Heredity (2007) 99, 331–339

www.nature.com/hdy

ORIGINAL ARTICLE

Inferring gene flow in coral reef fishes from different molecular markers: which loci to trust?

C Fauvelot[1,2,3,5], C Lemaire[3,4], S Planes[1,2] and F Bonhomme[3]

[1]*UMR 5244 EPHE/CNRS/UPVD, Université de Perpignan, Perpignan cedex, France;* [2]*UMS 2978 EPHE/CNRS, Centre de Recherches Insulaires et Observatoire de l'Environnement, Papetoai, Moorea, French Polynesia;* [3]*Département Biologie Intégrative, Institut des Sciences de l'Evolution, CNRS UMR 5554 Université Montpellier 2, Station Méditerranéenne de l'Environnement et du Littoral, Sète, France and* [4]*Laboratoire 'Paysages et Biodiversité' PPF-DS10, Université d'Angers, Angers cedex, France*

Contrasting results are usually reported in the literature regarding the factors influencing observed structuring of genetic variability. The goals of this study were, for five coral reef fishes in French Polynesia, (1) to infer the theoretical variance of single locus F_{ST} estimates expected under neutrality in order to exclude outlier loci before inferring gene flow and (2) to test thereafter whether species laying pelagic eggs effectively disperse more than species laying benthic eggs in this system. For this purpose, a total of 952 individuals from five species belonging to two families (Chaetodontidae and Pomacentridae) were screened among populations sampled within a 60–600 km spatial range for intron length polymorphism at 11 loci in order to illuminate contrasting results previously published on allozymes and mitochondrial DNA (mtDNA) control region polymorphisms. Statistically speaking, among the five species, four loci (three allozymes and one intron) were identified as outliers and discarded before interpretation of genetic differentiation in terms of effective dispersal. Biologically speaking, our results suggest that the observed genetic structure is not significantly related to the reproductive strategy of coral reef fish in the island system we analysed and that observed random genetic differentiation accommodates Wright's island model in all five species surveyed. Overall, our study emphasizes how cautious one has to be when trying to interpret present-day genetic structure in terms of gene flow while using a limited number of loci and/or different sets of loci.
Heredity (2007) **99,** 331–339; doi:10.1038/sj.hdy.6801005; published online 23 May 2007

Keywords: coral reef fishes; allozymes; introns; mtDNA; gene flow; neutrality

Introduction

Studying spatial partitioning of genetic variation is viewed as a robust method for inferring gene flow among discrete populations, conditional on taking into account all processes shaping the present genetic structure (Hewitt, 2000). For coral reef fishes, contrasting results have been found relative to the relationship between present genetic structure and larval duration (Waples, 1987; Doherty *et al.*, 1995; Shulman and Bermingham, 1995; Rocha *et al.*, 2002; Bay *et al.*, 2006), reproductive strategy (Shulman and Bermingham, 1995; Planes *et al.*, 1997), species swimming abilities (Waples, 1987; Riginos and Victor, 2001; Purcell *et al.*, 2006) or geographic distances (Planes *et al.*, 1993, 1996; Planes and Fauvelot, 2002; Purcell *et al.*, 2006). Fauvelot and Planes (2002) have shown, using allozymes, that the observed significant structuring of genetic variation among populations of eight coral reef fishes in a fragmented island system was not correlated with geographic distances at the studied spatial scale (ca 1500 km) nor associated with species reproductive strategy and larval duration. On average, only a third of the polymorphic loci contributed to the overall genetic differentiation. Conversely, the analysis of the genetic variation at the mitochondrial DNA (mtDNA) control region sequences in a subset of individuals revealed genetically homogeneous populations in all seven species under study (Fauvelot *et al.*, 2003).

Such puzzling discordant patterns of genetic variation obtained at different loci are increasingly described (Karl and Avise, 1992; Pogson *et al.*, 1995; Lemaire *et al.*, 2000; De Innocentiis *et al.*, 2001; Dufresne *et al.*, 2002; Arnaud-Haond *et al.*, 2003; Gomez-Uchida *et al.*, 2003; Nielsen *et al.*, 2006) and lead to question the interpretation made of genetic differentiation (F_{ST}) estimates in terms of effective dispersal. A key issue is that, under neutrality, all loci should theoretically be affected similarly by drift and migration (Lewontin and Krakauer, 1973; Slatkin, 1987), so that F_{ST} estimates should be homogeneous among loci, and multilocus F_{ST} estimates may be used to infer gene flow among populations. Yet, erroneous gene flow estimates may result from the interpretation of allelic distribution over multiple loci for two main reasons. One lies in the underlying high variance of the migration/drift process itself and its sensitivity to departure from equilibrium (Whitlock and McCauley, 1999). The other resides in selection or other non-neutral mechanisms (acting directly or through hitch-hiking) differentially affecting loci (Cavalli-Sforza, 1966; Slatkin, 1987). Indeed, disruptive selection will result in higher

Correspondence: Dr C Fauvelot, Scienze Ambientali, University of Bologna, Via S Alberto 163, I-48100 Ravenna, Italy.
E-mail: cecile.fauvelot@unibo.it
[5]*Current address: Scienze Ambientali, Università di Bologna in Ravenna, Via S Alberto 163, I-48100 Ravenna, Italy.*
Received 14 November 2006; revised 13 March 2007; accepted 16 April 2007; published online 23 May 2007

F_{ST} estimates than expected under neutrality, inflating the overall genetic differentiation. On the other hand, loci experiencing balancing selection will have allele frequencies more similar than expected under neutrality, reducing the F_{ST} estimates and further giving the false view of high gene flow.

In order to properly infer gene flow among populations, it appears therefore essential to (1) estimate genetic differentiation based on reasonable numbers of individuals (50–100 individuals; Ruzzante, 1998), loci (15–20 loci; Nielsen *et al.*, 2006) and alleles (either few loci with many alleles or many loci with a few alleles; Kalinowski, 2002) and (2) discard loci behaving non-neutrally before the estimation of genetic differentiation. Several neutrality tests are available in the current literature that may allow detection of selected loci at ecological timescales (reviews in Guinand *et al.*, 2004; Beaumont, 2005). Among these tests, those derived from the Lewontin–Krakauer method (that is, based on multilocus comparisons and the identification of F_{ST} outliers) appear robust when allele distributions are available at multiple loci (Beaumont and Nichols, 1996; Beaumont and Balding, 2004; Beaumont, 2005).

The goal of the present study is, for five species of coral reef fish, to infer gene flow among insular populations sampled within a 60–600 km spatial range and to test for the influence of reproductive strategies. Considering the problem of neutrality among loci, the theoretical neutral variance of single locus F_{ST} estimates in our insular system was first investigated in order to identify (and exclude from calculation) outlier loci before computing gene flow. For this purpose, the number of loci was increased to improve the power of the test. We reanalysed all previously assayed individuals from five species with a new set of loci with potentially higher mutation rates and gene diversities than allozymes, that is, intron length polymorphisms. Unlike exons, the nucleotide sequences of introns are little constrained so that introns accumulate mutations rapidly. Therefore, introns often display a high genetic variability (Palumbi, 1995), which make introns being targeted as markers of population variation and subdivision (Lessa, 1992; Borsa, 2003; Hassan *et al.*, 2003; Friesen *et al.*, 2005; Rohfritsch and Borsa, 2005; Berrebi *et al.*, 2006). Exon-primed, intron-crossing polymerase chain reaction (EPIC-PCR) amplification is an appreciable strategy for finding DNA polymorphism in several species of different families (Hassan *et al.*, 2002) and an alternative to the development of microsatellite loci. Length polymorphism at 11 intron loci was assessed in five coral reef fish species and the corresponding new data set was further added to the previous allozymes data set (Fauvelot and Planes, 2002) to conduct global analyses. The comparison of the distribution of genetic variation obtained from different markers using the same set of individuals from related species collected in the same area also provide a solid framework to test among mutation, migration and genetic drift as the main forces driving the distribution of genetic polymorphisms.

Materials and methods

Sampling

Sampling was carried out as described in Fauvelot and Planes (2002). Briefly, five different islands were sampled across French Polynesia based on a standard sampling effort of 50 individuals per island per species. Because the eight species selected for the original study were not all present in the five islands, and in order to compare effective dispersal among species at similar scales, we restricted further analysis to five species inhabiting four common islands: Rangiroa, Takapoto, Tetiaroa and Moorea (Figure 1). Geographic distances between islands ranged from 60 km (between Moorea and Tetiaroa) to 566 km (between Moorea and Takapoto). *Chaetodon citrinellus* is a butterflyfish strictly inhabiting lagoons while *C. quadrimaculatus* and *Forcipiger flavissimus* are both outer reef butterflyfishes. *Dascyllus aruanus* and *Pomacentrus pavo* are two damselfishes strictly inhabiting lagoons. Fishes were collected, tissue stored and DNA extracted as described in Fauvelot *et al.* (2003). A total of 952 individuals from five species (Table 1) were used for further genetic screening.

Genetic analysis – EPIC-PCR

Eleven introns loci were tested for length polymorphism for each species: intron 7 of creatine kinase; intron 2 of glyceraldehyde-3-phosphate dehydrogenase; introns 2 and 5 of growth hormone; introns 1, 2, 4 and 5 of aldolase B; introns 1, 2 and 3 of gonadotropin-releasing hormone (Hassan *et al.*, 2002). PCR amplification and electrophoresis follow protocols described in Hassan *et al.* (2002). Briefly, amplified PCRs (with forward primer labelled radioactively with ^{33}P) were loaded on a 6% polyacrylamide gel in TBE buffer. The gels were run for 3 h at 50 W. After drying, the gels were exposed against films of Biomax autoradiograph. Individual genotypes were directly scored from the autoradiograph films with

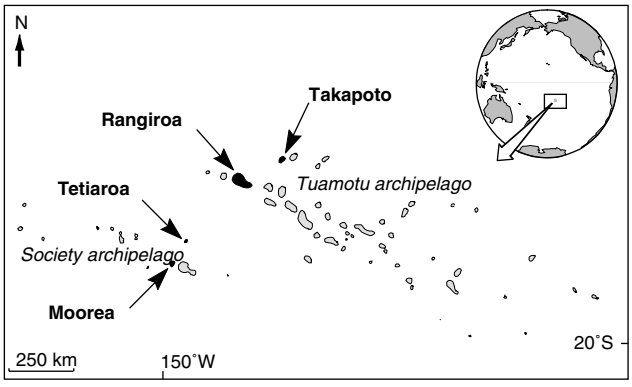

Figure 1 Location of sampling sites of the five fish species in French Polynesia (South Pacific Ocean). Details of sites are reported in Fauvelot and Planes (2002) and sampling effort is indicated in Table 1.

Table 1 Number of individuals scored at enzymatic and introns loci, indicated per species and per site

Species	*Rangiroa*	*Takapoto*	*Tetiaroa*	*Moorea*	*Total*
Chaetodon citrinellus	50	50	50	50	200
Chaetodon quadrimaculatus	50	48	49	49	196
Forcipiger flavissimus	50	50	49	39	188
Dascyllus aruanus	46	48	50	49	193
Pomacentrus pavo	27	50	48	50	175

alleles being identified according to their size relative to the most common allele, as for allozymes. To ensure scoring accuracy and consistency across gels and populations, a locus-specific size standard from a pool of individuals with known genotypes was run at both sides of each gel. With this method, we were able to segregate 1 bp length difference bands. Individuals for which no PCR products were visualized on gel were further re-amplified in a new PCR using a 2°C lower annealing temperature. Results of the amplification successes across species are presented in Hassan *et al*. (2002) and detailed PCR conditions for each species and locus are available from the authors.

Data analysis

The allozymes data set was reanalysed for genetic divergences in the overall populations following the removal of populations from the data set (see sampling). Allele frequencies, number of alleles per locus (N_A), observed heterozygosity (H_O) and unbiased gene diversity (H_E; Nei, 1987) were calculated within each population and over all populations for both allozymes and introns. Linkage disequilibrium between loci was tested using the procedure of Black and Krafsur (1985). Deviation from Hardy–Weinberg expectations (HWE) at introns loci was measured using Weir and Cockerham (1984) inbreeding coefficient f (equivalent to Wright's F_{IS}), and locus conformance to HWE was assessed by permuting alleles within populations (1000 permutations). The presence of null alleles was tested using the program MICRO-CHECKER (Van Oosterhout *et al*., 2004). Genetic divergences were estimated using the Weir and Cockerham (1984) estimator. For each species, overall single locus F_{ST} values were estimated by jackknifing over populations and overall global F_{ST} estimates were obtained by jackknifing over loci. Unless specified, all computations were conducted using GENETIX (Belkhir *et al*., 2002).

The theoretical distribution of single locus F_{ST} estimates expected under neutrality was computed through simulations and outlier loci identified using Beaumont and Nichols' (1996) test. The principle of this test is to estimate, conditional on gene diversity, the range of F_{ST} values expected under the combined effect of migration and genetic drift in a model approaching the studied system, and to compare the observed F_{ST} at this given locus to the simulated empirical distribution. The models for computer simulations were chosen as to fit at maximum the conditions of the studied system. Indeed, we chose the infinite allele model of evolution (Kimura and Crow, 1964), and the infinite island model of Wright (1943) as this is the best-suited available model for fragmented insular environments such as Pacific reefs. It should be noted that, for heterozygosities lower than 0.8 (which is our case), stepwise and infinite-allele models lead to similar simulated distributions of F_{ST} (Beaumont and Nichols, 1996), so that the choice of the mutation model has little impact on the simulated empirical distribution.

Coalescent simulations were performed to generate data sets with a distribution of F_{ST} close to the empirical distribution. Simulations were conducted assuming four sampled populations, with a metapopulation size of 100 populations (the approximate number of islands in French Polynesia), sample sizes of 50 individuals and an expected F_{ST} equal to the median estimate of F_{ST} calculated for all loci. Under the infinite allele mutation model, 20 000 H_E and F_{ST} paired values were generated and the P-values for each locus were then calculated using the procedure described in Beaumont and Nichols (1996). Loci with unusually low or high F_{ST} estimates (that is, out of the 95% distribution of simulated F_{ST}), conditional on heterozygosity, were regarded as outlier loci. We used the currently distributed version of the test, namely Fdist2 (http://www.rubic.reading.ac.uk/~mab/software/fdist2.zip), described in Beaumont and Balding (2004). Because the test conducts simulations based on initial expected F_{ST} value set up by the user, this value was first fixed equal to the global F_{ST} (estimated over all loci). However, because this median value may be biased owing to the presence of an outlier locus, we systematically conducted again the simulations with a new global F_{ST} estimated omitting the outlier locus (or loci), and so on until no more loci were identified as outliers.

Overall populations and pairwise F_{ST} were estimated for each species over all nuclear loci (allozymes and introns) once removing outliers detected by the Beaumont and Nichols' (1996) test. Significance of pairwise genetic divergence was tested using a permutation procedure (1000 permutations) in GENETIX (Belkhir *et al*., 2002). Mantel tests were performed to test for a correlation between geographical distance and multilocus F_{ST} (approach of Hutchison and Templeton, 1999). For each species, Mantel coefficient Z was computed and tested for significance through a permutation procedure with 100 permutations, under the null hypothesis that both matrices are not linearly related. To evaluate the level of correlation between genetic patterns exhibited by the different species, matrices of pairwise F_{ST} were compared with a Mantel test: Pearson correlation coefficient r was computed for each pair of species and its significance tested using a permutation procedure with 100 permutations. All tests were performed in GENETIX using the Mantel option.

Results

Out of the 11 intron primer pairs tested, five loci (*AldoB1*, *AldoB5*, *CK7*, *GnRH3-3* and *Gpd2*) could not be further used for the five species because they lead to multibanded (several loci amplified), monomorphic or not clearly interpretable patterns (Table 2). The remaining six loci (*AldoB2*, *AldoB4*, *GnRH3-1*, *GnRH3-2*, *GH2* and *GH2*) provided single locus amplifications with clearly interpretable length polymorphism, and finally, 2–3 polymorphic intron per species could be analysed (Table 2). Over all populations, the number of alleles per locus ranged from 2 to 12 (for *GH2* over all *F. flavissimus*-sampled populations). Gene diversity among populations in each species was more or less even and variations in gene diversity were locus specific, such as that observed at enzymatic loci (Fauvelot and Planes, 2002), with no congruent patterns among species. Over all four populations, gene diversity at intron loci reached 0.726 for *F. flavissimus* at *GH2*, and 0.501 at enzymatic loci for *F. flavissimus* at *Est1* (Table 3).

After standard Bonferroni correction of the 5% significance level, a single locus significantly deviated from HWE; *GnRH3-1* in *C. citrinellus* populations. The

Table 2 Amplifications results of the 11 tested

	Chaetodon citrinellus				Chaetodon quadrimaculatus				Forcipiger flavissimus				Dascyllus aruanus				Pomacentrus pavo			
	Ra	*Ta*	*Te*	*Mo*	*Ra*	*Ta*	*Te*	*Mo*	*Ra*	*Ta*	*Te*	*Mo*	*Ra*	*Ta*	*Te*	*Mo*	*Ra*	*Ta*	*Te*	*Mo*
AldoB1																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	Poly	Poly	Poly	Poly	Poly	Poly	Poly	Poly
AldoB2																				
H_E	0.115	0.078	0.099	0.098	M	M	M	M	M	M	M	M	0.145	0.082	0.217	0.136	M	M	M	M
H_O	0.120	0.080	0.102	0.061									0.152	0.083	0.160	0.143				
N_A	4	4	4	4									3	3	3	3				
AldoB4																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	0.472	0.428	0.421	0.467	0.492	0.494	0.494	0.481
H_O													0.421	0.478	0.409	0.511	0.444	0.567	0.483	0.460
N_A													2	2	2	2	2	2	2	2
AldoB5																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	nt	nt	nt	nt	nt	nt	nt	nt
CK7																				
H_E	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli	nt	nt	nt	nt	nt	nt	nt	nt
GnRH3-1																				
H_E	0.702	0.628	0.657	0.693	0.589	0.636	0.627	0.552	M	M	M	M	M	M	M	M	M	M	M	M
H_O	0.480**	0.240**	0.388**	0.449**	0.553	0.667	0.636	0.556												
N_A	8	5	7	8	5	5	5	6												
GnRH3-2																				
H_E	0.518	0.490	0.565	0.524	0.549	0.473	0.539	0.556	0.443	0.219	0.502	0.456	M	M	M	M	M	M	M	M
H_O	0.400*	0.460	0.510	0.347**	0.583	0.512	0.400*	0.575	0.396*	0.234	0.500	0.400								
N_A	4	4	4	4	3	3	3	4	7	6	7	5								
GnRH3-3																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	Poly	Poly	Poly	Poly	M	M	M	M
GH2																				
H_E	M	M	M	M	M	M	M	M	0.774	0.726	0.746	0.658	Muli	Muli	Muli	Muli	Muli	Muli	Muli	Muli
H_O									0.745	0.714	0.838	0.615								
N_A									11	11	8	8								
GH5																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	0.373	0.364	0.300	0.311
H_O																	0.333	0.400	0.353	0.260
N_A																	2	2	2	2
Gpd2																				
H_E	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M

Abbreviations: H_E, unbiased gene diversity (Nei, 1987); H_O, observed heterozygosity; M, monomorphic (no length polymorphism); Mo, Moorea; Multi, multiband pattern (more than one locus amplified); N_A, number of alleles; nt, not tested; Poly, length polymorphism but hardly interpretable; Ra, Rangiroa; Ta, Takapoto; Te, Tetiaroa.
Intron names are given as in Hassan *et al*. (2002).
Significant deviation from Hardy–Weinberg proportions is denoted by * for $P<0.05$ and ** for $P<0.01$.

Table 3 Gene diversity (H_E) over all four sampled populations and F_{ST} estimates for each scored locus (allozymes and introns) per species

	C. citrinellus		C. quadrimaculatus		F. flavissimus		D. aruanus		P. pavo	
	H_E	F_{ST}	H_E	F_{ST}	H_E	F_{ST}	H_E	F_{ST}	H_E	F_{ST}
Allozymes										
AAT-1	0.273	−0.005	0	—	0.091	0.000	0	—	0	—
AAT-2	0.216	−0.007	0.020	0.020	0.108	0.005	0.196	−0.009	0	—
ADA-1	0.511	0.012	0.005	0.000	0.019	0.030	0.191	0.198	0.005	−0.002
ADA-2							0.371	0.066		
ADH					0.005	−0.001				
CK	0	—	0	—	0.005	−0.001	0	—	0	—
Est1			0.326	0.025	0.501	0.012	0.358	0.019	0.254	0.047
Est2	0.481	0.014	0.050	−0.004	0.224	0.001			0.208	−0.004
Est3			0.473	0.070						
EstD					0.036	0.006				
GDA	0.345	0.036	0.375	0.025	0.444	0.088				
GDH					0.268	−0.007				
GPD-1	0.020	−0.001	0.015	0.002	0.015	0.001	0.214	0.030		
GPD-2					0.010	0.004	0.070	0.000	0.422	−0.009
GPI-1	0.078	0.002	0.134	0.002	0.010	−0.007	0.475	−0.004	0	—
GPI-2	0.020	−0.001	0.015	0.002	0.043	−0.001	0.031	−0.008	0	—
IDH-1	0.374	−0.001	0.005	0.000	0.012	−0.002	0.127	−0.008	0	—
IDH-2	0	—	0.020	−0.004	0.145	0.038	0.116	0.050	0.100	−0.011
LDH	0.121	0.019	0	—	0.017	0.001			0	—
MDH-1	0	—	0.005	0.000					0	—
MDH-2	0	—	0.010	0.010	0	—	0.103	0.000		
MEP-1	0.068	0.001	0.074	−0.009	0.015	−0.002	0.270	0.014	0.010	−0.004
MEP-2	0	—	0.264	0.026	0	—	0	—	0.381	0.042
MPI	0	—	0	—	0.036	−0.006			0.010	−0.004
PepB-1	0	—	0	—			0.390	0.019	0.005	−0.001
PepB-2	0.181	0.032	0.035	0.008	0.011	−0.005			0.005	−0.002
PepD-1	0.401	0.023	0.005	0.000	0.105	−0.003	0.294	0.077	0.014	0.000
PepD-2									0.322	0.028
PGAM	0	—	0.005	0.000					0.406	0.026
PGDH	0.098	0.027	0.390	0.016	0.021	0.000	0	—		
PGM	0.261	−0.008	0.005	0.000	0.405	0.048	0.108	−0.006	0.009	0.009
SDH							0.178	−0.002		
SOD-1	0	—	0.005	0.000					0	—
SOD-2			0.020	−0.003	0.005	−0.001				
XO									0.181	−0.009
Introns										
AldoB2	0.097	−0.009	0	—	0	—	0.447	−0.005	0	—
AldoB4	0	—	0	—	0	—	0.145	0.004	0.490	−0.014
GnRH3-1	0.670	−0.001	0.601	−0.004	0	—	0	—	0	—
GnRH3-2	0.524	−0.004	0.529	−0.004	0.405	0.017	0	—	0	—
GH2	0	—	0	—	0.726	0.002				
GH5	0	—	0	—	0	—	0	—	0.337	−0.011
Overall polymorphic loci										
Allozymes	0.230	0.012	0.103	0.030	0.106	0.031	0.218	0.031	0.156	0.018
Introns	0.430	−0.003	0.565	−0.004	0.566	0.008	0.296	−0.003	0.413	−0.014
Overall	0.263	0.008	0.141	0.018	0.142	0.023	0.227	0.026	0.186	0.010

cause of such departures is likely to be the presence of null alleles because the deviations resulted from significant heterozygote deficiencies, observed in all populations in this locus only. The occurrence of null alleles was further confirmed using MICRO-CHECKER and the average null allele frequency at *GnRH3-1* locus across populations was 0.165 (estimator of Brookfield, 1996). However, this locus was maintained for further analysis, and genotypes were not adjusted according to nulls to allow for multilocus analysis.

Significant ($P<0.01$) linkage disequilibria were found between introns originating from the same gene (that is, *GnRH3-1* and *GnRH3-2* for *C. citrinellus* and *C. quadrimaculatus*; between *AldoB2* and *AldoB4* for *D. aruanus*). For *C. citrinellus*, the association was significant between the two most common alleles at each locus, and between rare alleles at *GnRH3-1* with two of the four alleles at *GnRH3-2*. For *C. quadrimaculatus* and *D. aruanus*, the linkage involved rare alleles from one locus significantly associated with common alleles from the other locus. However, all loci were maintained for further analysis.

For each species, over all four populations, a wide range of single locus F_{ST} was observed among nuclear loci (allozymes and introns), from negative values up to 0.198 at *ADA-1* for *D. aruanus*. Multilocus F_{ST} estimates ranged from 0.008 for *C. citrinellus* to 0.026 for *D. aruanus* (Table 3). Discrepancies among classes of markers were observed because only at enzymatic loci, global F_{ST} estimates ranged from 0.012 for *C. citrinellus* to 0.031 for *F. flavissimus* and *D. aruanus*, while when looking at only

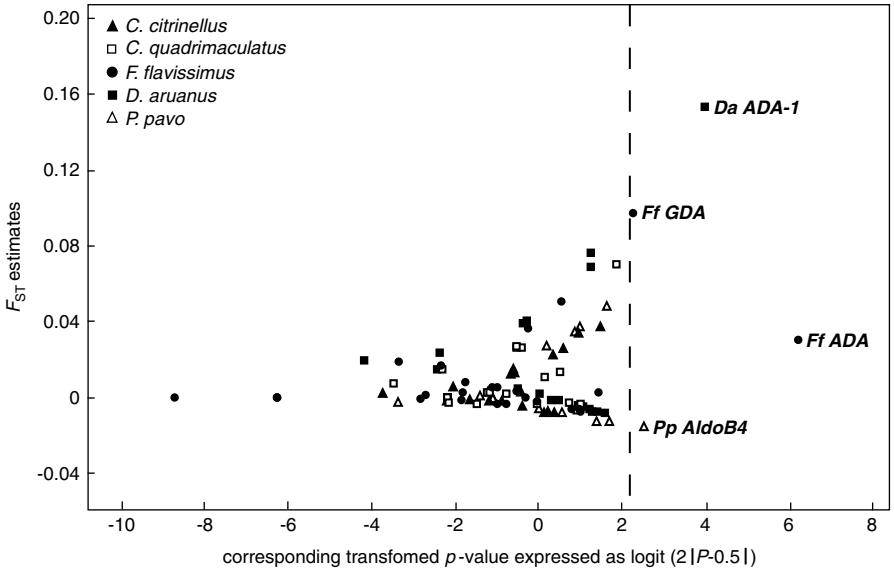

Figure 2 Summary of the results of the Beaumont and Nichols' (1996) neutrality test. For each locus, estimate of F_{ST} is plotted against its corresponding transformed *P*-value (expressed as logit(2|*P*–0.5|) as in Beaumont and Balding, 2004), corresponding to the rejection of the null hypothesis of neutrality. The vertical dotted line bar shows the critical transformed *P*-values used for identifying outlier loci at the 5% confidence level. Each outlier locus is denoted as follows: the two first letters refer to the species (*Da*, *Dascyllus aruanus*; *Ff*, *Forcipiger flavissimus*; *Pp*, *Pomacentrus pavo*), followed by the locus name.

Table 4 Pairwise genetic differentiation (F_{ST}) estimated among islands excluding outlier loci

	km	C. citrinellus	C. quadrimaculatus	F. flavissimus	D. aruanus	P. pavo
Ta/Mo	566	0.011	**0.018**	**0.030**	**0.016**	**0.024**
Ta/Te	520	0.011	0.005	**0.018**	**0.011**	0.000
Ra/Mo	365	0.000	**0.019**	**0.022**	**0.014**	**0.045**
Ra/Te	310	0.007	**0.016**	0.004	**0.026**	0.022
Ra/Ta	226	0.001	**0.012**	0.002	**0.011**	0.000
Te/Mo	60	**0.017**	**0.036**	0.005	**0.035**	0.006
Overall		**0.008**	**0.018**	**0.013**	**0.019**	**0.014**

Abbreviations: Mo, Moorea; Ra, Rangiroa; Ta, Takapoto; Te, Tetiaroa.
Significant values after sequential Bonferroni correction are indicated in bold.

introns, global F_{ST} estimates were zero for all species except one, *F. flavissimus* ($F_{ST} = 0.008$, $P > 0.05$). However, these discrepancies were not marker related but locus related because single locus F_{ST} estimates at introns were similar to those of most of allozymes within species (Table 3).

Beaumont and Nichols' (1996) test revealed that three loci demonstrated significantly higher F_{ST} values than expected under the null hypothesis of neutrality: *ADA-1* in *D. aruanus* ($P = 0.991$), and *ADA* and *GDA* in *F. flavissimus* ($P > 0.999$ and $P = 0.953$, respectively) (Figure 2). One additional locus, *AldoB4* in *P. pavo*, showed significantly lower F_{ST} values ($P = 0.037$) than expected under neutrality. Recalculating the global F_{ST} for each species without the outliers loci did not induce further rejection of neutrality for any other locus.

Excluding outlier loci detected by Beaumont and Nichols' (1996) test, global F_{ST} were significant for all five species and ranged from 0.008 ($P = 0.01$) for *C. citrinellus* to 0.019 ($P = 0.02$) for *D. aruanus* (Table 4). Pairwise F_{ST} estimated from all nuclear loci excluding outliers ranged from 0 to 0.045 (between Rangiroa and Moorea for *P. pavo*; Table 4). None of the five species expressed significant correlation between genetic divergences and geographic distances between populations (all $P > 0.05$). The comparison of pairwise F_{ST} matrices among the five species revealed that two species showed a significant correlation of their patterns of genetic structuring throughout the four sampled islands: *C. quadrimaculatus* and *D. aruanus* ($r = 0.843$, $P = 0.03$; see Table 4).

Discussion

Discrepancies in population differentiation estimates among genetic markers were found in all five species. While intron length polymorphisms (in the present analysis) and sequences of the mtDNA control region (in Fauvelot *et al*., 2003) revealed homogeneous gene frequency over the study area, allozymes showed, over all loci, two- to fivefold higher values of global differentiation estimates. When looking at F_{ST} estimates at nuclear loci, we found high variance among loci within each species, with specific allozymes showing much higher F_{ST} values than introns and other enzymatic loci.

Neutral single locus F_{ST} variance and assumptions of neutrality tests

Identifying the theoretical variance of F_{ST} estimates is a challenging task. Under neutrality, a large variance of F_{ST} estimates is observed among loci because of the stochasticity of genetic drift (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996; Baer, 1999) but other factors may also inflate such variance. While different distributions of F_{ST} may be expected from genetic markers that differ in the mode of inheritance (Buonaccorsi *et al.*, 2001), F_{ST} is theoretically insensitive to variation in mutation rate as long as mutation remains low compared to migration (Slatkin, 1995; but see Hedrick 1999). Meanwhile, the distribution of F_{ST} as a function of heterozygosity appears to be not strongly affected by the mutation rates typical of many nuclear genetic markers at mutation/drift equilibrium (Beaumont and Nichols, 1996). Discrepancies among loci may also originate from spurious typing errors and/or the presence of null alleles. We systematically reran several individuals from different gels side by side on new gels to minimize scoring errors. However, technical artefacts during allozymes migration and/or revelation cannot be totally discarded in generating scoring errors, and the occurrence of null alleles on specific gels may also increase variance in allele frequencies among gels/populations and inflate F_{ST} at particular loci.

Regarding the interpretation of Beaumont and Nichols' (1996) test, the problem of identifying statistically significant departures from neutrality is complicated by confounding effects on the simulated distribution itself. Indeed, the 95–99% confidence interval of the simulated distribution relies on the assumptions of the chosen population structure model, which may be violated. Even though in large populations (such as the ones we are dealing with) migration model has little effect on the distribution of expected F_{ST} (Beaumont and Nichols, 1996), one could ask if migration/drift equilibrium is attained, as assumed in neutrality models. In natural populations, equilibrium is reached after roughly $1/m$ generations (with m being the migration rate; Slatkin, 1985a, b) so that the higher the migration rate, the faster the equilibrium reached (Hardy and Vekemans, 1999). For our coral reef fishes, considering the high observed migration rates and a 2–3 years generation time, we can regard these populations as being at migration/drift equilibrium.

Nevertheless, two important caveats remain using this approach to detect outliers. The first one is that balancing selection is undetectable when the 95–99% confidence intervals of the expected distribution overlap zero (Beaumont and Balding, 2004). The second one is that simulations under neutrality are based on a specified global genetic differentiation supposed to be the 'true' F_{ST} expected overall loci (usually set equal to the observed genetic divergence over all loci for each species). However, this estimate may not be accurate if (1) a limited number of analysed loci is used and (2) outlier loci are present in the data set. One way to minimize the latter is to remove outliers and redo the test until no more outliers are detected. However, as marine organisms often exhibit high gene flow between populations, it may not be possible to detect loci under balancing selection because the confidence intervals of the expected distribution overlap zero (see Hoffman *et al.*, 2006).

Finally, for all the reasons mentioned, outcomes of Beaumont and Nichols' (1996) test regarding outlier loci being under selection should be interpreted with caution. In our study, the two allozyme loci identified as outliers in *F. flavissimus* and *D. aruanus* are both deaminases (*ADA*, adenosine deaminase; *GDA*, guanine deaminase). These loci may be under disruptive selection, but we cannot reject the hypothesis that these high F_{ST} estimates may be caused by mis-scoring of individuals. Contrary to *ADA* and *GDA*, *AldoB4* in *P. pavo* exhibits an F_{ST} estimate significantly smaller to what was expected under neutrality. However, the observed F_{ST} estimate at this locus is negative ($F_{ST} = -0.015$), and because FDIST estimates corresponding *P*-values as the probability of getting values as small as or smaller than the simulated data points, it is clearly impossible to reveal balancing selection when the lower confidence limit for F_{ST} falls below zero. Finally, although it is difficult to conclude that these particular loci are driven by selection, all outlier loci were further excluded before inferring gene flow.

Gene flow in coral reef fishes

Including introns and removing outliers, global F_{ST} estimates are not significantly different from those based only on allozymes presented in Fauvelot and Planes (2002) (paired *t*-test, $P = 0.49$). Geographic/genetic distance correlations for the five species remained insignificant, and they were still no significant correlative patterns of genetic structuring among species sharing similar reproductive strategy. Indeed, the only two species that show correlative pairwise F_{ST} matrix among islands differed in their reproductive strategies, *C. quadrimaculatus* releasing pelagic eggs and *D. aruanus* laying benthic eggs. However, it should be noted that these results may suffer from a low statistical power owing to the small number (four) of islands/populations compared, critical for the Mantel test.

While being significant, observed levels of genetic structuring for the five species are low (range 0.01–0.02), especially when compared to coral reef fishes lacking pelagic larval phase such as *Acanthochromis polyacanthus* and *Pterapogon kauderni* (Doherty *et al.*, 1994; Planes *et al.*, 2001; Hoffman *et al.*, 2005; Bay *et al.*, 2006). However, they are in the range described for related species with similar life histories and over comparable spatial scales (less than 1000 km) in the Pacific (Doherty *et al.*, 1995; Riginos and Victor, 2001; Planes, 2002; Bay *et al.*, 2006). Such rather high and chaotic levels of gene flow among islands accommodate Wright's island model in which each individual issued from one population is equally likely to move to any other population (Wright, 1931). This model may explain the lack of genetic structuring observed at mtDNA sequences. Indeed, such subtle genetic divergences may not be detected in high dispersive capability species using such molecular markers assessed on only 10 individuals per population (Fauvelot *et al.*, 2003).

An interesting result is the significant pairwise divergences revealed for all pairwise comparisons for *D. aruanus* (see Table 4), in contrast to what was reported previously for this species over similar spatial scales in

1990. Indeed, Planes *et al*. (1993) described genetically homogeneous populations across Polynesia (minor genetic differences were detected only when comparing remote archipelagos). However, Planes *et al*. (1993) used a set of 12 polymorphic allozymes from which 9 were in common with our survey (we used seven additional polymorphic loci). Retaining the nine loci common to both survey (that is, *AAT-2*, *Est1*, *GPD-1*, *GPD-2*, *GPI-1*, *GPI-2*, *IDH-1*, *IDH-2* and *MDH-2*), the global F_{ST} estimate over the four sampled islands decreased from 0.019 ($P = 0.002$) to 0.009 ($P = 0.017$), much more comparable to what was reported among archipelagos ($F_{ST} = 0.0076$; Planes *et al*., 1993). This shows that global estimates of gene flow among populations may vary depending of the sets of loci used (in this particular case, by a factor of 2), and that, in order to avoid a bias in global F_{ST} estimates, a reasonable number of analysed loci should be considered (15–20; Nielsen *et al*., 2006; but see also Kalinowski, 2002 regarding the number of alleles).

In conclusion, our study shows that neutrality was rejected at few loci within species, although we could not conclude these loci being under selection. Our study demonstrates how cautious one has to be when trying to interpret present-day genetic structure in terms of gene flow, especially when using a limited number of loci and/or different sets of loci. Indeed, selection may influence particular loci, but as well, the expected neutral range of single locus F_{ST} is wide for loci with low heterozygosities, which can lead to erroneous estimates of genetic differentiation when a limited number of loci are used.

Acknowledgements

We thank K Belkhir, P Borsa, M Hassan and J-D Durand for helpful discussions throughout this work. We appreciate comments of one anonymous referee on an early version of the manuscript. The new data set used in this article was subsidized by CNRS funding to FB. This is publication ISEM 2007-038.

References

Arnaud-Haond S, Bonhomme F, Blanc F (2003). Large discrepancies in differentiation of allozymes, nuclear and mitochondrial DNA loci in recently founded Pacific populations of the pearl oyster *Pinctada margaritifera*. *J Evol Biol* **16**: 388–398.

Baer CF (1999). Among-locus variation in Fst: fish, allozymes and the Lewontin–Krakauer test revisited. *Genetics* **152**: 653–659.

Bay LK, Crozier RH, Caley MJ (2006). The relationship between population genetic structure and pelagic larval duration in coral reef fishes on the Great Barrier Reef. *Mar Biol* **149**: 1247–1256.

Beaumont MA (2005). Adaptation and speciation: what can Fst tell us? *Trends Ecol Evol* **20**: 435–440.

Beaumont MA, Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**: 969–980.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc B* **263**: 1619–1626.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2002). Genetix 404 A windows program for population genetics. Laboratoire Génome, Populations, Interactions, CNRS UMR, 5000, Université de Montepellier II, Montepellier, France.

Berrebi P, Retif X, Fang F, Zhang CG (2006). Population structure and systematics of *Opsariichthys bidens* (Osteichthyes: Cyprinidae) in south-east China using a new nuclear marker: the introns (EPIC-PCR). *Biol J Linn Soc* **87**: 155–166.

Black WC, Krafsur ES (1985). A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor Appl Genet* **70**: 491–496.

Borsa P (2003). Genetic structure of round scad mackerel *Decapterus macrosoma* (Carangidae) in the Indo-Malay archipelago. *Mar Biol* **142**: 575–581.

Brookfield JFY (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol* **5**: 453–455.

Buonaccorsi VP, Mcdowell JR, Graves JE (2001). Reconciling patterns of inter-ocean molecular variance from four classes of molecular markers in blue marlin (*Makaira nigricans*). *Mol Ecol* **10**: 1179–1196.

Cavalli-Sforza LL (1966). Population structure and human evolution. *Proc R Soc B* **164**: 362–379.

De Innocentiis S, Sola L, Cataudella S, Bentzen P (2001). Allozyme and microsatellite loci provide discordant estimates of population differentiation in the endangered dusky grouper (*Epinephelus marginatus*) within the Mediterranean Sea. *Mol Ecol* **10**: 2163–2175.

Doherty PJ, Mather P, Planes S (1994). *Acanthochromis polyacanthus*, a fish lacking larval dispersal, has genetically differentiated populations at local and regional scales on the Great Barrier Reef. *Mar Biol* **121**: 11–21.

Doherty PJ, Planes S, Mather P (1995). Gene flow and larval duration in seven species of fish from the Great Barrier Reef. *Ecology* **76**: 2373–2391.

Dufresne F, Bourget E, Bernatchez L (2002). Differential patterns of spatial divergence in microsatellites and allozymes alleles: further evidence for locus-specific selection in the acorn barnacle, *Semibalanus balanoides*? *Mol Ecol* **11**: 113–123.

Fauvelot C, Bernardi G, Planes S (2003). Reductions in the mitochondrial DNA diversity of coral reef fish provide evidence of population bottlenecks resulting from Holocene sea-level change. *Evolution* **57**: 1571–1583.

Fauvelot C, Planes S (2002). Understanding origins of present-day genetic structure in marine fish: biologically or historically driven patterns? *Mar Biol* **141**: 773–788.

Friesen VL, Birt TP, Piatt JF, Golightly RT, Newman SH, Hébert PN *et al*. (2005). Population genetic structure and conservation of marbled murrelets (*Brachyramphus marmoratus*). *Conserv Genet* **6**: 607–614.

Gomez-Uchida D, Weetman D, Hauser L, Galleguillos R, Retamal M (2003). Allozyme and AFLP analysis of genetic population structure in the hairy edible crab *Cancer setosus* from the Chilean coast. *J Crustacean Biol* **23**: 486–494.

Guinand B, Lemaire C, Bonhomme F (2004). How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. *J Sea Res* **51**: 167–182.

Hardy OJ, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.

Hassan M, Harmelin-Vivien M, Bonhomme F (2003). Lessepsian invasion without bottleneck: example of two rabbitfish species (*Siganus rivulatus* and *Siganus luridus*). *J Exp Mar Biol Ecol* **291**: 219–232.

Hassan M, Lemaire C, Fauvelot C, Bonhomme F (2002). Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. *Mol Ecol Notes* **2**: 334–340.

Hedrick PW (1999). Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313–318.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.
Hoffman EA, Kolm N, Berglund A, Arguello JR, Jones AG (2005). Genetic structure in the coral-reef-associated Banggai cardinalfish, *Pterapogon kauderni*. *Mol Ecol* **14**: 1367–1375.
Hoffman EA, Schueler FW, Jones AG, Blouin MS (2006). An analysis of selection on a color polymorphism in the northern leopard frog. *Mol Ecol* **15**: 2627–2641.
Hutchison DW, Templeton AR (1999). Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution* **53**: 1898–1914.
Kalinowski ST (2002). How many alleles per locus should be used to estimate genetic distances? *Heredity* **88**: 62–65.
Karl SA, Avise JC (1992). Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. *Science* **256**: 100–102.
Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
Lemaire C, Allegrucci A, Naciri M, Bahri-Sfar L, Kara H, Bonhomme F (2000). Do discrepancies between microsatellite and variation reveal differential selection between sea and lagoon in the sea bass (*Dicentrarchus labrax*)? *Mol Ecol* **9**: 457–467.
Lessa EP (1992). Rapid surveying of DNA sequence variation in natural populations. *Mol Biol Evol* **9**: 323–330.
Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics* **74**: 174–195.
Nei M (1987). *Molecular Evolutionary Genetics*. Columbia Univ Press: New York.
Nielsen EE, Hansen MM, Meldrup D (2006). Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L): implications for inferring population structure in nonmodel organisms. *Mol Ecol* **15**: 3219–3229.
Palumbi SR (1995). Nucleic acids II: the polymerase chain reaction. In: Hillis D, Moritz C (eds). *Molecular Systematics* 2nd edn. Sinauer: Sunderland, MA. pp 205–247.
Planes S (2002). Biogeography and larval dispersal inferred from population genetic analysis. In: Sale PF (ed). *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem*. Academic Press: San Diego, CA. pp 201–220.
Planes S, Bonhomme F, Galzin R (1993). Genetic structure of *Dascyllus aruanus* populations in French Polynesia. *Mar Biol* **117**: 665–674.
Planes S, Bonhomme F, Galzin R (1996). A genetic metapopulation model for reef fishes in oceanic islands: the case of the surgeonfish, *Acanthurus triostegus*. *J Evol Biol* **9**: 103–117.
Planes S, Doherty PJ, Bernardi G (2001). Strong genetic divergence among populations of a marine fish with limited dispersal, *Acanthochromis polyacanthus*, within the Great Barrier Reef and the Coral Sea. *Evolution* **55**: 2263–2273.
Planes S, Fauvelot C (2002). Isolation by distance and vicariance drive genetic structure of a coral reef fish in the Pacific Ocean. *Evolution* **56**: 378–399.
Planes S, Parroni M, Chauvet C (1997). Evidence of limited gene flow in three coral reef fish species in the lagoon of New Caledonia. *Mar Biol* **130**: 361–368.
Pogson GH, Mesa KA, Boutilier RG (1995). Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. *Genetics* **139**: 375–385.
Purcell JFH, Cowen RK, Hughes CR, Williams DA (2006). Weak genetic structure indicates strong dispersal limits: a tale of two coral reef fish. *Proc R Soc B* **273**: 1483–1490.
Riginos C, Victor BC (2001). Larval spatial distributions and other early life-history characteristics predict genetic differentiation in eastern Pacific blennioid fishes. *Proc R Soc B* **268**: 1931–1936.
Rocha LA, Bass AL, Robertson DR, Bowen BW (2002). Adult habitat preferences, larval dispersal, and the comparative phylogeography of three Atlantic surgeonfishes (Teleostei: Acanthuridae). *Mol Ecol* **11**: 243–252.
Rohfritsch A, Borsa P (2005). Genetic structure of Indian scad mackerel *Decapterus russelli*: Pleistocene vicariance and secondary contact in the Central Indo-West Pacific Seas. *Heredity* **95**: 315–326.
Ruzzante DE (1998). A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. *Can J Fish Aquat Sci* **55**: 1–14.
Shulman MJ, Bermingham E (1995). Early life histories, ocean currents, and the population genetics of Caribbean reef fishes. *Evolution* **49**: 897–910.
Slatkin M (1985a). Rare alleles as indicators of gene flow. *Evolution* **39**: 53–65.
Slatkin M (1985b). Gene flow in natural populations. *Annu Rev Ecol Syst* **16**: 393–430.
Slatkin M (1987). Gene flow and the geographic structure of natural populations. *Science* **236**: 787–792.
Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.
Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004). Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes* **4**: 535–538.
Waples RS (1987). A multispecies approach to the analysis of gene flow in marine shore fishes. *Evolution* **41**: 385–400.
Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
Whitlock MC, McCauley DE (1999). Indirect measures of gene flow and migration: FST does not equal 1/(4Nm+1). *Heredity* **82**: 117–125.
Wright S (1931). Evolution in Mendelian populations. *Genetics* **16**: 97–159.
Wright S (1943). Isolation by distance. *Genetics* **28**: 114–138.