Genome-wide evidence for speciation with gene flow in Heliconius butterflies

Table S1. Sample information, whole-genome resequencing statistics and levels of variation

ID	taxon	sex	country	latitude	longitude	sequencing centre[1]	mean read depth	high quality calls, QUAL > 30, GQ>30	% genome	homozygous reference calls	% hom ref	heterozygous calls	% het	homozygous alternative calls	% hom alt	ts/tv[2]	nucleotide diversity
531	*H. melpomene rosina*	M	Panama	9.1206 N	79.6969 W	GenePool	26.9	173756009	63.5	168999083	97.3	3299296	1.9	1457630	0.8	1.34	
533	*H. melpomene rosina*	M	Panama	9.1206 N	79.6969 W	GenePool	26.7	169196141	61.8	164616728	97.3	3218505	1.9	1360908	0.8	1.35	0.011
546	*H. melpomene rosina*	M	Panama	9.1206 N	79.6969 W	GenePool	26.5	168777709	61.6	164233376	97.3	3201240	1.9	1343093	0.8	1.35	
2071	*H. melpomene rosina*	M	Panama	9.1206 N	79.6969 W	GenePool	36.7	192118461	70.2	186349747	97.0	3907084	2.0	1861630	1.0	1.29	
9315	*H. melpomene melpomene*	M	French Guiana	4.9632 N	52.4200 W	GenePool	24.1	172131172	62.9	166649482	96.8	2944742	1.7	2536948	1.5	1.35	
9316	*H. melpomene melpomene*	M	French Guiana	4.9632 N	52.4200 W	GenePool	23.1	165556522	60.5	160449755	96.9	2826210	1.7	2280557	1.4	1.36	0.010
9317	*H. melpomene melpomene*	M	French Guiana	4.9632 N	52.4200 W	GenePool	35.0	188558223	68.9	181987552	96.5	3331861	1.8	3238810	1.7	1.31	
13435	*H. melpomene melpomene*	M	French Guiana	4.9151 N	52.3755 W	GenePool	35.8	189676258	69.3	183014648	96.5	3403983	1.8	3257627	1.7	1.31	
1	*H. melpomene melpomene*	?	Panama	8.6136 N	78.1398 W	GenePool	23.0	205457597	75.0	204003629	99.3	1451816	0.7	2152	0.0	1.26	
18038	*H. melpomene melpomene*	F	Panama	8.6136 N	78.1398 W	GenePool	62.0	200772018	73.3	194535256	96.9	4107267	2.0	2129495	1.1	1.28	0.011
18097	*H. melpomene melpomene*	M	Panama	8.2797 N	77.8098 W	GenePool	15.6	128839980	47.1	125364711	97.3	2463134	1.9	1012135	0.8	1.38	
11-48	*H. melpomene amaryllis*	M	Peru	6.0960 S	76.9774 W	FAS	55.6	203382502	74.3	195316567	96.0	4548273	2.2	3517662	1.7	1.25	
11-160	*H. melpomene amaryllis*	F	Peru	6.4685 S	76.3533 W	FAS	44.0	202419813	73.9	194682847	96.2	4264767	2.1	3472199	1.7	1.26	0.013
09-216	*H. melpomene amaryllis*	M	Peru	5.6756 S	77.6747 W	FAS	32.6	199850729	73.0	192459782	96.3	4221690	2.1	3169257	1.6	1.27	
11-293	*H. melpomene amaryllis*	F	Peru	6.4703 S	76.3473 W	FAS	53.3	203132880	74.2	195213283	96.1	4346893	2.1	3572704	1.8	1.25	
09-108	*H. melpomene*	M	Peru12	5.9103 S	76.2258 W	FAS	36.6	196980199	71.9	189754332	96.3	4138437	2.1	3087430	1.6	1.28	0.013

	aglaope																
09-112	*H. melpomene aglaope*	M	Peru	5.9103 S	76.2258 W	FAS	38.9	201571827	73.6	193931834	96.2	4355468	2.2	3284525	1.6	1.27	
11-569	*H. melpomene aglaope*	M	Peru	5.9458 S	76.2453 W	FAS	44.4	202076594	73.8	194329380	96.2	4403693	2.2	3343521	1.7	1.27	
11-572	*H. melpomene aglaope*	M	Peru	5.9458 S	76.2466 W	FAS	37.4	200353281	73.2	192856619	96.3	4297728	2.1	3198934	1.6	1.27	
553	*H. cydno chioneus*	M	Panama	9.1714 N	79.7573 W	GenePool	35.8	187496457	68.5	180531193	96.312	3971128	2.1	2994136	1.6	1.28	
560	*H. cydno chioneus*	M	Panama	9.1714 N	79.7573 W	GenePool	35.3	188841274	69.0	181878294	96.3	3904927	2.1	3058053	1.6	1.28	0.013
564	*H. cydno chioneus*	M	Panama	9.1714 N	79.7573 W	GenePool	39.2	188356346	68.8	181303275	96.3	4038599	2.1	3014472	1.6	1.28	
565	*H. cydno chioneus*	M	Panama	9.1714 N	79.7573 W	GenePool	46.0	193728169	70.8	186243283	96.1	4223905	2.2	3260981	1.7	1.26	
09-57	*H. timareta thelxinoe*	M	Peru	6.4550 S	76.2983 W	FAS	45.8	201925567	73.7	194366548	96.3	3546884	1.8	4012135	2.0	1.34	
09-84	*H. timareta thelxinoe*	M	Peru	6.4550 S	76.2983 W	FAS	32.0	195759965	71.5	188772959	96.4	3365674	1.7	3621332	1.8	1.27	0.010
09-86	*H. timareta thelxinoe*	M	Peru	6.4550 S	76.2983 W	FAS	45.6	202568221	74.0	194926501	96.2	3602687	1.8	4039033	2.0	1.25	
09-313	*H. timareta thelxinoe*	M	Peru	6.4531 S	76.2886 W	FAS	42.0	202371330	73.9	194848222	96.3	3478024	1.7	4045084	2.0	1.25	
09-371	*H. pardalinus ssp. nov.*	M	Peru	8.3425 S	74.5922 W	FAS	42.0.	190677937	69.6	180602761	94.7	5747677	3.0	4327499	2.3	1.27	
09-202	*H. pardalinus sergestus*	M	Peru	6.4778 S	76.3517 W	FAS	41.6	191285431	69.9	183387825	95.9	2325748	1.2	5571858	2.9	1.26	
09-67	*H. ethilla aerotome*	M	Peru	6.4667 S	76.3347 W	FAS	45.5	192825749	70.4	184093324	95.5	1776222	0.9	6956203	3.6	1.25	
09-273	*H. hecale felix*	F	Peru	5.9717 S	76.2319 W	FAS	44.2	191658803	70.0	182719189	95.3	3278707	1.7	5660907	3.0	1.25	

[1] FAS: FAS Center for Systems Biology, Harvard University; GenePool: The GenePool, University of Edinburgh

[2] The ratio of transitions to transverions with respect to the reference. The differences between individuals are consistent with theoretical expectations given differences in coverage. In coding regions, the nature of the genetic code ensures that transversions are more often non-synonymous, and thus more likely to be selected against. In this dataset, Ts/Tv tends to be higher in genomes with lower coverage, which should contain proportionally more coding sequence ($r^2 = 0.67$).

Table S2. Numbers of genomic windows supporting each of four topologies

window size (kb)	species tree			geography tree			control tree			unresolved	
	count	% all trees	% resolved trees	count	% all trees	% resolved trees	count	% all trees	% resolved trees	count	% all trees
Dataset 1: cydno,rosina,melpomene[FG],Outgroups											
10	7753	38.0	53.8	6230	30.6	43.2	438	2.1	3.0	5960	29.2
20	5350	45.1	54.0	4343	36.6	43.9	211	1.8	2.1	1971	16.6
50	2715	50.8	54.2	2214	41.4	44.2	77	1.4	1.5	338	6.3
100	1510	53.0	55.0	1201	42.2	43.8	32	1.1	1.2	105	3.7
200	794	55.8	56.8	589	41.4	42.2	14	1.0	1.0	25	1.8
Dataset 2: timareta,amaryllis,melpomene[FG],Outgroups											
10	5345	27.2	72.1	1724	8.8	23.3	341	1.7	4.6	12260	62.3
20	3946	34.4	70.7	1374	12.0	24.6	263	2.3	4.7	5883	51.3
50	2664	47.8	72.5	878	15.8	23.9	132	2.4	3.6	1898	34.1
100	1304	53.2	71.5	454	18.5	24.9	67	2.7	3.7	628	25.6
200	697	61.8	73.9	215	19.1	22.8	31	2.8	3.3	184	16.3

Table S3. Mean and standard errors for F_{ST} values of non-overlapping 100 kb windows

relationship	population pair	WG	autosomes	Z chromosome
parapatric races	*amaryllis* : *aglaope*	0.009 +- 0.001	0.010 +- 0.000	0.002 +- 0.001
	rosina : *melpomene* (Pan)	0.038 +- 0.001	0.040 +- 0.000	0.048 +- 0.002
allopatric races	*amaryllis* : *melpomene* (FG)	0.226 +- 0.002	0.227 +- 0.001	0.339 +- 0.006
	rosina : *melpomene* (FG)	0.350 +- 0.002	0.349 +- 0.001	0.443 +- 0.005
	rosina : *amaryllis*	0.294 +- 0.002	0.295 +- 0.001	0.379 +- 0.005
sympatric species	*amaryllis* : *timareta*	0.287 +- 0.003	0.282 +- 0.002	0.672 +- 0.004
	rosina : *cydno*	0.292 +- 0.003	0.286 +- 0.001	0.515 +- 0.004
allopatric species	*cydno* : *timareta*	0.357 +- 0.001	0.358 +- 0.001	0.442 +- 0.003
	melpomene (FG) : *timareta*	0.419 +- 0.003	0.415 +- 0.002	0.716 +- 0.004
	rosina : *timareta*	0.393 +- 0.003	0.385 +- 0.001	0.702 +- 0.004
	amaryllis : *cydno*	0.374 +- 0.002	0.377 +- 0.001	0.440 +- 0.004
	melpomene (FG) : *cydno*	0.439 +- 0.002	0.440 +- 0.001	0.540 +- 0.003

“*aglaope*”: *H. m. aglaope*, “*amaryllis*”: *H. m. amaryllis*, “*rosina*”: *H. m. rosina*, “*melp*.”: *H. m. melpomene*, “*cydno*”: *H. c. chioneus*, “*timareta*”: *H. t. thelxinoe*, “Pan”: Panama, “FG”: French Guiana”

Supplementary Figures

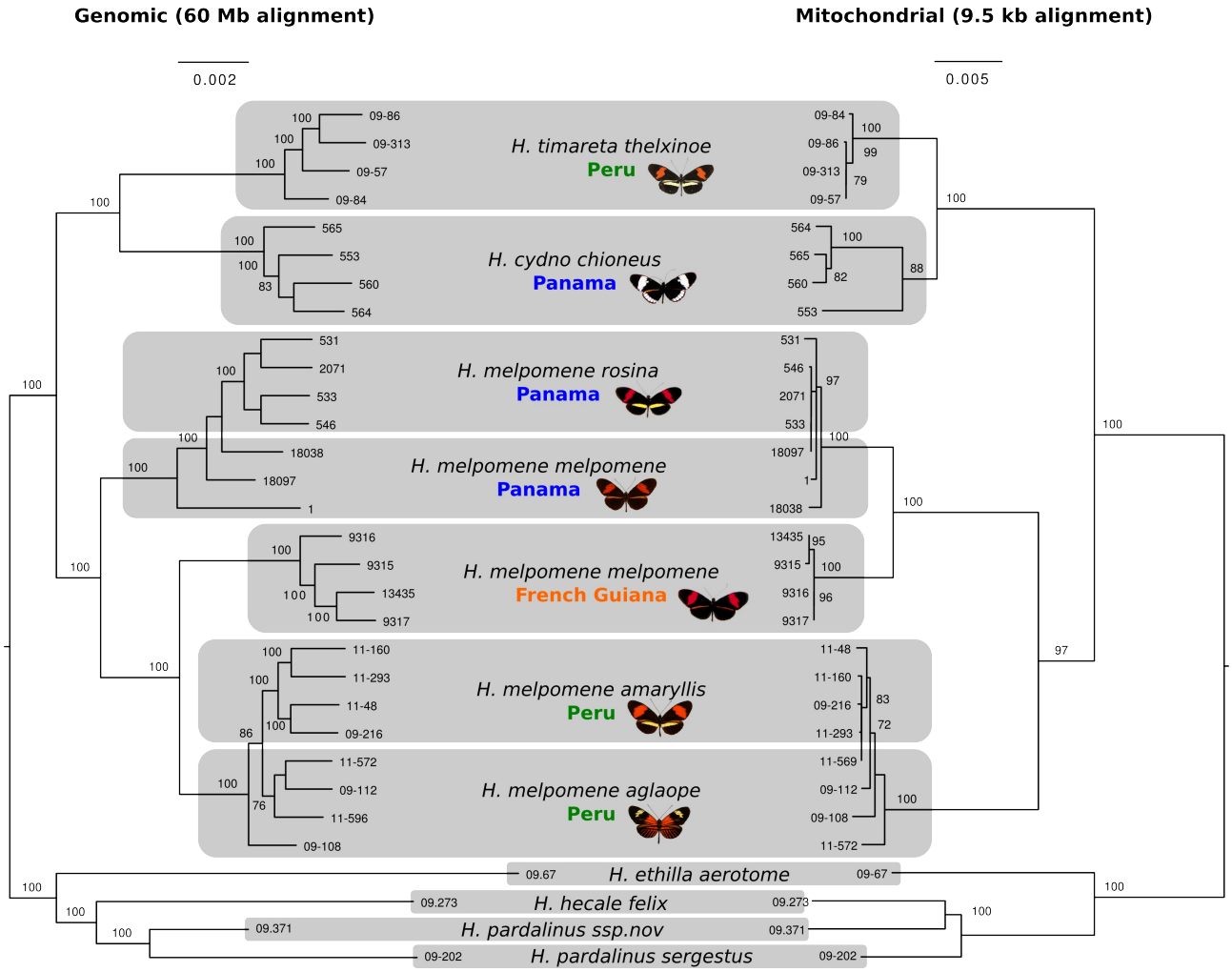

Figure S1. Genome-wide and mitochondrial maximum-likelihood phylogenies

Both trees were generated using RAxML, using GTRGAMMA model, with 100 bootstrap replicates for the whole-genome tree and 1000 for the mitochondrial tree. Bootstrap supports ≥ 90% are displayed. The alignments consisted of all sites that had a high quality genotype call for all 31 genomes analysed.

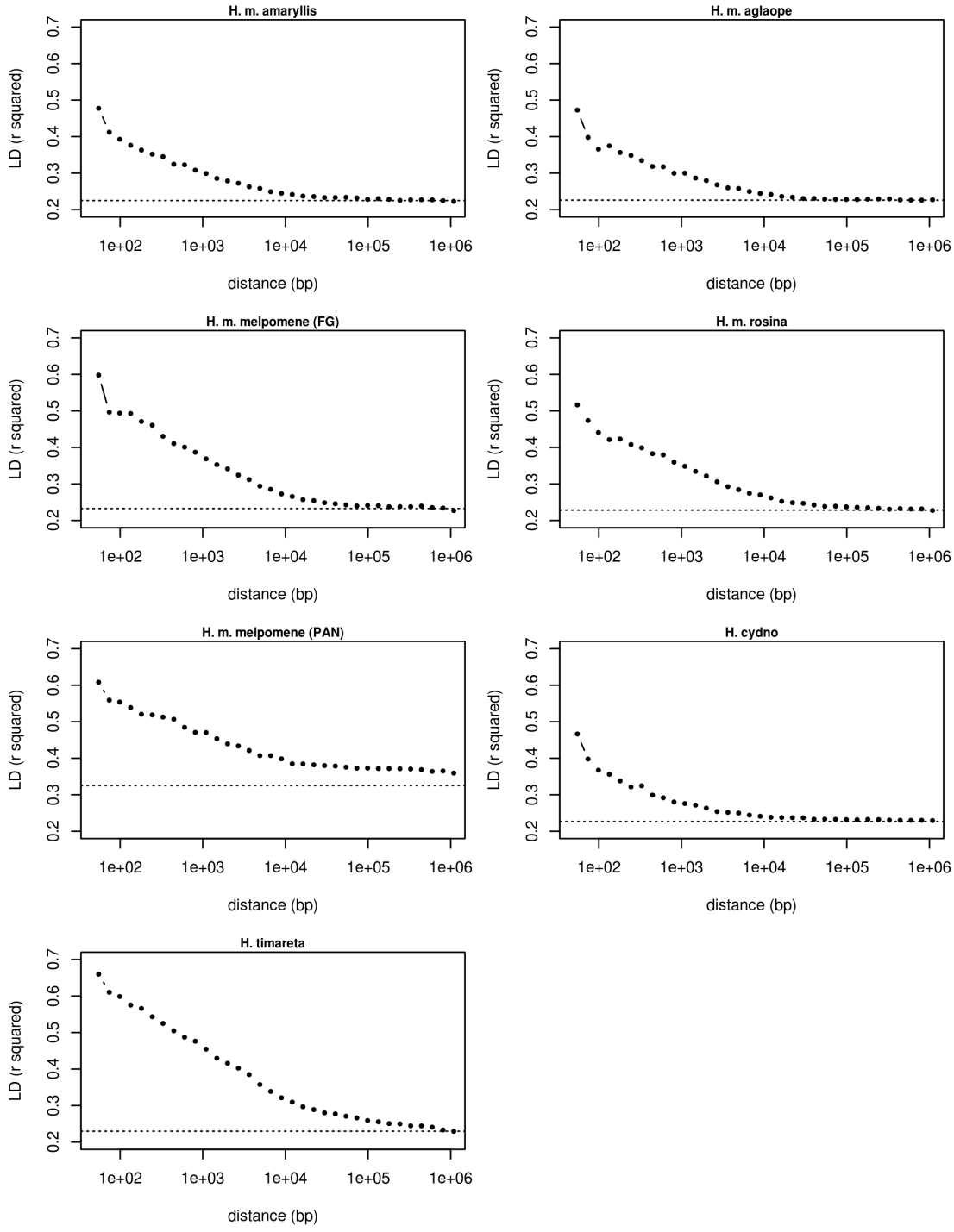

Figure S2. Decay of genome-wide linkage disequilibrium (LD) with distance

r^2 values were averaged in bins by distance, with bin size increasing logarithmically. Distance is plotted on a logarithmically scaled axis. The dashed line indicates the genomic background LD between unlinked SNPs on separate chromosomes. This level is expected to be non-zero due to small sample size. 95% confidence intervals were all too narrow to display (<0.002). In most populations, LD drops to background between 10 kb and 100 kb . The exceptions are *H. timareta*, where the decline is somewhat slower and *H. m. melpomene* (PAN), where LD does not reach the background level due to the presence of a highly inbred individual in this sample.

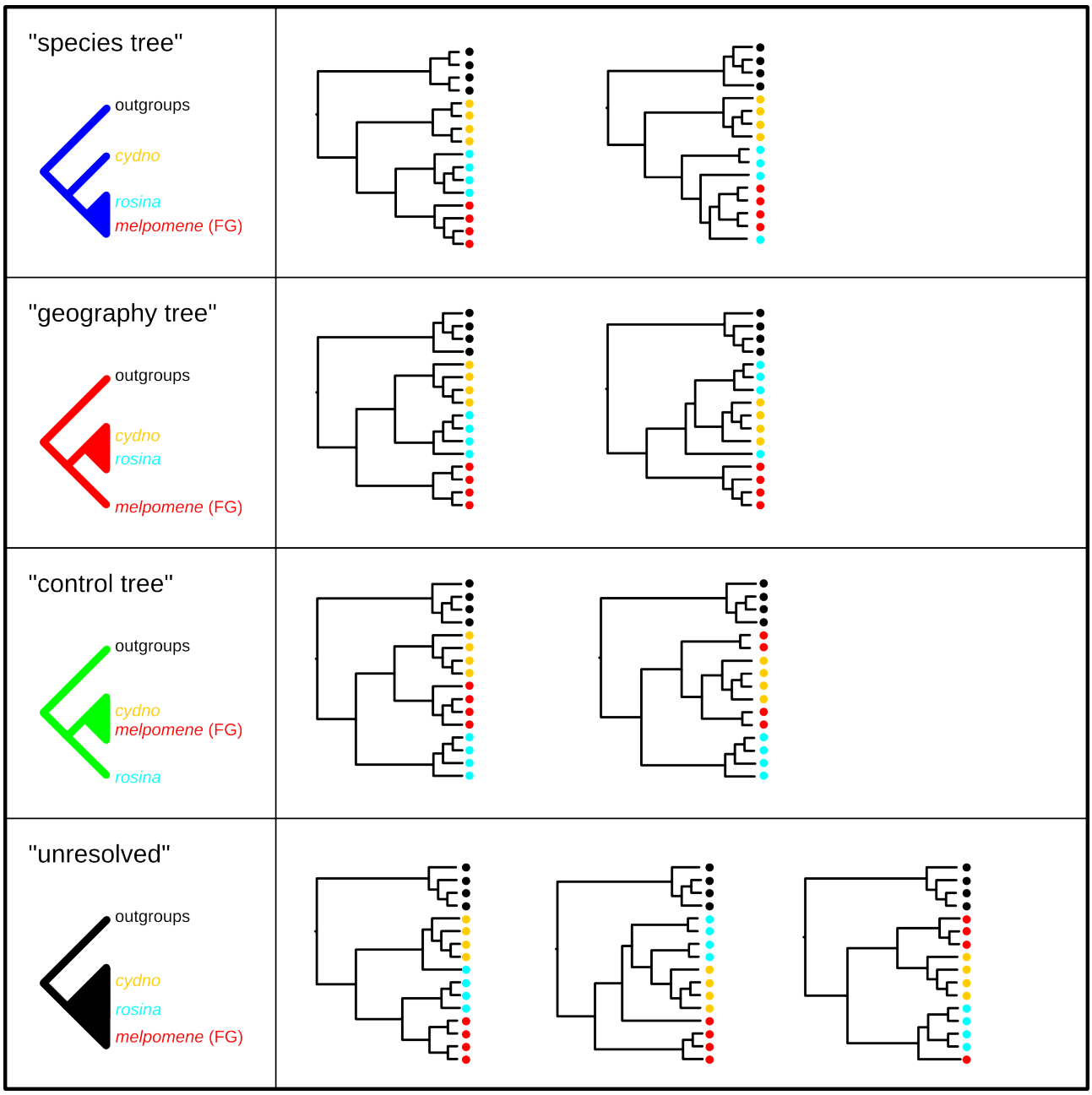

Figure S3. Examples of trees matching the four types of topologies

This figure serves to clarify how trees were assigned to one of the four possible topologies described in Fig. 2 and Table S2. See the Results section of the main paper for a verbal explanation.

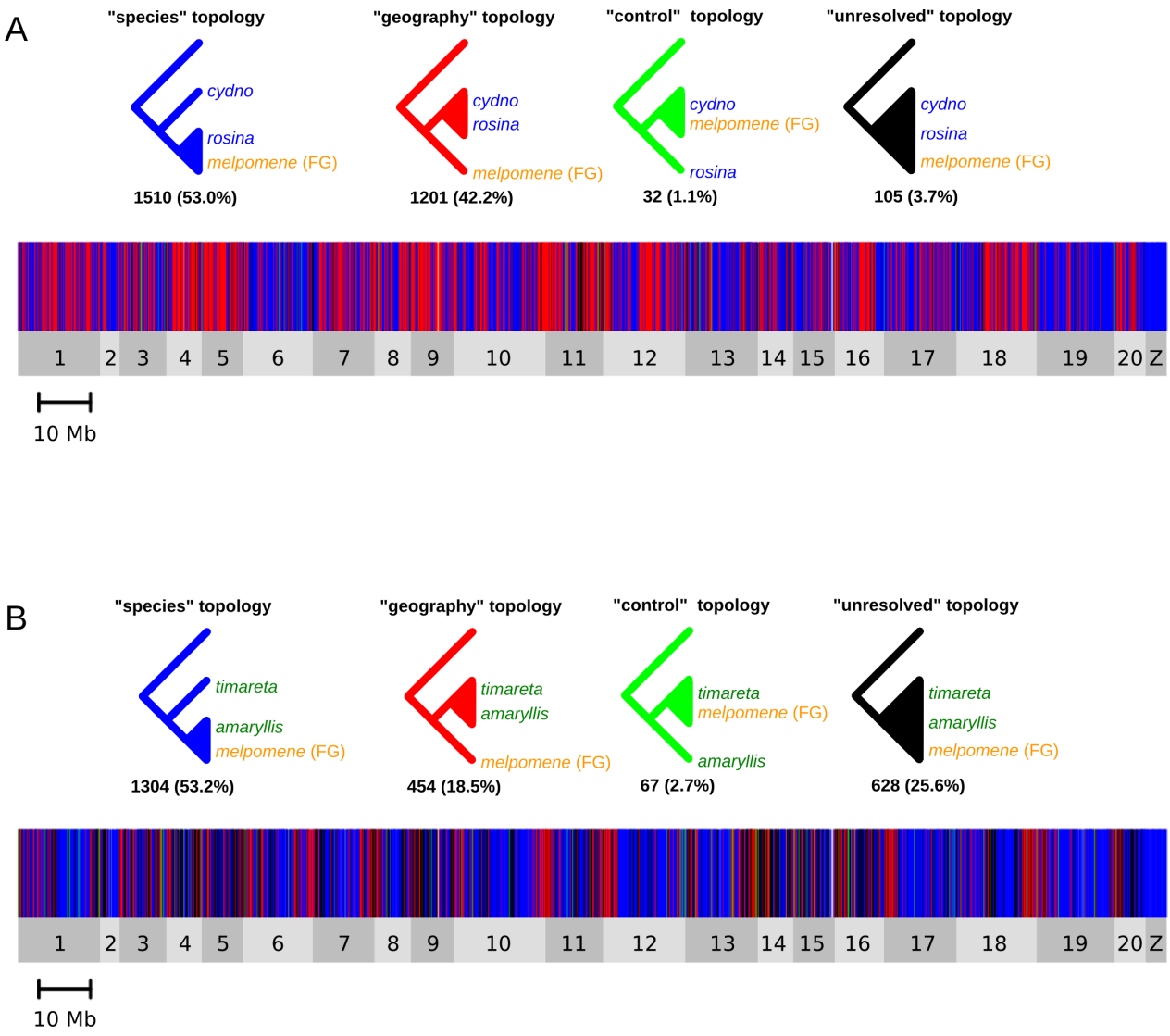

Figure S4. Distribution of the four types of topologies across the genome.

Vertical lines represent non-overlapping 100 kb windows, coloured according to the maximul likelihood topology for that window (see Fig. S3 for details).. Chromosomes are indicated with light and dark ahding. Scaffolds were ordered according to the *Heliconius* melpomene linkage map (Dasmahapatra et a. 2012).

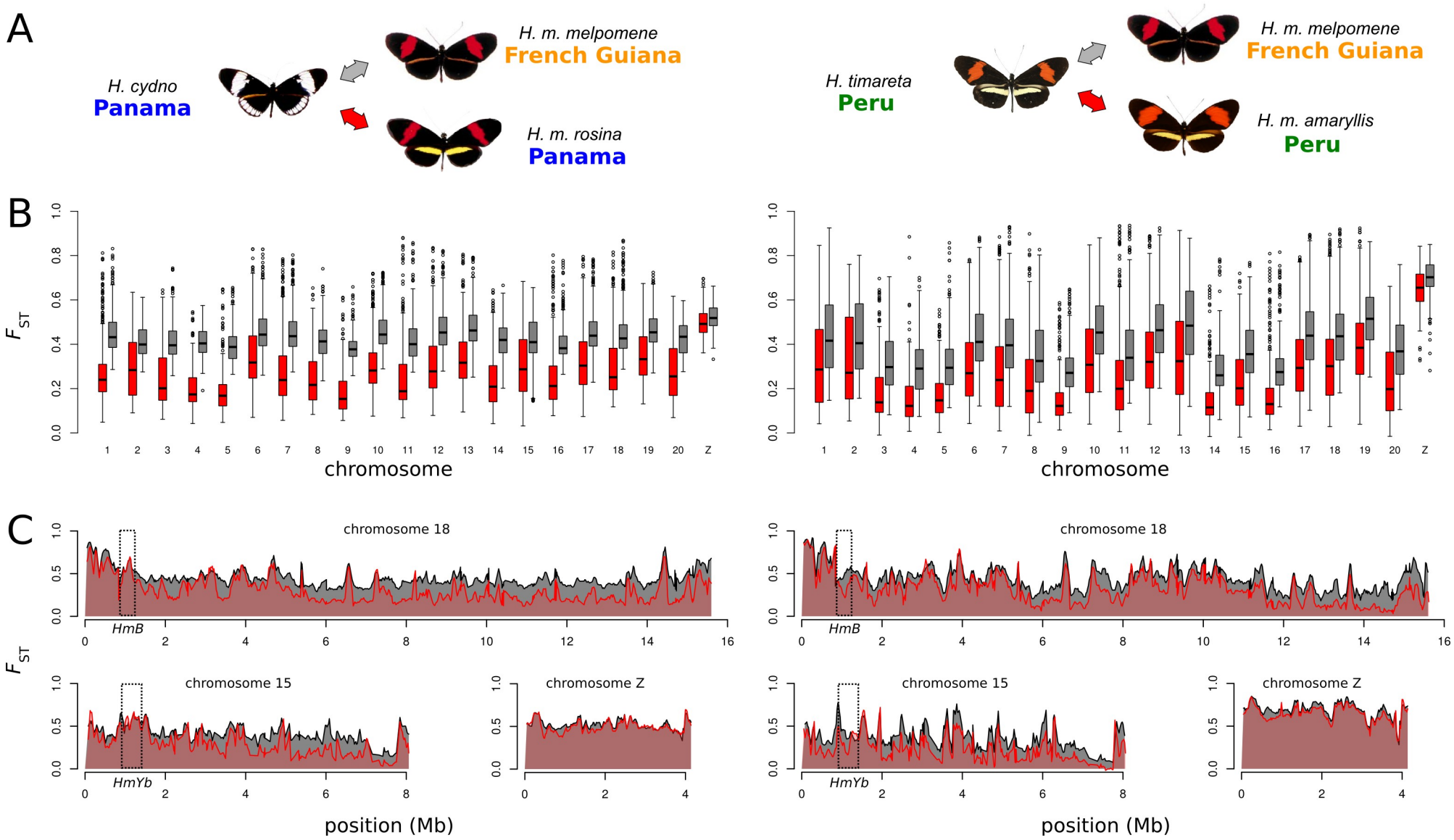
A
H. cydno
Panama
H. m. melpomene
French Guiana
H. m. rosina
Panama
H. timareta
Peru
H. m. melpomene
French Guiana
H. m. amaryllis
Peru
B
F_{ST}
chromosome
C
chromosome 18
chromosome 15
chromosome Z
HmB
HmYb
position (Mb)

Figure S5. Comparing F_{ST} in sympatry and allopatry

(A) Sympatric population pairs from Panama (right) and Peru (left) are indicated by red arrows, with the equivalent allopatric comparison indicated by grey arrows. In both cases the allopatric comparison is with *H. melpomene* from French Guiana. **(B)** Box plots of F_{ST} for all non-overlapping 100 kb windows, grouped by chromosome, with values for sympatric and allopatric pairs shown in red and grey, respectively. **(C)** F_{ST} plotted across three selected chromosomes (see Figs S5 & S6 for all chromosomes). The locations of the pattern loci *HmB* (chrom. 18) and *HmYb* (chrom. 15), are indicated by boxes.

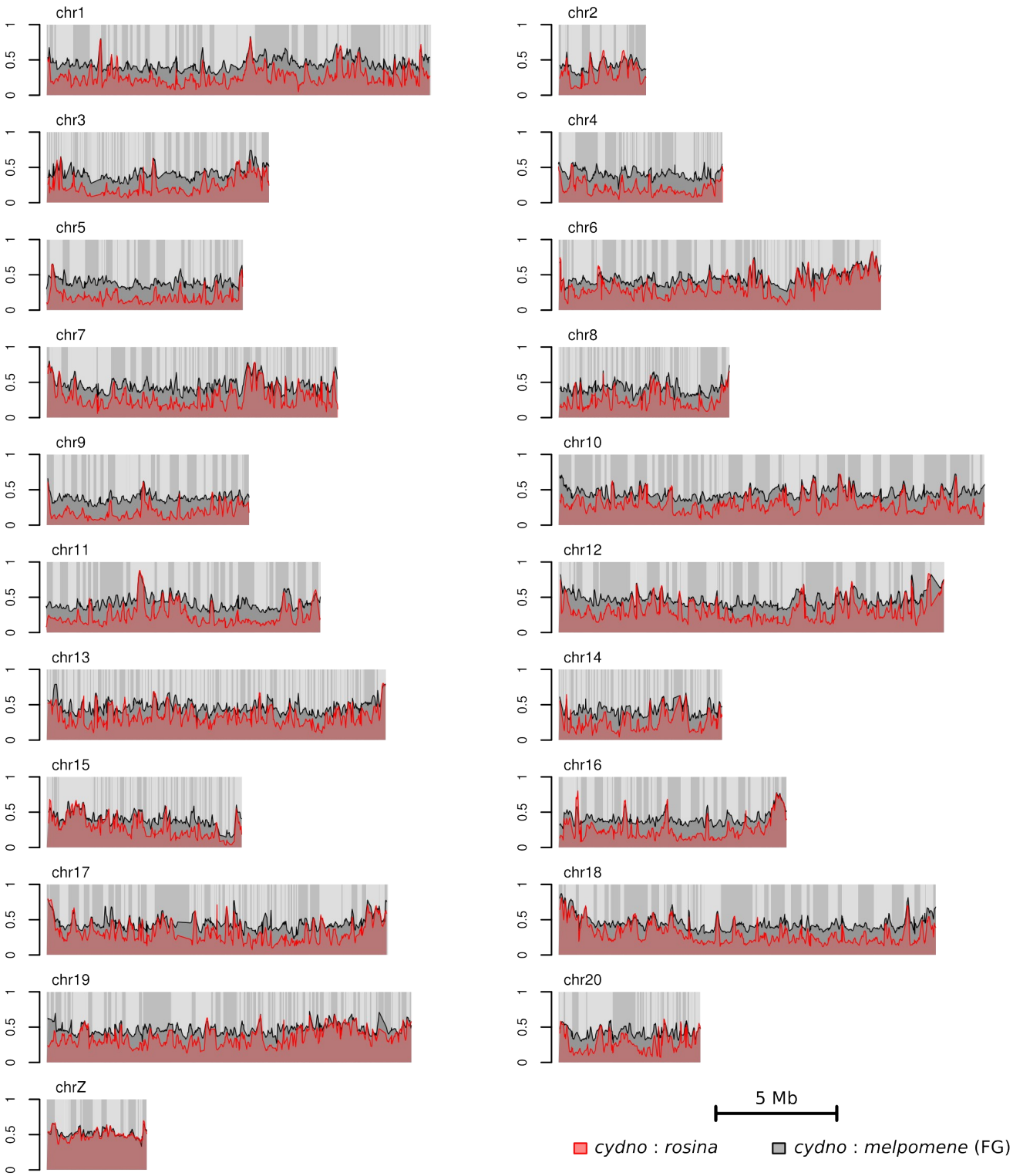

Figure S6 F_{ST} between *H. cydno* and sympatric and allopatric *H. melpomene* populations

F_{ST} is averaged over 100 kb windows, sliding in increments of 10 kb. F_{ST} between *H. cydno* and *H. m. rosina* (sympatric) is shown in red, and between *H. cydno* and *H. m. melpomene* from French Guiana (allopatric) is shown in grey. Scaffolds are indicated by alternating dark and light shading.

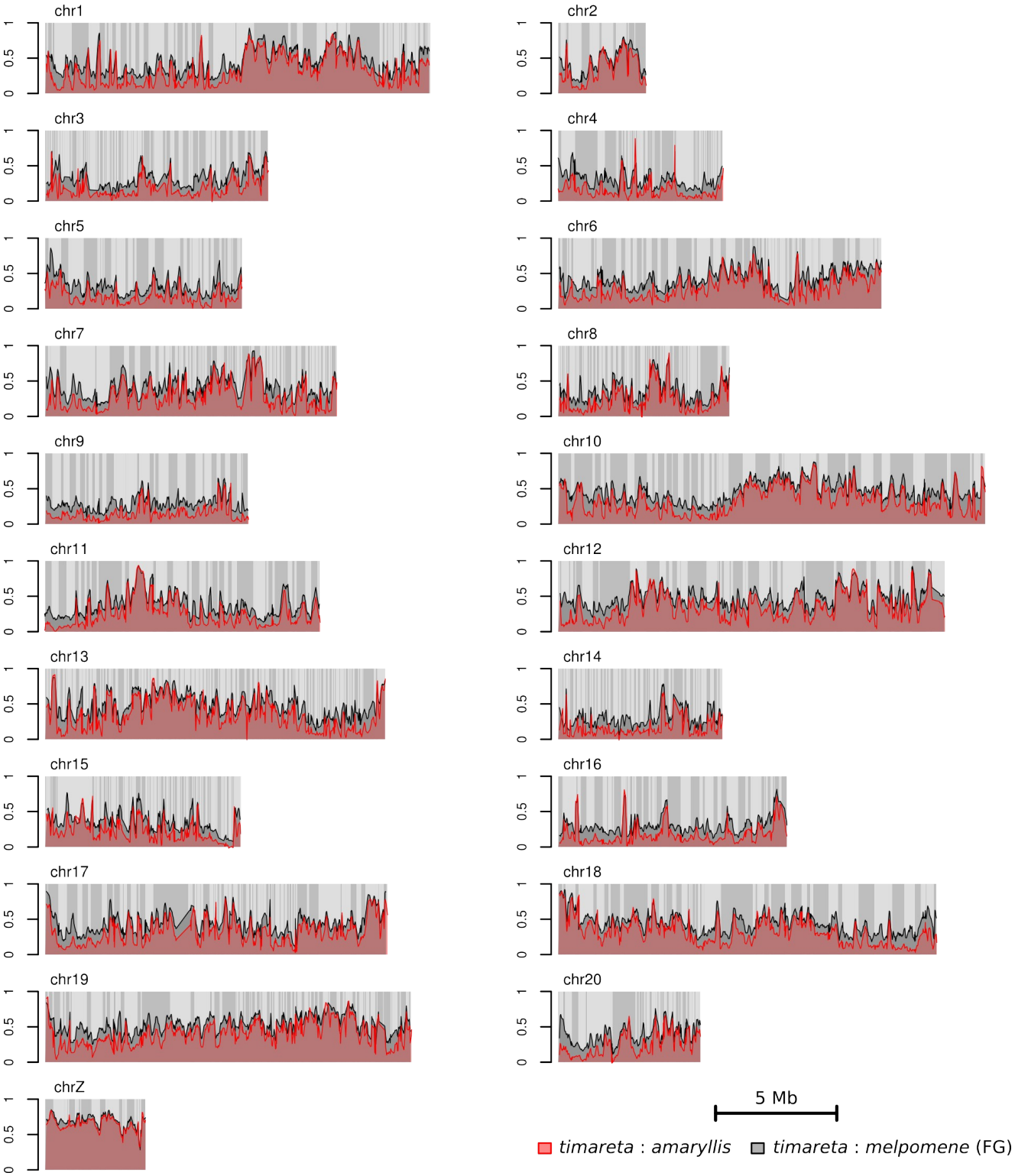

Figure S7. F_{ST} between *H. timareta* and sympatric and allopatric *H. melpomene* populations

F_{ST} is averaged over 100 kb windows, sliding in increments of 10 kb. F_{ST} between *H. timareta* and *H. m. amaryllis* (sympatric) is shown in red, and between *H. timareta* and *H. m. melpomene* from French Guiana (allopatric) is shown in grey. Scaffolds are indicated by alternating dark and light shading.

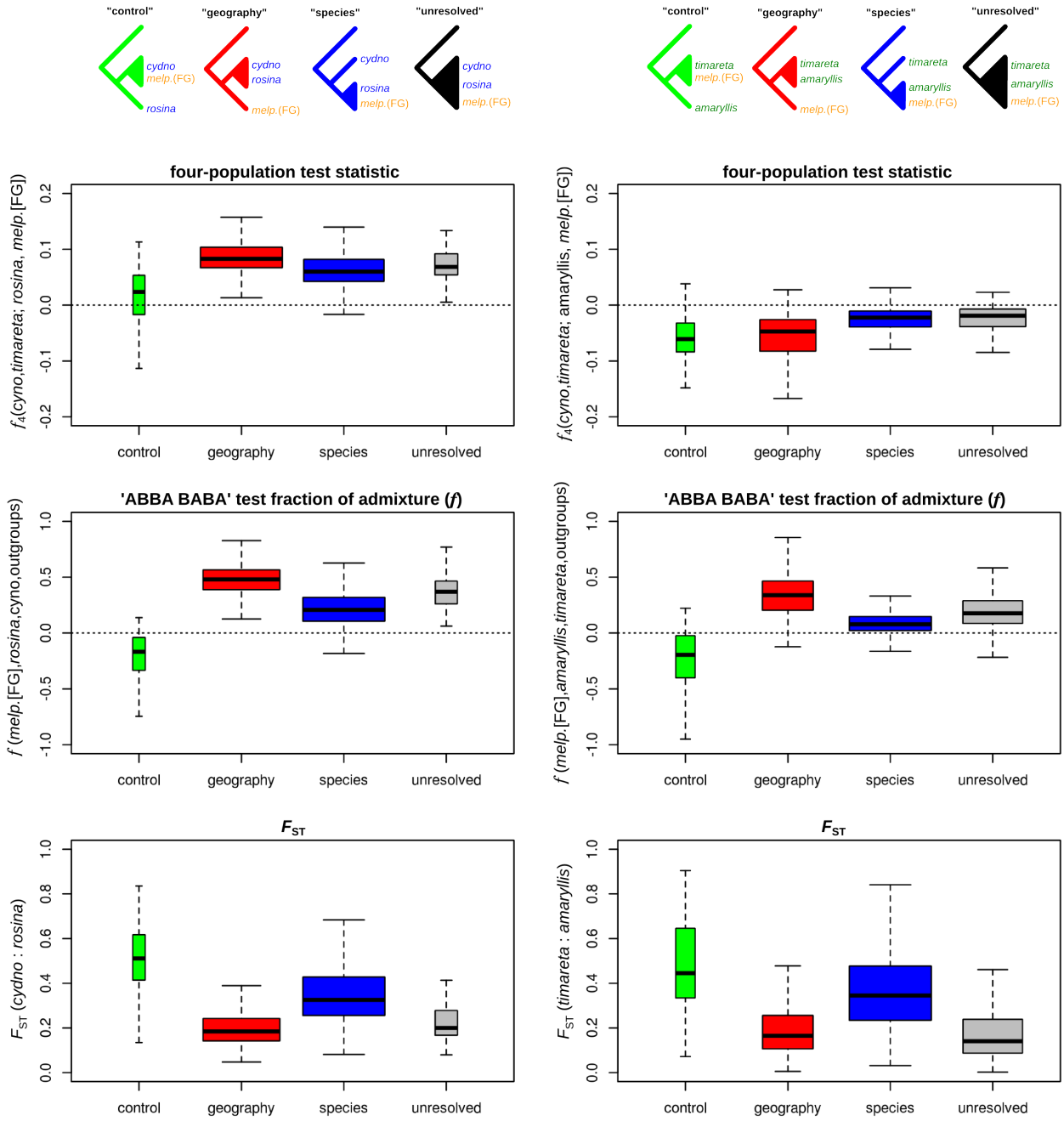

Figure S8. Concordance between phylogenetic analyses and population genomic estimators

f_4 values from the four population test, fraction of admixture (*f*) estimates from the ABBA BABA analyses and F_{ST} values between sympatric species for all 100 kb windows are all box-plotted in groups according to the topology supported by the window (see Fig. 2). Box widths were scaled according to the square root of the number of windows in each category.

Appendix A

Identification of Z-chromosomal regions

Differences in Illumina sequencing coverage were used to identify Z-linked scaffolds as well as scaffolds which represented Z-autosomal chimeras due to incorrect genome assembly. Since males are diploid for the Z chromosome, but females are haploid, the expected sequencing coverage of Z-linked genomic regions in females should be one half that of males. Median depth of coverage was calculated for each of two male and two female *H. melpomene amaryllis* individuals. Genomic regions masked as repetitive elements and regions without any aligned reads were excluded from the calculation. These point-estimates of coverage depth per scaffold were median normalized within each sample and for each scaffold the mean female and male coverages were compared.

Plotting the log-transformed female:male coverage ratios by scaffold length reveals a few distinct patterns (Fig. A1). First, variance in coverage ratios decreases with scaffold length. This likely reflects a combination of statistical and biological phenomena. Statistically, larger scaffolds represent larger samples of basepairs collecting coverage and therefore will more accurately reflect the 'true' sequencing coverage (i.e., the *central limit theorem of probability* in action). Biologically, smaller scaffolds tend to be disproportionately composed of repetitive elements, often collapsed during genome assembly. If repeats are incompletely masked, differences in repeat copy number between individuals will skew estimates of coverage and inflate variance in the ratios between individuals or sexes.

A second general pattern is that scaffolds assigned to autosomes via linkage mapping tend to exhibited equal coverage between sexes, as expected (ref 9 of main paper). Similarly, scaffolds assigned to the Z via linkage mapping typically showed the expected 50% reduction in female coverage. Several more scaffolds previously unassigned to chromosomes showed this same two-fold difference between sexes, indicating they are Z-linked. Curiously, at least a dozen large (> 100 kb) autosomally assigned scaffolds yielded intermediate female:male coverage ratios that fell between the expected values for autosomes or the Z (Fig. A1). One likely explanation for such intermediate values is that these scaffolds are actually Z-autosome chimeras resulting from errors in the genome assembly.

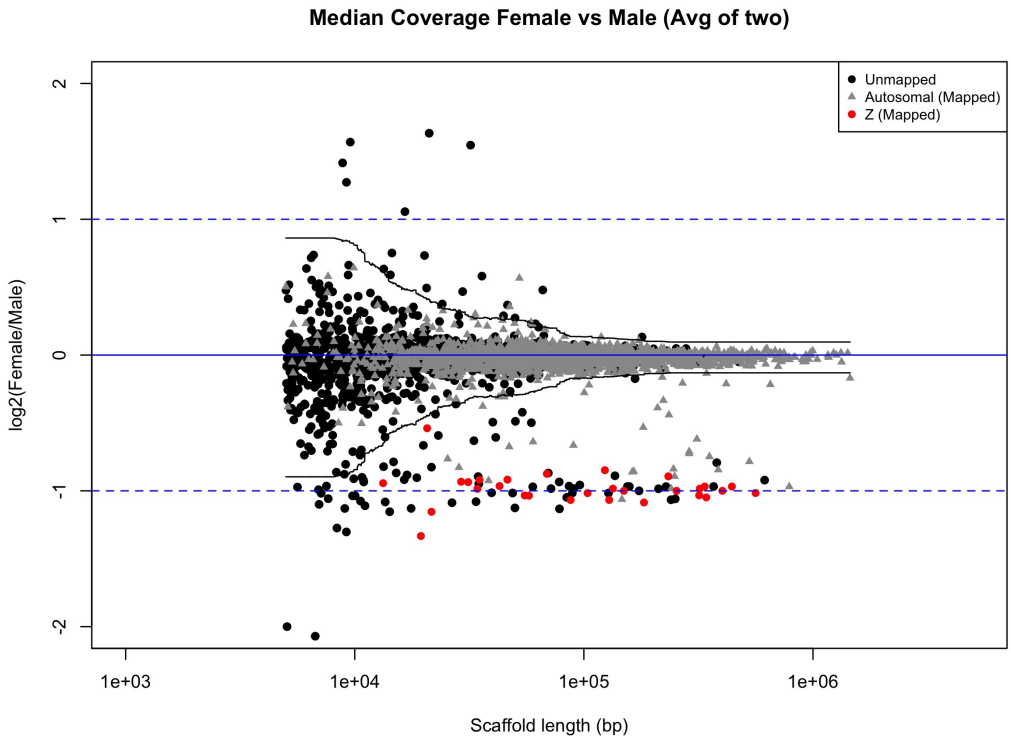

Figure A1. Relative female:male sequencing coverage. Relative female:male sequencing coverage of the ~1700 *H. melpomene* genome scaffolds greater than 5 kb in length. Chromosomal assignments (Z or autosomal) reflect linkage mapping results. Black lines indicate two standard deviations away from the median female/male coverage ratio. In this case, the standard deviation was calculated for each scaffold using within-sex comparisons and a window +/- 250 length-ordered scaffolds.

It should be possible to identify such chimeric scaffolds (and putative break points between Z and autosomal sections) by identifying distinct shifts in sequencing coverage in females. A sharp transition from equal to one-half coverage relative to males is expected within a scaffold containing a Z-autosome "fusion". We identified outlier scaffolds and examined them in detail to search for such Z-autosome chimeras and to confirm a uniform pattern of 50% reduced female coverage for putatively Z-linked scaffolds or regions.

Outlier scaffolds were delineated using a sliding-window analysis of variation in coverage ratios between individuals of the same sex. These within-sex comparisons should capture variation in the data due to technical noise, sampling effects, and biological variation not associated with differences between sexes. Within-sex measures of variation in sequencing coverage can be applied to female-male comparisons to identify scaffolds with extreme differences in coverage between sexes.

We calculated log-transformed coverage ratios between the two males and also between the two

females for all scaffolds greater than 5 kb (Fig. A2). Scaffolds, ordered by length, were grouped in "windows" of 500 scaffolds and the standard deviation of within-sex coverage log-ratios was calculated in each window, advancing one scaffold at a time. This provided a length-appropriate measure of variation in coverage for each scaffold. Values from the largest and smallest terminal windows were applied uniformly to the 250 largest and smallest scaffolds, respectively. Scaffolds were considered outliers when the female:male coverage log-ratio was greater than two standard deviations away from the median of all scaffolds in the female:male comparison. The standard deviation value applied to each scaffold was the within-sex window centered on that scaffold (Fig. A1). 158 outlier scaffolds were identified in this manner, the vast majority of which showed reduced female coverage.

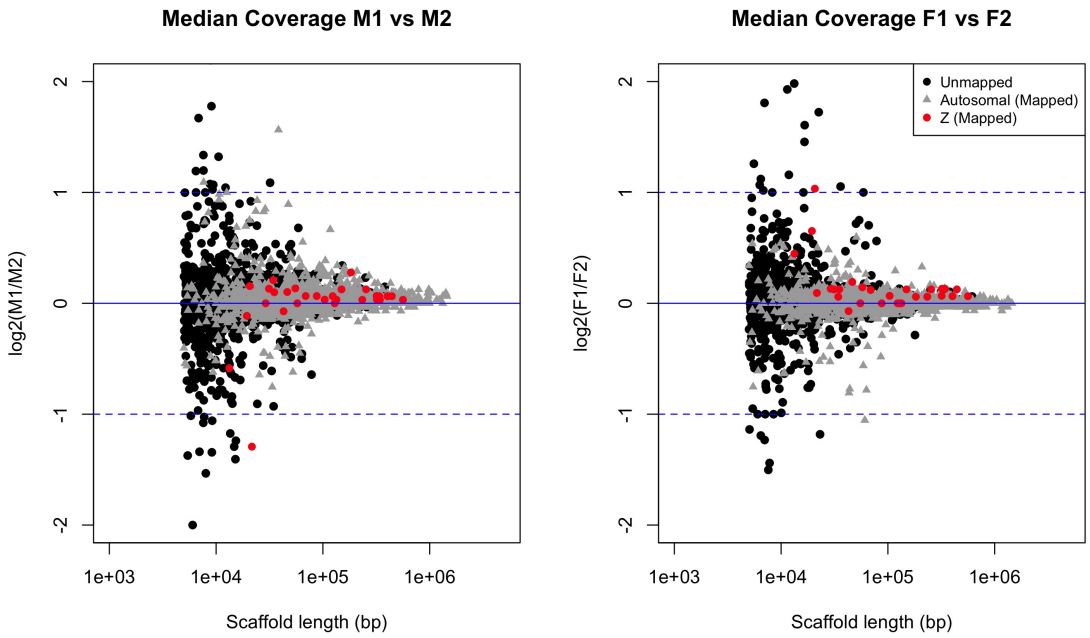

Figure A2. Within-sex ratios of sequencing coverage for all scaffolds longer than 5kb. Chromosomal assignments (Z or autosomal) reflect linkage mapping results.

A detailed investigation of sequencing coverage was conducted for these 158 outlier scaffolds. For each scaffold, a high-resolution plot of sequencing coverage was generated for each of the four *H. melpomene amaryllis* samples, this time without masking repeats. Figure A3 shows one such plot. Additionally, a sliding window of mean female:male coverage log-ratio helped visualize and pinpoint shifts in sequencing coverage along the scaffold. The window size considered was 10 kb for scaffolds longer than 100kb and was 10% of the scaffold length for those shorter than 10 kb. The window was shifted by increments of 100 bp. Putative breakpoints between Z-linked and autosomal regions were inferred from the absolute value of the difference between adjacent

windows. In theory, this value should be maximized at the fusion point of Z-linked and autosomal scaffold regions.

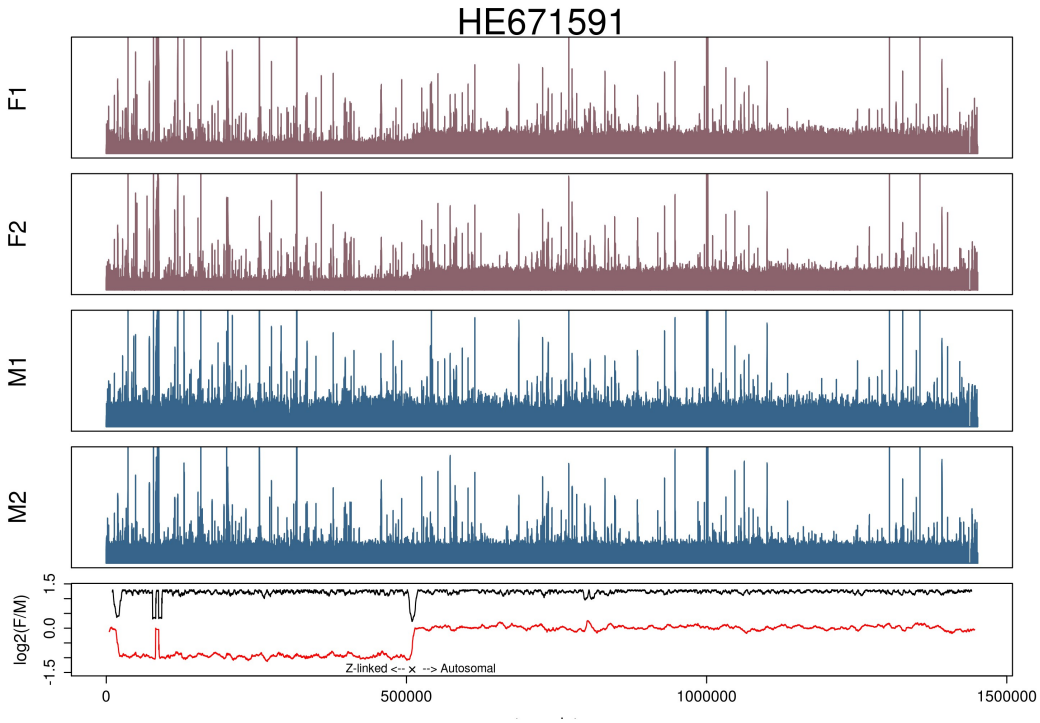

Figure A3. Coverage for scaffold HE671591. High-resolution plot of coverage for scaffold HE671591 demonstrating the Z-autosome chimeric breakpoint near 500 kb. The top four panels show coverage depths across the scaffold for the two female (F) and male (M) *H. melpomene amaryllis*. Coverage is normalized to arbitrary units to facilitate direct comparison between samples. In the bottom panel, the red line shows a 10kb sliding window of log2 (female:male) coverage. The black line, which is not drawn to scale vertically, reflects the negative absolute value of the difference between adjacent sliding windows. The "X" indicates the inferred breakpoint between Z-linked and autosomal sections of this scaffold.

Plots of each of the outlier scaffolds were individually inspected and scaffolds were judged as being entirely Z-linked, entirely autosomal, or chimeric. In addition to the 158 outlier scaffolds identified among scaffolds longer than 5kb, we inspected coverage plots for five scaffolds smaller than 5kb. These five shorter scaffolds received special attention because they contained coding sequences and also yielded female:male log-ratios below -0.6.

Thirty-two scaffolds were judged to be chimeric and were split at putative breakpoints for subsequent genomic analyses. In cases where chimeric scaffolds had been assigned to an autosome via linkage mapping, that assignment was retained for the autosomal section of the split scaffold. After manual inspection, splitting chimeras, and assigning Z-linkage, the total amount of Z-linked

sequence was 13.05 mb spread across 96 scaffolds. Assignment of Z-linkage from coverage was quite consistent with previous results from linkage mapping. While four Z-mapped scaffolds contained some autosomal regions, none of the other 34 Z-mapped scaffolds were incorrectly assigned to autosomes.

Revised plots of female:male log-ratio by scaffold length reflecting the coverage-based analysis of Z-linkage show much more consistent clustering of scaffolds around the expected log-ratios of -1 for Z-linked and 0 for autosomes (Fig. A4). This is especially true for scaffolds longer than 100 kb. Several scaffolds shorter than 100kb still exhibit intermediate log-ratios that fall between the expected values. There are even a few cases where autosomally mapped scaffolds cluster with otherwise Z-linked scaffolds. These cases arise from substantial variation in sequencing coverage within or between individuals apparently arising from copy-number variation that is not consistent with Z-linkage. For example, in the autosomally mapped scaffold HE670616, coverage is comparable across the scaffold for both females and one male sample (Figure A5). The remaining male sample has a distinct increase in coverage which causes the overall female:male log-ratio to be -0.92. Such a pattern is much more consistent with an autosomal CNV than sex-linkage.

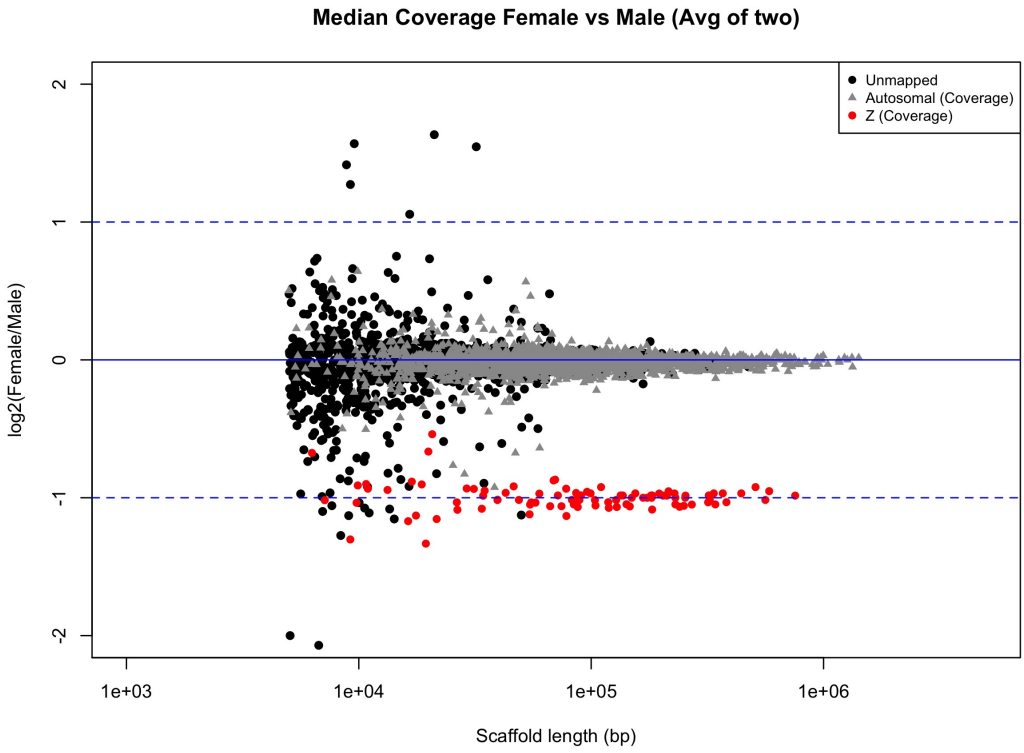

Figure A4. Relative female:male sequencing coverage. Relative female:male sequencing coverage of the ~1700 H. melpomene genome scaffolds greater than 5 kb in length after evaluating Z-linkage based on coverage. Red dots indicate Z-linkage based on coverage analysis while grey triangles indicate scaffolds assigned to autosomes via linkage mapping.

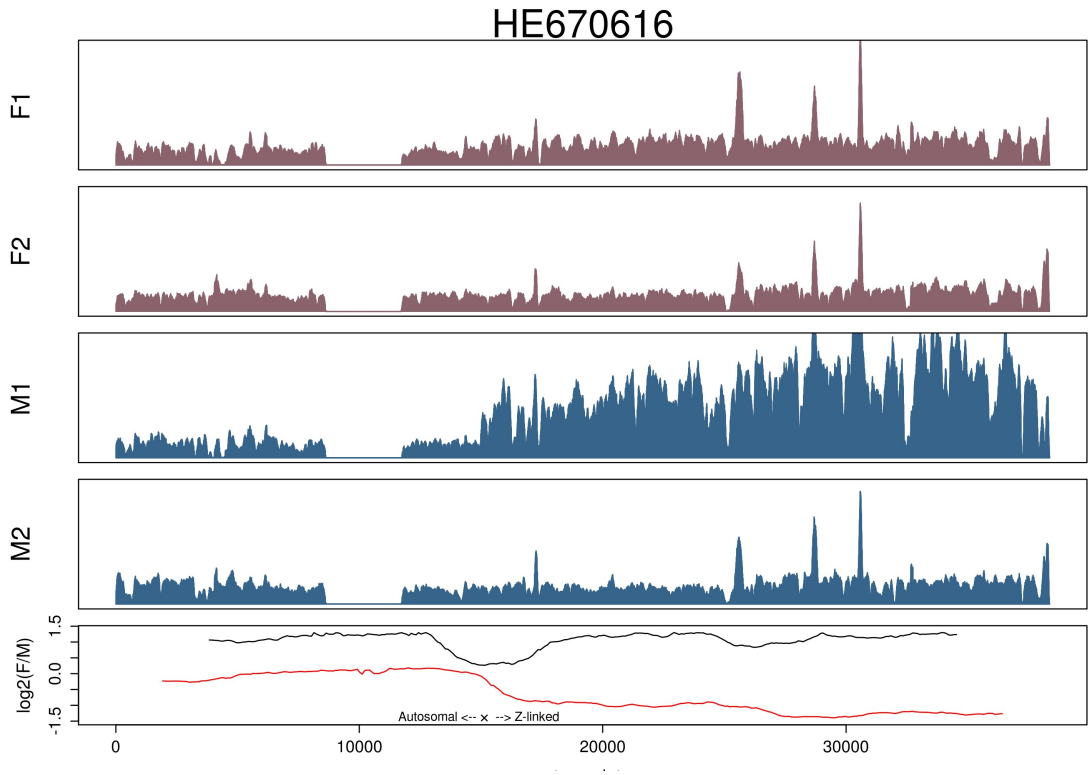

Figure A5. Coverage for scaffold HE67016. Scaffold HE67016 was mapped to an autosome but shows a scaffold-wide coverage log-ratio of -.9 when male & female samples are averaged. However the anomalously high coverage in the first male sample strongly skews the average value, while the second male's coverage is comparable to females. Plot details are as described in Fig. A3

In conclusion, we believe that this analysis of sequencing coverage between males and females has greatly increased resolution and confidence in Z-linked portions of the *H. melpomene* genome assembly. This result is an important foundation in future functional and evolutionary studies of sex-chromosome in this species and other lepidopterans.