MOLECULAR ECOLOGY

Local inter-species introgression is the main cause of outlying levels of intra-specific differentiation in mussels

Journal:	*Molecular Ecology*
Manuscript ID:	MEC-15-0353
Manuscript Type:	Original Article
Date Submitted by the Author:	31-Mar-2015
Complete List of Authors:	Fraisse, Christelle; CNRS-Université de Montpellier 2, ; University of Cambridge, Genetics Belkhir, Khalid; CNRS-Université de Montpellier 2, Welch, John; University of Cambridge, Genetics Bierne, Nicolas; CNRS-Université de Montpellier 2,
Keywords:	Genome scan, Adaptation, Selection, Outlier loci, Target enrichment sequencing, BAC assembly

Original Articles

Local inter-species introgression is the main cause of outlying levels of intra-specific differentiation in mussels

Christelle Fraïsse[1,2,3], Khalid Belkhir[1], John J. Welch[3], Nicolas Bierne[1,2]

1. Université Montpellier 2, Institut des Sciences de l'Évolution, UMR 5554, Montpellier Cedex 05, France.

2. CNRS, Institut des Sciences de l'Évolution, UMR 5554, Station Méditerranéenne de l'Environnement Littoral, CNRS, Sète, France.

3. Department of Genetics, University of Cambridge, Downing Street, Cambridge, UK.

Keywords:

Genome scan, adaptation, selection, outlier loci, target enrichment sequencing, BAC assembly

Author for correspondence:

Christelle Fraïsse, e-mail: Christelle.Fraisse@univ-montp2.fr

Running title:

Local introgression in mussels.

Abstract

Structured populations, and replicated zones of contact between species, are an ideal opportunity to study regions of the genome with unusual levels of differentiation; and these can illuminate the genomic architecture of species isolation, and the spread of adaptive alleles across species ranges. Here, we investigated the effects of gene flow on divergence and adaptation in the *Mytilus* complex of species, including replicated parental populations in quite distant geographical locations. We used target enrichment sequencing of 1269 contigs of a few Kb each, including some genes of known function, to infer gene genealogies at a small chromosomal scale. We show that geography is an important determinant of the genome-wide patterns of introgression in *Mytilus*, and that gene flow between different species, with contiguous ranges, explained up to half of the intra-specific outliers. This suggests that local introgression is both widespread and tends to affect larger chromosomal regions than purely intraspecific processes. We argue that this situation might be common, and this implies that genome scans should always consider the possibility of introgression from sister species, unsampled differentiated backgrounds, or even extinct relatives, e.g. Neanderthals in humans. The hypothesis that reticulate evolution over long periods of time contributes widely to adaptation, and to the spatial and genomic reorganisation of genetic backgrounds, needs to be more widely considered in order to make better sense of genome scans.

Introduction

The literature now contains many genome-wide surveys of differentiation, in a wide variety of systems (e.g. *Anopheles,* Turner *et al.* 2005; *Ficedula,* Ellegren *et al.* 2012; *Heliconius,* Martin *et al.* 2013; *Helianthus,* Renaut *et al.* 2013; *Corvus,* Poelstra *et al.* 2014; *Timena,* Soria-Carrasco *et al.* 2014; *Gasterosteus,* Jones *et al.* 2012). One of the most striking and consistent results is the heterogeneity of differentiation across the genome, including highly differentiated regions, sometimes called "genomic islands of differentiation" (Turner *et al.* 2005).

Several theories have been proposed to explain this pattern. The most prominent involves speciation with gene flow, driven by local adaptation (Via & West 2008; Nosil *et al.* 2009), but simulation models suggest that genomic islands arise only in restricted biological conditions (Feder & Nosil 2010; Feder *et al.* 2012; Flaxman *et al.* 2013). Other theories invoke background selection and hitchhiking in closely related species (Noor & Bennett 2009; Roesti *et al.* 2012; Renaut *et al.* 2013; Yeaman 2013; Cruickshank & Hahn 2014), the sorting of shared ancestral variation (Nielsen & Wakeley 2001), or variable persistence after secondary contact of differences accumulated in allopatry (e.g. Fraïsse *et al.* 2014a). Distinguishing between these scenarios is difficult, because the origins of semipermeable genetic barriers to gene flow are intrinsically difficult to trace (Endler 1977; Barton & Hewitt 1985; Harrison 1986).

Interpreting of regions of increased differentiation is also linked to questions regarding the origins of adaptive genotypes (Bierne *et al.* 2013; Roesti *et al.* 2014; Welch & Jiggins 2014). Barrier loci will delay the introgression of neutral alleles in proportion to their linkage (Barton & Bengtsson 1986; Charlesworth *et al.* 1997), but universally advantageous alleles can usually cross species barriers without much delay (Piálek & Barton 1997). As such, hybridization could lead to adaptive introgression of complex co-adapted haplotypes, and candidates have been reported in plants (Arnold 2004) and animals (Hedrick 2013), including humans (Mendez *et al.* 2012, 2013; Huerta-Sanchez *et al.* 2014).

In all of these cases, inferences are strengthened by the fact that the adaptation was not species-wide, allowing the researchers to focus on local introgression, and the abnormal differentiation of single populations (e.g. Europeans *vs* Africans, or Tibetans *vs* Hans in the study of human introgression from extinct relatives, Huerta-Sanchez *et al.* 2014). Nevertheless, even in such cases it is usually necessary to complement genome scans with surveys of genetic differentiation along small-scale chro-

mosomal regions (e.g. in mussels, Bierne 2010), reconstructions of the historical divergence of alleles at candidate genes (e.g. in mice, Domingues *et al.* 2012) and ultimately experimental tests (e.g. rodent poison resistance in mice, Song *et al.* 2011).

Here, we performed target enrichment sequencing in the *Mytilus* complex of species, focussing on long anonymous regions obtained from BAC sequencing and cDNAs from databases and a transcriptome survey (Romiguier *et al.* 2014). The *Mytilus* complex includes three incompletely isolated species of marine mussels, *Mytilus edulis*, *Mytilus galloprovincialis* and *Mytilus trossulus*. These species have experienced a complex history of divergence with periods of gene exchange during the Quaternary (Roux *et al.* 2014). They lie along a gradient of genetic divergence: *M. edulis* and *M. galloprovincialis* diverged 2.5 million years ago (Roux *et al.* 2014), while their divergence from *M. trossulus* is estimated at 3.5 million years (Rawson & Hilbish 1995). Today, these species are in contact in several places in the northern hemisphere (Figure 1), and replicated parental populations are found in quite distant geographical locations.

We took advantage of the original spatial genetic structure of the *Mytilus* mussels to explore the consequences of local inter-species introgression on patterns of differentiation within and between species. In Europe, a mosaic hybrid zone between *M. edulis* and *M. galloprovincialis* extends from the Mediterranean Sea to the North Sea (Figure 1A, Bierne *et al.* 2003; Hilbish *et al.* 2012; Quesada *et al.* 1995b), enabling the investigation of varying levels of inter-species introgression along a geographical gradient. Equally important, natural replication of contact zones between *M. edulis* and *M. trossulus* (one in Europe, Figure 1B, and one in North America, Figure 1C, Riginos & Cunningham 2005) gave us the opportunity to study how the genomic architecture of species isolation varies in space. Finally, we gain further insights into the history of adaptation within species by studying regions of the genome with abnormal levels of differentiation. Careful analyses of genetic differentiation and gene genealogies at a small chromosomal scale revealed that local introgression contributes significantly to intra-specific outliers in mussels. Such outliers would have been misinterpreted had the analysis considered only a single species. As it is common to address the question of the origin of adaptations by scanning genomes, we argue that introgression from sister species or unsampled differentiated backgrounds should be considered whenever possible.

Materials and Methods

Sampling

The *Mytilus edulis* species complex comprises three species that hybridize at several places in the northern hemisphere. We sampled individuals from eleven locations from both sides of the Atlantic Ocean (Figure 1 and TableS1). Sampling took place outside hybrid zones in recognized patches of panmictic populations. To investigate how patterns of genetic differentiation evolved along a gradient of genetic divergence, we made use of the original genetic structure of the mussel complex of species. Within *M. galloprovincialis,* previous studies have reported a genetic break either side of the Almeria-Oran front between Atlantic and Mediterranean populations (Quesada *et al.* 1995a,b). Samples from two localities along the eastern Atlantic coast were obtained (the Iberian Coast and Brittany) as well as samples either side of the Sicilo-Tunisian strait in the Mediterranean Sea (in Sete, France, for the western basin and Crete for the eastern basin). The two closest species of the mussel complex, *M. edulis* and *M. galloprovincialis,* meet along the French coast in a well-studied mosaic hybrid zone characterized by three successive transition zones (Bierne *et al.* 2003; Hilbish *et al.* 2012). To investigate gradients of introgression between them, we sampled populations enclosed within the mosaic zone (the Bay of Biscay for *M. edulis* and Brittany for *M. galloprovincialis*) together with external populations (the North Sea for *M. edulis* and the Iberian Coast for *M. galloprovincialis*). At the extreme of the gradient of divergence, *M. edulis* and *M. trossulus* met on two independent occasions in the northern Atlantic (in Europe and in North America, Riginos & Cunningham 2005), giving us the opportunity to study the outcomes of replicated contacts. In Europe, the two species meet in a clinal hybrid zone at the entrance of the Baltic Sea (Väinölä & Hvilsom 1991). Individuals of *M. trossulus* were sampled at the bottom of the Baltic Sea, in the gulf of Finland. In North America, the zone of contact between the two species extends from Maine to Nova Scotia (Koehn *et al.* 1984). *M. trossulus* mussels were obtained from the Saint-Lawrence river in Canada. *M. edulis* mussels were sampled in Rhode Island (USA). In total, eight individuals per sample were examined, except for the American *M. edulis* sample which comprised 11 individuals (see TableS1), the *M. trossulus* sample from Saint-Lawrence river in which a *M. edulis* individual was found (see TableS3), and the *M. galloprovincialis* sample from the Iberian Coast for which two individuals had to be removed for technical reasons (low coverage, see below and TableS3). Genomic DNA was extracted from adults using the DNeasy Blood

and Tissue Kit (Qiagen) following the manufacturer's protocol.

BAC sequencing and assembly

BAC sequencing

A BAC library was constructed by Rx Biosciences (Rockville, MD, USA) from whole genomic DNA of three *M. edulis* individuals; and using pCC1BAC BamHI as vectors. BAC sequencing was performed in three different experiments. First, 32 clones were prepared for Roche 454 pyrosequencing. Indexed libraries were pooled in equimolar proportions and sequenced on a Roche (Branford, CO, USA) GS FLX instrument that generated single reads of 600 bp on average. The sequencing of eight libraries failed, and they were sequenced again on an Illumina MiSeq platform (San Diego, CA, USA) that produced paired-end reads of 250 bp. Because the BAC inserts proved to be shorter (25Kb on average) than expected (100 Kb) and because of assemby difficulties with the AT-rich mussel genome, we conducted a third sequencing experiment, in which a single pool of 192 anonymous clones was sequenced on a single lane of an Illumina HiSeq2000 instrument that generated paired-end reads of 101 bp. Reads were trimmed for index sequences and low-quality terminal bases. They were deposed in the NCBI Short Read Archive [[XXX]].

de novo BAC assembly

In the absence of a sequenced *Mytilus* genome at the time of the experiments, we conducted *de novo* assemblies of BAC sequences in two steps for each sequencing run. We first assembled reads into contigs with different programs depending on the sequencing technology. 454 reads were assembled with *Newbler v1.0.1* (Margulies *et al.* 2005), a *de novo* DNA assembler designed for pyrosequencing, with the following parameters: seed length of 6 bp, minimal read length of 40 bp, minimal overlap length of 40 bp, minimal overlap identity of 90%. We also used the program *MIRA v3.4* (Chevreux 2005) with 454 and accurate settings. MiSeq and HiSeq reads were assembled using *ABySS v1.2.1* (Simpson *et al.* 2009) and *SOAPdenovo v2.0.4* (Luo *et al.* 2012), two *de novo* short-reads assemblers, using a k-mer size of 61 bp and an insert size of 500 bp for the MiSeq paired-reads and of 200 bp for the HiSeq paired-reads. Secondly, contigs generated in the first step were assembled into longer fragments with the program *CAP3* (Huang 1999). The resulting genomic assemblies for each

sequencing run were compared with the *DC-MEGABLAST* algorithm. We retained the longest contig of each assembly as well as uniquely assembled contigs. DNA contaminants (BAC vectors, bacteria, etc.) were removed with *DC-MEGABLAST*. Contigs of length < 1 Kb were discarded, except for the HiSeq assembly in which we used a 5 Kb threshold. Finally, a comparison of the 454, MiSeq and HiSeq genomic assemblies was performed with *DC-MEGABLAST* to produce a final contig set of 378 sequences (average length 8.5 Kb, maximal length 26.4 Kb) without duplicates.

Target enrichment sequencing

Targets

We enriched genomic DNA for 3 Mb of target regions using a SureSelectXT Custom system (Agilent Technologies, Santa Clara, CA) comprising $\sim$ 55,000 RNA probes of 120 bp (Mamanova *et al.* 2010). First, probes were designed from our *M. edulis* BAC assembly totaling 2 Mb of filtered sequences. Second, we designed probes from cDNA contigs. We used a random panel of 338 *M. galloprovincialis* cDNA contigs of $\sim$ 1.6 Kb on average (0.5 Mb in total) previously generated for a transcriptome sequencing project (Romiguier *et al.* 2014). BAC and cDNA reference sequences were annotated with the genome annotation program *MAKER v2.31* (Cantarel *et al.* 2008) using *ab-initio* gene predictions as well as local alignements onto *Mytilus* GenBank collections. In addition, we designed an additional set of probes from publicly released expressed sequence tags (ESTs) databases by focusing on genes with functions of interest. We targeted 553 ESTs averaging $\sim$ 1 Kb in length (0.5 Mb in total): 262 immunity genes (from *Mytibase*, Venier *et al.* 2009 and another *Mytilus* repertoire of immune genes, Philipp *et al.* 2012); 133 genes involved in cytonuclear interactions (identified from *mitodrome*, D'Elia *et al.* 2006, and *mitores*, Catalano *et al.* 2006, in *Drosophila*); 30 reproduction-related genes, 20 habitat-related genes (because of the known association with wave action, proteins from the foot, the byssus filament and adhesive plaques are suspected to be involved with habitat specialization) and 6 nucleoporines (following Nolte *et al.* 2012) directly recovered from GenBank. To these genes of known function, we added a panel of control genes composed of 102 genes known to be single gene orthologous for phylogenomics analysis (*OrthoDB* in vertebrates (Kriventseva *et al.* 2008) and genes used in molluscan phylogenomics Kocot *et al.* 2011). We eliminated redundancy between the two coding gene sets (from RNA-seq and EST libraries) by a local alignement analysis

(*DC-MEGABLAST*). To maximize capture of unique sequences, we identified and masked repetitive and low-complexity regions with *WindowMasker* (Morgulis *et al.* 2006) using *Mytilus* GenBank collections and our own BAC sequences as references. We designed 120 bp probes (2X tilling) covering the final masked genomic data with *OligoTiler* ($http://tiling.gersteinlab.org$). Orphan probes were duplicated and low GC-content probes (< 10 *GC%*) were quadruplicated. The production of the probe library was performed by Agilent SureDesign services.

Capture and sequencing

Illumina paired-end sequencing libraries with insert sizes of 300 - 600 bp were prepared for each individual. The standard Illumina TruSeq DNA Sample Preparation ($http://support.illumina.com$) was followed, except that we used custom paired-end adaptors incorporating a unique 6-bp index as well as Indexed Blocking Reagent to perform pre-capture multiplexing (Kenny *et al.* 2010). A total of 88 libraries were produced (additional samples not discussed here were also included). Individuals of the same population were pooled in equimolar proportions prior to being subjected to TruSeq Custom Enrichment ($http://support.illumina.com$). Several enrichment protocols were attempted in pilot runs on Miseq and GA2X instruments in order to increase capture specificity (TableS3). We sequenced all libraries on a full flow-cell of HiSeq2000 producing 101 bp paired-end reads. Reads were trimmed for index sequences and low-quality terminal bases; low-quality reads were discarded. They were deposed in the NCBI Short Read Archive [[XXX]]. The capture step and pilot runs were subcontracted to the Plateforme Génome Transcriptome (CGFB, Université Bordeaux Segalen, France) and HiSeq sequencing was performed at the sequencing plateforme of the "Génomique et maladies métaboliques" laboratory (CNRS UMR 8199, Lille, France).

Genomic enrichment mapping to reference contigs and SNP calling

Read mapping

We conducted mapping with *bwa-mem v0.7.5a* (Li & Durbin 2009), a fast aligner that use the Burrows-Wheeler algorithm. Two challenges had to be overcome. First, because our reference genomic assembly consisted of sequences from the three closely-related *Mytilus* species, a trade-off between stringency and specificity had to be found. Second, because masking was necessarily incomplete given the

relatively little genomic information available in mussels, sequencing depth was highly variable along BAC sequences. Typically, depth in repeated regions was above 1000X whereas regions of interest were often < 50X. As a consequence, we evaluated the mapping performance of different settings in one individual of each species based on the proportion of mapped reads and sliding window analysis of depth on BAC sequences. Ultimately, *bwa-mem* was run with the following non-default parameters: a clipping penality of 3, a mismatch penality of 2, a gap open penality of 3, and a minimum seed length of 10. Mapping was independently performed for each sequencing run. We then generated final mapping for each individual by merging the sorted alignements (either GA2X or MiSeq, with HiSeq) using *SAMtools v0.1.19* (Li *et al.* 2009).

SNP and genotype calling

SNP calling was performed following successive steps to obtain a dataset of high-quality SNPs across the three species. We used the *mpileup* tool of *SAMtools* to pileup the merged alignements of each individual. We then called variant candidates from all individuals with *bcftools* (Li 2011) to produce an initial database of SNPs. This database was subsequently used for multisample variant calling, performed separately in each population and applying various quality filters with *VCFtools v0.1.12a* (Danecek *et al.* 2011). To reduce as much as possible bias in allele frequencies, we excluded calls from positions with an averaged depth across individuals below 10 reads. We also applied an upper depth limit to exclude unmasked repeated regions. As populations varied in terms of sequencing depth, we used a maximal value set at the 98.5th percentile of the depth distribution across all positions (see TableS3). To filter out paralogous regions, we excluded sites deviating from Hardy-Weinberg equilibrium (*p-value* < 0.05) using an exact test implemented in *VCFtools*.

Only variants that passed filters across all populations were retained for subsequent analysis. According to the algorithm implemented in *bcftools*, a variant was called when the posterior probability of non-reference allele counts was above 50%, assuming a standard allele frequency spectrum. For each variant detected, the maximum *a posteriori* genotype was assigned to each individual assuming Hardy-Weinberg proportions in genotype prior probabilities. Only genotype calls with a quality score above 10 were retained, otherwise missing data was applied. For genealogical analyses, we used a higher genotype quality threshold set at 30. Any position with more than 20% of missing data were discarded. For outlier gene analysis, contigs with fewer than 20 SNPs were discarded.

Data analysis

The final data set consisting of 1269 reference sequences and polymorphism data is available on *http://www.scbi.uma.es/mytilus/index.php*. Population genetic analyses were performed using custom scripts in R (R Core Team 2012).

Patterns of genetic differentiation

We initially explored the data using a Principal Component Analysis implemented in the R package *ade4* (Dray & Dufour 2007) based on genotypes and excluding any position with missing data and singletons. FST values (Weir & Cockerham 1984) were calculated using the R package *hierfstat* (Goudet 2005) for each SNP between pairs of populations. Data were analysed at the level of the contig to limit pseudoreplication of closely-linked SNPs. Because we were interested in genomic regions that were highly differentiated between populations, we calculated the 90th percentile of the FST distribution of each contig (FST_{90}), as well as its maximal FST value (FST_{max}). Joint distributions of interspecific FST_{90} values were analysed by Standardised Major Axis regressions, slope and elevation were estimated and tested using the R package *smatr* (Warton *et al.* 2012). For outlier identification, we fitted a null empirical FST_{90} and FST_{max} distributions across all contigs. Contigs in the upper 2.5% of the empirical FST_{90} and/or FST_{max} distributions were categorized as outliers. They were further analysed by estimating the allele frequency variation along the sequence. Open reading frames were predicted with *ORF finder* (*http://www.ncbi.nlm.nih.gov/projects/gorf*) and non-synonymous changes were identified using *BioEdit v7.2.5* (Hall 1999). We also evaluated the proportion of exclusively shared SNPs in a given set of populations, after removing singletons.

Phylogenomic analysis

Genotype data were phased with *BEAGLE v3.3.2* (Browning & Browning 2007) using genotype likelihoods provided by *bcftools*. All individuals were included in the analysis to maximize linkage disequilibrium and 20 haplotype pairs were sampled for each individual during each iteration of the phasing algorithm to increase accuracy. FASTA haplotype sequences were then generated using a custom perl script. A phylogenetic network analysis was conducted with *SplitsTree4 v4.12.6* (Huson & Bryant 2006) to get insight into the population relationships across the three hybridizing species. All hap-

lotype sequences were compiled to create an artificial chromosome of 51,878 variable positions and analysed using the neighbour-net method. Haplotype sequences of each candidate locus were also individually analysed with the neighbour-net method. Finally, we quantified the degree of exclusive ancestry between populations by computing a Genealogical Sorting Index (Cummings *et al.* 2008) for each locus based on allelic genealogies inferred with the R package *ape* (Paradis *et al.* 2004) using a neighbour-joining algorithm with F84 distances (Felsenstein & Churchill 1996).

Results

BAC assembly, target enrichment performance and variant calling

Screening and assembly results for each BAC clone are summarized in TableS2. The final BAC assembly contained 378 contigs, with an average length of 8.5 Kb, including 5% above 18 Kb (maximum 26.4 Kb), and a total assembly length of 3.2 Mb. Performance of DNA target enrichment sequencing and mapping are reported in TableS3. A total of 1269 contigs (378 BAC and 891 cDNA contigs) were captured in 75 individuals spanning the three species (see Material and Methods for details). Reads were aligned against a single reference made of sequences from the three species. On average, 55% of the reads mapped to the reference and all three species had a similar proportion of reads aligned. The performance of target enrichment mainly resulted from a combined effect of library quality and capture protocol (the second alternative protocol led to higher capture specificity). 85% of targeted sequences (1079) were successfully captured and assembled with mean read depth of 35X. After quality filtering (see Material and Methods for details), variant calling produced a total of 122,144 SNPs across all individuals. Among populations, the proportion of variant sites (13,827 on average) and observed heterozygosity ($h_o = 0.032$ on average) were roughly similar. However, *M. trossulus* populations tended to be more polymorphic ($h_{o\_trossulus} = 0.042$) than populations from the other species ($h_{o\_galloprovincialis} = 0.030$ and $h_{o\_edulis} = 0.029$), mainly due to private SNPs (Figure 2).

Genome-wide species relationships

A genome-wide network of genetic relationships (Figure 2A) was built from a subset of 51,878 high-quality SNPs genotyped in 72 individuals (3 individuals were discarded due to misidentification or sequencing failure; Table S3). We observed that each species formed a distinct cluster suggesting that

on average a high proportion of SNPs supports the "species tree" topology. Given their more recent divergence, *M. edulis* and *M. galloprovincialis* were less differentiated from each other than from *M. trossulus*. This is also apparent in the multivariate analysis of genotypes (Figure 2B) in which *M. trossulus* individuals were clearly separated from individuals of the other species in the first axis. The second axis differentiated *M. edulis* from *M. galloprovincialis* individuals. In the following axes (not shown), the American and European populations of *M. edulis* were separated as well as the Atlantic and Mediterranean populations in *M. galloprovincialis*; then the east and west Mediterranean Sea and finally the enclosed patch from the peripheral parental population in *M. galloprovincialis* (Brittany *vs* Iberian Coast) and *M. edulis* (Bay of Biscay *vs* North Sea). A last line of evidence comes from Figure 2C showing that the majority of exclusively shared SNPs stands between populations of the same species ($S_{trossulus} = 8245$, $S_{galloprovincialis} = 1673$ and $S_{edulis} = 417$) or between genetic clusters within species ($S_{edulisEU} = 555$ and $S_{edulisAM} = 358$; $S_{galloprovincialisATL} = 387$ and $S_{galloprovincialisMED} = 1011$). This is even clearer when considering the subset of shared and fixed SNPs (Table1) which were nearly all species-specific (i.e. polymorphic or fixed in one species but absent in others). We noted in Figure 2A that a *M. trossulus* individual from America ranked at the bottom of its cluster suggesting significant levels of recent introgression, which is consistent with the close geographical proximity of *M. edulis* populations there. This is further supported by Figure 2B in which this individual appears to group outside of its population, shifted toward the *M. edulis* clusters. Together with the *M. edulis* individual found in this population (excluded from analysis), this reveals that admixed groups coexist in the Saint-Lawrence river.

Genetic differentiation: geography

Alongside the "species tree" topologies, discrepant gene histories were also clearly identified. These may be due to shared ancestral polymorphism, or gene flow experienced by populations during their history. We characterized patterns of differentiation between populations at different levels of divergence ("intraspecific" *vs* "interspecific"), and geographical isolation ("parapatric" *vs* "allopatric") using the upper decile of the FST distribution for each contig (FST_{90}, Figure 3). The intraspecific FST_{90} distribution was L-shaped, with most loci undifferentiated between populations and a few loci highly differentiated. With increasing geographical isolation, the genome-wide average FST_{90} value increased from 0.076 to 0.131 in *M. edulis* and from 0.086 to 0.107 in *M. galloprovincialis* (Table1).

316 The two *M. trossulus* populations were by far the most differentiated, with an average FST_{90} value of
317 0.362 (also visible in Figure 2). The genome-wide variance in FST_{90} increases with increasing lev-
318 els of divergence and tends to become bimodal in interspecific comparisons, including both a higher
319 proportion of highly and lowly differentiated loci.

320 *M. edulis* and *M. galloprovincialis* populations range along a gradient of geographical distances
321 from pure allopatry with American *M. edulis* populations, to different degrees of parapatry in Eu-
322 rope (Figure 1). As expected based on geography, average FST_{90} values between *M. edulis* and *M.*
323 *galloprovincialis* were higher in allopatric populations ($FST_{90\ galloprovincialis-edulisAM} = 0.41$) than in
324 parapatric populations ($FST_{90\ galloprovincialisMED-edulisEU} = 0.39$ and $FST_{90\ galloprovincialisATL-edulisEU} =$
325 0.31). The joint distribution of interspecific FST_{90} in allopatric and parapatric populations further
326 confirmed the effects of gene flow in reducing differentiation (Standardised Major Axis regression,
327 $elevation = 0.05;\ p < 0.0001$, Figure 4A). Similarly, the joint distribution of interspecific FST_{90} in
328 the Atlantic (closer to *M. edulis*) and Mediterranean (further from *M. edulis*) populations confirmed the
329 genome-wide difference in introgression rates (SMA regression: $elevation = 0.03;\ p < 0.0001$, Fig-
330 ure 4B). However this also revealed that some outlier loci showed the opposite pattern of introgression
331 (i.e. highly differentiated in the Atlantic but lowly differentiated in the Mediterranean Sea), suggesting
332 a more complex history between *M. edulis* and *M. galloprovincialis*. Interspecific FST_{90} correlations
333 between Mediterranean populations showed no differences in introgression rates between the Eastern
334 and Western basins (non-significant elevation, Figure 4C); whereas the internal Atlantic population
335 was significantly more introgressed than its external counterpart ($elevation = 0.02;\ p = 0.001$, Figure
336 4D). Overall, this is consistent with the genome-wide gradient of increasing *M. edulis* introgression
337 observed in Figure 2, from the enclosed patch in Brittany to the Iberian Coasts, the West and the East
338 of the Mediterranean Sea. With regard to *M. edulis* in Europe, the joint distribution of interspecific
339 FST_{90} showed that the enclosed patch in the Bay of Biscay was not more introgressed than its external
340 reference (non-significant $elevation$, Figure 4E).

341 Genetic differentiation: replicated contacts

342 *M. edulis* and *M. trossulus* are currently in contact in Europe and in America (Figure 1). To evaluate the
343 degree of genetic parallelism between contacts, we assessed the level of interspecific FST_{90} correlation
344 between the European and American replicates (Figure 5A). Despite being lower than the correlation

in *M. edulis* and *M. galloprovincialis* comparisons (Figure 4), it was still significant ($r_{pearson} = 0.57$) showing that the outcomes were to some extent similar in the two contacts (highly/lowly differentiated regions partially overlap). More importantly, Figure 5A showed that genome-wide differentiation between European species was lower than their American counterparts ($elevation = 0.17;\ p < 0.0001$) reflecting genome-wide asymmetry in *M. edulis* introgression rates between the two species; the European *M. trossulus* population being more introgressed by *M. edulis* alleles at the genomic level. Outlier loci were also asymmetric with a deficit of loci both highly differentiated in Europe and lowly differentiated in America. Figure 5B further confirmed the asymmetry of introgression. It shows that loci characterized by a strong phylogenetic separation between American *M. trossulus* and other species (high GSI values) were less exclusive to *M. trossulus* in Europe ($elevation = 0.03;\ p < 0.01$). Also, this is clear across Figure 2 in which the European *M. trossulus* individuals were closer to *M. edulis* individuals; and from Figure 3 in which lower genome-wide differentiation was observed in comparisons involving the European *M. trossulus* population.

Within-species outliers: examples of the different categories

Making clear the evolutionary relationships between species of the *Mytilus* complex allowed us to identify the causes of outlying levels of differentiation within species and to reconstruct the evolutionary histories of outlier loci. Based on analyses of variation in differentiation at small chromosomal scales, as well as allele frequencies and allele genealogies (Figure 6), candidate outliers were placed in different categories depending on whether they were most plausibly due to differentiation of intraspecific alleles, or introgression of heterospecifc alleles. Figure 6A and 6B illustrate representative cases of candidate loci for local introgression (see Figure S1 for additional examples). Figure 6A represents a complete introgression of *M. edulis* haplotypes of several Kb in length, into the European *M. trossulus* population; while American *M. trossulus* remained unaffected by *M. edulis* introgression. In most cases, a major part of the heterospecific genetic diversity has introgressed into the recipient species. In a few cases however (3/83), a small proportion of the heterospecifc diversity has swept and introgressed between species. A clear example is a female-specific transcript for which a single edulis haplotype has introgressed into Atlantic populations of *M. galloprovincialis* (Figure 6B). The chromosomal footprint shows a peak of FST at the 3' side of the sequence; and several radical amino acid changes have occured in the protein. In addition, a distinct group of haplotypes are found in the

Mediterranean Sea and in the American *M. trossulus* populations. Together, these patterns are consistent with the hypothesis of adaptive introgression across the mosaic hybrid zone in Europe. Figure 6C illustrates an example of an outlying level of differentiation that did not involve introgression from another species. It is a classical sweep in the external Mediterranean population of *M. galloprovincialis*, characterized by a star-shape clade and a high frequency of a new variant at the maximal FST value (see Figure S1 for additional examples).

Within-species outliers: major causes

Table1 reports the total number of intraspecific outliers from within-species comparisons. The number of such outliers increased steadily with the average level of genomic differentiation, from ∼15 in parapatric comparisons, to 40 between the allopatric *M. trossulus* populations. Outlier tests were not possible between species because the distribution of FST values was overdispersed. At the contig level, some loci were repeatedly involved in outlying levels of differentiation ("shared outliers"; TableS4), but their number was not significantly different than would be expected by chance.

Overall, outlier analysis revealed that local introgression of heterospecific variants contributes to a significant part of the within-species differentiation (Table1 and TableS5 for details). In *M. galloprovincialis*, introgression of *M. edulis* variants explains between 55% and 80% of the outliers. Both basins of the Mediterranean Sea (East and West) presented highly introgressed variants; whereas outlying levels of introgression were much more asymmetric between the two Atlantic populations, corroborating the genome-wide trend. Between European and American *M. edulis* populations, more than half of the outliers were due to local introgression. While heterospecific variants came from *M. trossulus* in America, the European populations were introgressed by *M. galloprovincialis* variants. Notably, outlying levels of differentiation between the European *M. edulis* populations were never due to introgression, in agreement with the genome-wide picture. Moreover the number of outliers was small, their level of differentiation low and restricted to a small region of the contigs. In *M. trossulus*, local introgression from *M. edulis* explains more than 90% of the outliers. As is the case genome-wide, outlying levels of introgression were highly asymmetric: European populations being more permeable to *M. edulis* introgression than were their American counterparts.

Remarkably, the different sources of within-species differentiation often led to subtle chromosomal footprints, which would be difficult to identify in a large-scale survey of genome-wide differen-

tiation. For example, introgression candidates generally involved several haplotypes from the donor species leading to a "soft sweep" patterns (Figure 6 and Figure S1). Similarly, variants that swept to high frequency in a single population often segregated as ancestral polymorphism in other populations, instead of arising as new mutations (TableS5). The chromosomal scale of the footprint sometimes appeared to extend beyond the scale investigated here (a few Kb). This was particularly the case for local introgression candidates, because heterospecific haplotypes are initially introgressed from one species to another. Nevertheless, in some cases, a single peak of a few hundred base pairs was detected (Figure 6 and Figure S1) allowing the identification of candidate causal variants (TableS5). When direct selection was suspected (i.e. when non-synonymous variants were identified in open reading frames), amino acid changes were rarely ever differentially fixed between populations, suggesting a multigenic architecture for the traits implicated in differentiation. We also investigated whether any functional category was overrepresented among the list of EST outliers (TableS6). Across comparisons, there was no category represented more often that would be expected by chance. However, when considering EST outliers differentially fixed between species ("diagnostic loci"), the most likely candidates to represent direct barrier to gene flow, we noted a slight excess of immune genes together with a slight deficit of single copy orthologous genes.

Discussion

Early empirical studies of the build up of reproductive isolation, emphasized the role of gene flow in shaping genome-wide patterns of differentiation. But depite this, the evolutionary importance of gene flow for adaptation and speciation remains unclear (Hedrick 2013; Seehausen *et al.* 2014). This is partly because variation in levels of genetic differentiation are also influenced by variation in local genomic parameters (such as mutation and recombination rates), which determine the speed of divergence tuned by selective interference (including background selection, hitchhiking, and Hill-Robertson effects), which implies that candidate regions may not represent interspecific barriers to gene flow (Cruickshank & Hahn 2014). Indeed, genome scans in regions of low recombination struggle to disentangle the relative contribution of reduced gene flow and selection at linked sites in producing high level of differentiation (Nachman & Payseur 2012). Moreover, even if gene flow and hybridization occur during speciation, they are not expected to greatly impede divergence (Endler

1977). In contrast, the acquisition of beneficial variants by migration from another population or species (i.e. adaptive introgression) may be crucial in adaptation (Arnold 2004; Hedrick 2013). This shows the importance of considering the history of related lineages when asking how organisms adapt to their environment.

Here, we used targeted enrichment sequencing to perform fine-scale empirical genome scans of differentiation between closely related species of mussels in different geographical contexts. Given that detection of loci influenced by divergent selection using genome-scans carries a significant risk of false positives (extreme values of FST might not necessarily be due to selection) and that multifarious processes can lead to FST outliers, we further investigated candidate outliers by analyzing allele frequencies and gene genealogies along contigs. A phylogenetic network (Figure 2) suggested discordant genealogical histories across the genome, i.e. genome regions that do not follow the "species-tree" topology. Incomplete lineage sorting, in addition to introgression, can lead to interspecies shared polymorphism. This is especially the case for recently-diverged lineages with large effective population sizes, such as *Mytilus* mussels. We were able to control for this, by using an allopatric population (e.g. *M. edulis* in America) as a reference for the level of shared ancestral polymorphism, and thereby showed that interspecific differentiation between parapatric populations was lower than between their allopatric counterparts (Figure 3 and Figure 4). This result matches other studies highlighting the role of gene flow in eroding genetic differentiation in neutral regions (e.g. in *Heliconius,* Martin *et al.* 2013 and in *Helianthus,* Renaut *et al.* 2013). If selection against hybrids/immigrants is sufficiently strong compared to the migration rate, genomic regions involved in the interspecific genetic barrier are not expected to introgress between species. As such, outlier loci, highly differentiated in both allopatric and parapatric comparisons, are good candidates for reproductive isolation genes (Figure 4).

In Europe, a genome-wide gradient of introgression, in agreement with current range distributions, was observed in *M. galloprovincialis* populations from the Mediterranean Sea to the internal Atlantic population (Figure 2). The high level of introgression detected in the internal Atlantic population (Figure 4) is well explained by its enclosed position within the mosaic hybrid zone, and concords with a previous study using few markers (Bierne *et al.* 2003). While this pattern is consistent with ongoing neutral introgression between geographically proximate populations, it may also be explained by migration of heterospecific variant favoured in the internal *M. galloprovincialis* population due to similarities in selective pressures. Such candidates should be weakly differentiated from *M. edulis*

in the internal population, but highly differentiated elsewhere (Figure 4). Another line of evidence supporting cases of non-neutral introgression is that some of the outliers of differentiation within *M. galloprovincialis* retained the signature of ancient introgression into the Mediterranean Sea (Figure 4 and Table1). In addition to candidates previously reported (Gosset & Bierne 2012; Fraïsse *et al.* 2014b), many new candidates for adaptive introgression have been found in this study (TableS5). However, demonstrating that introgression was initially selectively-driven is tricky, because the pattern left by adaptation from recurrent migration may be quite different from the classical hitchhiking signature. Specifically, if adaptive introgression involves several haplotypes of the same beneficial mutation, most of the heterospecific diversity is expected to introgress into the foreign genetic background leading to a "soft sweep" pattern (Pennings & Hermisson 2006). Such a pattern was often found among within-species outliers when local introgression was observed (Figure S1). This emphasizes a neglected hypothesis in which genetic hitchhiking is not involved: a local change in the permeability of a barrier to gene flow (Gagnaire *et al.* 2013; Fraïsse *et al.* 2014b), so that in some genomic regions, some populations are more permeable to introgression than others.

In contrast to other comparisons, the two European *M. edulis* populations shared similar level of differentiation with *M. galloprovincialis* even though the internal population is enclosed within the mosaic hybrid zone (Figure 2 and Figure 4). This suggests that the genome of *M. edulis* resists current introgression from *M. galloprovincialis* and that the two European populations share a similar genetic architecture with respect to interspecifc gene flow. Accordingly, all cases of outlying differentiation involved homospecific processes (Table1). For example, a candidate "global sweep", identified previously (Faure *et al.* 2008) was characterized by fixed haplotypes in the external population and non-fixed haplotypes in the internal population (Figure S1-B4). In fact two domes of differentiation are expected on either side of the benefical mutation because recombination breaks down its association with the haplotype initially entering the internal population (Bierne 2010). Unfortunately, we failed to assemble the targeted BAC to the region directly under selection and so we cannot confirm this pattern. Other outliers showed a classical signature of a selective sweep, with a sharp single peak of FST between the European and American *M. edulis* (Figure S1-B2). Even when non-synonymous changes were identified - suggesting direct selection, variants were generally not differentially fixed between them (TableS5). This may reveal pervasive polygenic adaptation which implies slight allele frequency differences at a large number of loci underlying the selected traits (Pritchard & Di Rienzo

2010). Also swept variants often segregated as low-frequency polymorphisms in other populations (TableS5), suggesting that selection from standing genetic variation may be widespread, although it leaves a less pronouced signature at linked sites than does a sweep from a single new mutation (e.g. in mice, Domingues *et al.* 2012; in sticklebacks, Hohenlohe *et al.* 2010; Roesti *et al.* 2014; and in mussels, Gosset *et al.* 2014).

While most of the genome is resistant to current introgression in European *M. edulis*, identification of within-species outliers between American and European *M. edulis* populations showed a putative case of adaptive introgression in Europe. A gene expressed solely in females (H5, Anantharaman & Craft 2012) has swept through *M. edulis* and *M. galloprovincialis* across the hybrid zone, while being structured between American *M. edulis* and Mediterranean *M. galloprovincialis* (Figure 6B). In addition, the different haplotypes carry several non-synonymous changes suggesting that the gene is rapidly evolving (TableS5). Similar patterns have been documented in *Mytilus* mussels in a gene expressed in male gametes (M7 lysin) which dissolves the egg vitelline coat (Riginos *et al.* 2006; Springer & Crespi 2007). This surprising observation may suggest that a locus *a priori* implicated in reproductive isolation may have swamped the species barrier between specific genetic backgrounds while remaining impermeable between others. Generally speaking, recent reviews of the molecular basis of species formation have not shown any specific biochemical pathway preferentially involved in incipient reproductive isolation (Presgraves 2010). Instead the main feature of isolating genes is their rapid evolution, often attributable to evolutionary conflicts (cyto-nuclear interactions, host-pathogen interactions or sexual conflicts). For example, Burton & Barreto (2012) emphasize the role of the mitochondrial genome in producing incompatible interactions with nuclear genes. Immune genes are also known to evolve faster than others (e.g. in humans, Fumagalli *et al.* 2011 and in flies, Obbard *et al.* 2009) and so they may be implicated in species barriers. However, our functional comparison of diagnostic ESTs between species did not reveal any obvious enrichment for these genes, although immune genes were slightly more common than expected by chance (TableS6).

The European and American contacts between *M. edulis* and *M. trossulus* (Figure 1) allowed us to investigate the extent to which differentiation is parallel in replicated contacts. Interestingly, we found incomplete parallelism in the genome-wide patterns of differentiation (Figure 5). The independent history of divergence between the two pairs is likely to be responsible for this pattern. Specifically, the genetic differences accumulated in allopatry may either scatter, or couple with each other when the

divergent lineages come into contact in locations stabilized by a natural barrier to dispersal (Barton & Hewitt 1985) or environmental boundaries (Bierne *et al.* 2011). This leads to a genomically localized breakdown or strengthening of the barrier to gene flow (Barton & de Cara 2009). Therefore, independent outcomes of the coupling process following secondary contact between the same two lineages (*M. edulis* and *M. trossulus*) is expected to produce partially different genetic architectures of reproductive isolation. Supporting this hypothesis, changes in the permeability of the species barrier have also been reported in replicated pairs of whitefish species (Gagnaire *et al.* 2013). However, we observed in mussels a striking pattern of genome-wide asymmetry in introgression level between the European and American contacts. More importantly, outlying levels of differentiation were also highly asymmetric; almost all cases of local introgression from *M. edulis* were found in the European *M. trossulus* lineage (Table1). Our study consolidates previous similar findings with nuclear DNA markers (Borsa *et al.* 1999) and mitochondrial genomes, which have been completely replaced in the Baltic Sea (Rawson & Hilbish 1998; Quesada *et al.* 1999). This asymmetry is not an expected outcome of the coupling process, because independent incompatibilities can couple in either direction (Barton & de Cara 2009) leading to overall symmetric reproductive isolation. Riginos & Cunningham (2005) and Väinölä & Strelkov (2011) have proposed that the secondary contact in Europe may be older than that in America; so that an extended period of gene flow could have led to the erosion of isolating barriers still functioning in America. To assess whether or not this hypothesis is valid will require a rigorous inference of the demographic history of the two lineages. Asymmetrical introgression may also simply reflect differences in effective population size or in the environmental landscape in Europe, which influence the direction of gene flow and the level of genetic drift. It worth emphasising that if the *M. trossulus* ancestry of Baltic mussels is still clearly evident, a more permeable barrier would have easily resulted in a more general swamping of the Baltic genome by the North Sea genome. In this case, the reconstructed history of populations from the neutral fraction of the genome could easily result in the mistaken inference that differentiation was recent (Bierne *et al.* 2013). This could have happened in other species exhibiting a genetic break at the entrance of the Baltic Sea (Bierne *et al.* 2011) and explained the astonishing number of outliers sometimes found between Baltic and North Sea populations (Lamichhaney *et al.* 2012).

In conclusion, this study has revealed the complexity of the relationships among populations of *Mytilus* species which are both influenced by current and past demogeography. Based on qualita-

tive analysis of genetic polymorphism, we highlighted the importance of introgression as a neglected source of adaptive variation to be considered in genome scans.

Acknowledgments

Analyses largely benefited from the Montpellier Bioinformatics Biodiversity computing cluster platform. It is a pleasure to thank Rémy Dernart for his support on the computing platform; Christophe Hubert and Véronique Dhennin for their technical help in producing data; Baptiste Faure and Pierre-Alexandre Gagnaire for sampling. We also thank Nicolas Galtier, Pierre-Alexandre Gagnaire, Patrice David and François Bonhomme for discussions. Molecular data were produced through the ISEM platform Génomique marine at the Station Méditerranenne de l'Environnement Littoral (OSU OREME (Observatoire de Recherche Méditerranéen de l'Environnement)) and the platform Génomique Environnementale of the LabEx CeMEB (Laboratoire d'Excellence Centre Méditerranéen de l'Environnement et de la Biodiversité). This work was funded by the Agence Nationale de la Recherche (HYSEA project, ANR-12-BSV7- 0011) and the project Aquagenet (SUDOE, INTERREG IV B). This is article 2015-XXX of Institut des Sciences de l'Evolution de Montpellier.

References

- Anantharaman S, Craft JA. (2012) Annual variation in the levels of transcripts of sex-specific genes in the mantle of the common mussel, *Mytilus edulis*. *PLoS One,* **7**.
- Arnold ML. (2004) Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? *Plant Cell*, **16**, 562–570.
- Barton N, Bengtsson BO. (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57**, 357–376.
- Barton NH, de Cara MAR. (2009) The evolution of strong reproductive isolation. *Evolution,* **63**, 1171–1190.
- Barton NH, Hewitt GM. (1985) Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.,* **16**, 113–148.
- Bierne N. (2010) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution,* **64**, 3254–3272.
- Bierne N, Borsa P, Daguin C, Jollivet D, Viard F, Bonhomme F, David P. (2003) Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. *Mol. Ecol.,* **12**, 447–461.
- Bierne N, Gagnaire PA, David P. (2013) The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr. Zool.,* **59**, 72–86.
- Bierne N, Welch J, Loire E, Bonhomme F, David P. (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.,* **20**, 2044–2072.
- Borsa P, Daguin C, Ramos Caetano S, Bonhomme F. (1999) Nuclear-DNA evidence that northeastern atlantic M*ytilus trossulus* mussels carry *M. edulis* genes. *J. Mollus. Stud.,* **65**, 504–507.
- Browning SR, Browning BL. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.,* **81**, 1084–1097.
- Burton RS, Barreto FS. (2012) A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? *Mol. Ecol.*, **21**, 4942–4957.
- Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B *et al.* (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.,* **18**, 188–196.
- Catalano D, Licciulli F, Turi A, Grillo G, Saccone C, D'Elia D. (2006) MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. *BMC Bioinformatics*, **7**, 36.
- Charlesworth B, Nordborg M, Charlesworth D. (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.,* **70**, 155–174.

- Chevreux B. (2005) MIRA: An Automated Genome and EST Assembler. Duisburg: The Ruprecht-Karls-University.
- Cruickshank TE, Hahn MW. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.*, **23**, 3133–3157.
- Cummings MP, Neel MC, Shaw KL. (2008) A genealogical approach to quantifying lineage divergence. *Evolution,* **62**, 2411–2422.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA *et al.* (2011) The variant call format and VCFtools. *Bioinformatics,* **27**, 2156–2158.
- Domingues VS, Poh YP, Peterson BK, Pennings PS, Jensen JD, Hoekstra HE. (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution,* **66**, 3209–3223.
- Dray S., Dufour A. (2007) The ade4 package: implementing the duality diagram for ecologists. *J. stat. softw*., **22**, 1–20.
- D'Elia D, Catalano D, Licciulli F, Turi A, Tripoli G, Porcelli D *et al.* (2006) The MitoDrome database annotates and compares the OXPHOS nuclear genes of *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. *Mitochondrion,* **6**, 252–257.
- Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature,* **491**, 756–760.
- Endler JA. (1977) Geographic Variation, Speciation and Clines. New Jersey: Princeton University Press.
- Faure MF, David P, Bonhomme F, Bierne N. (2008) Genetic hitchhiking in a subdivided population of M*ytilus edulis*. *BMC Evol. Biol.,* **8**, 164.
- Feder JL, Gejji R, Yeaman S, Nosil P. (2012) Establishment of new mutations under divergence and genome hitchhiking. *Phil. Trans. R. Soc. B,* **367**, 461–474.
- Feder JL, Nosil P. (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.
- Felsenstein J, Churchill GA. (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93–104.
- Flaxman SM, Feder JL, Nosil P. (2013) Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution,* **67**, 2577–2591.
- Fraïsse C, Elderfield JD, Welch JJ. (2014a) The genetics of speciation: are complex incompatibilities easier to evolve? *J. Evol. Biol.,* **27**, 688–699.
- Fraïsse C, Roux C, Welch JJ, Bierne N. (2014b) gene flow in a mosaic hybrid zone: Is local introgression adaptive? *Genetics,* **197**, 939–951.
- Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.,* **7**, e1002355.

- Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. (2013) The genetic architecture of reproductive isolation during speciation-with-gene flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution,* **67**, 2483–2497.
- Gosset CC, Bierne N. (2012) Differential introgression from a sister species explains high FST outlier loci within a mussel species. *J. Evol. Biol.,* **26**, 14–26.
- Gosset CC, Do Nascimento J, Augé MT, Bierne N. (2014) Evidence for adaptation from standing genetic variation on an antimicrobial peptide gene in the mussel *Mytilus edulis*. *Mol. Ecol.,* **23**, 3000–3012.
- Goudet J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes,* **5**, 184–186.
- Hall T. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.,* **41**, 95–98.
- Harrison RG. (1986) Pattern and process in a narrow hybrid zone. *Heredity,* **56**, 337–349.
- Hedrick PW. (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.,* **22**, 4606–4618.
- Hilbish TJ, Lima FP, Brannock PM, Fly EK, Rognstad RL, Wethey DS. (2012) Change and stasis in marine hybrid zones in response to climate warming. *J. Biogeogr.,* **39**, 676–687.
- Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.,* **6**, e1000862.
- Huang X. (1999) CAP3: A DNA sequence assembly program. *Genome Res*., **9**, 868–877.
- Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature,* **512**, 194–197.
- Huson DH, Bryant D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.,* **23**, 254–267.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature,* **484**, 55–61.
- Kenny EM, Cormican P, Gilks WP, Gates AS, O'Dushlaine CT, Pinto C *et al.* (2010) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res.* (pp. dsq029).
- Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A *et al.* (2011) Phylogenomics reveals deep molluscan relationships. *Nature,* **477**, 452–456.
- Koehn RK, Hall JG, Innes DJ, Zera AJ. (1984) Genetic differentiation of *Mytilus edulis* in eastern North America. *Mar. Biol.,* **79**, 117–126.
- Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. (2008) OrthoDB: the hierarchical catalog of Eukaryotic orthologs. *Nucleic Acids Res.,* **36**, D271–D275.

- Lamichhaney S, Martinez Barrio A, Rafati N, Sundström G, Rubin C-J, Gilbert ER *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19345–19350.
- Li H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* **27**, 2987–2993.
- Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* **25**, 1754–1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* **25**, 2078–2079.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* **1**, 18.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* **437**, 376–380.
- Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.,* **23**, 1817-1828.
- Mendez FL, Watkins JC, Hammer MF. (2012) A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.*, **91**, 265–274.
- Mendez FL, Watkins JC, Hammer MF. (2013) Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Mol. Biol. Evol.,* **30**, 798–801.
- Morgulis A, Gertz EM, Schäffer AA, Agarwala R. (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics,* **22**, 134–141.
- Nachman MW, Payseur BA. (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Phil. Trans. R. Soc. B,* **367**, 409–421.
- Nielsen R, Wakeley J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
- Nolte V, Pandey RV, Kofler R, Schlötterer C. (2012) Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Res.,* **23**, 99–110.
- Noor MAF, Bennett SM. (2009) Islands of speciation or mirages in the desert? examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.
- Nosil P, Funk DJ, Ortiz-Barrientos D. (2009) Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.,* **18**, 375–402.

- Obbard DJ, Welch JJ, Kim KW, Jiggins FM. (2009) Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet.*, **5**, e1000698.
- Paradis E, Claude J, Strimmer K. (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics,* **20**, 289–290.
- Pennings PS, Hermisson J. (2006) Soft sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.,* **23**, 1076–1084.
- Philipp EER, Kraemer L, Melzner F, Poustka AJ, Thieme S, Findeisen U *et al.* (2012) Massively parallel RNA sequencing identifies a complex immune gene repertoire in the Lophotrochozoan *Mytilus edulis*. *PLoS ONE,* **7**, e33091.
- Piálek J, Barton NH. (1997) The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes. *Genetics,* **145**, 493–504.
- Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1410–1414.
- Presgraves DC. (2010) The molecular evolutionary basis of species formation. *Nat. Rev. Genet.,* **11**, 175–180.
- Pritchard JK, Di Rienzo A. (2010) Adaptation - not by sweeps alone. *Nat. Rev. Genet.,* **11**, 665–667.
- Quesada H, Beynon CM, Skibinski DO. (1995a) A mitochondrial DNA discontinuity in the mussel *Mytilus galloprovincialis* Lmk: pleistocene vicariance biogeography and secondary intergradation. *Mol. Biol. Evol.*, **12**, 521–524.
- Quesada H, Zapata C, Alvarez G. (1995b) A multilocus allozyme discontinuity in the mussel *Mytilus galloprovincialis*: The interaction of ecological and life-history factors. *Mar. Ecol. Prog. Ser.,* **116**, 99–115.
- Quesada H, Wenne R, Skibinski DO. (1999) Interspecies transfer of female mitochondrial DNA is coupled with role-reversals and departure from neutrality in the mussel *Mytilus trossulus*. *Mol. Biol. Evol.*, **16**, 655–665.
- Rawson PD, Hilbish TJ. (1995) Distribution of male and female mtDNA lineages in populations of blue mussels, *Mytilus trossulus* and *M. galloprovincialis*, along the Pacific coast of North America. *Mar. Biol.*, **124**, 245–250.
- Rawson PD, Hilbish TJ. (1998) Asymmetric introgression of mitochondrial DNA among European populations of blue mussels (*Mytilus spp.*). *Evolution,* **52**, 100–108.
- Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.*, **4**, 1827.
- Riginos C, Cunningham CW. (2005) INVITED REVIEW: Local adaptation and species segregation in two mussel (*Mytilus edulis* x *Mytilus trossulus*) hybrid zones. *Mol. Ecol.,* **14**, 381–400.
- Riginos C, Wang D, Abrams AJ. (2006) Geographic variation and positive selection on M7 lysin, an acrosomal sperm protein in mussels (*Mytilus spp.*). *Mol. Biol. Evol.* **23**, 1952–1965.

- Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D. (2014) The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.,* **23**, 3944–3956.
- Roesti M, Hendry AP, Salzburger W, Berner D. (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.,* **21**, 2852–62.
- Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature,* **515**, 261–263.
- Roux C, Fraïsse C, Castric V, Vekemans X, Pogson GH, Bierne N. (2014) Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? a case study in a *Mytilus* hybrid zone. *J. Evol. Biol.,* **27**, 1662–75.
- Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA *et al.* (2014) Genomics and the origin of species. *Nat Rev Genet,* **15**, 176–192.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.,* **19**, 1117–1123.
- Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH *et al.* (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr. Biol.,* **21**, 1296–1301.
- Soria-Carrasco V, Gompert Z, Comeault A., Farkas TE, Parchman TL, Johnston JS *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science,* **344**, 738–742.
- Springer SA, Crespi BJ. (2007) Adaptive gamete-recognition divergence in a hybridizing *Mytilus* population. *Evolution,* **61**, 772–783.
- Turner TL, Hahn MW, Nuzhdin SV. (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.,* **3**, e285.
- Venier P, De Pittà C, Bernante F, Varotto L, De Nardi B, Bovo G *et al.* (2009) MytiBase: a knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences. *BMC Genomics,* **10**, 72.
- Via S, West J. (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.,* **17**, 4334–4345.
- Väinölä R, Hvilsom MM. (1991) Genetic divergence and a hybrid zone between Baltic and North Sea *Mytilus* populations (Mytilidae: Mollusca). *Biol. J. Linn. Soc.,* **43**, 127–148.
- Väinölä R, Strelkov P. (2011) *Mytilus trossulus* in Northern Europe. *Mar. Biol*., **158**, 817–833.
- Warton DI, Duursma RA, Falster DS, Taskinen S. (2012) smatr 3– an R package for estimation and inference about allometric lines. *Methods Ecol. Evol.,* **3**, 257–259.
- Weir BS, Cockerham CC. (1984) Estimating F-statistics for the analysis of population structure. *Evolution,* **38**, 1358–1370.

- Welch JJ, Jiggins CD. (2014) Standing and flowing: the complex origins of adaptive variation. *Mol. Ecol.,* **23**, 3935–3937.
- Yeaman S. (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U.S.A.,* **110**, E1743–E1751.

Data Accessibility

Sequencing data were deposed in NCBI Short Read Archive, accession numbers [[XXX]] for BAC sequences and [[XXX]]] for target-enrichment sequences.

Author Contributions

C.F. and N. B. designed the research. C.F. and K.B. performed the *in silico* work. C.F., J.W. and N.B. analysed the data. C.F. wrote the article, N.B., J.W. and K.B. revised and commented the article.

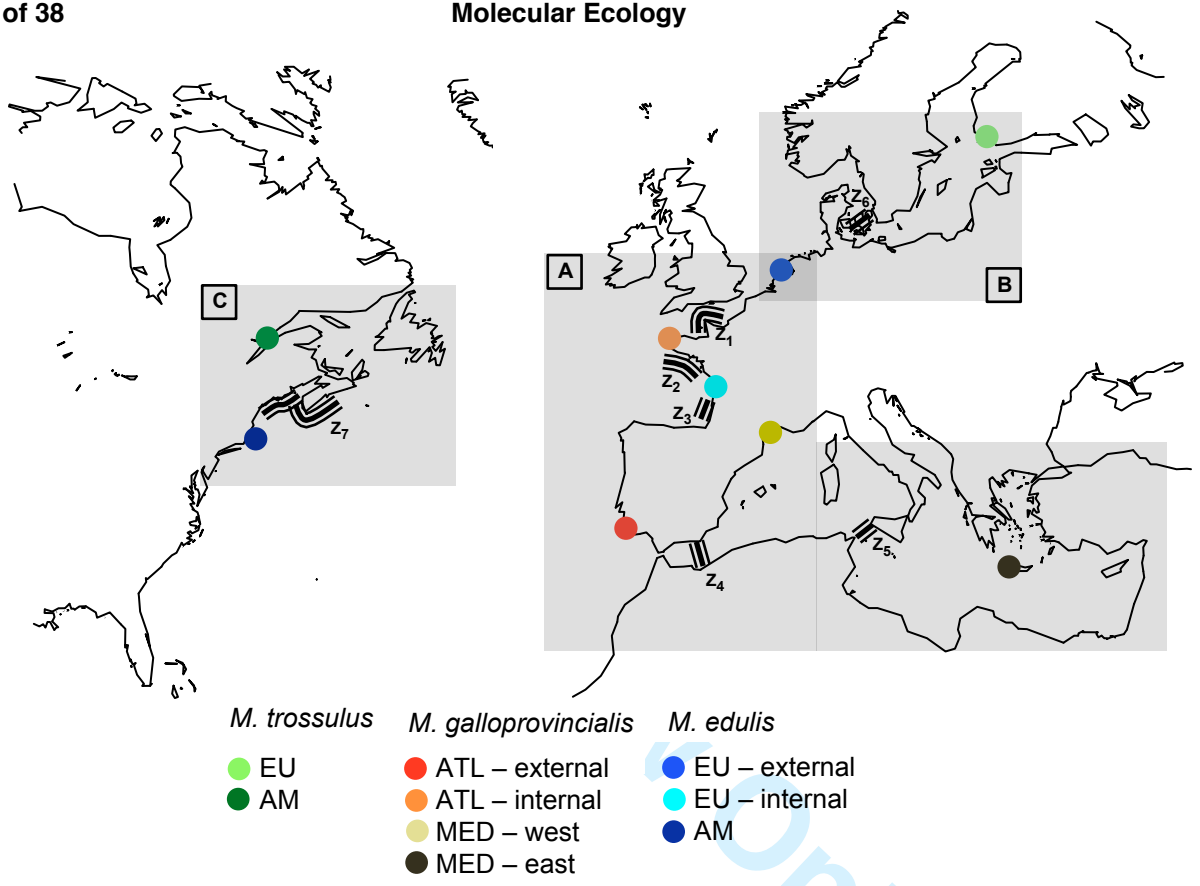


Figure 1 Localities of *Mytilus spp*. samples. We studied three geographical areas (shaded in grey) characterized by transitional zones, denoted z_i , between genetic backgrounds. (A) Mosaic hybrid zone between *M. edulis* and *M. galloprovincialis* of the Atlantic Coast (z_1 : Normandy, z_2 : South of Brittany, z_3 : Landes; Bierne 2003); followed by the transition with the *M. galloprovincialis* of the two Mediterranean basins (z_4 : Almeria-Oran Front, z_5 : Siculo-Tunisian Strait). (B) Clinal hybrid zone between *M. edulis* and *M. trossulus* (z_6 : Danish Straits; Väinölä & Hvilsom 1991). (C) Mosaic hybrid zone between *M. edulis* and *M. trossulus* (z_7 : Maine and Nova Scotia; Koehn 1984). *M. trossulus* samples are (1) Tvarminne (EU, light green) in the European population of the Baltic Sea and (2) Tadoussac (AM, dark green) in the American population of the Saint-Lawrence River. *M. galloprovincialis* samples are (1) Faro (ATL - external, red) in the peripheral Atlantic population of Iberian Coast, (2) Guillec (ATL - internal) in the enclosed Atlantic population of Brittany, (3) Sete (MED - west, yellow) in the Occidental Mediterranean basin and (4) Crete (MED - east, black) in the Oriental Mediterranean basin. *M. edulis* samples are (1) Holland (EU - external, light blue) in the peripheral European population of the North Sea, (2) Lupin/Fouras (EU - internal, cyan) in the enclosed European population of the Bay of Biscay and (3) Quonochontaug (AM, dark blue) in the American population of Rhode Island.

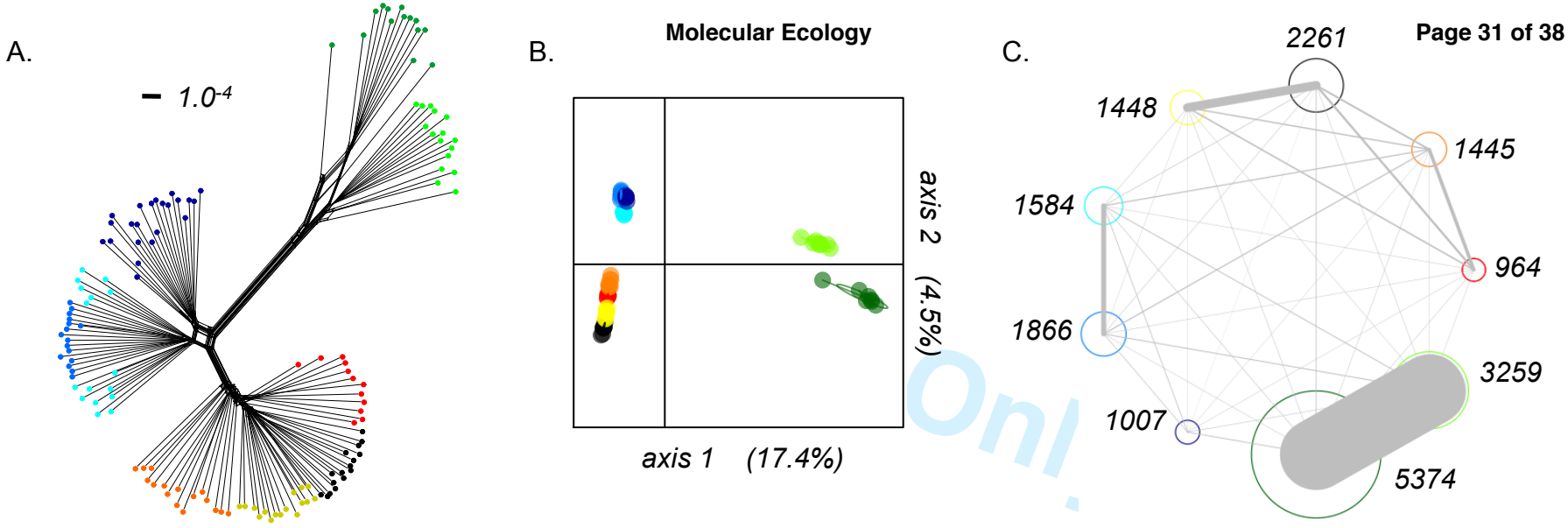


Figure 2 Genome-wide relationships between populations. (A) Phylogenetic network, produced with the neighbour-net method (*SplitsTree4*) based on 51,878 high quality SNPs genotyped in 72 individuals. Haplotype sequences were statistically phased with *BEAGLE*. (B) Principal Component Analysis. Results of the first two factorial components are shown. Sites with missing data were removed. (C) Pairwise privately shared SNPs. The size of lines is proportional to the logarithm of the number of SNPs shared. Non-shared private SNPs are indicated to the side of the corresponding population. Singletons were removed. The color code matches Figure 1.

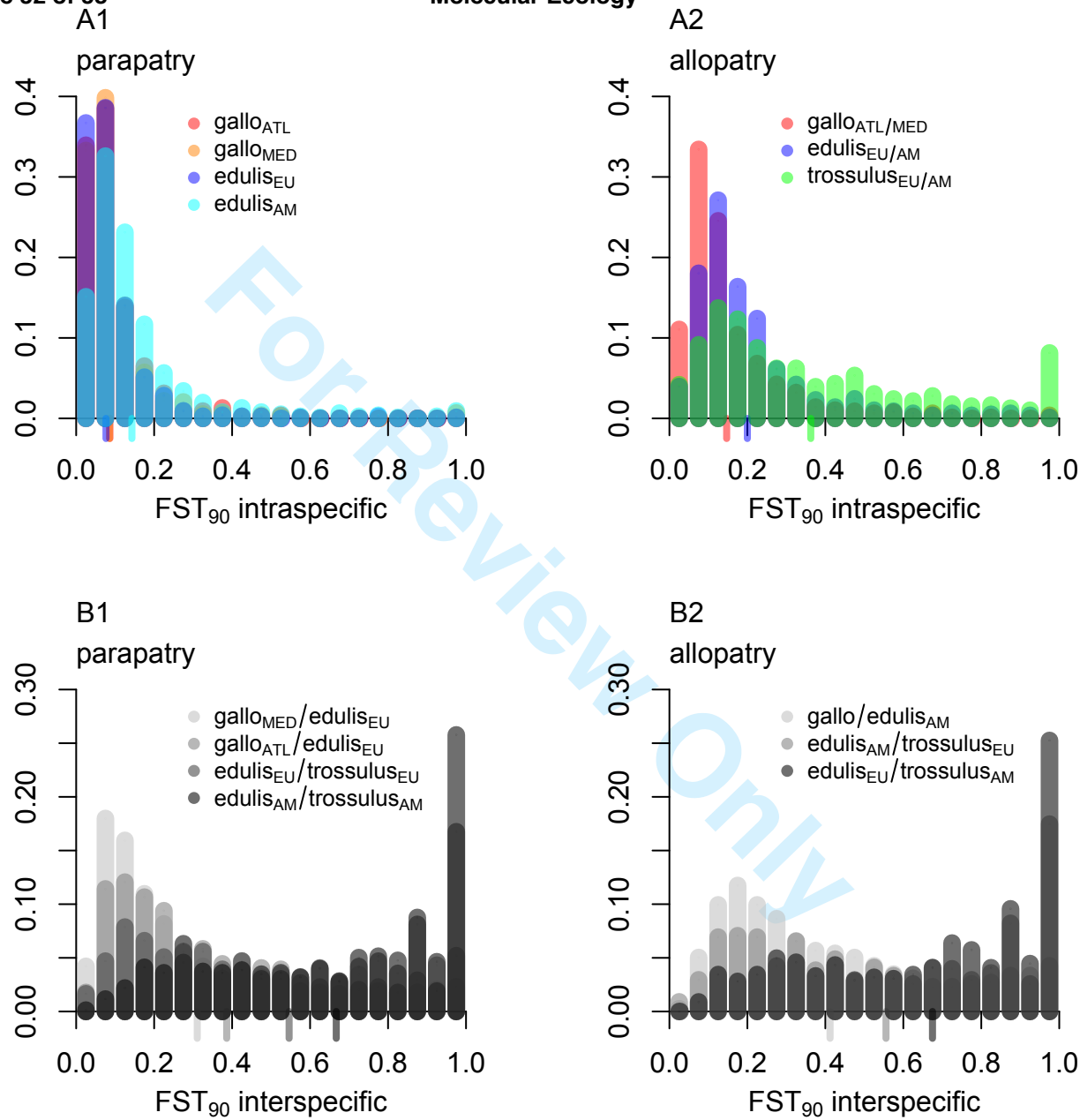


Figure 3 Distributions of FST_{90} values between (A) intraspecific and (B) interspecific populations in (1) parapatry and (2) allopatry. Mean values are indicated with vertical stripes underneath the x-axis. Differentiation levels were measured with the 90th percentile of the FST distribution of each contig (FST_{90}). Names match Figure 1.

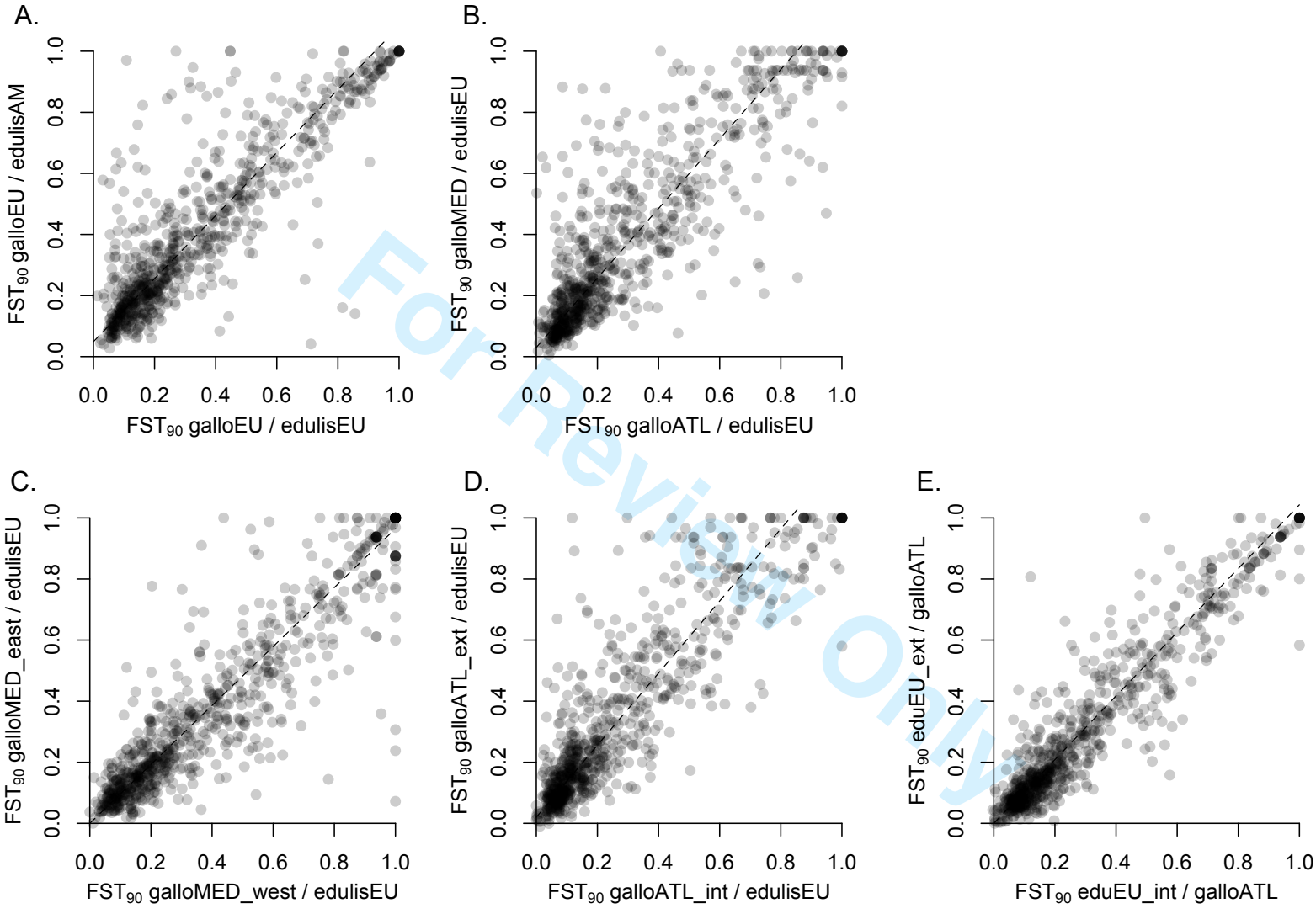


Figure 4 Joint distributions of interspecific FST_{90} values between *M. edulis* and *M. galloprovincialis* pairs in different geographical contexts. Allopatry: (A) European *M. galloprovincialis* with "European *vs* American" *M. edulis* (slope=1.032; elevation=0.049). Parapatry in Europe: (B) *M. edulis* with "Atlantic *vs* Mediterranean" *M. galloprovincialis* (slope=1.138; elevation=0.030); (C) *M. edulis* with "West *vs* East" Mediterranean *M. galloprovincialis* (slope=0.965; elevation=0); (D) *M. edulis* with "Internal *vs* External" Atlantic *M. galloprovincialis* (slope=1.18; elevation=0.020) and (E) *M. galloprovincialis* with "Internal *vs* External" European *M. edulis* (slope=1.043; elevation=0). The Standardised Major Axis regression is indicated in dashed line. All other details match Figure 3.

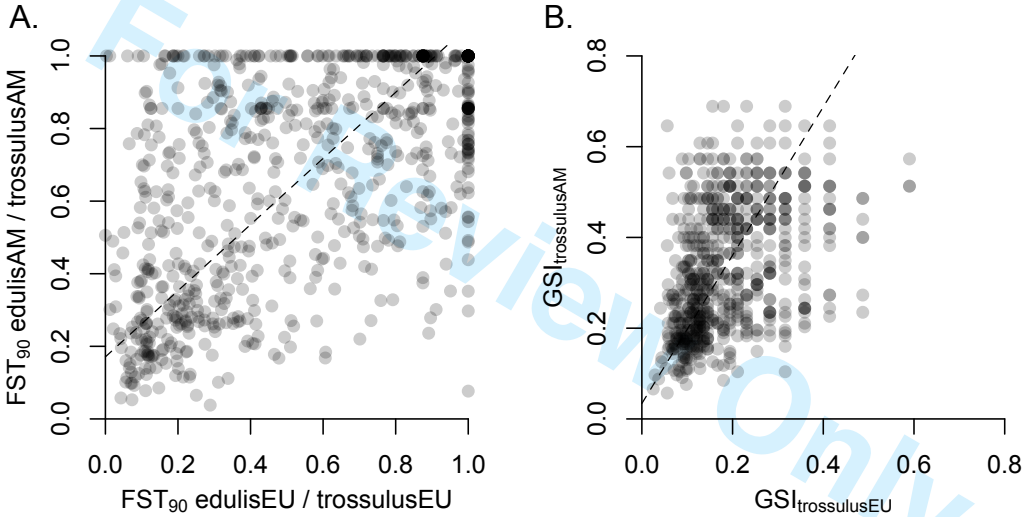
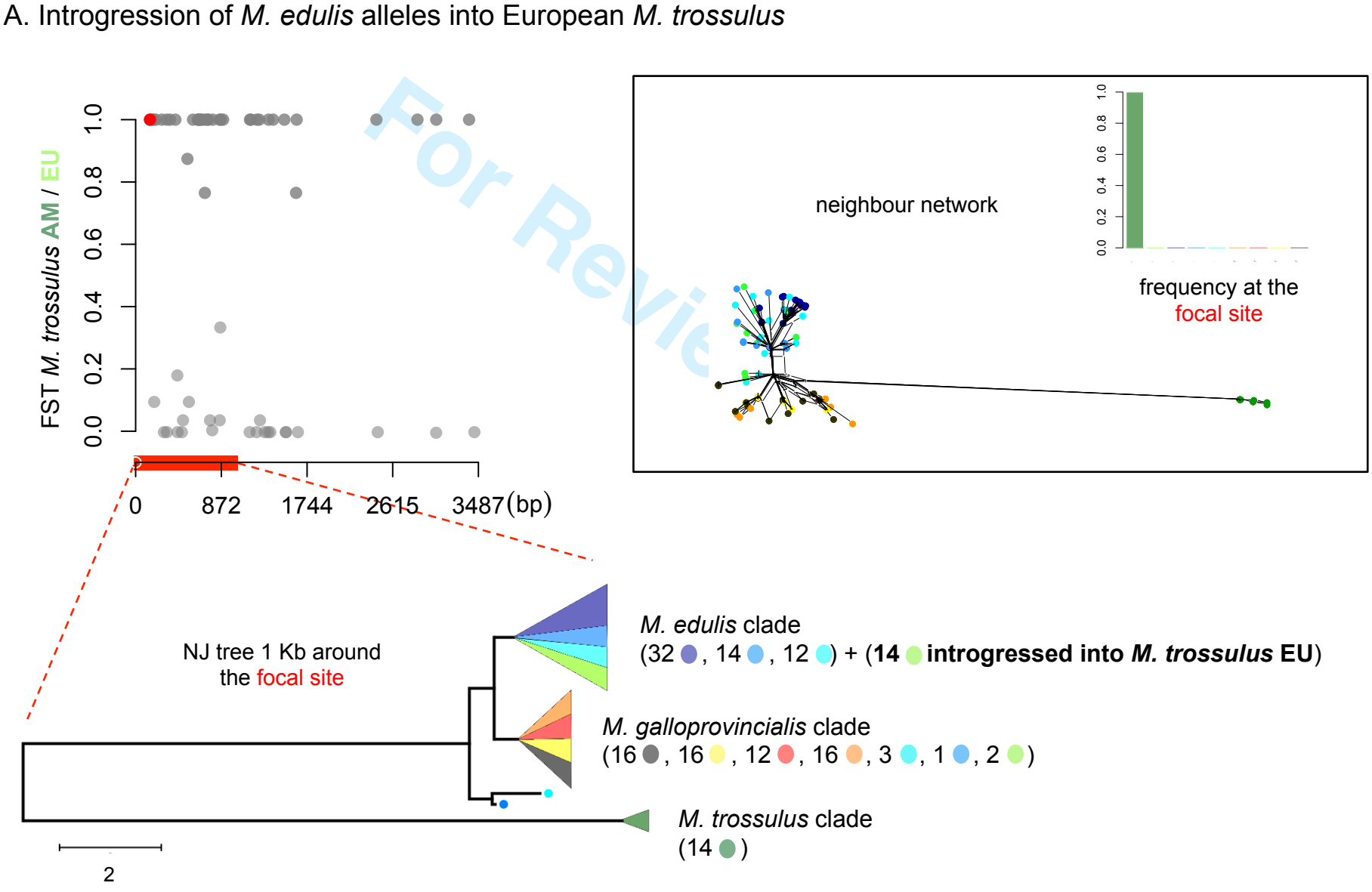
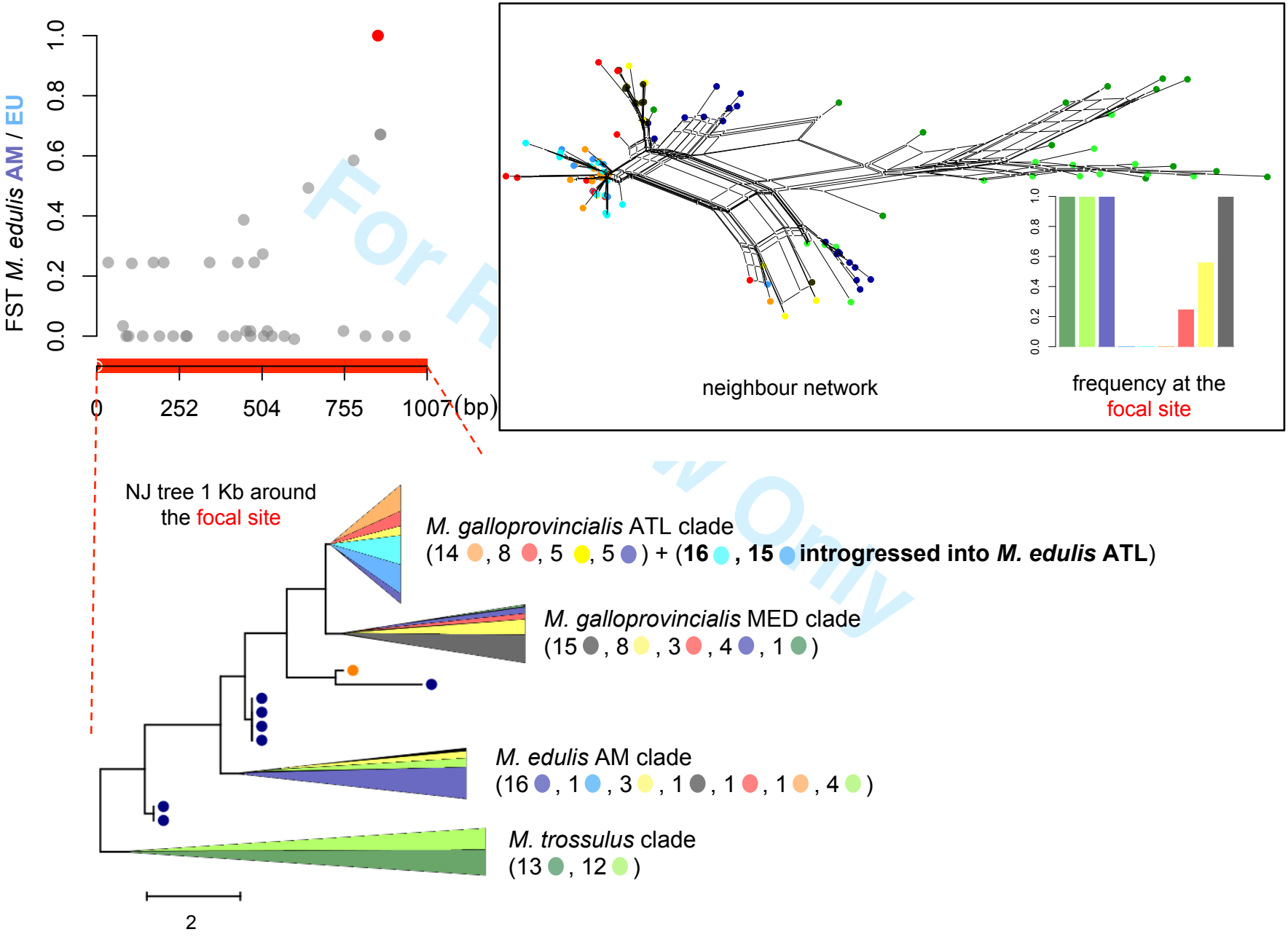


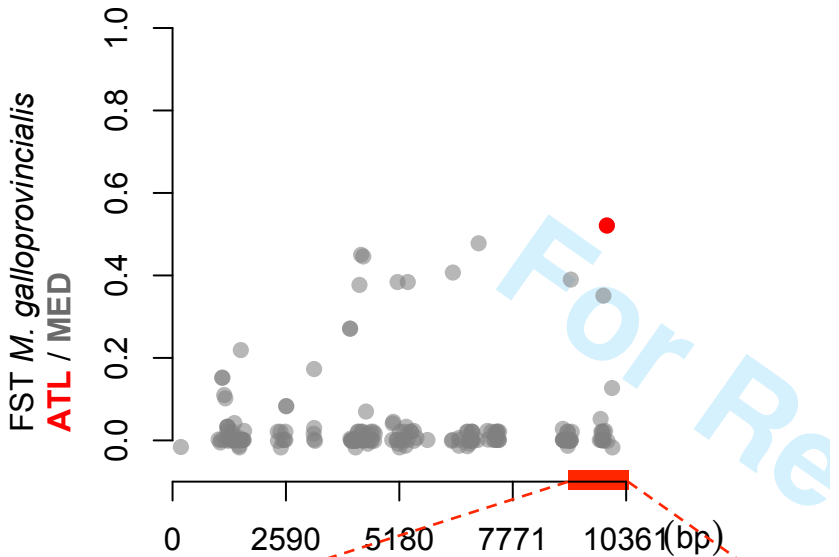
Figure 5 Parallelism between the European and American contacts of *M. edulis* and *M. trossulus.* (A) Joint distribution of interspecific FST_{90} values (slope=0.911; elevation=0.171). (B) Joint distribution of genealogical sorting index (GSI) values (slope=1.634; elevation=0.034). All other details match Figure 4.

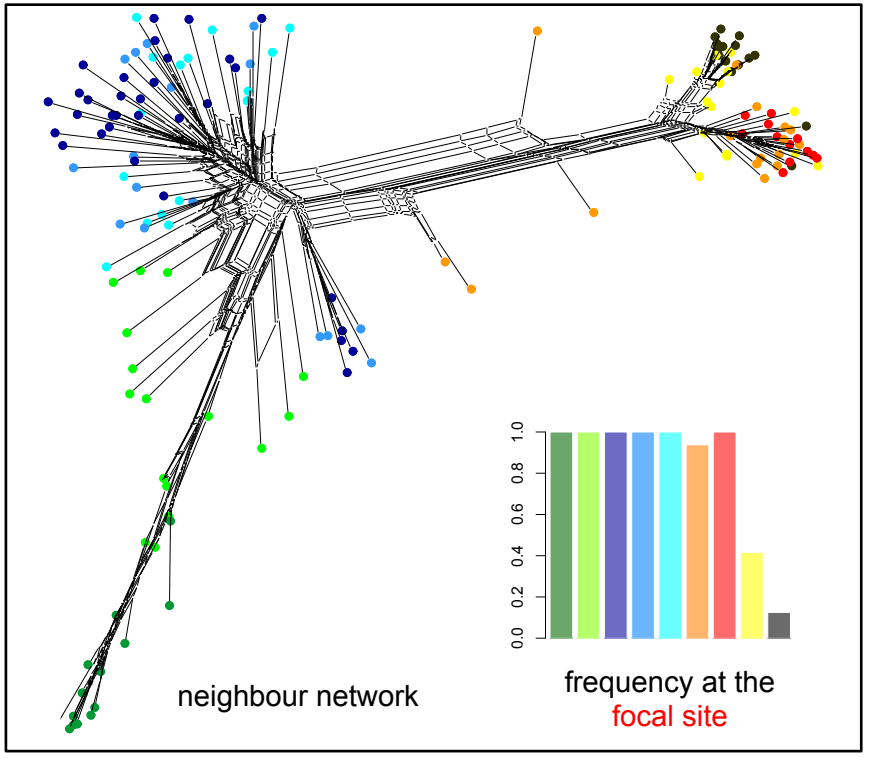
Figure 6 Examples of three candidate outliers in intraspecific comparisons. Top left panel: variation in FST level along the contig; the red point shows the focal site (i.e. maximal FST value); the red rectangle marks off 1 Kb around the focal site. Top right panel: neighbor network of the contig (*SplitsTree4*) and allele frequency at the focal site in all samples. Bottom panel: neighbour-joining tree 1 Kb around the focal site (R package *ape*). (A) Introgression of *M. edulis* alleles into the European *M. trossulus* population ("Contig54420_GA36A"). (B) Introgression of *M. galloprovincialis* alleles into the Atlantic *M. edulis* populations ("gi_403238785_gb_JX297444"). (C) Sweep of *M. galloprovincialis* Mediterranean specific alleles ("H_L1_abyss_Contig783"). The contig name is given in brackets (see TableS5 for details). The color code matches Figure 1.



Introgression of *M. galloprovincialis* alleles into Atlantic *M. edulis*







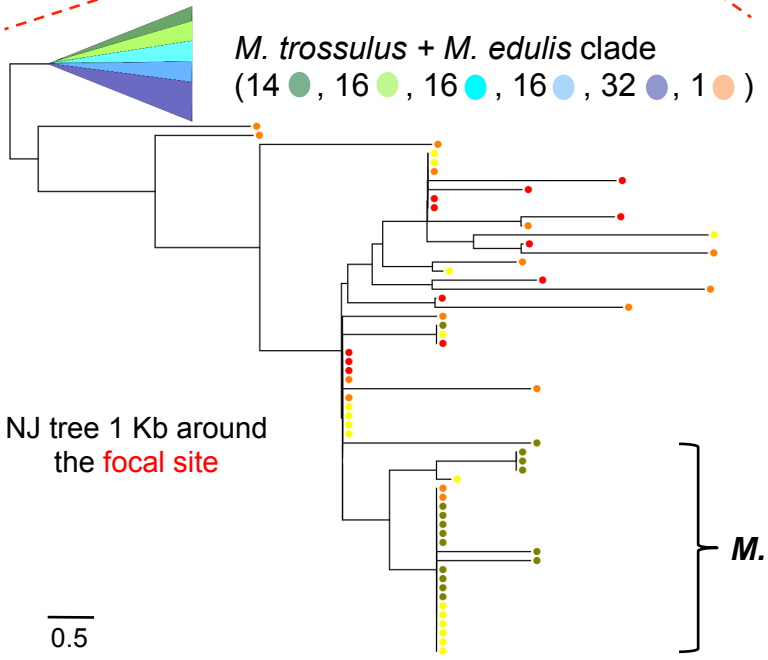


TABLE 1. Intraspecific outliers.

species	population$_1$	population$_2$	n_1	n_2	nb_{SNP} shared		FST_{90}		FST_{max}		$nb_{contigs}$ outliers				
											introgression				
					total	fixed	genomic	outliers	genomic	outliers	into population$_1$	into population$_2$	population-specific	na	TOTAL
M. trossulus	troAM	troEU	16	16	8245	318	0.362	0.693	0.521	1	2	35	3	0	40
M. edulis	eduAM	eduEU	22	32	417	15	0.131	0.544	0.211	0.802	10	3	10	0	23
M. galloprovincialis	galATL	galMED	28	32	1673	113	0.107	0.435	0.167	0.625	11	2	6	3	22
M. galloprovincialis	galATL-external	galATL-internal	12	16	387	0	0.086	0.321	0.139	0.51	1	10	3	0	14
M. galloprovincialis	galMED-east	galMED-west	16	16	1011	4	0.085	0.333	0.135	0.574	6	3	7	0	16
M. edulis	eduEU-external	eduEU-internal	16	16	555	0	0.076	0.185	0.134	0.464	0	0	10	4	14

n_1 and **n_2**: the number of haplotypes sampled; **nb_{SNP} shared**: the number of SNPs shared between population$_1$ and population$_2$ (without singletons); **FST_{90}** (and **FST_{max}**): the 90th percentile (and maximal value) of the FST distribution of each contig, averaged across all contigs ("genomic") or outlier contigs ("outliers")**; $nb_{contigs}$ outliers**: total number of outlier contigs, i.e. in the upper 2.5% of the FST_{90} or FST_{max} distributions. Outliers were categorized according to the cause of differentiation: (1) introgression of foreign alleles into population$_1$ or population$_2$; (2) differentiation of population-specific alleles between population$_1$ and population$_2$; "**na**" stands for outliers that we were unable to classify without ambiguity. In total, data include 1269 contigs; 122,144 SNPs; 471 shared and fixed SNPs.