.069	0.04	0.065 / 0.071	0.322 / 0.433
JVD	3343	0.957	0.067	0.041	0.063 / 0.069	0.3 / 0.399
PI	3997	0.952	0.075	0.049	0.070 / 0.076	0.245 / 0.354
STC	3949	0.947	0.083	0.053	0.078 / 0.085	0.271 / 0.37
STJ	3724	0.952	0.076	0.047	0.072 / 0.078	0.282 / 0.404
STT	3827	0.956	0.07	0.042	0.066 / 0.072	0.298 / 0.414
TOR	4789	0.95	0.079	0.05	0.075 / 0.081	0.265 / 0.386
VGB	3707	0.961	0.063	0.038	0.059 / 0.064	0.292 / 0.408
VGP	4390	0.948	0.08	0.058	0.077 / 0.081	0.212 / 0.29

Table S3. Pairwise F_{ST}-values as calculated by STACKS (above the diagonal) and GENODIVE (below the diagonal). Although the absolute values differ, reflecting differences in how the programs calculate F_{ST} (e.g., whether they only include polymorphic sites and how they treat missing data), the relative differences among populations are qualitatively similar (i.e. the estimated values by the two programs are highly correlated; R^2=0.87, p-value < 0.0001). All values were significant at the 0.005 level, as tested with 10,000 permutations in GENODIVE and a Bonferroni correction for multiple comparisons.

	ANEG	BI	JVD	PI	STC	STJ	STT	TOR	VGB	VGP
ANEG	--	0.128	0.170	0.127	0.133	0.134	0.155	0.107	0.123	0.068
BI	0.311	--	0.095	0.076	0.096	0.090	0.117	0.061	0.073	0.068
JVD	0.379	0.215	--	0.081	0.079	0.068	0.107	0.063	0.099	0.095
PI	0.331	0.176	0.170	--	0.089	0.076	0.117	0.072	0.081	0.072
STC	0.321	0.245	0.158	0.217	--	0.059	0.083	0.082	0.112	0.090
STJ	0.334	0.228	0.125	0.175	0.092	--	0.089	0.067	0.095	0.079
STT	0.394	0.331	0.276	0.338	0.208	0.236	--	0.092	0.118	0.105
TOR	0.303	0.125	0.137	0.177	0.202	0.147	0.273	--	0.071	0.064
VGB	0.300	0.158	0.240	0.205	0.308	0.254	0.344	0.178	--	0.044
VGP	0.194	0.177	0.280	0.201	0.260	0.231	0.336	0.173	0.076	--

Table S4. Results of divergence time estimation with FASTSIMCOAL2 under three alternative models. Model 1: Isolation without migration (I); model 2: Isolation after Migration (IaM) i.e., assuming an initial phase of symmetric migration following population divergence, but lack of migration since the last connection of the islands (8 ky ago); model 3: Isolation with constant symmetric Migration (IM). Note that model 2 (IaM) was not applied to the St. John – St. Croix population pair, as these islands have been disconnected since the Pliocene or Miocene, so the change in migration at 8ky would not have any biological significance. The effective population size of one descendant population (N_1) was calculated from nucleotide diversity estimates and fixed, while the ancestral population size (N_A), effective population size of second descendant population (N_2), divergence time (T_{DIV}) and migration rate (m) were estimated based on the joint site frequency spectrum. Composite Maximum likelihood estimates of divergence times are presented as the number of generations (i.e. number of years ago, with 1 generation per year) and as a function of the effective population size in parentheses. The number of loci that were used for the calculation of the joint site frequency spectrum for each population pair is indicated. Population codes are the same as in Table 1 and Fig. 1.

			model 1 (I)			model 2 (IaM)				model 3 (IM)			
pop pair	SNPs	N_1 (fixed)	N_A	N_2	T_{DIV}	N_A	N_2	T_{DIV}	m	N_A	N_2	T_{DIV}	m
PR-STT	1768	97399	176436	291346	479756 (4.9N)	3299	308116	698873 (7.2N)	8.1×10^{-8}	6027	315552	711426 (7.3N)	6.9×10^{-8}
PR-STJ	4330	97399	118032	410960	566035 (5.8N)	2769	469234	802696 (8.2N)	5.8×10^{-8}	2508	456741	787143 (8.1N)	5.1×10^{-8}
STT-TOR	2979	122241	155295	414207	47880 (0.4N)	110107	403407	110145 (0.9N)	3.2×10^{-6}	99850	391575	107328 (0.9N)	2.7×10^{-6}
STJ-VGP	4891	127297	88126	198968	34742 (0.3N)	69368	239212	109741 (0.9N)	5.7×10^{-6}	63357	205687	102336 (0.8N)	4.1×10^{-6}
TOR-VGB	3299	134881	87268	73904	15831 (0.1N)	53595	103923	64210 (0.5N)	1.5×10^{-5}	52068	103468	80914 (0.6N)	7.9×10^{-6}
STJ-STC	3141	127297	76728	179395	18899 (0.1N)	NA	NA	NA	NA	118862	617850	113880 (0.9N)	1.2×10^{-5}

Supplementary Figure Legends

Figure S1. Number of reads per individual before and after the quality filtering step, individuals are ordered alphabetically by population code. The cumulative stacked bars represent the number of raw reads per individual. Within each bar the light grey color represents the reads that were discarded due to low quality, adapter contamination or ambiguous barcode when trimming to 50bp, while the medium gray color represents the additional reads that were discarded when using the full 100bp reads. The number of high-quality reads used in the analyses is thus represented by the cumulative dark and medium grey bars. Individuals with less than 1 million high-quality reads are marked with X and were removed from the analyses.

Figure S2. Average numbers of a) polymorphic loci and b) SNPs per individual, for each population, when the 100bp and 50bp reads were analysed using STACKS.

Figure S3. Comparison of observed (H_{obs}; left box) and expected heterozygosity (H_{exp}; right box) per population averaged across loci, for 10 random samples of 6 individuals (i.e. the smallest sample of individuals across populations) based on 5,558 SNPs (see Fig. 1 for population color codes). The median, the first and third quartiles, standard deviation, and range across random samples are shown with the box-and-whisker plots. A single value is presented for Anegada, as 6 individuals were sampled for this population.

Figure S4. Procrustes-transformed PCA plot of genetic variation for the analysis without the St. Croix and Anegada populations (procrustes similarity score, t_0=0.772). PC1 and PC2 axes of the genetic data (explaining 7% and 5% of the genetic variation respectively) are shown by the solid lines and x and y axes of the geographic data are shown by dashed lines.

Figure S5. Coalescent-based tree of 136 *A. sanctaecrucis* and 12 outgroup individuals based on 82,440 SNPs analyzed using SVDQUARTETS. Population membership of each individual is denoted using the same color code as in Fig. 1. Grey asterisks indicate the 2 individuals from St. John's island and one from Tortola that did not cluster with the other individuals from the same island.

Figure S6. Maximum likelihood tree based on a concatenated matrix of 5,558 SNPs analyzed in RAXML v. 8, using the ASC_GTRGAMMA model, which corrects the likelihood calculations for ascertainment bias when analysing SNP matrices not containing invariant sites. Values on the branches indicate clade support, based on 1,000 bootstrap pseudoreplicates.

Figure S7. Map of the Virgin Island PAIC, with bathymetric data. Grey shading indicates bathymetric lines between -20 and -50m, white areas > -15m and black < -50m.

Figure S1

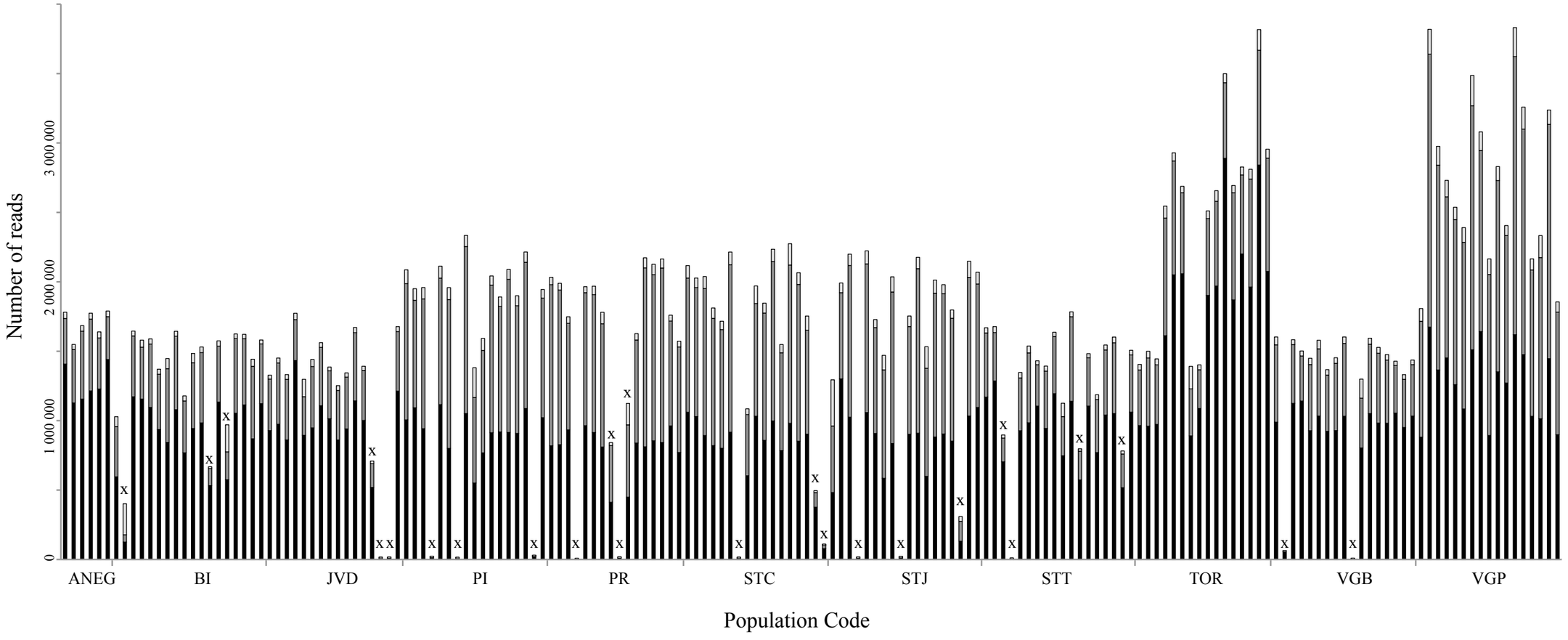
Number of reads
0
1 000 000
2 000 000
3 000 000
ANEG
BI
JVD
PI
PR
STC
STJ
STT
TOR
VGB
VGP
Population Code

Figure S2

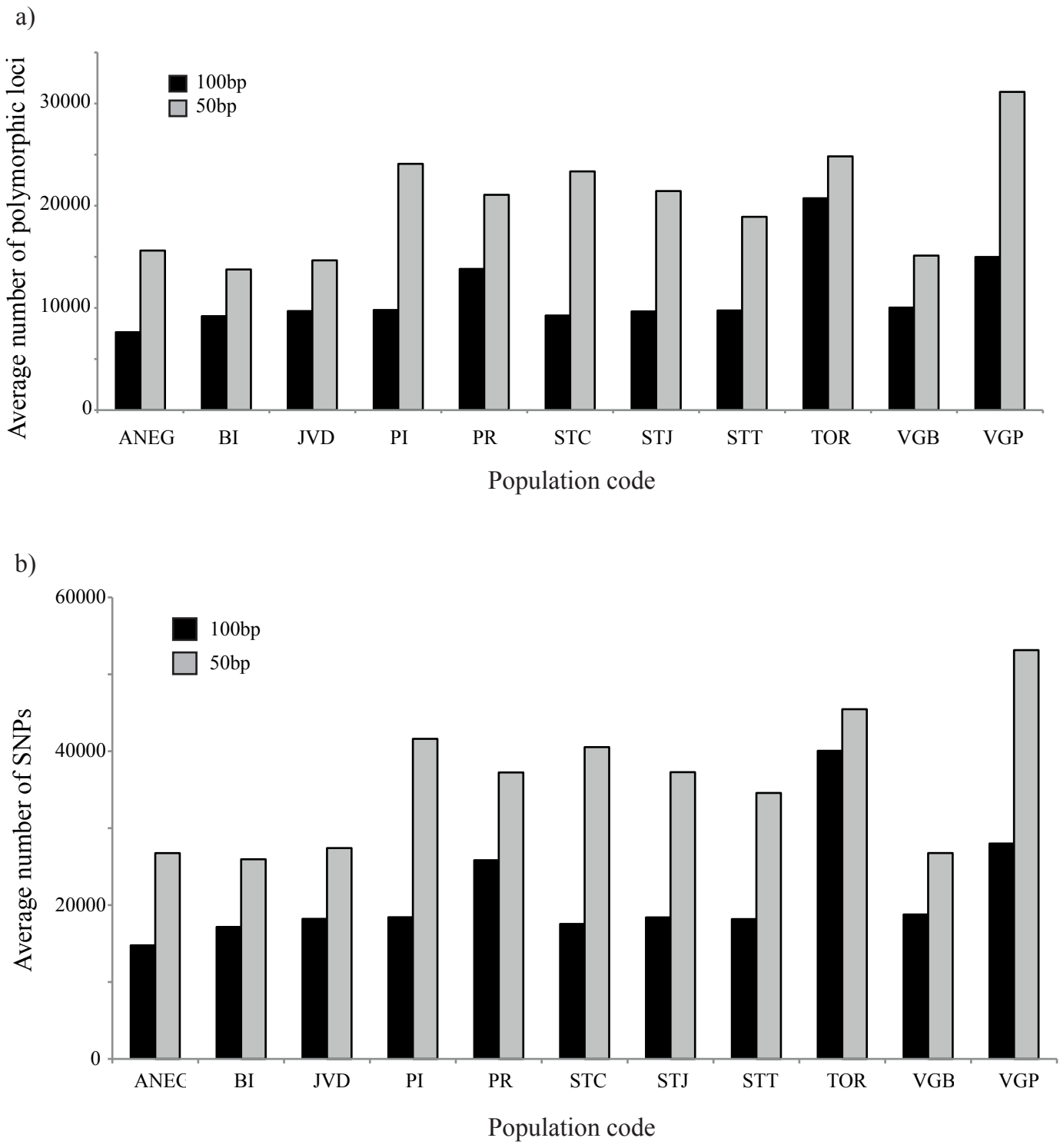
a)
Average number of polymorphic loci
100bp
50bp
30000
20000
10000
0
ANEG
BI
JVD
PI
PR
STC
STJ
STT
TOR
VGB
VGP
Population code
b)
Average number of SNPs
100bp
50bp
60000
40000
20000
0
ANEC
BI
JVD
PI
PR
STC
STJ
STT
TOR
VGB
VGP
Population code

Figure S3

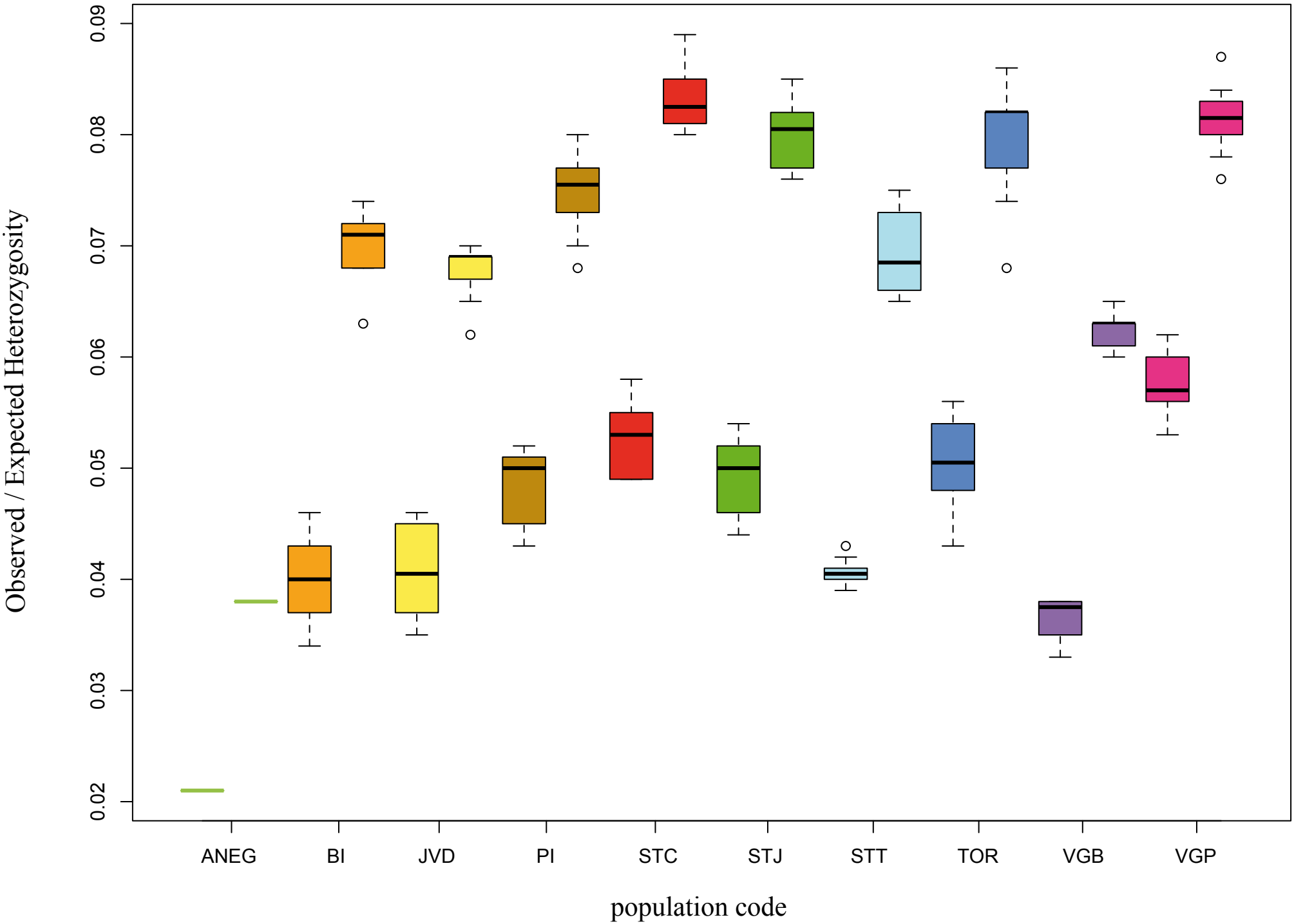
Observed / Expected Heterozygosity
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
ANEG
BI
JVD
PI
STC
STJ
STT
TOR
VGB
VGP
population code

Figure S4

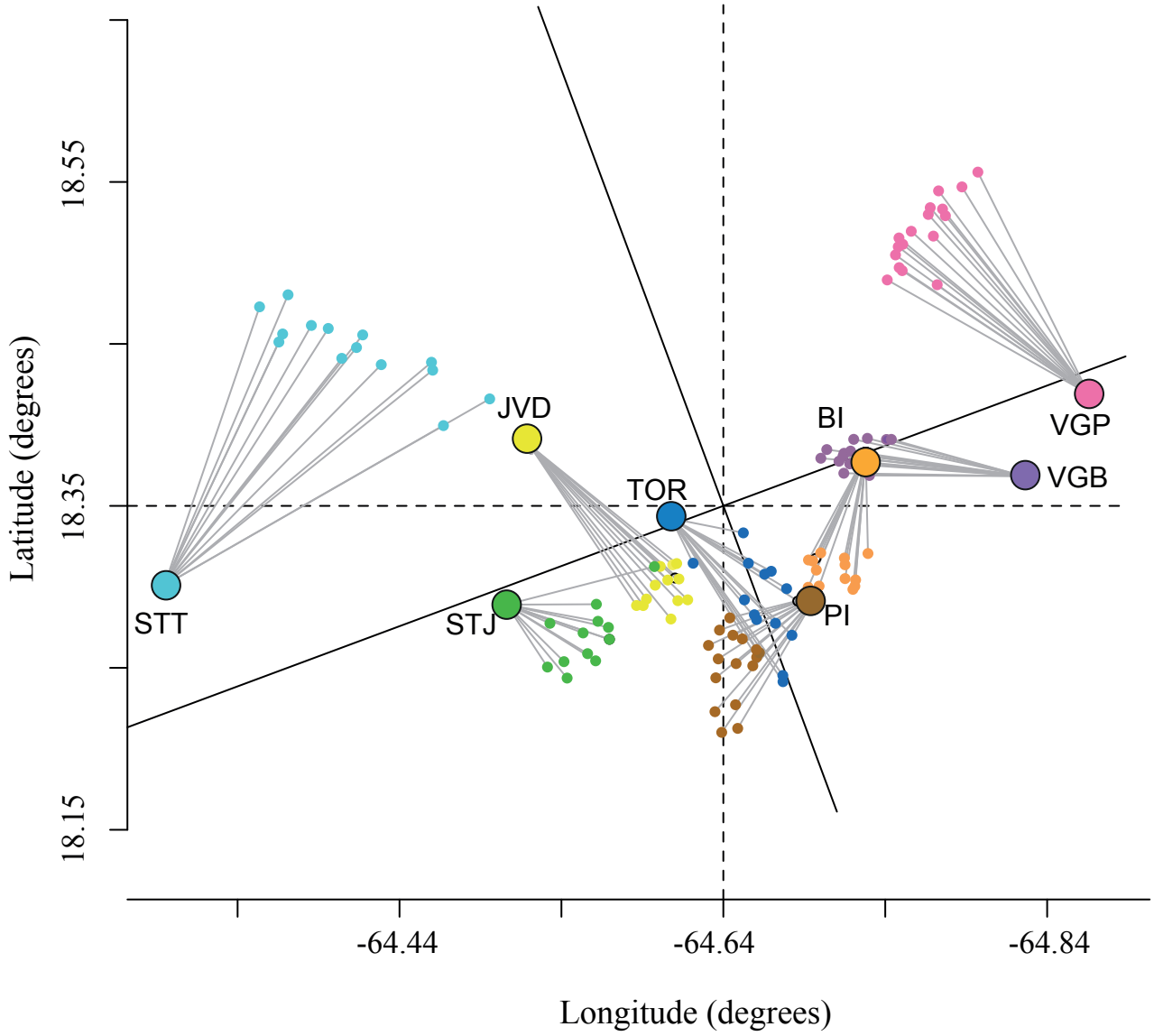
Latitude (degrees)
18.55
18.35
18.15
-64.44
-64.64
-64.84
Longitude (degrees)
STT
JVD
STJ
TOR
BI
PI
VGP
VGB

Figure S5

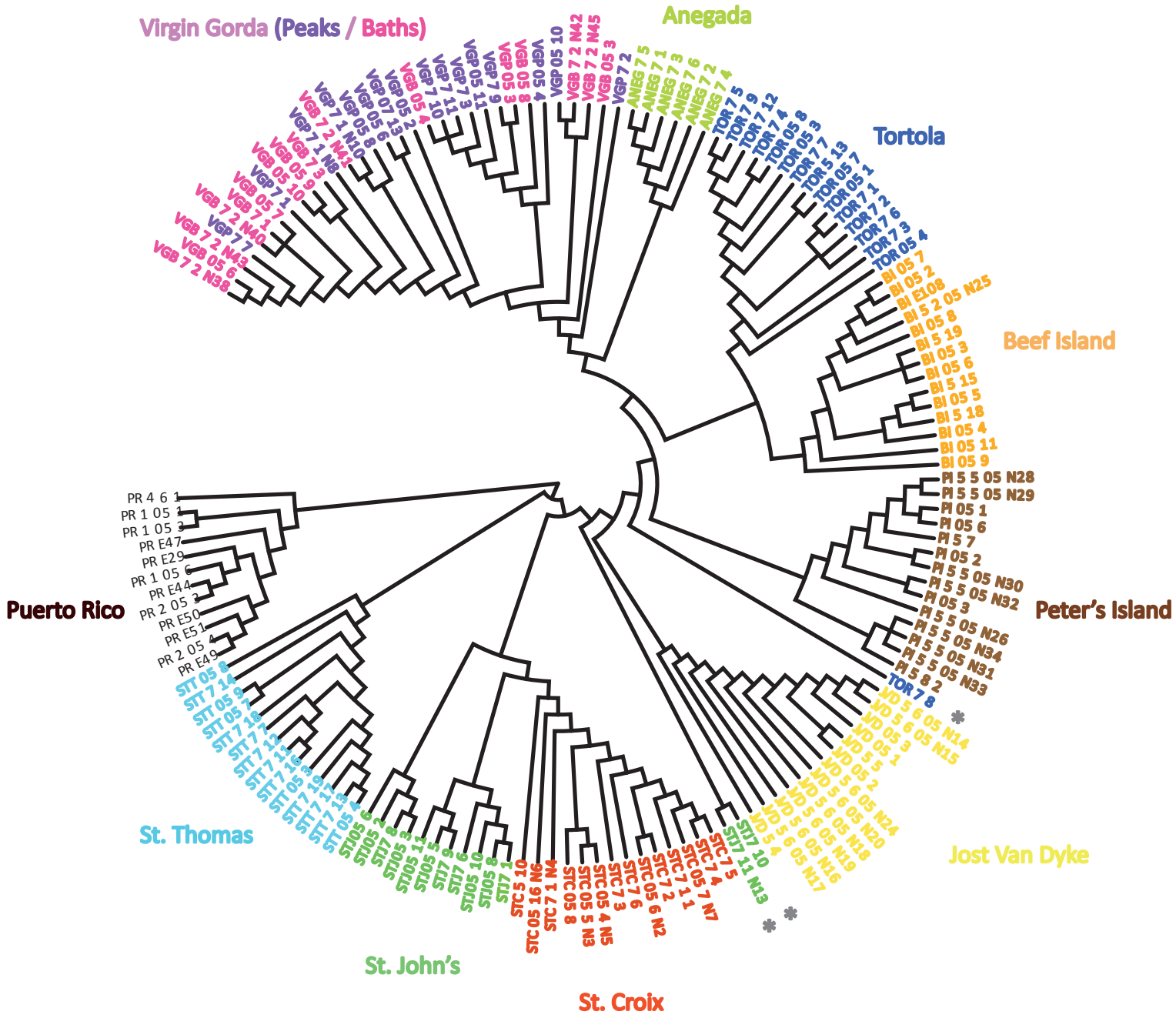
Virgin Gorda (Peaks / Baths)
Anegada
Tortola
Beef Island
Peter's Island
Jost Van Dyke
St. Thomas
St. John's
St. Croix
Puerto Rico

Figure S6

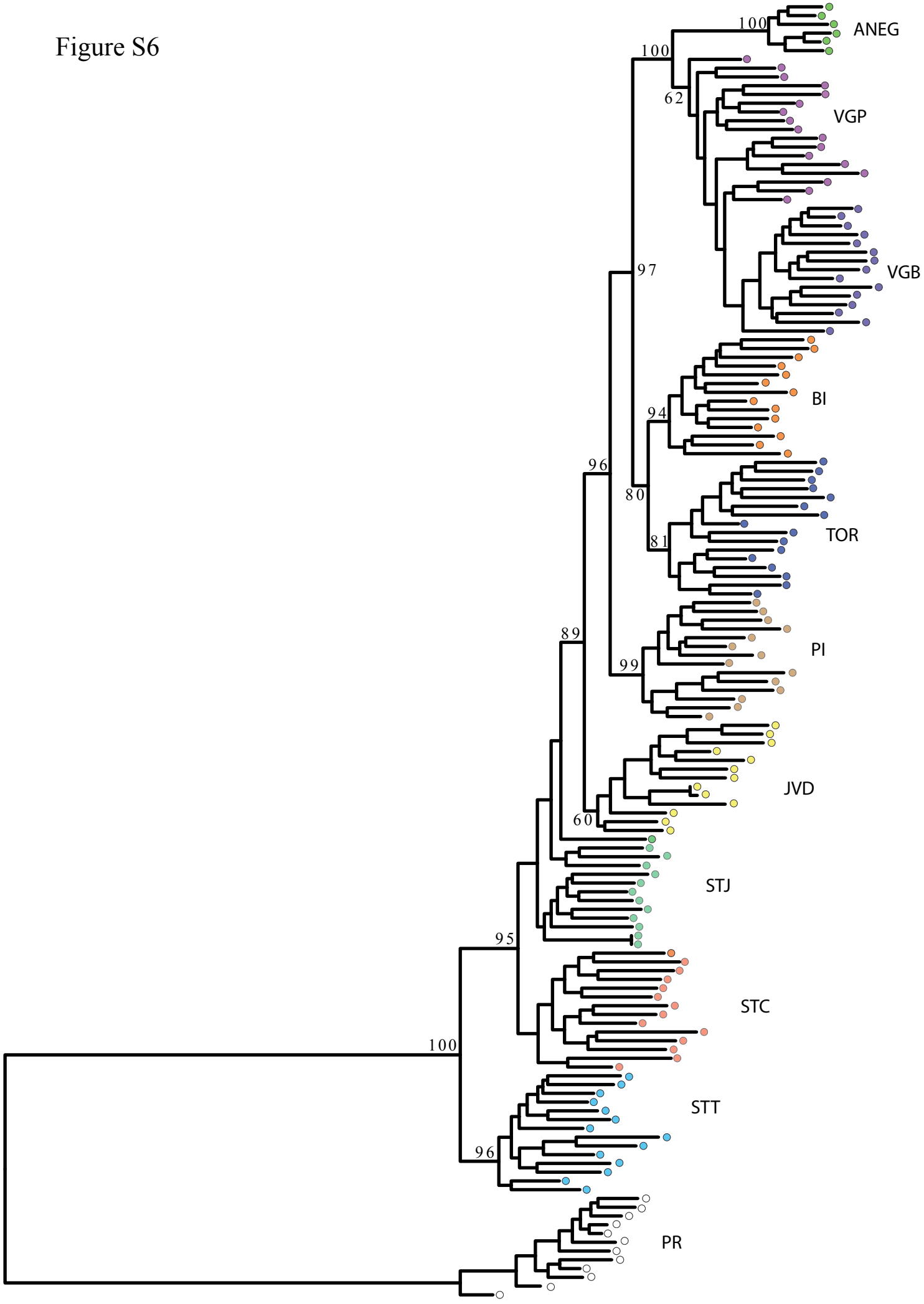
100
ANEG
100
62
VGP
97
VGB
BI
94
96
80
TOR
81
89
PI
99
JVD
60
STJ
95
STC
100
STT
96
PR

Figure S7