Supplementary File 3

A linkage disequilibrium perspective on the genetic mosaic of speciation in two hybridizing Mediterranean white oaks

Pablo G Goicoechea[1*], Ana Herrán[1], Jerome Durand[2.3], Catherine Bodénès[2,3], Christophe Plomion[2,3], Antoine Kremer[2,3]

[1]Department of Biotechnology, NEIKER-Tecnalia, P.O. Box 46, 01080 Vitoria-Gasteiz, Spain

[2]INRA, UMR 1202 BIOGECO, F-33610 Cestas, France

[3]Univ. Bordeaux, BIOGECO, UMR 1202, F-33170 Talence, France

Keywords: genetic mosaic, outliers, linkage disequilibrium, divergence hitchhiking, genomic hitchhiking, *Quercus*.

Running Title: LD and the genetic mosaic of speciation in oaks

Null Alleles, Population Structure and Genetic Diversity and Differentiation

Results

Thirty-five markers showed evidence of null alleles in at least one out of the sixteen sampled populations (Table S3-1). However, most of these markers detected null alleles in just one or two populations (19 markers), and only a few showed null alleles in a large number of populations. On the other hand, none of the 16 populations was free of null alleles for at least one marker, although the number of markers with null alleles per population also varied considerably (from two to twelve).

Homozygote's surplus in the bulked *Q. faginea* and *Q. pyrenaica* populations affected nearly two thirds of the markers (Supplementary. File 2). Even though other causes could account for such excess in synthetic populations (i.e., Wahlund's effect), we followed a conservative approach and estimated the effects of null allele frequencies on the fixation index (F_{ST}) for all markers with significant F_{IS} values. Very few markers showed any significant effect of the *ENA* dataset on the fixation index estimates (Figure S3-1). However, it is noteworthy that the two markers that showed the largest differences between both estimates of the differentiation coefficient (VIT031.1 and PIE202) were detected as outlier loci by the ln*RH* method (see main text). The overall lack of influence over the F_{ST} estimates argue against bias introduced by null alleles in the F_{ST}-based clustering of samples and in the diversity and differentiation estimates. However, it is not clear how null alleles would influence LD estimates, and hence the score tests for association, a subject that warrants further investigation.

We used the full dataset to cluster the studied genotypes from both species. The Evanno *et al.*, (2005) *ad-hoc* maximizations supported the existence of two groups (Fig. S3-2) that closely matched the two species; i.e., differentiation was much larger between species than among populations within species. The intra-individual ancestries plot (Fig. S3-3) showed

that most populations were composed of pure-bred individuals, but some trees also showed variable amounts of mixed ancestry, indicating possibly the existence of admixture. Hybridization between *Q. faginea* and *Q. pyrenaica* occurred mainly in the central part of Spain (latitudes 39-40º; Cabañeros National Park and Talayuela), as deduced both from field observations and from the trees ancestry coefficients. Trees with mixed ancestry from Monasterio de la Sierra and from Montejo de la Sierra probably indicate admixture between *Q. pyrenaica* and *Q. petraea*, as they were sampled from mixed forests of these two species and the closest *Q. faginea* trees are far from those locations. These results suggest that reproductive isolation between the two Mediterranean oaks is not complete yet.

Genetic differentiation among populations (F_{ST}), within species, was very low (Table S3-2). However, it was significant among the *Q. faginea* populations when the AMOVA results were calculated as a weighted average over loci (i.e., having into account the different numbers of missing data for the different markers). This analysis should be viewed with caution, as the low number of individuals per population creates a large sampling variance and coefficient of variation in the F_{ST} estimates.

Genetic diversity was similar in the two species (Supplementary. File 2) in spite of some large differences at particular marker loci. Mean allelic richness was a little lower in *Q. faginea* than in *Q. pyrenaica* (12.40 *vs.* 12.85), while gene diversities were alike in both species (0.83 and 0.85, respectively). Di-nucleotide repeat motifs showed larger allelic richness and heterozygosities than tri- and hexa-nucleotide repeat motifs, but the same trends were observed in both groups.

Genetic differentiation between the two species (Jost's D) ranged from 0 (non-differentiation) to 0.95 (no alleles in common for D = 1) among the 98 markers (Fig. S3-4). Single-marker differentiation was general over the genome, as only nine markers distributed on four LG showed non-significant differentiation; i.e., their lower confidence intervals contained zero.

However, there was considerable variation across the genome and within LGs too. For instance, LG2 concentrated 7 highly differentiated markers (D > 0.5) and 8 markers with very low differentiation (D < 0.1). On the other hand, LG12 presented all 12 markers with intermediate differentiation (0.2 < D < 0.5), and LG4 displayed a clustering of non-differentiated markers, with four out of the nine non-differentiated markers mapping to this LG. Noteworthy, two markers from LG7 presented the highest differentiation (D > 0.95), while coalescent simulations indicated they were not outliers (see main text).

The variability for G_{ST} (and therefore, for the fixation index F_{ST}) was not as pronounced as for D (Fig. S3-5), a result expected on the basis of the maximum possible G_{ST} dependency on the observed homozygosity (Hedrick, 2005). Its major feature was the large values for several markers with low allelic richness. A pair-wise comparison between D and G_{ST} values indicated a few similarities but also several differences. Most important similarities were the non-significant/small differentiation in LG4, LG5 and LG9, and the outstanding large differentiation for markers PIE127 and PIE137 in LG7. The main discrepancies were: (i) G_{ST} was much more uniform than D, only 5 markers outstanding for their large G_{ST} values (> 0.075), (ii) mean differentiation in LGs#1, 2, 3, 8 and 10 was comparatively much lower measured by G_{ST} than by D, (iii) individual locus differences were common, although a large disagreement between G_{ST} and D was more evident at two loci in LG2 (FIR032 and ZQP119), one locus in LG3 (GOT021), LG8 (PIE054), LG9 (PIE081) and LG11 (PIE202), and two loci in LG12 (ZQR30 and VIT050).

References

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, **14**, 2611-2620.

Hedrick PW (2005). A standardized genetic differentiation measure. *Evolution*, **59**, 1633-1638.

Table S3-1: Estimates of null allele frequencies for each of the 98 markers in the 16 sampled populations. Empty cells indicate that no null alleles were detected.

Locus	Anc-Py	Cab-Fg	Cab-Py	Fer-Fg	FRo-Fg	Izki-Fg	Izki-Py	Lli-Py	MSi-Py	Mtj-Py	RLo-Fg	SCu-Fg	SNe-Fg	SNe-Py	Tal-Fg	Tal-Py	Nb of Pops with Null Alleles
FIR073		0.18	0.15		0.17				0.21								4
VIT026			0.24					0.30					0.25				3
POR040																	
VIT007											0.17						1
WAG011																	
FIR040																	
VIT020																	
FIR065																	
ZQP104																	
ZQR87																	
FIR112			0.34														1
ZQP36											0.36						1
WAG019																	
FIR048																	
PIE136																	
POR041																	
WAG054																	
ZQP119				0.28	0.20	0.44											3
WAG032											0.23						1
POR038																	
PIE197																	
FIR016																	
FIR032																	
PIE135		0.22				0.33											2
GOT021																	

PIE021								0.15							0.26	2
PIE198																
FIR075	0.28							0.33	0.32	0.20			0.28			5
POR011			0.32													1
PIE051																
VIT084																
VIT107																
PIE088		0.29	0.38		0.29	0.31		0.24	0.38		0.26	0.24	0.37	0.26	0.32	0.21	12
VIT043																
PIE082																
POR029																
VIT003				0.20		0.17	0.21	0.25					0.22	0.20	0.18	0.20	8
VIT017																
PIE002																
WAG043									0.23							1
POR014																
GOT065																
VIT023					0.24				0.23	0.23					0.24	4
POR039																
VIT021																
POR024																
POR034																
VIT028				0.32				0.22		0.24	0.30	0.23				5
POR027								0.26								1
PIE071												0.22		0.21		2
PIE137																
VIT046																
FIR030										0.20						1
PIE127																
POR025											0.22			0.14		2
VIT009																

PIE011		0.29								0.20	0.20					3
PIE023																
PIE054									0.24							1
POR008																
PIE101					0.27				0.22	0.32						3
FIR046						0.19		0.23	0.23			0.25	0.18			5
VIT013			0.32												0.31	2
PIE090												0.35			0.35	2
WAG014																
PIE249																
VIT031.1	0.41	0.26	0.36	0.28			0.41	0.36				0.26		0.26	0.35	9
VIT031.2																
PIE155																
VIT099																
PIE259	0.39							0.25							0.18	3
PIE059																
FIR005																
PIE081																
POR028	0.29	0.24	0.27					0.23		0.28	0.15			0.34	0.21	0.20	9
ZQR96																
PIE148																
ZQR11		0.29		0.28		0.18		0.19		0.35	0.24			0.38	0.25	0.15	9
PIE186																
PIE187																
PIE144																
PIE257																
WAG079																
FIR104																
PIE102																
PIE202								0.34								1
PIE089																

GOT009					0.23												1
VIT010		0.27															1
PIE126																	
GOT032																	
PIE196																	
VIT037						0.23						0.17	0.32				3
PIE236																	
ZQR112																	
POR020																	
ZQR30									0.29								1
VIT050																	
Nb Markers with Null Alleles	4	8	8	5	6	7	2	9	7	8	12	5	9	6	7	10	

Table S3-2: AMOVA results and design, as a weighted average over 98 loci.

	Quercus faginea				*Quercus pyrenaica*			
Source of variation	Degrees of freedom[1]	Sum of squares	Variance components	Percentage variation	Degrees of freedom	Sum of squares	Variance components	Percentage variation
Among populations	6	363.57	1.025 Va	2.87	7	379.23	0.691 Va	1.89
Among individuals within populations	70	2649.93	3.676 Vb	11.28	77	3007.00	4.095 Vb	11.19
Within Individuals	77	2359.00	31.047 Vc	86.85	85	2649.00	31.800 Vc	86.92
Total	153	5269.01	34.595		153	6035.22	36.586	
		$F_{ST} = 0.029$	$p = 0.027$			$F_{ST} = 0.019$	$p = 1.000$	

1: The Talayuela population was not included as it contained only 6 purebred individuals

Figure S3-1: Comparison between the F_{ST} estimates obtained with the original data assuming no null-alleles were present (F_{ST}) and with the dataset excluding null alleles (ENA-F_{ST}), for the 60 markers showing significant inbreeding values in the bulked samples of the two oak species.

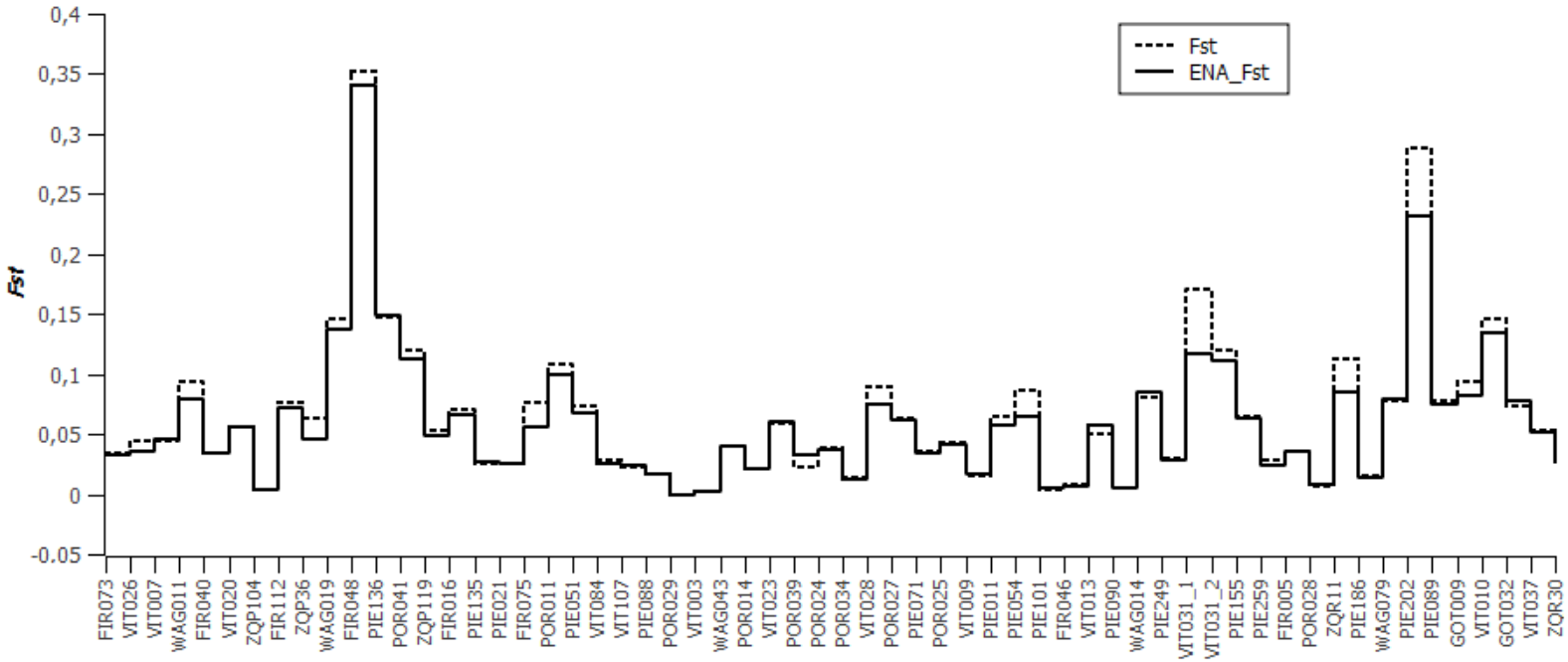

Figure S3-2: Model probabilities [LnP(D)] and its maximization (Δ, DeltaK) for Bayesian clustering with different number of groups (*K* = 1-17)

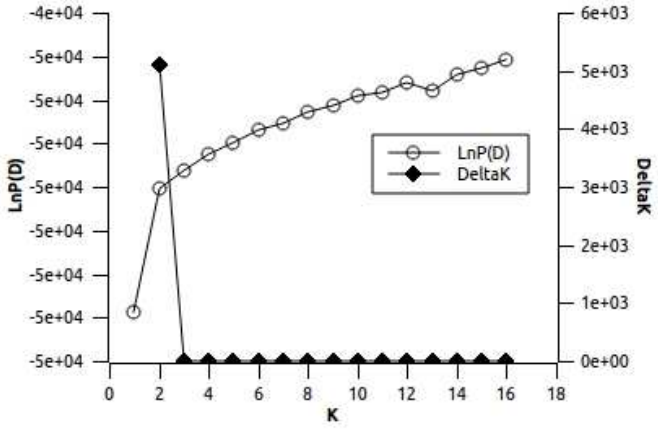

Figure S3-3: Individual ancestry coefficients (*K*=2) for the 192 trees analyzed in this study.

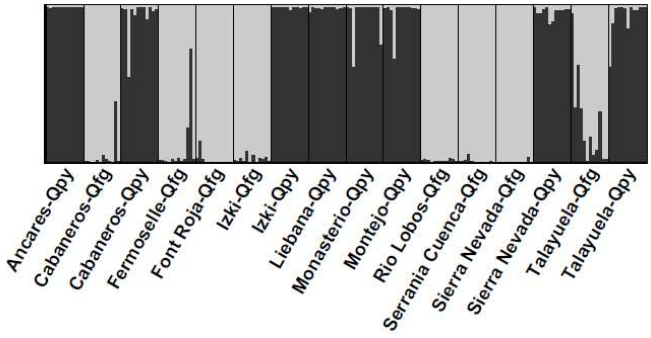

Figure S3-4: Inter-specific differentiation estimates, together with 95% confidence intervals, using Jost's D

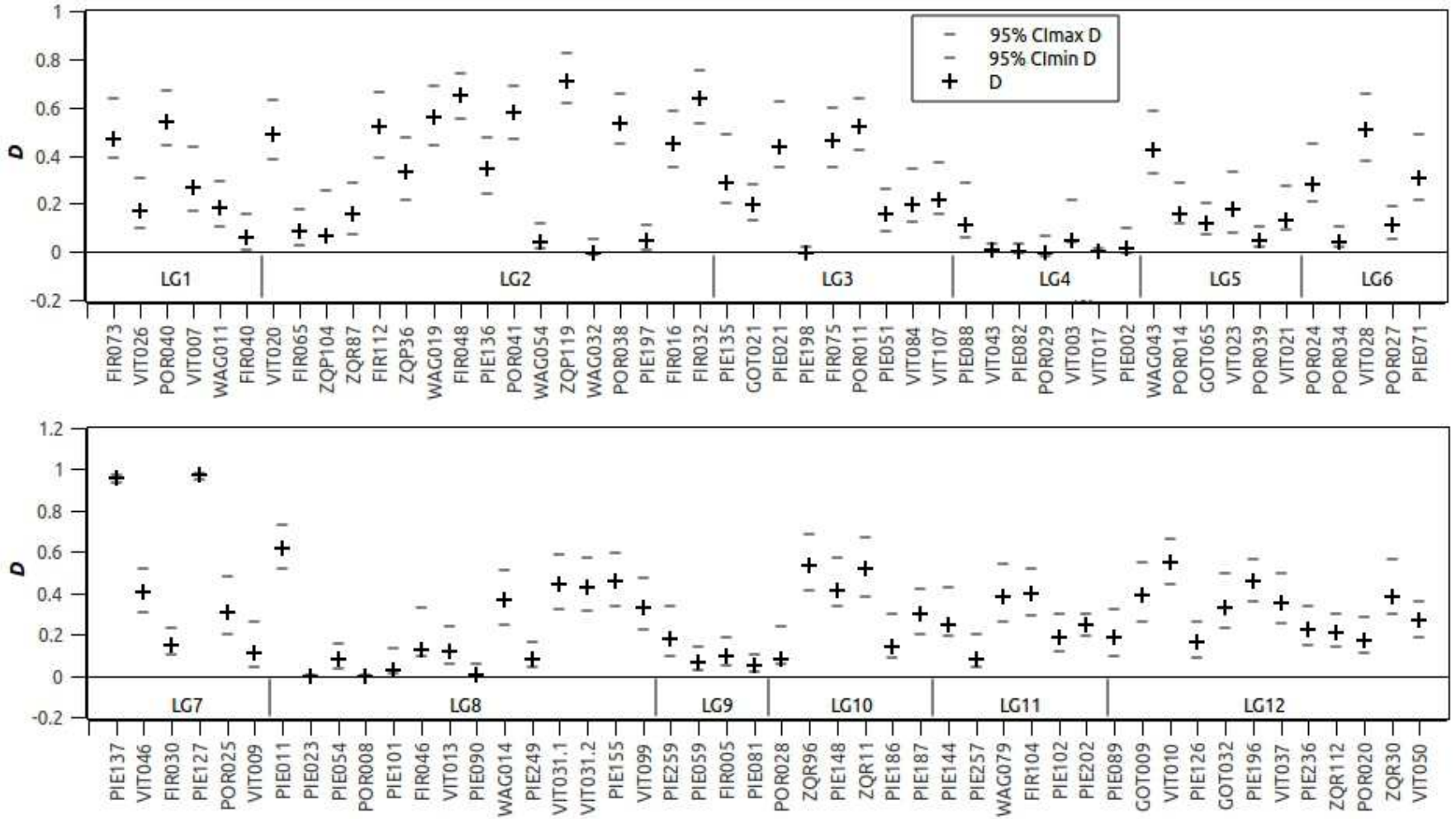

Figure S3-5: Inter-specific differentiation estimates, together with 95% confidence intervals, using G_{ST}.

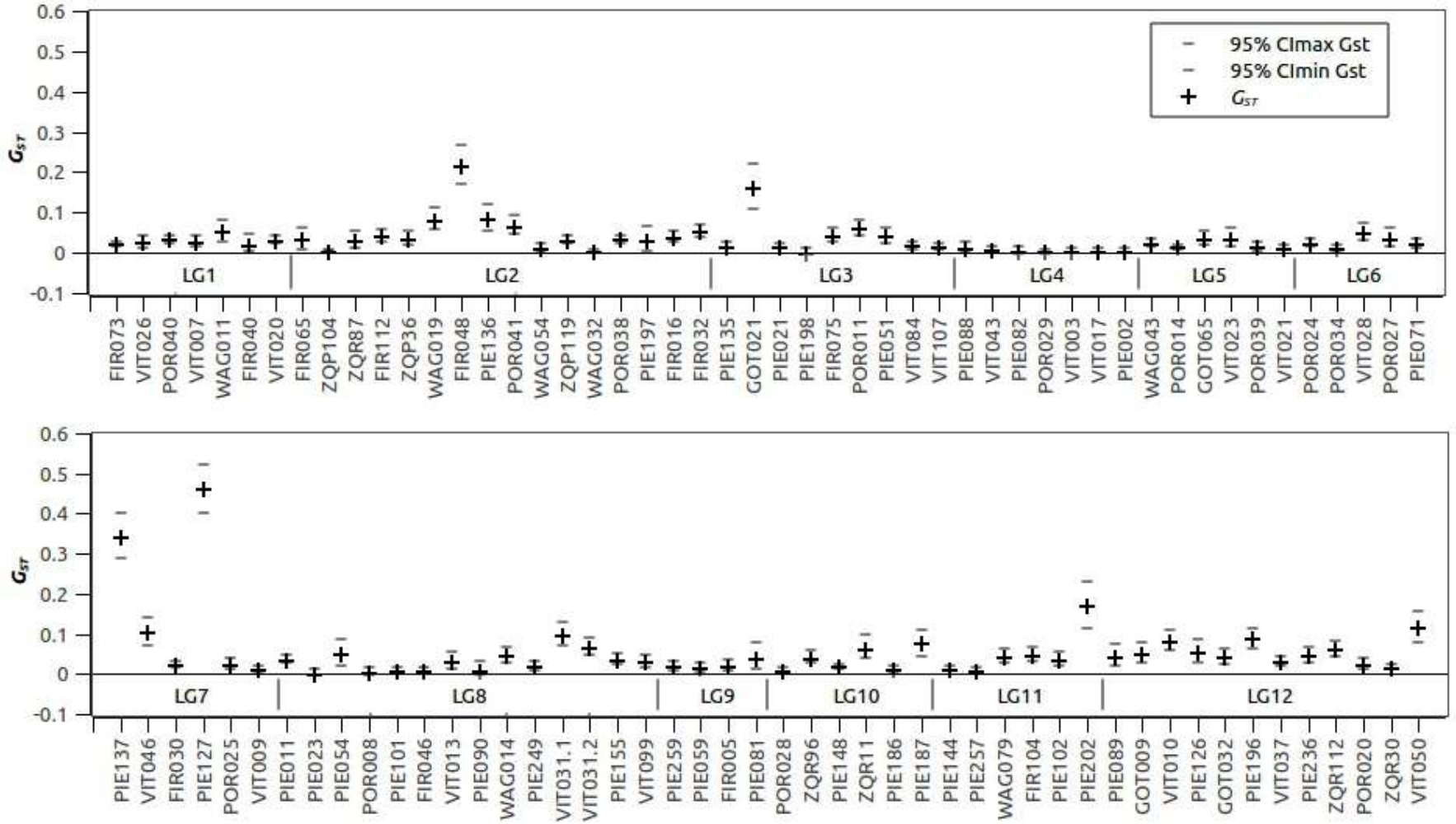

Figure S3-6: Pair-wise comparisons between the D and Gst estimates of inter-specific differentiation.