Population genomic footprints of fine-scale differentiation between habitats in Mediterranean blue tits

M. Szulkin, P.-A. Gagnaire, N. Bierne, A. Charmantier

MOLECULAR ECOLOGY

Population genomic footprints of fine-scale differentiation between habitats in Mediterranean blue tits

Journal:	*Molecular Ecology*
Manuscript ID	MEC-15-0774.R3
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Szulkin, Marta; CEFE, UMR 5175 CNRS Gagnaire, Pierre-Alexandre; ISEM, UMR 5554 CNRS Bierne, Nicolas; Institut des Sciences de l'Evolution, Integrative Genomics Charmantier, Anne; CEFE, UMR 5175 CNRS
Keywords:	blue tit, Population Genetics - Empirical, Landscape Genetics, local adaptation, RAD sequencing, genomic differentiation

Population genomic footprints of fine-scale differentiation between habitats in Mediterranean blue tits

M. Szulkin[1§*], P.-A. Gagnaire[2,3*], N. Bierne[2,3] & A. Charmantier[1]

[1] Centre d'Ecologie Fonctionnelle et Evolutive, UMR 5175 Campus CNRS, 1919 Route de Mende, 34293 Montpellier cedex 5, France

[2]Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

[3]ISEM - CNRS, UMR 5554, SMEL, 2 rue des Chantiers, 34200 Sète, France

* Joint first authors

§ corresponding author

Keywords: blue tit, population genomics, landscape genetics, RAD sequencing, local adaptation, genetic differentiation

Address of corresponding author:

Marta Szulkin

CEFE CNRS

1919, route de Mende

34293 Montpellier cedex 5, France

Fax number: +33 4 67 61 33 36

Email: marta.szulkin@zoo.ox.ac.uk

Running title: population genomics of wild blue tits

Abstract

Linking population genetic variation to the spatial heterogeneity of the environment is of fundamental interest to evolutionary biology and ecology, in particular when phenotypic differences between populations are observed at biologically small spatial scales. Here, we applied restriction-site associated DNA sequencing (RAD-Seq) to test whether phenotypically differentiated populations of wild blue tits (*Cyanistes caeruleus*) breeding in a highly heterogeneous environment exhibit genetic structure related to habitat type. Using 12106 SNPs in 197 individuals from deciduous and evergreen oak woodlands, we applied complementary population genomic analyses, which revealed that genetic variation is influenced by both geographical distance and habitat type. A fine-scale genetic differentiation supported by genome- and transcriptome-wide analyses was found within Corsica, between two adjacent habitats where blue tits exhibit marked differences in breeding time while nesting less than 6 km apart. Using redundancy analysis (RDA), we show that genomic variation remains associated with habitat type when controlling for spatial and temporal effects. Finally, our results suggest that the observed patterns of genomic differentiation were not driven by a small proportion of highly differentiated loci, but rather emerged through a process such as habitat choice, which reduces gene flow between habitats across the entire genome. The pattern of genomic isolation-by-environment closely matches differentiation observed at the phenotypic level, thereby offering significant potential for future inference of phenotype-genotype associations in a heterogeneous environment.

Introduction

The evolutionary tug-of-war between local adaptation and counteracting gene flow is a fascinating biological process which plays a key role in shaping genetic and phenotypic diversity of natural populations. In the absence of gene flow and other evolutionary constraints (such as genetic correlations, or a lack of adaptive genetic variation), divergent selection should cause each local population to evolve traits providing an advantage in its local habitat (Kawecki & Ebert 2004). However, local adaptation may be limited by gene flow, especially if habitat patch size is small relative to the scale of dispersal (Lenormand 2002; Slatkin 1973, 1987).

Genetic evidence for local adaptation is usually inferred indirectly by searching for molecular signatures of selection, with the implicit expectation that selection varies across environments (Barrett and Hoekstra 2011). Another important prediction is that genetic differentiation should correlate with environmental variables independently of geographic distance - a pattern commonly referred to as Genetic-Environment Association (GEA; Hedrick *et al.* 1976) or Isolation-by-Environment (IBE; Wang & Bradburg 2014).

Gene flow mediated through dispersal is a key element opposing the effect of local adaptation. Consequently, gene flow across habitats has long been assumed to preclude adaptive differentiation, thereby preventing the evolution of marked intraspecific phenotypic differences at small spatial scales in highly mobile organisms such as birds (Garant *et al.* 2007; Slatkin 1987) or marine species (Palumbi 1994). More recently, the importance of non-random gene flow through matching habitat choice has received increased theoretical and empirical attention (see Edelaar *et al.* 2012; Edelaar & Bolnick 2012; Edelaar *et al.* 2008; Ravigne *et al.* 2009). However, the extent to which individuals choose to settle in the habitat that maximizes their fitness with respect to their phenotype remains

63 poorly understood, and so are the consequences of matching habitat choice on the evolution of local
64 adaptation (Edelaar *et al.* 2008).

65 The long-term monitoring of several populations of a small passerine bird, the blue tit *Cyanistes*
66 *caeruleus* breeding in a highly heterogeneous habitat in Southern France (Blondel *et al.* 2006)
67 revealed multiple lines of evidence offering scope for non-random dispersal and habitat-dependent
68 selection: marked phenotypic differences in lay date, clutch size, number of fledglings and
69 morphometric traits can be observed not only between the Southern French mainland and Corsica
70 (Figure 1), but also between two Corsican populations residing 27 km apart (Blondel *et al.* 1999;
71 Blondel *et al.* 2006; Lambrechts *et al.* 1997) (Figure 1). Even more strikingly, these differences were
72 also observed at a finer scale between two Corsican populations located only 5.6 km apart within the
73 same valley (Blondel *et al.* 2006, Figure 1). In addition, quantitative genetic models revealed that all
74 of these traits harbour significant genetic variation (reviewed in Charmantier *et al.* in press, Blondel
75 et al. 2006).

76 The phenotypic differences observed in this study system are expected to be driven by habitat
77 heterogeneity, and in particular by the type of oak species dominating the habitat where blue tits
78 breed (Blondel *et al.* 2001; Blondel *et al.* 2006; Lambrechts *et al.* 2004). Indeed, Mediterranean
79 habitats are interspersed with distinct patches of either evergreen (holm oak *Quercus ilex*) or
80 deciduous (downy oak *Quercus pubescens*) oak populations. Oak type influences the entire food
81 chain blue tits depend on to feed their young: First, the *ca.* one month time-lag in leaf development
82 between evergreen and deciduous oaks translates in a time-lag in oak leaf-feeding caterpillar hatch
83 dates. Second, temporally contrasted caterpillar availability (the primary food source of blue tit
84 nestlings) triggers shifts in the distribution of blue tit breeding time between habitat patches. As a
85 result, blue tit populations breeding 27 km from each other in a heterogeneous environment
86 including evergreen and deciduous habitat patches, with no clear-cut boundaries limiting dispersal

(such as open spaces generated by large crop fields), start to breed on average at a one month difference from each other (Blondel *et al.* 2006). This temporal breeding shift recorded between "early" deciduous habitats and "late" evergreen habitats at a small geographical scale is recurrently noted at larger, but also at smaller spatial scales when blue tit populations from several evergreen and deciduous oak habitats are compared (Blondel *et al.* 2001; Blondel *et al.* 2006; Lambrechts *et al.* 1997; Porlier *et al.* 2012a; Szulkin *et al.* in press). This metapopulation blue tit study system thus offers a particularly suitable model to test for Isolation-by-Environment (IBE) over short geographical distances.

Available evidence for a genetic basis to differences in habitat-specific laying date was originally deduced from common-garden experiments (Blondel *et al.* 1990; Lambrechts *et al.* 1997). In nature, genetic differences between habitats were also found over large (Corsica *vs* French mainland) and small (27 km in Corsica) spatial scales using microsatellite markers (Porlier *et al.* 2012b). However, no differentiation could be evidenced between habitat patches located 5.6 km apart within the same Corsican valley. This suggests that either gene flow at such a fine spatial scale homogenises allelic frequencies within the valley, or that genetic inference made from a limited number of neutral markers was underpowered to detect biologically significant fine-scale population structure visible at the phenotypic level (Figure 1). In this context, the potential of large single nucleotide polymorphisms (SNPs) datasets may be of particular interest to increase the power and resolution in the detection of fine-scale genetic structure and IBE.

Here, we used Restriction-site associated DNA sequencing (RAD seq) to generate a high density SNP dataset covering the entire blue tit genome and characterise genetic patterns of diversity in blue tit populations from Southern France (mainland and the island of Corsica). First, we present a general overview of the *de novo* strategy to obtain genome-scale polymorphism data in this wild passerine bird with no reference genome. We emphasise the usefulness of reporting checkpoints for data

validation throughout the analysis pipeline, by matching genome-wide estimates of relatedness with field and microsatellite-based pedigrees, and by including blind-sequencing control samples to estimate genotyping repeatability. Second, we investigated whether the previously described habitat-specific phenotypic differentiation is corroborated by genetic differentiation between populations at the genomic and transcriptomic levels. We evaluated the robustness of our results by controlling for the effect of rare variants, family relatedness, sample size and variation in individual birth year. We also took advantage of earlier molecular work in the study population to test whether microsatellite and SNP data concur in population genetic estimates of population differentiation. In particular, we aimed to confirm or refute (i) earlier reports suggesting a lack of genetic differentiation between two phenotypically contrasted blue tit populations located 5.6 km apart with no dispersal barrier between them, as well as (ii) the general role of the environment in creating habitat-dependent genetic structuring that is independent of geographical distance.

Materials & Methods

Study system and data collection

The blue tit *Cyanistes caeruleus* is a small resident passerine bird of the tit (Paridae) family, breeding throughout temperate Europe and western Asia in deciduous or mixed woodlands (Snow 1954). In this study, we sampled 197 blue tits breeding in nestboxes as part of a long-term monitoring survey (Blondel *et al.* 2006; Charmantier *et al.* in press). Study sites include a forest in southern French mainland near Montpellier ("D-Rouviere", with "D" for deciduous habitat), where blue tits belong to the continental nominal subspecies (*C. caeruleus caeruleus*), and 3 locations on Corsica ("E-Muro", "D–Muro" and "E–Pirio", with "E" for evergreen habitat – see Figure 2 and Table 1). Corsican blue tit populations belong to the subspecies *C. caeruleus ogliastrae*, which is *ca.* 15% smaller compared to

its mainland relative (Martin 1991). All four populations breed in a mosaic of heterogeneous habitats containing in majority, among other tree species, interspersed patches of deciduous downy oak (*Quercus pubescens*) and evergreen holm oak (*Quercus ilex*) (Figure 2, Table 1). In Corsica, habitat type is known to be associated with marked differences in the timing of breeding and reproductive success at a small geographic scale (Blondel *et al.* 1999; Blondel *et al.* 2006; Porlier *et al.* 2012a) (Figure 1). These populations are described in further detail in previous studies (Blondel *et al.* 2006; Charmantier *et al.* in press; Porlier *et al.* 2012b).

Birds were captured in the nestboxes when offspring were between 9 and 15 days old; their identity and morphometric measurements were recorded, and 7-30 µl of blood was taken from a small neck vein – a method deemed safer relative to the risk of hematomas and flight impairment caused by sampling in the wing. A total of 197 birds were selected; all were residents - i.e. they were born and later recruited as breeding individuals in one of the 4 locations of interest (Table 1). Birth year varied between 1991 and 2008, with an average birth year in 2002. Maternal and paternal identities were obtained from field observations, which enabled us to identify 93.5% and 95% of social fathers and mothers of birds in our dataset, respectively. GPS coordinates were measured for most nestboxes using a handheld GPS device (Garmin GPSMAP 62S). Missing nestbox coordinates were retrieved using annotated maps of the study sites.

RAD libraries construction and sequencing

Blue tit blood samples used in this study were stored in Queen's buffer, and DNA extraction was performed using Qiagen DNeasy Blood & Tissue kits. DNA extractions were quantified using a NanoDrop ND8000 spectrophotometer and a Qubit 2.0 fluorometer with the DNA HS assay kit (Life Technologies), and checked for DNA quality after migration on agarose gel to select samples with appropriate DNA concentration (>20ng/ µl) and molecular weight (>10 000bp).

In order to assess the repeatability of library construction and to evaluate the rate of genotyping errors, DNA from 5 of the 197 individuals were replicated as follows: DNA from 4 individuals were extracted twice and the DNA extract of one individual was split into two samples which were analysed independently. In total, 202 DNA extracts were sent to Floragenex Inc. for library preparation and single-end sequencing according to the original protocol (Baird *et al.* 2008). We used the restriction enzyme *Sbf*I, which targets an 8-bp cutting site (5'CCTGCAGG3'). Each individual RAD library was ligated to a unique molecular identifier (a 6bp DNA barcode) before sample multiplexing was performed in equimolar proportions by groups of 29 individuals per pool. Each pool was then sequenced on one lane of an Illumina HiSeq 2000 instrument, generating 101-bp single reads which were further automatically trimmed to 91bp reads. As recommended by Meirmans *et al.* (2015), samples were assigned to sequencing lanes in a randomised fashion, and samples from each of the 4 populations were present in all 7 or 8 (out of 8) sequencing lanes used in the study.

Bioinformatics

Short sequence reads were quality filtered and demultiplexed using individual barcode information. We used the *Stacks* pipeline (Catchen *et al.* 2013; Catchen *et al.* 2011) to identify loci *de novo*, discover SNPs and infer individual genotypes. Preliminary runs were performed to determine the most appropriate parameter combination for *UStacks*. SNPs were detected at each locus using the maximum likelihood approach under the 'snp' model. We empirically determined an optimal minimum depth of coverage of 5 reads per allele (m = 5) and a maximum of 3 nucleotide mismatches between alleles (M = 3). Increasing the number of mismatches between alleles did not allow to retrieve many more SNPs while increasing the risk of merging paralogs, as detected by HWE tests. A catalog of loci found across all individuals was then built using *CStacks*, allowing a maximum number of 3 mismatches between two homozygous individuals at a same locus (n = 3). *De novo* loci

constructed with *UStacks* were then searched against the catalog of loci using *SStacks*. Finally, we used the *Stacks* module *populations* to retain only loci that were successfully genotyped in at least 50% of the individuals from at least 2 populations. Individual genotypes were outputted as a VCF file which was submitted to further downstream filtering.

The increasing sequencing error rate toward the end of reads produced an elevation in the total number of SNPs being called from position 85 to 91bp. Therefore, all variable sites located after position 84 of the reads were removed from the VCF file. We further filtered the SNP dataset based on several quality and population-genetic criteria to only retain highly reliable SNPs using *VCFtools* (Danecek *et al.* 2011) (Table 2). After removing the five individual replicates, we excluded SNPs showing strong deviations to Hardy-Weinberg equilibrium (HWE) within at least one of the three locations (D-Rouviere, Muro and E-Pirio) using a p-value threshold of 0.01 (D-Muro and E-Muro were pooled together due to low sample size in E-Muro and close physical proximity). This filtering step aimed to remove poor-quality SNPs and artefactual variation due to the merging of paralogous sequences, but was insensitive to small deviations from HWE resulting from subtle within-population structure. The dataset was then filtered to only retain loci that were genotyped in at least 90% of all samples (90% call rate), and with a global Minor Allelic Frequency (MAF) of at least 2%. Details on the number of SNPs retained at subsequent bioinformatics filtering steps are presented in Table 2.

Genome-wide relatedness between individuals

To infer SNP-based relatedness structure within populations, we calculated pairwise identity-by-state (IBS) coefficients between all possible pairs of individuals within the mainland (D- Rouviere) and within Corsica, as well as for the 5 pairs of individual replicates using the R package *SNPRelate* (Zheng *et al.* 2012). Contrasting RAD-seq derived pairwise IBS estimates with independently acquired information about the relatedness between any two given indviduals can be used as a valuable data

processing check point throughout the analysis pipeline. For example, such contrasts can be used (1) as a quality control check to confirm that samples were not mixed-up in the lab preparation and sequencing stages (this complements the repeatability values of extracts sequenced independently), (2) to confirm that RAD-seq data, microsatellite data (if available) and pedigree data concur, and finally (3) to verify that RAD-seq derived IBS values are linearly related to independently established relatedness values. Genome-wide relatedness values estimated from IBS coefficients were therefore compared with (i) expected relatedness values for 5 full siblings (r = 0.5) confirmed with microsatellite genotyping data (Charmantier *et al.* 2004), (ii) 10 mother-offspring pairs (r=0.5) and (iii) 5 maternal half-siblings (r=0.25) established using social pedigree information. Individual replicates were further used to evaluate the overall genotyping repeatability of the RAD marker dataset. Because absolute (unscaled) values of IBS are influenced by the population's allele frequency spectrum, we provide the folded allele frequency spectrum along with the genome-wide relatedness distributions of mainland and Corsica separately (Figure 3).

Analyses of population genomic structure

We used a combination of complementary population and landscape genetics analyses to evaluate the extent of genetic structure within and among Mediterranean blue tit populations. Table 2 summarises the different nested datasets used for analysis and their associated number of markers.

We first evaluated the extent of population differentiation between the French mainland and Corsica, as well as between Corsican populations. Principal Component Analysis (PCA) implemented in the R package SNPRelate (Zheng *et al.* 2012) was used to illustrate population structure at different scales (*i.e*. Mainland-Corsica, between samples within Corsica, between habitats within the Muro valley).

We estimated the genome-wide average genetic differentiation between each pair of populations and habitats by computing Nei's pairwise F_{ST} using the *adegenet* R package (Jombart 2008; Jombart & Ahmed 2011) after applying an MAF threshold of 5% (n=3159 SNPs, see Table 2). The significance of pairwise F_{ST} values was tested through 500 random permutations of the genotypes among populations.

To further investigate the spatial scale of genetic variation within populations, we performed spatial PCA analyses, a method for detecting spatial patterns that are not always associated with the principal components of genetic variation detected in standard PCA. Spatial PCA uses connection networks to separate the product of the genetic variance between individuals and their spatial autocorrelation into negative and positive components (Jombart *et al.* 2008). For instance, global structures, such as two spatial groups, or a cline, will display positive spatial autocorrelation (Moran's I, Moran 1950) that can be inferred from allelic frequency data. We applied the "Global Test" (Jombart *et al.* 2008) to test for global structures against the null hypothesis of no genetic structure in the population. Spatial PCA was performed using the *adegenet* package in R (Jombart 2008; Jombart & Ahmed 2011), with a connection network of 15 nearest neighbours for Muro, Pirio and Rouviere (and 12 neighbours in the "no family ties dataset"). Because (i) the power of principal component based methods usually scales with the product between the number of individuals and the number of markers (Patterson *et al.* 2006), and because (ii) the detection of fine-scale population structures may benefit from the inclusion of rare variants (O'Connor *et al.* 2015), we used an MAF threshold of 2% to maximise dataset size for this analysis (n= 12 106 SNPs, see Table 2).

Genetic variation among blue tit populations may be confounded by several factors including habitat type, geography, individual birth year and individual sequencing depth (expected to reflect genotyping accuracy). We therefore used constrained ordination to specifically test the marginal effect of each factor on the distribution of samples genotypes. Here, we used redundancy analysis

(RDA), a constrained ordination method implemented in the *Vegan* package (Oksanen *et al.* 2014) in *R* to infer the extent to which available environmental, but also experiment-dependent variables are influencing SNP genotypic variation in the dataset (see Meirmans *et al.* 2015). A key strength of this analysis is to provide a statistical means for inferring the effect of partially confounded variables separately. The following initial model was used: Y (individual genotype) ~ Latitude + Longitude + Habitat + Birth year + Number of Reads. To assess whether the different variables significantly influenced allele frequencies, we first used permutation tests to assess the global significance of the RDA by performing 1000 permutations where the genotypic data were permuted randomly and the model was refitted. Second, the significance of each individual variable was tested by running an RDA marginal effects permutation test (with 1000 permutations) where we removed each term one by one from the model containing all other terms. Non-significant effects were removed from the final model. This procedure was implemented both for all birds in the dataset and for Corsican birds only.

To establish the role of habitat independently from other sources of genetic variation (i.e. the remaining explanatory variables in the final model), we performed conditioned (partial) RDA where the effects of all significant explanatory variables but habitat were removed from the ordination by using the *condition* function: Y ~ habitat + *condition*(remaining significant variables in the final model).

Finally, the distribution of SNP contributions to the single RDA habitat axis after conditioning on remaining variables was compared to that obtained for conditioned RDA estimating the specific effect of geography or birth year (equivalently conditioned by all other significant variables in the final model). It is expected that directional selection on loci conferring adaptation to habitat type will generate outlier SNPs in the distribution of SNP contributions to the effect of habitat. Therefore, the distribution of SNP contributions to the conditioned effect of habitat should differ from the conditioned effect of geography or birth year if the habitat-dependent IBE pattern is mainly driven by

277 directional selection.

278

279 **Controlling for family relatedness**

280 The genetic sampling of free-living animals is frequently made without considering the underlying
281 family genetic structure, which is often unknown during sampling. Moreover, it is often difficult to
282 specify what constitutes a bias when sampling relatives (Szulkin *et al.* 2013): while relatives are part
283 of a null distribution of genetic relatedness for a given animal system, the distribution of relatedness
284 in a sampled population may become skewed because of field work protocols, for example due to
285 site and nestbox fidelity. To control for a possible bias induced by relatedness in our inference of blue
286 tit genetic structuring in Corsica and mainland France, we reran PCA, sPCA and Fst analyses using a
287 "no family ties" dataset. Out of the 197 individuals in the original dataset, we removed closely
288 related individuals (offspring or siblings), reducing the original dataset to 119 individuals. In the case
289 of full siblings, we conserved the sibling with the largest number of reads (e.g. of best genotyping
290 quality). The "no family ties" sample sizes per population are as follows: D-Muro = 36, E-Muro = 8, E-
291 Pirio = 43, D-Rouviere = 32; see also Part II of Sup. Mat. for more details).

292

293 **Controlling for inequalities in sample size, sex composition and birth year**

294 Because sample size varied among the four sampled populations, reaching the lowest value of 9
295 individuals in E-Muro, we generated a "symmetrical minimal dataset" where all individuals from E-
296 Muro were complemented by 9 individuals from each of the remaining 3 populations, matching E-
297 Muro individuals in terms of birth year (± 1 year difference) and sex. This resulted in a dataset of 36
298 individuals in which each population was equally represented and was homogeneous in terms of
299 birth year and sex composition (see also Part III of Sup. Mat for more details).

Transcriptomic variation analyses

To estimate what proportion of genomic RAD tags could be identified as transcriptomic sequences, we used the full length (91 bp-long) consensus sequence of each polymorphic RAD locus to perform Blast searches against transcriptome databases. To maximise the number of annotated sequences, we used 19 760 RAD loci with a global call rate > 80% and a MAF > 1%. All loci were blasted against the three following transcriptomes with *Blastx*, using an e-value threshold of 10^{-7} to retain significant matches:

(1) the blue tit *Cyanistes careuleus* transcriptome. RNA from blood of ten blue tits (including 4 *Cyanistes c. ogliastrae* individuals from Corsica, 3 *Cyanistes c. careuleus* from D-Rouviere) was used to synthesise and sequence cDNA fragments on 454 and Illumina sequencers using a previously described protocol (Cahais *et al.* 2012; Romiguier *et al.* 2014). (2) the great tit *Parus major* transcriptome (Santure *et al.* 2011); divergence time from *Cyanistes caeruleus*: 19 million years (onezoom.org). (3) the zebra finch *Taeniopygia guttata* transcriptome, where both (a) ab-initio predicted genes and (b) cDNA transcripts, available at www.ensembl.org, were inspected. Divergence time from *Cyanistes caeruleus*: 72 million years (Hedges *et al.* 2006).

To determine the percentage of genomic RAD tags mapping to the transcriptome, we estimated the proportion of blue tit RAD sequences matching (i) either a blue tit or great tit transcriptomic sequence, and (ii) a blue tit, great tit or zebra finch transcriptomic sequence. In cases where more than one tag matched the same transcriptomic contig, we selected the RAD tag with the lowest E-value. Transcriptomic SNPs used to compute pairwise Fst values were extracted from the previously described dataset (Table 2), using only SNPs derived from RAD sequences that fully matched the blue tit transcriptome on at least 80bp (Table 2).

Results

Sequencing RAD-Tags from blue tit populations generated an average of *ca.* 4,7 million sequences per sample (median 4.5 million reads). Overall, sequencing quality control reports revealed mostly uniform, high quality sequencing across samples. One sample had a lower than expected number of reads, which translated into a lower number of RAD tags (Figure S1). On the other hand, over-sequencing resulted in an increased number of variable RAD-tags produced by sequencing errors (Figure S1). Applying HWE tests, genotyping call rate and MAF thresholds efficiently removed poorly sequenced tags and artefactual SNPs originating from sequencing errors or paralogous tags. The resulting dataset is characterised by a minimum of 90% genotype call rate (and an average of 96%), a 2% MAF threshold, and contains 12 106 SNPs with an average sequencing depth of 73X per individual.

Relatedness distribution and repeatability of control samples

The analysis of the identity-by state (IBS) matrix calculated for both Corsican and mainland birds revealed unimodal distributions of IBS coefficients flanked by right-hand tails of high IBS values (Figure 3A&B). These distributions were further annotated with independently confirmed family links (full siblings and mother-offspring pairs), which showed that the right tails reflect the presence of close relatives in the dataset. The genotyping repeatability of RAD loci assessed with sample replicates averaged 97%, a value which was well above the range of IBS values observed in the dataset (Figure 3A&B). Genome-wide relatedness measured with IBS coefficients increased linearly with expected relatedness inferred from microsatellite and pedigree data (Figure S2). The folded allele frequency spectrum differed between the mainland and Corsica, revealing a deficit of rare variants (<10%) on the mainland (Figure 3C) compared to Corsica (Figure 3D).

Strong genome- and transcriptome-wide signals of between-population differentiation

Large scale as well as small scale genetic differentiation was confirmed by pairwise Fst analyses (Table 4). At the genomic level, we detected highly significant differentiation between Corsican populations and Southern French mainland. No significant genetic differentiation was detected between Pirio and Muro sites (27 km apart; Table 4). At the same time, a clear signal of genetic differentiation was found between Muro Evergreen and Muro Deciduous sites, two sites with contrasted vegetation cover located 5.6 km apart (Table 1). These Fst values were qualitatively and quantitatively similar in the "no family ties" dataset albeit 23% higher on average than in the entire dataset (Table 4). This is not surprising, since removing close relatives inflates total genetic variance more strongly than it increases within-population variance, which causes the Fst to increase. Importantly, Fst measures applied to the "no family ties" and "symmetrical minimal" datasets yielded qualitatively the same results as the full dataset containing 197 individuals (Table 4, Table S4). Fst values derived from microsatellites (data from Porlier et al. 2012b) were 11% lower on average relative to those measured from genome-wide data, which is an intrinsic consequence of higher polymorphism in microsatellite markers (Edelaar *et al.* 2011; Jakobsson *et al.* 2013).

We further measured genetic differentiation values at the transcriptomic level. A summary of matched RAD sequences against each of the three transcriptomes (blue tit, great tit, zebra finch) is presented in Table 3. Overall, 6.1% of the RAD sequences (1202 out of 19760 sequences) in this study were matched to either great tit or blue tit transcriptome sequences, and 11.4% (2251 out of 19760 sequences) aligned to either one of four available transcriptomes (blue tit, great tit, zebra finch cDNA, zebra finch *ab initio* genes). Out of the 326 RAD loci that aligned on the blue tit transcriptome (1.65%, Table 3), 179 SNPs were retained for pairwise Fst tests. Transcriptomic Fst values strongly corroborated those found genome-wide, although they were stronger on average by 24% (Table 4).

PCA and sPCA analyses of genetic distinctiveness

Genetic distinctiveness between Corsican birds (*C. caeruleus ogliastrae*) and Southern French blue tits from the mainland (*C. caeruleus caeruleus*) was evidenced by the projection of individuals on PCA axis 1, which encompasses 6.68% of the entire genetic variance (Figure 4). Genetic differentiation between the Corsican sites of Muro and Pirio was explained by PCA axis 2, which captured 1.61% of the genetic variance. PCA analyses within Corsica showed a fine scale genetic differentiation with both spatial and habitat components (Figure S3).

Inferring spatially explicit within-population structure using the spatial PCA method corroborated Fst and PCA results since it revealed a significant differentiation within Muro reflecting habitat structure (Global P-value: 0.028, N=57, Table S2). In addition, we found no evidence for spatial structure within the Muro deciduous habitat (Global P-value: 0.214, N=48), the Pirio evergreen habitat (Global P-value: 0.103, N=83), or within the Rouviere habitat (Global P-value: 0.116, N=57). PCA and sPCA analyses applied to the "no family ties" dataset yielded qualitatively the same results as in the full dataset containing 197 individuals (Figure S5, Table S1, Table S2).

Redundancy analysis reveals significant habitat and spatial components of differentiation

When the 4 sites were analysed together (n=197), the proportion of constrained variance explained by the redundancy analysis (RDA) was highly significant (Table 5), thus confirming the informativeness of the constraining variables used in the full RDA model. After removing the single non-significant term (i.e. the number of reads), four constrained axes explained 9.6% of the total genotypic variance and the first two RDA axes received a large contribution of both habitat and spatial variables (Table 5, Table 6 & Table S3). Geographical location (latitude and longitude) was largely represented by RDA axis 1, whereas habitat type (deciduous or evergreen) was mainly

captured both by RDA axis 1 and 2, and birth year mainly by RDA axis 2 (Figure 5A).

Because geographical coordinates explained the largest amount of variance among individuals, we further restricted the RDA analysis to Corsican birds only (n=140) to test for habitat effects independently of the geographical distance between continent and Corsica. With the 3 Corsican sites included in the analysis (D-Muro, E-Muro, E-Pirio), the RDA was highly significant (Table 5). After removing correlated and non-significant terms (longitude and number of reads, respectively), the three constrained axes explained 3.6% of the total genotypic variance and the first two RDA axes received a large contribution of habitat and latitude (Table 6, Figure 5B).

The partial habitat RDA conditioned on geography and birth year revealed a significant effect of habitat after removing variation caused by the other significant factors, both for the full dataset (Table 5), the Corsican dataset (Table 5) and for the "symmetrical minimal" dataset (Table S3). Thus, habitat type (deciduous vs. evergreen oaks) was a significant predictor of genotypic variation independently of geographical distance and birth date. Interestingly, when Corsican genotypes were projected on the single habitat RDA axis conditioned for other variables (the direction of the habitat vector, Figure S4), we observed greater genetic distinctiveness between D-Muro and E-Muro than between D-Muro and E-Pirio, which reflected the Fst values between these populations presented in table 4. Also, a non-explained source of genotypic variance in E-Pirio was captured by the first principal component of the three partial RDAs (Figure S4). These analyses revealed the genetic distinctiveness of 17 individuals from E-Pirio (those with the most negative coordinates on PC1), which already occupied extreme positions on the axis 2 of the PCA (Fig. 4).

Finally, we compared the distributions of SNP contributions to the single conditioned RDA axis of 3 partial RDAs that independently captured the effect of habitat, latitude and birth date in Corsica. We found that the three distributions largely overlapped, and that none of them was driven by SNPs showing large contributions (Figure S4).

Discussion

Dense SNP genotyping obtained by RAD sequencing and applied to a phenotypically well characterised study system of free living birds revealed significant fine-scale genetic structuring at a small spatial scale (5.6 km), a distance considered to be within the natal dispersal range of continental blue tits as estimated by Tufto et al. (2005) and discussed by Charmantier et al. (in press). Moreover, RDA analyses confirmed that genomic differentiation between populations was significantly driven by the type of oaks blue tits reproduced in, independently of geographical distance. A previous study (Porlier *et al.* 2012b) using microsatellite genetic characterisation of the two populations inhabiting the Muro valley did not reach enough power to detect a significant differentiation signal between D-Muro and E-Muro, although the order of pairwise Fst values was the same both here as in Porlier *et al*. (2012b). Here, we demonstrated that the two populations significantly differ from each other in terms of their allelic frequencies at the genomic and transcriptomic level, independently of the presence of individuals with close family ties in the dataset or other confounding effects such as sample size or temporal variation in birth year.

The possibility to identify genetic distinctiveness at such a small spatial scale in a highly mobile species unequivocally suggests that the large number of SNPs identified through RAD sequencing brings unprecedented explanatory power in elucidating weak yet distinct genetic signals harboured by populations breeding in heterogeneous habitats. The fine-scale genetic structuring coincides with evergreen and deciduous habitat patchiness in the valley of Muro, and agrees with earlier reports identifying habitat type to be instrumental in creating structure at the genetic (Porlier *et al.* 2012b) and phenotypic (Blondel *et al.* 1999; Blondel *et al.* 2006; Charmantier *et al.* in press; Lambrechts *et al.* 1997) level. At the same time, the population genetic landscape of Corsican blue tit populations was found to be more complex than expected, and required the use of complementary analytical methods to unravel the potential of habitat-dependent genetic structuring. In this context, RDA

analysis proved particularly suitable to identify and test the effect of individual variables influencing genomic variability, while also offering the potential to detect collinearity between them. Below we discuss in detail both methodological aspects as well as key biological findings of the study.

Genetic diversity, relatedness distribution and repeatability

RAD-sequencing and subsequent bioinformatic analyses resulted in identifying c. 12 000 SNPs with a 2% MAF (and c. 6500 SNPs with a 5% MAF) genotyped on average in 96% of birds. Overall, single-end RAD sequencing confirmed its suitability for dense genotyping in a natural bird population with no available reference genome.

Allele frequency spectra differed markedly between Corsica and the mainland, due to high-levels of low-frequency polymorphisms on Corsica contrasting with a deficit of low-frequency variants on the mainland. These differences in the distributions of allele frequencies probably reflect contrasting demographic histories between mainland and Corsica. By contrast, analysing the distribution of genome-wide similarity between birds in the dataset revealed highly similar population relatedness composition on the mainland and in Corsica. These relatedness structures were characterised by a mode of unrelated individual and a right-hand tail of close relatives, that were already validated by parentage analysis based on microsatellite data from Charmantier *et al.* (2004) and field observations. The presence of family members in the dataset, sometimes associated with field sampling limitations in time and space, always constitutes an inherent yet unknown fraction of populations sampled at random when no pedigree is available. These unknown family links can thus be straight-forwardly revealed when dense SNP genotyping is available, without the need for computationally intensive genetic pedigree reconstruction. Importantly, the population genetic patterns observed in this study were confirmed using concurrent “no family ties” and “symmetrical

463 minimal" datasets, yielding qualitatively and quantitatively comparable results (see Part II and III of
464 supplementary material).

465 Finally, genotyping repeatability scores not only provided a useful analytical control step in sample
466 processing, but it also generated insight into the limits of genotype-by-sequencing accuracy. While
467 next-generation sequencing repeatability scores have been reported at intermediate analytical
468 stages (Sharma *et al.* 2012), genotyping error rates estimated at the final stages of bioinformatic
469 analyses are rarely reported in RAD-seq studies (but see Mastretta-Yanes *et al.* 2015). Here, our
470 strategy for inferring individual genotypes *de novo* allowed to keep the genotyping error rate below
471 3%, a value that is close to the lowest error rates estimated by Mastretta-Yanes and colleagues
472 (2015) in RAD-seq studies.

473

474 **Population genetic structure**

475 When Corsica and the mainland population of Rouvière were compared, Fst values observed in this
476 study were well within the range of genetic differentiation observed in other blue tit and great tit
477 (*Parus major*) populations. Indeed, these studies reported Fst indexes that can be as low as 0.01 for
478 the great tit between a Dutch and a UK population (Van Bers *et al.* 2012), and as high as 0.79
479 between two insular blue tit populations in the Canary Island system (Hansson *et al.* 2014). Genetic
480 differentiation between Corsican populations and the mainland site of Rouviere was nearly 4 times
481 stronger than Fst values within Corsica, thereby confirming the genetic distinctiveness of the
482 Corsican blue tit sub-species, and the very limited gene flow between the island and the mainland (as
483 in Porlier *et al.* 2012b).

484 Within the island of Corsica however, genetic differentiation does not only scale with geographical
485 distance: the strongest, and highly significant genetic differentiation between Corsican populations

was found between two populations inhabiting two different oak habitats within the same Muro valley, with an average nestbox distance between the two populations of 5.6 km, and 4 km between the closest nestboxes from each habitat. While the signal of genetic differentiation between the two populations is surprising, this result is particularly robust since it was not only confirmed by four complementary population genomic analyses (Fst , PCA, sPCA and RDA), but also when tested at genome-wide and transcriptome-wide levels, and when controlling for family structure, sample size and birth year (Table 4, Supplementary material part II and III). It is worth noting that the strength of genetic differentiation (pairwise Fst) was strongest using transcriptome-derived SNPs, followed by genomic SNPs, and weakest when using microsatellites (Table 4). High-dimensional SNP datasets undoubtedly provide an increased precision and a more powerful detection of small Fst values than those derived from a small number of markers (Waples 1998). At the same time, higher transcriptome Fst values relative to those calculated genome-wide likely reflect the effect of selection at linked sites, causing local reductions in effective population size in coding regions due to purifying selection (Charlesworth *et al.* 1993). Further analyses will be required to gain better insight into which genomic, and in particular transcriptomic regions co-vary with the phenotypic differences in the study system.

Spatial PCA and RDA analysis explicitly tested and provided support for small-scale differentiation and the possible role of habitat in generating genomic structuration (Figure 5, Table 5, Table S2). Concurrent efforts to include a greater number of sampling sites with replicated oak habitats and nestbox-specific indicators of environmental heterogeneity (Szulkin *et al.* in press) would be valuable to fully validate the role of IBE in generating genomic structuring in this study system. Moreover, the Mediterranean blue tit study system offers decades of individual life-history and fitness measures in the 4 sites studied here, thus offering considerable potential for complementary analyses of covariation between genomic, phenotypic and environmental data.

510

511 **Isolation-by-Environment despite high capacities for gene flow**

512 One important question that needs to be further addressed by integrating ecological, behavioural
513 and genetic data, is whether isolation-by-environment results from reduced dispersal through
514 habitat choice, local selection against maladapted genotypes, or a combination of both. The average
515 natal dispersal distance of blue tits on the continent is crudely estimated to range between ~330m
516 and 4 km (depending on population and dispersal distance estimation method, (Ortego *et al.* 2011;
517 Tufto *et al.* 2005)). However, there is variation around these average values, and there are known
518 records of much larger blue tit natal dispersal distances (see Charmantier et al. in press and
519 discussion therein). The scale of natal dispersal in Corsica is currently unknown and may be smaller
520 than in the rest of the species range due to the insular nature of the population.

521 Given that there is no barrier to dispersal in the Corsican landscape (such as important mountain
522 ridges or open spaces birds would be reluctant to fly over (Blondel *et al.* 2006; Porlier *et al.* 2012b)),
523 a significant habitat-driven genetic differentiation at a 5.6 km scale in such a highly mobile species as
524 the blue tit suggests either strong local selection capable of counteracting gene flow, or non-random
525 dispersal of genotypes with respect to habitat type. The fact that we could not detect loci with
526 extreme contributions to fine-scale differentiation between habitats in our partial RDA analysis
527 (Figure S4) suggests that the signal of differentiation is genome-wide rather than driven by a subset
528 of loci strongly influenced by selection. Admittedly, polygenic selection acting on a large suite of
529 complex quantitative traits could generate correlated but minor allele frequency changes (Latta
530 1998; Le Corre & Kremer 2012), but it remains unclear whether polygenic selection in the face of
531 gene flow could translate into a detectable IBE pattern in the RDA analysis. Thus, it is possible that
532 the fine-scale genetic structure is more likely to be the outcome of a migration-drift balance (here
533 associated with habitat choice) than that of a migration-selection balance (concurrent with local

adaptation). This interpretation may also account for the surprising finding of small and non-significant Fst values between D-Muro and E-Pirio (located 25 km apart), contrasting with the significant differentiation found between D-Muro and E-Muro over a much smaller spatial scale. Indeed, it is likely that the blue tit population from E-Muro has an overall smaller population effective size than E-Pirio. The E-Muro population also represents an isolated patch of evergreen habitat surrounded by deciduous habitat populations while E-Pirio is well connected to other evergreen habitat populations; hence the possibly higher rate of drift in E-Muro. Therefore, for similar migration rates between D and E habitats, we expect a stronger genetic differentiation at equilibrium between D-Muro and E-Muro than between D-Muro and E-Pirio, which is what we observed. Our results thus suggest that there is limited dispersal across habitats resulting most probably from matching habitat choice (Edelaar *et al.* 2008); but currently on-going cross-fostering between habitats, coupled with larger spatial sampling schemes will be instrumental to understand in greater detail the relative importance of local adaptation and habitat choice.

While genetic structuring at small spatial scales is known to occur in some vertebrate species, three avian studies (Garcia-Navas *et al.* 2014; Postma *et al.* 2009; Senar *et al.* 2006) mirror the fine-scale pattern of genetic differentiation reported here. However, the reported genetic differences are not always supported by evident environmental differences, and in all three cases comparisons are based on 2 populations. Interestingly, Postma and van Noordwijk (2005) and Postma et al. (2009) found that phenotypic differences in great tit clutch size on Vlieland (an island in The Netherlands 19km long, 2km wide and 25km from the mainland) were coupled with microsatellite genetic differentiation at a similar geographical scale as in this study. The authors argued that such genetic differences could arise thanks to highly restricted gene flow to some parts of the island and selection against immigrants. While there are no physical barriers to dispersal for Corsican blue tits (Blondel *et al.* 2006; Porlier *et al.* 2012b), limited dispersal may act as a component of the "insular syndrome" (Adler & Levins 1994; Blondel *et al.* 2006); see also Bertrand et al. (2014) and Komdeur *et al*. (2004)),

559 contributing to enhance the genetic differentiation and contributing to local adaptation at small
560 spatial scales in island settings.

561

562 **Conclusions and Perspectives**

563 It is undisputable that ongoing improvements in high-throughput sequencing, SNP chip development
564 and genotyping-by-sequencing approaches facilitate the creation of a rapidly growing number of
565 large population genomic datasets in wild animal populations, and are in consequence impacting our
566 understanding of the factors influencing genomic structuration of natural populations (Ellegren *et al.*
567 2012; Poelstra *et al.* 2014). Here we have validated the potential for RAD sequencing to study small
568 scale genomic differentiation in an avian system (see also Bertrand et al. (2014)). Results in this
569 study, but also in those of Bertrand *et al*. (2014) and Postma *et al*. (2009) contribute to undermine
570 generally held assumptions regarding the homogenising effect of gene flow at small spatial scales in
571 terrestrial vertebrates and birds in particular. Moreover, our study suggests that habitat may play a
572 key role in generating genome-wide IBE patterns, which is concomitant to habitat-dependent
573 phenotypic variation reported earlier (Figure 1). In the next phase of genomic exploration of the
574 Mediterranean blue tit study system, we plan to apply finer-scale axes of environmental variation
575 (Garroway *et al.* 2013), in particular by focusing on high-resolution satellite imagery (Szulkin *et al.* in
576 press) and quantitative genetic analyses of phenotypic trait variation, thereby providing a robust
577 framework to test hypotheses of habitat-dependent adaptation at the genetic level.

578

579

580

Acknowledgements

This study was funded by an IEF Marie Curie Fellowship to M.S., an ANR BioAdapt grant (ANR-12-ADAP-0006-02-PEPS) to A.C., an APEGE funding to A.C. & M.S., and by OSU-OREME funding for the long-term monitoring of tits in Corsica and La Rouviere sites. We thank Pascal Marrot for discussion and for providing nestbox GPS coordinates and oak data, and Marie-Pierre Dubois and Max Galan for help in the lab. We thank Nicolas Galtier for access to blue tit transcriptome data and Anna Santure for the great tit transcriptome. Bioinformatic analyses were conducted at the ISEM platform PGPM7 at the Station Méditerranéenne de l'Environnement Littoral (OSU OREME) and through access to the Montpellier LabEx CeMEB computation facilities. We thank Jason Boone for advice on RAD analyses, and Jacques Blondel ,Dany Garant and Charles Perrier for valuable feedback on the manuscript. We also whole-heartedly thank the many generations of researchers, students and field assistants who contributed to blue tit population monitoring and sampling and who made this study possible.

References

Adler GH, Levins R (1994) The island syndrome in rodent populations. *Quarterly Review of Biology* **69**, 473-490.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Bertrand JAM, Bourgeois YXC, Delahaie B*, et al.* (2014) Extremely reduced dispersal and gene flow in an island bird. *Heredity* **112**, 190-196.

Blondel J, Dias PC, Ferret P, Maistre M, Lambrechts MM (1999) Selection-based biodiversity at a small spatial scale in a low-dispersing insular bird. *Science* **285**, 1399-1402.

Blondel J, Perret P, Dias PC, Lambrechts MM (2001) Is phenotypic variation of blue tits (Parus caeruleus L.) in Mediterranean mainland and insular landscapes adaptive? *Genetics Selection Evolution* **33**, S121-S139.

Blondel J, Perret P, Maistre M (1990) On the genetic basis of the laying date in an island population of blue tits. *Journal of Evolutionary Biology* **3**, 469-475.

Blondel J, Thomas DW, Charmantier A*, et al.* (2006) A thirty-year study of phenotypic and genetic variation of blue tits in Mediterranean habitat mosaics. *Bioscience* **56**, 661-673.

Cahais V, Gayral P, Tsagkogeorga G*, et al.* (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources* **12**, 834-845.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3-Genes Genomes Genetics* **1**, 171-182.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303.

Charmantier A, Blondel J, Perret P, Lambrechts MM (2004) Do extra-pair paternities provide genetic benefits for female blue tits Parus caeruleus? *Journal of Avian Biology* **35**, 524-532.

Charmantier A, Doutrelant C, Dubuc Messier G, Fargevieille A, Szulkin M (in press) Mediterranean blue tits as a case study of local adaptation. *Evolutionary Applications*.

Danecek P, Auton A, Abecasis G*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Edelaar P, Alonso D, Lagerveld S, Senar JC, Bjorklund M (2012) Population differentiation and restricted gene flow in Spanish crossbills: not isolation-by-distance but isolation-by-ecology. *Journal of Evolutionary Biology* **25**, 417-430.

Edelaar P, Bolnick DI (2012) Non-random gene flow: an underappreciated force in evolution and ecology. *Trends in Ecology & Evolution* **27**, 659-665.

Edelaar P, Burraco P, Gomez-Mestre I (2011) Comparisons between Q(ST) and F-ST-how wrong have we been? *Molecular Ecology* **20**, 4830-4839.

Edelaar P, Siepielski AM, Clobert J (2008) Matching habitat choice causes directed gene flow: a neglected dimension in evolution and ecology. *Evolution* **62**, 2462-2472.

Ellegren H, Smeds L, Burri R*, et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**, 756-760.

Garant D, Forde SE, Hendry AP (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation. *Functional Ecology* **21**, 434-443.

Garcia-Navas V, Ferrer ES, Sanz JJ, Ortego J (2014) The role of immigration and local adaptation on fine-scale genotypic and phenotypic population divergence in a less mobile passerine. *Journal of Evolutionary Biology* **27**, 1590-1603.
Garroway CJ, Radersma R, Sepil I*, et al.* (2013) Fine-scale genetic structure in a wild bird population: the role of limited dispersal and environmentally based selection as causal factors. *Evolution* **67**, 3488-3500.
Hansson B, Ljungqvist M, Illera J-C, Kvist L (2014) Pronounced Fixation, Strong Population Differentiation and Complex Population History in the Canary Islands Blue Tit Subspecies Complex. *Plos One* **9**.
Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972.
Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics* **7**, 1-32.
Jakobsson M, Edge MD, Rosenberg NA (2013) The Relationship Between FST and the Frequency of the Most Frequent Allele. *Genetics* **193**, 515-528.
Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.
Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.
Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**, 92-103.
Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* **7**, 1225-1241.
Komdeur J, Piersma T, Kraaijeveld K, Kraaijeveld-Smit F, Richardson DS (2004) Why Seychelles Warblers fail to recolonize nearby islands: unwilling or unable to fly there? *Ibis* **146**, 298-302.
Lambrechts MM, Blondel J, HurtrezBousses S, Maistre M, Perret P (1997) Adaptive inter-population differences in blue tit life-history traits on Corsica. *Evolutionary Ecology* **11**, 599-612.
Lambrechts MM, Caro SP, Charmantier A*, et al.* (2004) Habitat quality as a predictor of spatial variation in blue tit reproductive performance: a multi-plot analysis in a heterogeneous landscape. *Oecologia* **141**, 555-561.
Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist* **151**, 283-292.
Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology* **21**, 1548-1566.
Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution* **17**, 183-189.
Martin JL (1991) Patterns and significance of geographical variation in the blue tit (Parus caeruleus). *Auk* **108**, 820-832.
Mastretta-Yanes A, Arrigo N, Alvarez N*, et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* **15**, 28-41.
Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* **37**, 17-23.
O'Connor TD, Fu W, Turner E*, et al.* (2015) Rare variants facilitates inferences of fine scale population structure in humans. *Molecular Biology and Evolution* **32**, 653-660.
Oksanen J, Blanchet GG, Kindt R*, et al.* (2014) vegan: Community Ecology Package.
Ortego J, Garcia-Navas V, Ferrer ES, Sanz JJ (2011) Genetic structure reflects natal dispersal movements at different spatial scales in the blue tit, Cyanistes caeruleus. *Animal Behaviour* **82**, 131-137.
Palumbi SR (1994) Genetic divergence, reproductive isolation and marine speciation. *Annual Review of Ecology and Systematics* **25**, 547-572.

688 Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *Plos Genetics* **2**, 2074-
689 2093.
690 Poelstra JW, Vijay N, Bossu CM*, et al.* (2014) The genomic landscape underlying phenotypic integrity
691 in the face of gene flow in crows. *Science* **344**, 1410-1414.
692 Porlier M, Charmantier A, Bourgault P*, et al.* (2012a) Variation in phenotypic plasticity and selection
693 patterns in blue tit breeding time: between- and within-population comparisons. *Journal of*
694 *Animal Ecology* **81**, 1041-1051.
695 Porlier M, Garant D, Perret P, Charmantier A (2012b) Habitat-Linked Population Genetic
696 Differentiation in the Blue Tit Cyanistes caeruleus. *Journal of Heredity* **103**, 781-791.
697 Postma E, Den Tex R-J, Van Noordwijk AJ, Mateman AC (2009) Neutral markers mirror small-scale
698 quantitative genetic differentiation in an avian island population. *Biological Journal of the*
699 *Linnean Society* **97**, 867-875.
700 Postma E, van Noordwijk AJ (2005) Gene flow maintains a large genetic difference in clutch size at a
701 small spatial scale. *Nature* **433**, 65-68.
702 Ravigne V, Dieckmann U, Olivieri I (2009) Live Where You Thrive: Joint Evolution of Habitat Choice
703 and Local Adaptation Facilitates Specialization and Promotes Diversity. *American Naturalist*
704 **174**, E141-E169.
705 Romiguier J, Gayral P, Ballenghien M*, et al.* (2014) Comparative population genomics in animals
706 uncovers the determinants of genetic diversity. *Nature* **515**, 261-U243.
707 Santure AW, Gratten J, Mossman JA, Sheldon BC, Slate J (2011) Characterisation of the transcriptome
708 of a wild great tit Parus major population by next generation sequencing. *Bmc Genomics* **12**.
709 Senar JC, Borras A, Cabrera J, Cabrera T, Bjorklund M (2006) Local differentiation in the presence of
710 gene flow in the citril finch Serinus citrinella. *Biology Letters* **2**, 85-87.
711 Sharma R, Goossens B, Kun-Rodrigues C*, et al.* (2012) Two Different High Throughput Sequencing
712 Approaches Identify Thousands of De Novo Genomic Markers for the Genetically Depleted
713 Bornean Elephant. *Plos One* **7**.
714 Slatkin M (1973) Gene flow and selection in a cline. *Genetics* **75**, 733-756.
715 Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* **236**, 787-
716 792.
717 Snow DW (1954) The habitats of eurasian tits (Parus spp.). *Ibis* **96**, 565-585.
718 Szulkin M, Stopher KV, Pemberton JM, Reid JM (2013) Inbreeding avoidance, tolerance, or
719 preference in animals? *Trends in Ecology & Evolution* **28**, 205-211.
720 Szulkin M, Zelazowski P, Marrot P, Charmantier A (in press) Application of high temporal and spatial
721 resolution satellite imagery to characterise individual-based environmental heterogeneity in
722 a wild bird. *Remote Sensing*.
723 Tufto J, Ringsby TH, Dhondt AA, Adriaensen F, Matthysen E (2005) A parametric model for estimation
724 of dispersal patterns applied to five passerine spatially structured populations. *American*
725 *Naturalist* **165**, E13-E26.
726 Van Bers NEM, Santure AW, Van Oers K*, et al.* (2012) The design and cross-population application of
727 a genome-wide SNP chip for the great tit Parus major. *Molecular Ecology Resources* **12**, 753-
728 770.
729 Wang IJ, Bradburg GS (2014) Isolation by environment. *Molecular Ecology* **23**, 5649-5662.
730 Waples RS (1998) Separating the wheat from the chaff: Patterns of genetic differentiation in high
731 gene flow species. *Journal of Heredity* **89**, 438-450.
732 Zheng XW, Levine D, Shen J*, et al.* (2012) A high-performance computing toolset for relatedness and
733 principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328.

734

735

Data Accessibility

Demultiplexed RAD sequencing read data is available at the NCBI Short Read Archive under accession number SRP065946 . The raw VCF file (209 860 SNPs), the filtered VCF file (12 106 SNPs), a list of consensus RAD sequences and sample details are available on Dryad (doi:10.5061/dryad.713v1).

Author contributions

M.S. and A.C. designed the study; M.S. performed preliminary lab work; M.S. and P.-A. G. analysed the data with input from N.B.; M.S. wrote the manuscript with input from P-A. G., N. B. and A. C.

Tables and Figures

Table 1.

Location & coordinates (long, lat)	Breeding site	Predominant Oak species	N birds (N females)	Average geographical distances between (in km):
Corsica (42.5893, 8.9667)	E-Muro	Evergreen holm oak *Quercus ilex* (100%)	9 (3)	E-Muro and D-Muro: 5.6
Corsica (42.5509, 8.9233)	D-Muro	Deciduous downy oak *Quercus pubescens* (100%)	48 (19)	
Corsica (42.3763, 8.7497)	E-Pirio	Evergreen holm oak *Quercus ilex* (100%)	83 (42)	E-Muro and E-Pirio: 29.6 D-Muro and E-Pirio: 24.1
Continental France (43.6639, 3.6658)	D-Rouviere	Deciduous downy oak *Quercus pubescens* (81%)	57 (32)	Corsican sites and D-Rouviere: 441.2

Table 1. Spatial and vegetation-based habitat characteristics of the different sampling sites in Southern French mainland and on Corsica Island. For oak species, values in brackets indicate the mean proportion (in %) of predominant oak species (Evergreen "E" vs. Deciduous "D") within a 50m radius of each nestbox in the dataset. Sampling site coordinates were calculated as the average of nestbox coordinates for all breeding birds sampled in the study.

Table 2.

Step in the bioinformatic pipeline	Number of SNPs	Analysis applied to that dataset
Raw VCF file from *Stacks*	209 860	Quality filtering
Trimming bases ≥ to position 85 of reads	189 356	Quality filtering
Excluding loci that are not in within-population HWE at p<0.01 within D-Rouvière D, E&D-Muro or E-Pirio.	166 410	Quality filtering
Working dataset MAF 2%, 90% call rate	12 106	IBS analysis, Spatial PCA, RDA
Working dataset MAF 5%, 90% call rate	6 555	PCA
Working dataset MAF 5%, 95% call rate	3 159	Genome-wide pairwise Fst Transcriptome pairwise Fst

Table 2. RAD sequencing bioinformatic analysis pipeline, starting from the VCF file generated with *Stacks*, and detailing the filtering steps undertaken in *VCFtools*. The resulting changes in the number of SNPs are indicated at each step. The genotype call rate threshold represents the minimum proportion of genotypes called for each locus across all birds in the dataset. The MAF represents the minimum allele frequency threshold applied to the rare variant for each locus. Number of birds sampled: 197.

Table 3.

transcriptome	N of sequences present in the transcriptome	N of aligned RAD tags	% of mapped RAD tags	Average E value	% identical
Blue tit	7120	326	1.6	1.02 E-9	97
Great tit	95979	1007	5.1	1.56 E-8	96
Zebra finch *cDNA*	59816	722	3.7	4.89 E-10	96
Zebra finch *ab-initio*	18610	1135	5.7	2.14 E-9	95

Table 3. Blue tit *Blastx* outputs obtained from searches against blue tit, great tit and zebra finch transcriptomes. Total number of consensus RAD tags blasted = 19 760 (global call rate > 80%, MAF > 1%); all E-values assessing the significance of blast matches were ≤1E-7.

Table 4.

	D-Rouviere	D-Muro	E-Muro	E-Pirio
D-Muro	**0.0541 (p≤0.002**)** 0.0572 (p≤0.002**) 0.0591 (p≤0.002**) *0.0487*	-	-	-
E-Muro	**0.0335 (p≤0.002**)** 0.0480 (p≤0.002**) 0.0379 (p≤0.002**) *0.0415*	**0.0156 (p=0.004**)** 0.0182 (p=0.006**) 0.0246 (p≤0.002**) *0.0065*	-	-
E-Pirio	**0.0520 (p≤0.002**)** 0.0531 (p≤0.002**) 0.0600 (p≤0.002**) *0.0413*	**0.0099 (p=0.347)** 0.0111 (p=0.762) 0.0102 (p=0.333) *0.0043*	**0.0102 (p=0.403)** 0.0160 (p=0.168) 0.0151 (p=0.015*) *0.0050*	-

Table 4. Fst values for SNPs retained after filtering with 5% MAF and a 95% call rate. In bold: all SNPs *n*=197 individuals, 3159 SNPs. Underlined: all SNPs in a "no family ties" dataset", *n*=119 individuals, 2816 SNPs. Normal type: transcriptome-derived SNPs, *n*=197 individuals, 179 SNPs. Empirical p-values were computed using 500 permutations (lowest p-values are therefore bounded by 0.002). After Bonferroni correction, a significant signal of genetic differentiation in RADseq derived Fst values was found in all Corsica-mainland comparisons and between D-Muro and E-Muro in each of the inspected datasets. In italic: Fst values from Porlier *et al.* 2012 (data averaged across several years, *n*=607 adults (see manuscript for sampling details), 6-10 microsatellite markers).

784 # Table 5.

785

	Full dataset (continent+Corsica) N=197	**Conditioned RDA full dataset N=197**	**Corsican birds only N=140**	**Conditioned RDA Corsican birds N=140**
	Variance (D.f), p-value	Variance (D.f.), p-value	Variance (D.f), p-value	Variance (D.f), p-value
Global Analysis	56.86 (4), 0.001	-	18.15 (3), 0.001	-
Residual	534.87 (192)		485.27 (136)	
Marginal Test				
Latitude	3.83 (1), **0.020**	3.83 (1), **0.002**	5.41 (1), 0.001	5.41 (1), 0.001
Longitude	4.21 (1), **0.010**	4.21 (1), **0.001**	-	-
Habitat	3.50 (1), **0.035**	3.50 (1), **0.002**	4.89 (1), 0.001	4.89 (1), 0.001
Birth Year	3.48 (1), **0.040**	3.48 (1), **0.001**	4.52 (1), 0.001	4.52 (1), 0.001
Residual	534.87 (192)	534.87 (192)	485.27 (136)	485.27 (136)

786

787 **Table 5.** Results of RDA significance tests (variance, d.f. and p-values obtained through 1000
788 permutations; significant p-values are in bold), detailed for the global RDA analysis (model with non-
789 significant terms removed) and the marginal effect of each constraining variable in the model. RDA
790 were performed on the full dataset (*n*=197) and on Corsican birds only (*n*=140). The marginal effect
791 of each constraining variable was tested trough permutation tests by removing each term one by one
792 from the model containing all other terms. The second and fourth results column report partial RDA
793 significance tests (variance, d.f. and p-values) for each term, after conditioning on other constraining
794 variables to remove their confounding effects. Longitude was not used as a constraining variable in
795 Corsican RDA due to its strong correlation with Latitude.

Table 6.

	Full dataset (continent+Corsica)				Corsican birds only		
RDA axis	RDA1	RDA2	RDA3	RDA4	RDA1	RDA 2	RDA 3
% of variance explained	7.39%	1.06%	0.58%	0.57%	1.77%	0.96%	0.88%
Constraining variables							
Latitude	-0.9926	0.1146	0.0404	0.0041	-0.9647	-0.2371	0.1144
Longitude	0.9988	0.0476	0.0003	0.0091	-	-	-
Habitat	0.6084	-0.7227	0.1196	0.3053	0.9119	-0.3577	-0.2014
Birth year	0.0212	-0.4572	0.7455	-0.4846	0.5012	-0.2634	0.8243

Table 6. Summary of RDA analysis for the full dataset and Corsican birds only. The proportion of genotypic variance explained by each RDA axis is provided, along with the vector coordinates of each constraining variable in the RDA space (these vectors are represented in blue colour, Figure 5). For each RDA axis, the longest vector projection indicates the most important variable explaining variation along that axis.

Figure Legends

Figure 1. Phenotypic trait values (means and 95% confidence intervals) for (A) fitness traits and (B) morphological traits for the four study sites. All traits depicted have a significant genetic basis established with quantitative genetic models (see Charmantier et al. in press for more details). All traits show high similarity within habitat type (deciduous or evergreen oaks – see illustrative tab below each graph), or (non-exclusively) carry continental distinctiveness (D-Rouviere), followed by intermediate values of E-Muro relative to D-Muro and E-Pirio. All traits were recorded annually between 1991 and 2014 for D-Rouviere, 1993 and 2014 for D-Muro, 1998 and 2014 for E-Muro and 1976 and 2014 for E-Pirio.

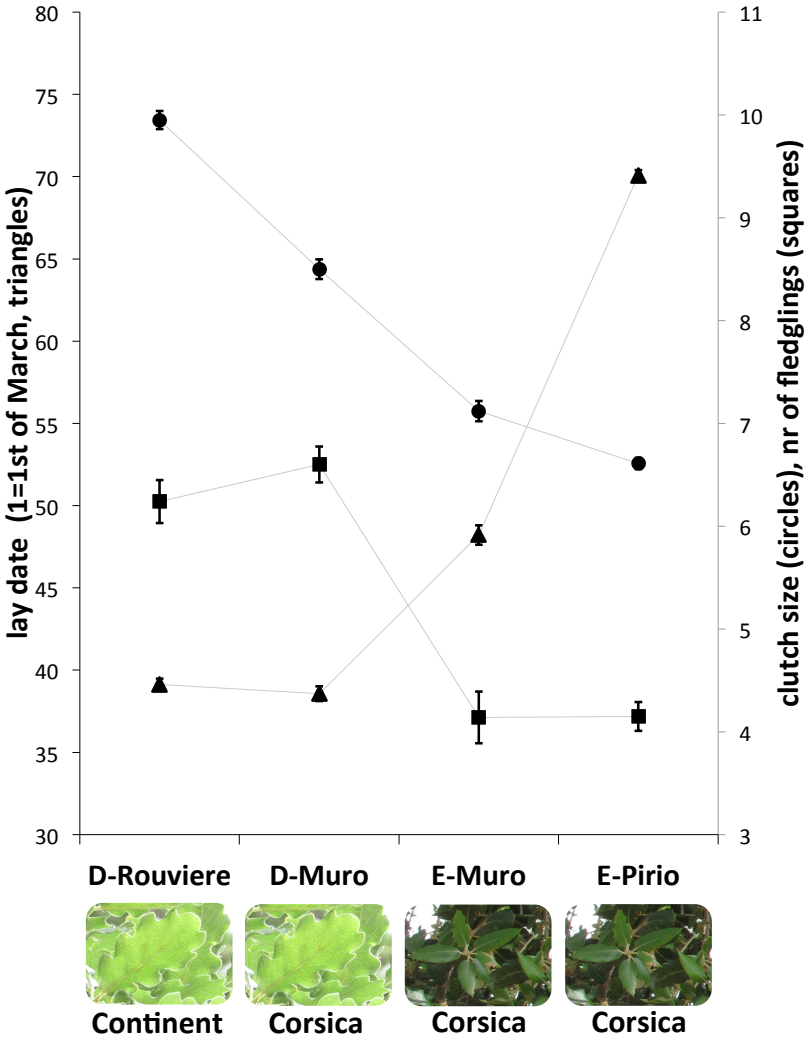
(A) Fitness traits: egg laying date is represented as filled triangles(1=1st March), clutch size (from first broods only) as circles and the number of fledglings as squares (total n=5566, 5555, 4367, respectively).

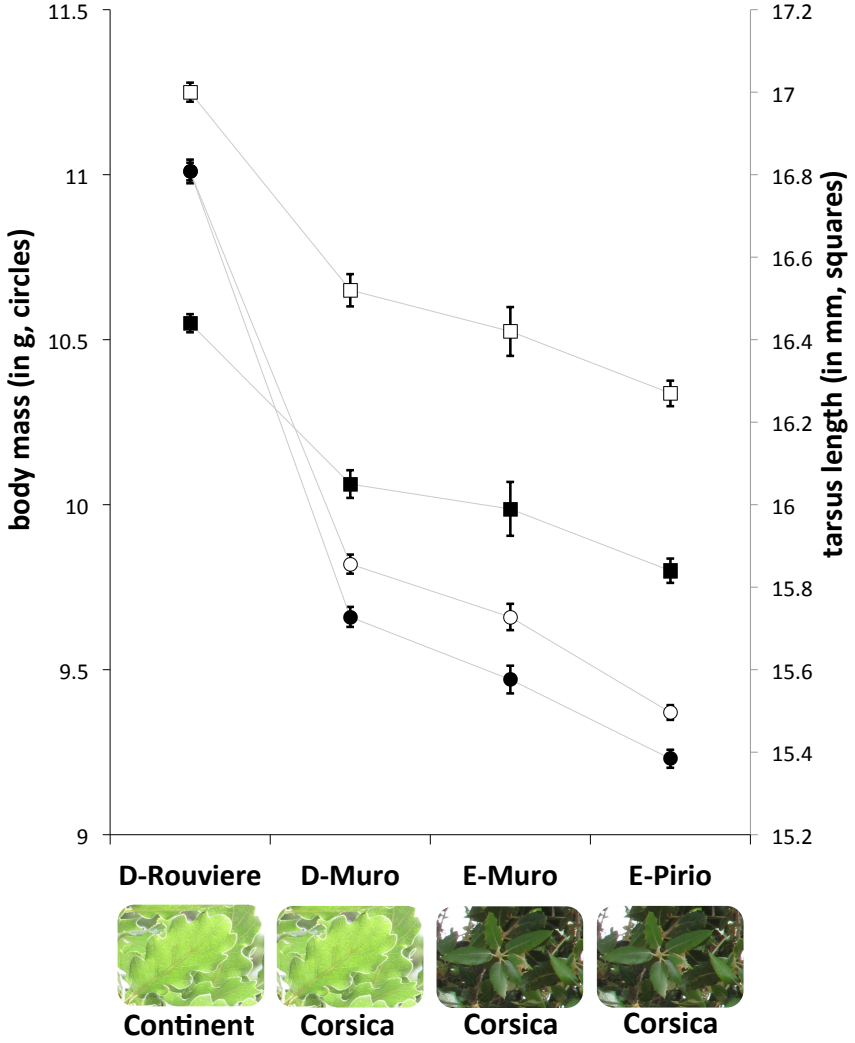
(B) Morphological traits: female and male body mass are represented as filled and open circles, female and male tarsus length are represented as filled and open squares (total n=4962, 4559, 3068, 2792 respectively).

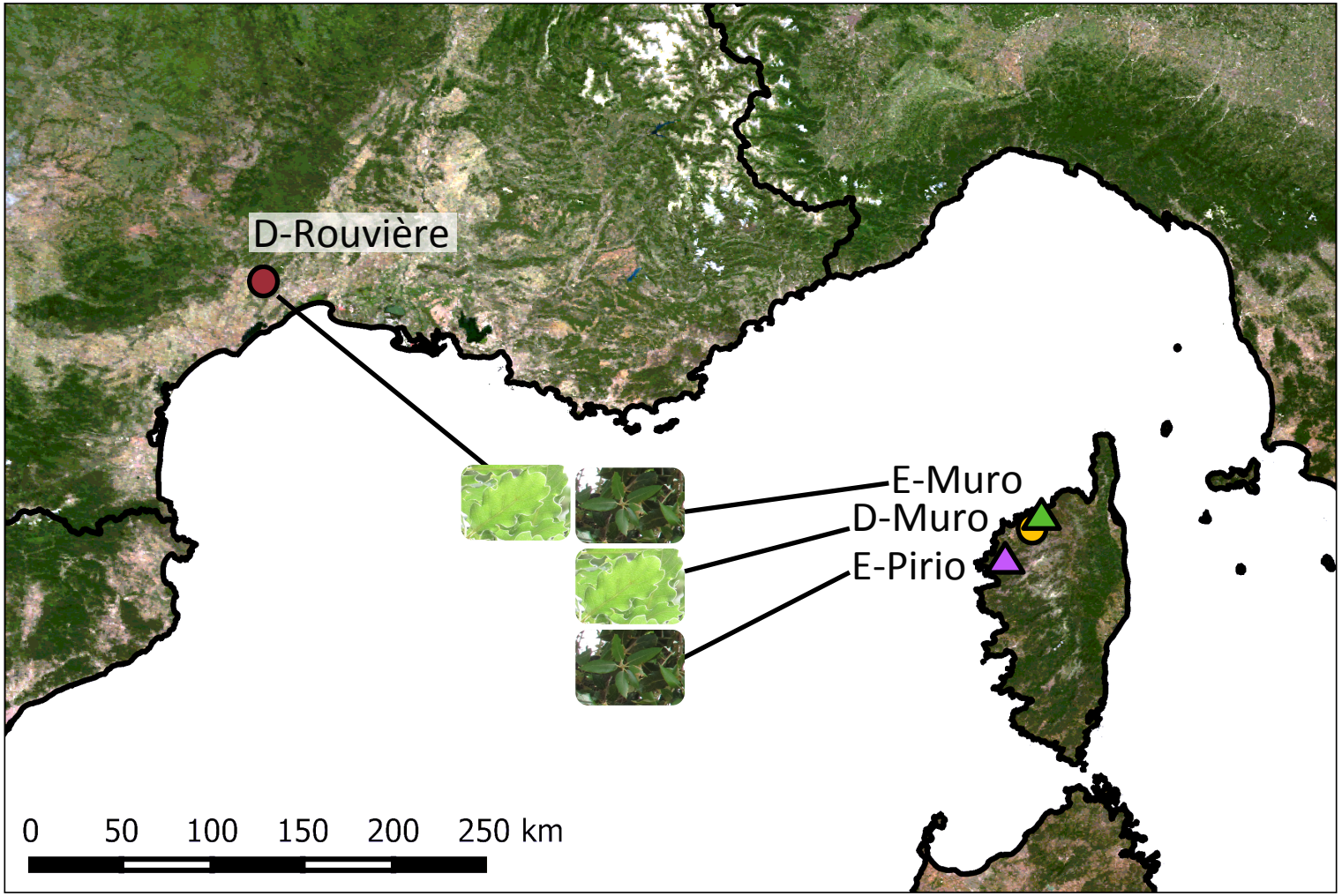
Figure 2. Map of the 4 study sites: D-Rouviere on the French mainland, E-Muro, D-Muro and E-Pirio in Corsica, complemented with an illustration of predominant oak species (Deciduous or Evergreen) for each site.

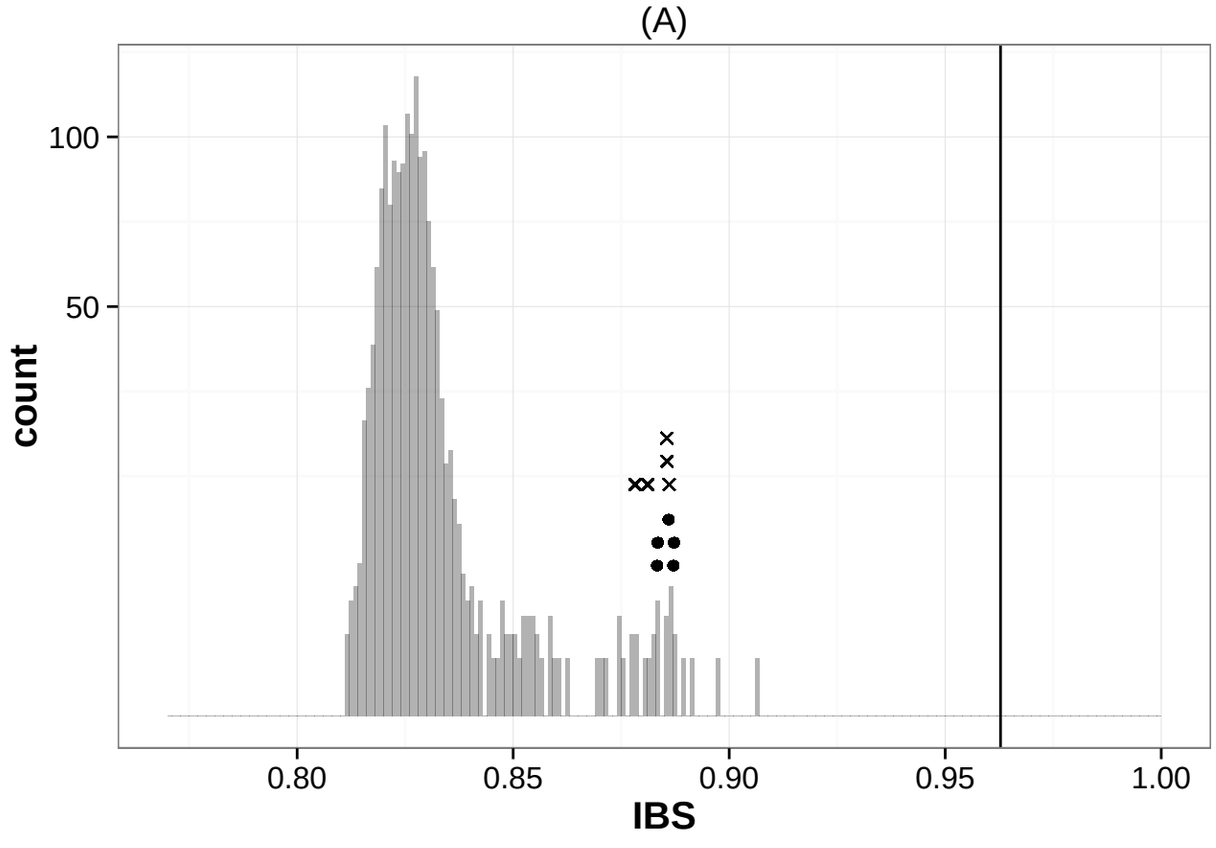
828 **Figure 3.** Population-specific distribution of pairwise relatedness coefficient and folded-allele
829 frequency spectrum.
830
831 **(A&B):** Identity-by-state (IBS) pairwise relatedness distribution for (A) the mainland (D-Rouviere) and
832 (B) Corsica (D-Muro, E-Muro and E-Pirio). The right hand tails of both continental and Corsican
833 populations reflect family structure in the datasets, further validated with IBS values of full siblings
834 (black crosses, established with microsatellite data from Charmantier et al. 2003) and mother-
835 offspring pairs (black diamonds, established with pedigree data). Black vertical lines are IBS values of
836 sample replicates (one replicate from the mainland, four replicates from Corsica). (**C&D):** Minor allele
837 frequency (MAF) spectrum for (C) the continent and (D) Corsica.
838
839 **Figure 4.** Principal Component Analysis (Axes 1 & 2, explaining 6.7% and 1,6% of the variance,
840 respectively – see also Table S1) of the 4 blue tit populations (*n*=197 individuals, 6555 SNPs, MAF 5%,
841 90% call rate). Colour legend: D-Rouviere (red circles), D-Muro (orange circles), E-Muro (green
842 triangles), E-Pirio (violet triangles).
843
844 **Figure 5.** RDA analysis (12106 SNPs, MAF 2%, 90% call rate) for **(A)** the entire dataset or **(B)** Corsican
845 birds only. Data points correspond to the projection of individual genotypes on RDA axes 1 and 2,
846 explaining cumulatively 8.5% of the total genotypic variance in (A) and 2.7% in (B). Vectors of
847 constraining variables (latitude, longitude, habitat, birth year) are projected on RDA axes 1 and 2;
848 their arrows points to the direction of strongest gradient of variation, and their projected lengths
849 indicate the strength of their contribution to each axis. The projection of a factor's vectors was
850 rescaled (right and upper blue scales) to facilitate their interpretation. The value of each individual

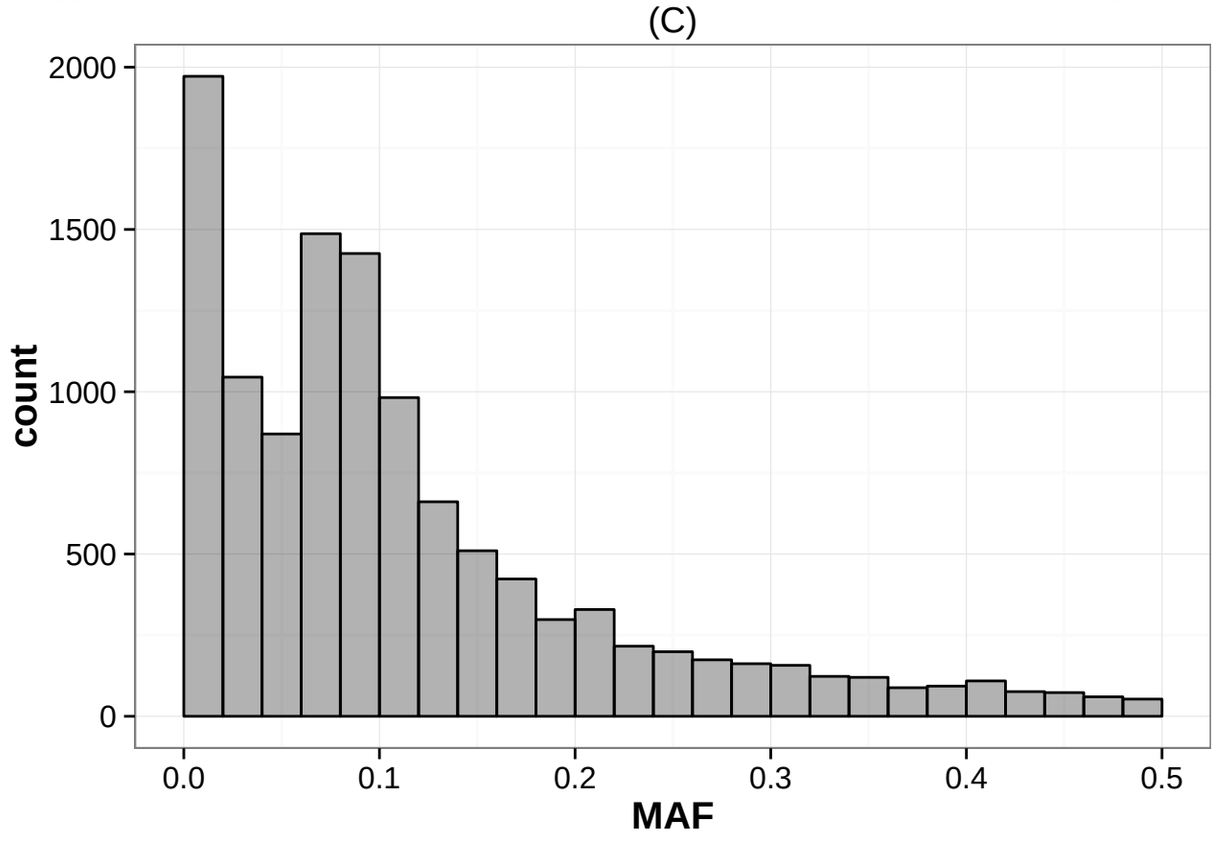
851 data point on any factor vector can be inferred by performing an orthogonal projection of that point
852 on any chosen vector (for example in Figure 4B, E-Muro and E-Pirio data points project on the same
853 space of the habitat vector, while D-Muro data points are shifted to the left for that vector. On the
854 latitude vector, E-muro data points are grouped mostly to the left, D-Muro values have intermediate
855 values, and E-Pirio data points are grouped to the right – this projection of points on the latitude
856 vector is concordant with their geographical positioning (Figure 2).

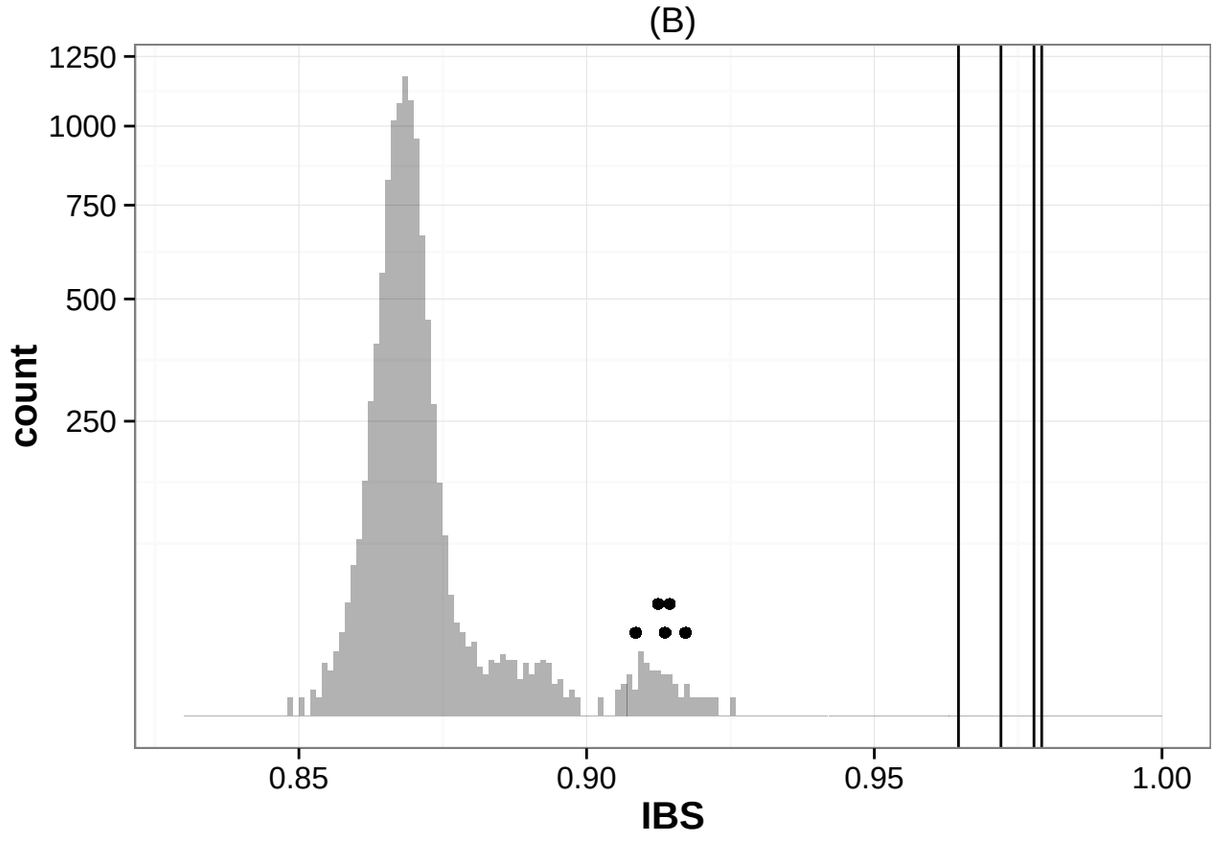

lay date (1=1st of March, triangles)
clutch size (circles), nr of fledglings (squares)
80
75
70
65
60
55
50
45
40
35
30
11
10
9
8
7
6
5
4
3
D-Rouviere
D-Muro
E-Muro
E-Pirio
Continent
Corsica
Corsica
Corsica

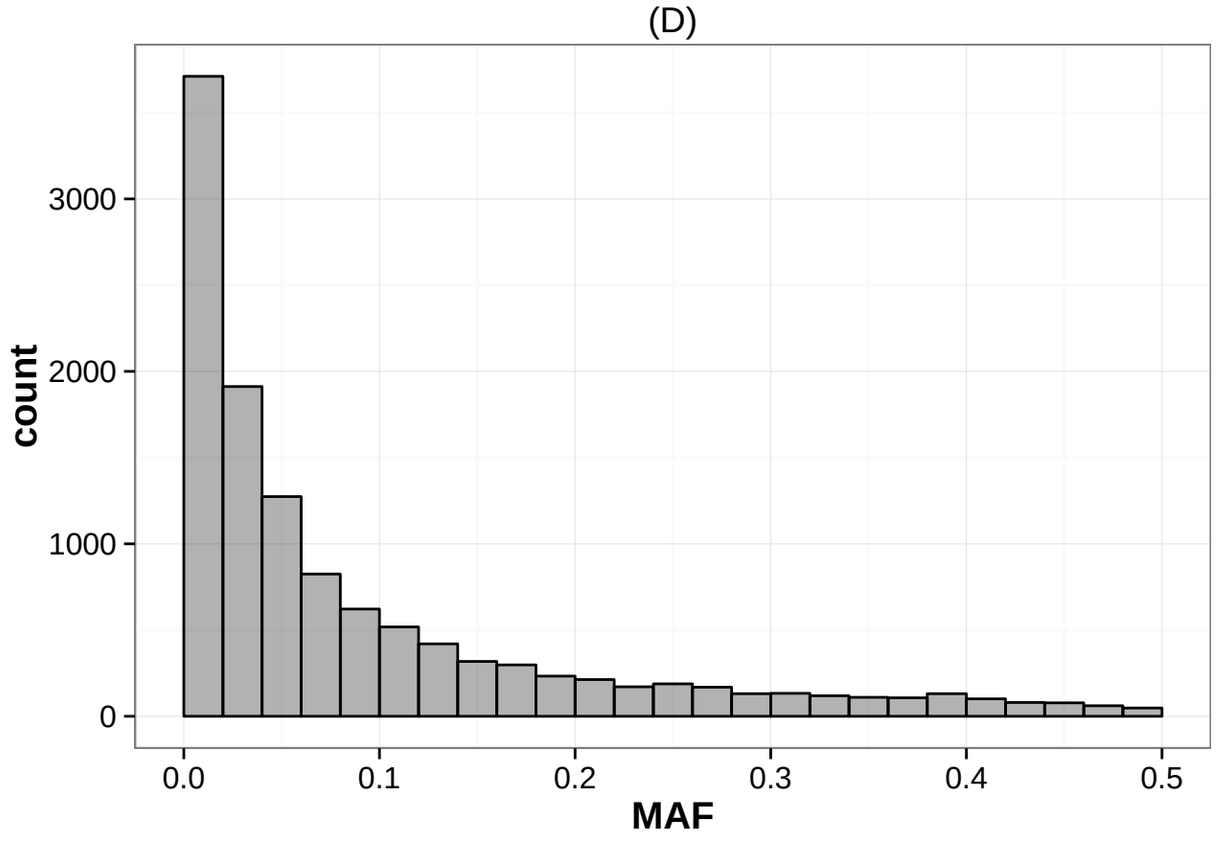

body mass (in g, circles)
tarsus length (in mm, squares)
11.5
11
10.5
10
9.5
9
17.2
17
16.8
16.6
16.4
16.2
16
15.8
15.6
15.4
15.2
D-Rouviere
D-Muro
E-Muro
E-Pirio
Continent
Corsica
Corsica
Corsica

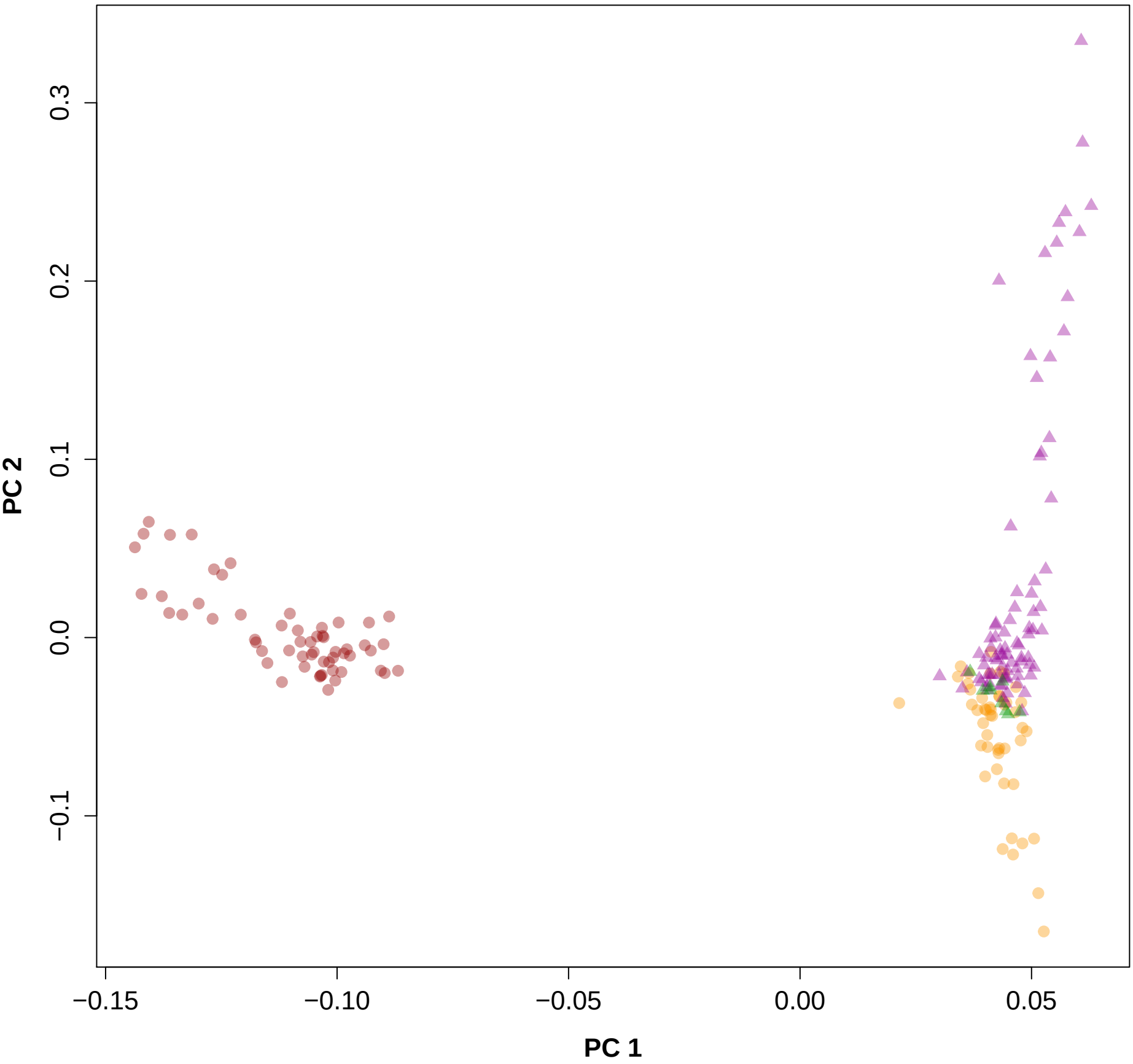

D-Rouvière
E-Muro
D-Muro
E-Pirio
0 50 100 150 200 250 km


(A)
count
100
50
0.80
0.85
0.90
0.95
1.00
IBS

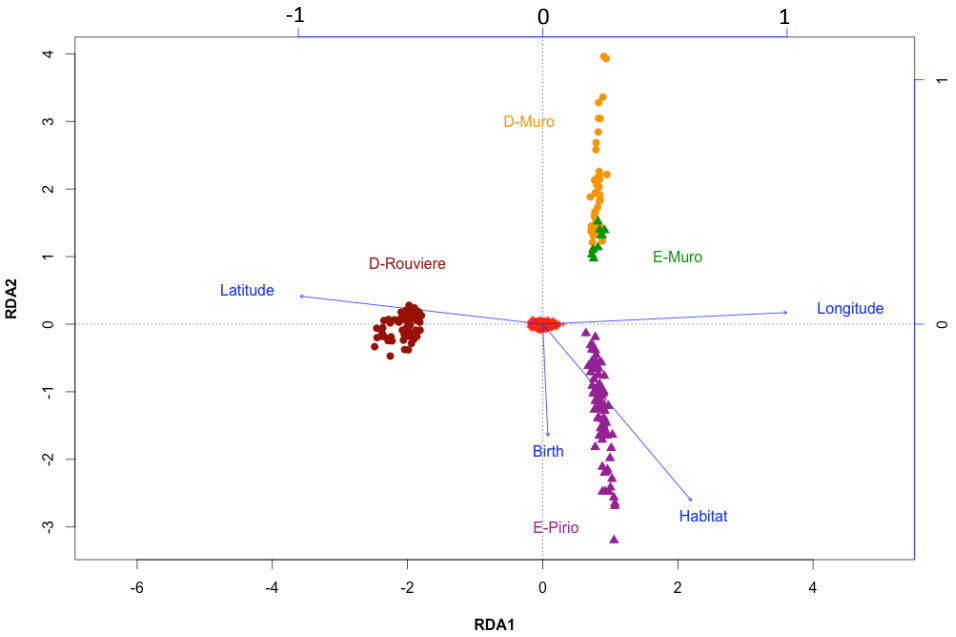

(C)
count
2000
1500
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
MAF


(B)
count
1250
1000
750
500
250
0.85
0.90
0.95
1.00
IBS

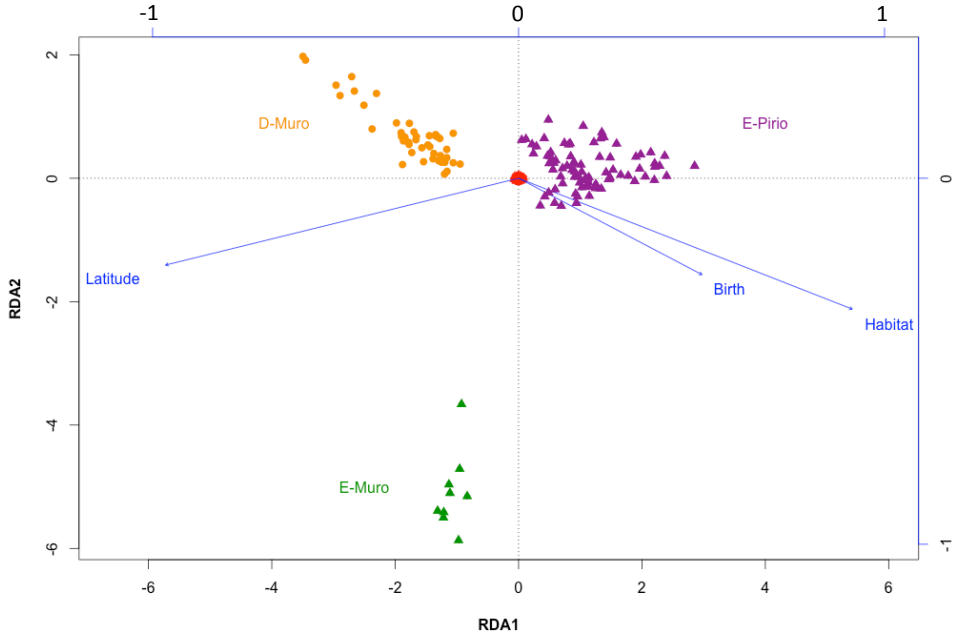

(D)
count
3000
2000
1000
0
0.0
0.1
0.2
0.3
0.4
0.5
MAF


PC 2
0.3
0.2
0.1
0.0
−0.1
−0.15
−0.10
−0.05
0.00
0.05
PC 1

(A)


D-Muro
E-Muro
D-Rouviere
Latitude
Longitude
Birth
Habitat
E-Pirio
RDA1
RDA2

(B)


D-Muro
E-Pirio
Latitude
Birth
Habitat
E-Muro
RDA1
RDA2