



Institute of Science and Technology

Institute of Science and Technology
Am Campus 1
3400 Klosterneuburg, Austria

Klosterneuburg, April 11, 2018

Dear Marianne Elias,

We are grateful for the detailed and constructive comments to our work. We believe that the valuable inputs from Thomas Broquet and Tatiana Giraud greatly helped improving our manuscript. Thank you all for the time spent on our manuscript.

We have addressed each of the referees' comments below (printed in blue, interspersed with the original email from the editor). The new version has been uploaded onto the bioRxiv website, and we also uploaded a revised version with the modifications in "Track Changes" mode on the PCI Evol Bio website.

With best regards,

A handwritten signature in black ink, appearing to be 'CF' with a long horizontal stroke extending to the left.

Christelle Fraïsse

Postdoctoral FWF Fellow,

+33 (0)6 69 55 20 85

christelle.fraisse@ist.ac.at

On behalf of all authors.

Decision

by Marianne Elias, 2018-02-20 14:42

Manuscript version: <https://doi.org/10.1101/239244>

Fraisse et al.: revisions requested

Dear Dr Fraïsse,

Your manuscript has now been assessed by two reviewers, Tatiana Giraud and Thomas Broquet. While both of them found the paper interesting, they also have important comments that need to be addressed before the paper can be reassessed for recommendation. Notably, Tatiana Giraud questions the ability of the methods used to discriminate between introgression and ILS, and Thomas Broquet is not convinced that the pattern of differentiation detected indicates local adaptation, rather than a physical barrier to dispersal. In line with this, Thomas Broquet also suggests changes in the focuses and structures of introduction and, to a lesser extent, discussion. Both reviewers also have other minor points, particularly about clarifying the methods, which need to be addressed.

We have substantially altered the Introduction and Discussion to respond to individual comments. We thank Thomas Broquet to have highlighted we pushed to strongly on the side of the hypothesis of adaptation from admixture variation. This was clearly not what we intended to do and we were very pleased to endeavoured making clearer that Gagnaire et al.'s hypothesis of introgression clines was our preferred hypothesis. We now illustrate this hypothesis with simulations, that can reproduce the pattern observed with true data. This is, we hope, now much better highlighted in the revised manuscript. We perfectly understand and share Tatiana Giraud's scepticism about our ability to evidence secondary introgression. Although we believe the f_3 statistic (computed with *TreeMix*) is now commonly recognised as a valid support for secondary admixture, we now perform in addition a new analysis (inference of the divergence history with *∂a∂i*) to explicitly show that introgression contributed to the genetic composition of the Kerguelen population. This new analysis supports genomewide heterogeneous gene flow between the Kerguelen lineage and North Hemisphere species, yet in the past. At any rate we made clearer we do not deny a high rate of ILS in this system. In addition, the main objective of this study is to explain why ancestry-informative loci between North Hemisphere species reveal fine-grained genetic structure in the Kerguelen. We can only think about hypotheses that imply secondary admixture but we would very welcome an alternative hypothesis we would have missed.

On the BioRxiv site, as noted by Thomas Broquet, the supplementary files do not correspond to the supplementary files mentioned in the MS (file S1, S2 and S3): a pdf file called 'Supplementary Information bioRxiv' and a xlsx file called 'Supplementary Information table'. This xlsx file cannot be opened. Please make available all necessary information with a clear correspondence between the MS and the supplementary files, and please provide a correct table. In addition, make sure that:

- all the data are available to readers.
- all details of the quantitative analyses (e.g. data treatment and statistical scripts in R, bioinformatic pipelines scripts, etc.) are available to the readers, as appendices or supplementary online materials

The Supplementary Information are provided in four parts: (i) Supplementary Information M&M, (ii) Supplementary Information Figures (from S1 to S7), (iii) Supplementary Information Tables (from S1 to S12), and (iv) Supplementary Information Texts, including the raw data and the script for $\partial a \partial i$ (from S1 to S8).

Good luck with the manuscript revision,
Marianne Elias

Reviews

Reviewed by Thomas Broquet, 2018-01-10 11:18

This paper presents an empirical study of mussel populations (*Mytilus spp*) in the Kerguelen island, exploring the role of past admixture events on the local genetic structure of mussel populations around the island. Using a combination of genotyping-by-sequencing and target SNP genotyping, the authors dissect the complex divergence/admixture history of worldwide mussel lineages, infer the most likely origin of the Kerguelen population, and show that the local genetic structure is linked with past admixture events between native and non-native mussels.

How complex histories of reticulated evolution may affect current population genetic processes is a very interesting question that bears upon many important research topics (local adaptation, speciation in general and ecological speciation in particular, molecular analyses of connectivity...). The *Mytilus* complex is a good model species to tackle this question, and the dataset used here is adequate (the combination of worldwide and local SNP data is most interesting). The subset of SNPs used to look at the local genetic structure is small (33 loci), but this data set convincingly demonstrates that past admixture events have effects on the local micro-geographic structure. This result is, in my opinion, very interesting for a large audience in evolutionary biology. The paper is clearly written (although the methodology used is complex and could be better explained in places). For all these reasons I believe that this paper will make a very useful contribution in evolutionary biology.

I found, however, that there is perhaps room for improvement in two aspects (first concerning the interpretation of the results, and second how the introduction and discussion are structured). As detailed below, my comments call essentially for some clarification and perhaps a better structure of the presentation of scientific arguments and rationale.

(1)

The authors not only conclude that past admixture has effects on local genetic patterns but, more precisely, that it facilitated or at least impacted in some way local adaptation processes. The methods employed to reach this conclusion are thorough, and the investigation is interesting, but I don't think that the results are strong enough to reach a firm conclusion concerning local adaptation (yet). Alternatively, as recognized by the authors in several places, the genetic break observed at four loci could be due to a genome-wide reduction in gene flow, with consequences visible only on a subset of markers that are linked with

reproductive isolation genes (Fig. 3 in Gagnaire *et al.* 2015). In that case, the genetic data indicate a possible physical barrier to dispersal between sites RdA and PAF.

We are much concerned about the difficulty to actually demonstrate that local adaptation is the process that caused the local genetic patterns we observed, and we thank you to make us realize that our thoughts were not clearly presented. Actually, we deliberately used the phrase “genetic-environmental association” to stress the fact that there is an association of genotypes with habitats in the Kerguelen Islands, but this could be either due to (i) local adaptation *per se* (i.e. adaptation to local habitat opposed by gene flow between them), or (ii) introgression between two genetic backgrounds that generates gradients in allele frequencies better correlated with habitat variation than geographical distance. Moreover, we strongly concur that finding selected regions involved in local adaptation must be hard, and so it is more likely that the “genetic-environmental association” observed is due to the second hypothesis exposed above. We hope we made this clearer in the revised version of the manuscript. Below, we reply point by point to your thorough comments.

This seems consistent with several observations:

i) the geographical structure of allelic frequencies at the four loci after the RdA/PAF break seems nearly linear, i.e. compatible e.g. with isolation-by-distance inside the Gulf of Morbihan. There is no obvious reason for local adaptation to result in such a clinal distribution, unless habitat conditions themselves vary along a geographic gradient. Can the authors comment on this?

We agree that the clinal genetic pattern observed outside the Gulf of Morbihan does not follow habitat conditions closely. Basically, all sites from RdA to BdS are characterized by the presence of *Macrocystis* kelps (except RdA and AS), the absence of sandy bottoms (except RdA and FPN) and oceanic waters (except RdA and FPN). As of the slope and exposure, they vary between sites without clear spatial pattern. Therefore, instead of markers linked to locally adapted genes, the four loci may better reveal gradient in neutral allele frequencies due to a genetic connectivity break (a density trough or a barrier to dispersal) between sites PAF and RdA. More precisely, it is possible that a genetic background have been trapped in RdA, and introgression with the other background (everywhere else) produces the clinal pattern. The physical barrier between PAF and RdA would make the genetic break more visible on the West side of the contact than on the other side, where alleles can flow more easily between the two backgrounds. Moreover, based on allele frequencies at the four candidate loci (see our updated Figure 4B), the “RdA genetic background” may be more related to *M. edulis* while the second genetic background may be more related to the Chilean mussels.

To explore this idea, we performed a supplementary admixture analysis (with the program ADMIXTURE) of the 35 sites in the Kerguelen by fixing *M. edulis* and the Chilean mussels as two reference populations. Individual ancestries are provided in the new Text S5 and average ancestries for each locality are represented with pie charts in Figure 4B. These results indicate that RdA has the largest *M. edulis* ancestry (69%) compared to other localities (>81% *M. edulis* ancestry), supporting the idea that two genetic backgrounds (*M. edulis*-like and *M. platensis* Chili-like) segregate in the Kerguelen Islands. This may further suggest that all the mussels in the island arrived recently and that most of the admixture we described in the first part of the paper was already present in the Chilean mussels.

In addition, we illustrated this hypothesis by performing simulations (following the methods of Gagnaire *et al.* 2015) of a secondary contact between *M. edulis* mussels located

at RdA and Chilean mussels present everywhere else. We simulated a circular stepping stone model with a barrier to dispersal between sites PAF and RdA, and we followed the allele frequencies at two neutral loci: one linked to the barrier gene and another unlinked. The simulated patterns are presented in Figure 4C, and they fit the observations quite well. We also changed the text in many places to better highlight this hypothesis we also believe to be more likely.

ii) The RdA site, which shows the lowest foreign allele frequency, is not occupied by *Macrocystis* kelp. This seems to oppose directly the local adaptation hypothesis, which is based on the observation that foreign alleles are more frequent in habitats characterized by the absence of *Macrocystis* (text p.22 and Fig. 4B, grey dots). Moreover the authors mention p. 18 that "these two sites [i.e. RdA and PAF] differ at all five ecological variables". While I could not open Table S3 (format not recognized, for some reason), it seems from Fig. 4B that these two sites indeed differ in presence/absence of *Macrocystis*, but NOT in the direction predicted by the local adaptation hypothesis at most other sites. If this is true for the other environmental variables as well, then the fact that the two most differentiated sites differ in all ecological variables goes quite strongly against local adaptation.

You are totally right, RdA and PAF differ at all five ecological factors that were measured, but not in the expected direction. Basically, RdA has an habitat characteristic of the Gulf of Morbihan (sheltered site, flat sandy bottoms, low salinity and absence of *Macrocystis*) while being on the North-East coast and having the lowest foreign allele frequency. This is because, RdA is a river at the bottom of a bay, and so it is little influenced by marine waters. On the other hand, PAF has an opposite type of habitat while being located inside the Gulf of Morbihan. So there is not a perfect correlation between genotypes and environmental factors, which suggests that enhanced genetic drift and intense gene flow in the island blurred the signal at our markers. We now more clearly say that this pattern weakens the hypothesis of local adaptation, and we make more room for the second hypothesis, i.e. introgression between two genetic backgrounds. We changed the text lines 722-724 "*Nevertheless, the RdA site, which has the lowest foreign allele frequency, is not occupied by Macrocystis kelp, weakening this local adaptation hypothesis.*".

iii) "water masses between Gulf of Morbihan and North coast do not mix well" (p. 18), suggesting that a physical barrier to dispersal may exist, even at such a small spatial scale.

Indeed, there is potentially an oceanic barrier between these two sites: PAF is at the entrance of the "Gulf of Morbihan" and RdA is located to the other side, behind the "Presqu'île du Prince de Galles". At the entrance of the Gulf, there is an oceanic threshold that impede exchanges with water masses from outside the Gulf. Moreover, the Gulf is overall less salty because of the numerous rivers that flow into it. At a larger scale, the Antarctic circumpolar current moves the water masses from West to East, and this causes gyres and turbulences on the Northern and Eastern coasts (including RdA) that pushes water masses off the coast. We now mention this explicitly in the Results on lines 594-598 : "*Accordingly, there is an oceanic threshold at the entrance of the Gulf of Morbihan that impedes exchanges with water masses from outside; and at a larger scale, the Antarctic circumpolar current moves the water masses from West to East causing gyres and turbulences on the North-Eastern coast and pushing water masses far to the East (Karin Gerard, pers. comm.)*"; and Discussion on lines 731-732: "*However, the physical barrier to dispersal between sites PAF and RdA produces a clear genetic break on the West side of the contact.*".

iv) I did not find table S6 in sup mat, but from the text it seems that there is some genome-wide differentiation between several sites (e.g. PCu), compatible with short-scale spatial heterogeneity in neutral gene flow (this is not a strong argument, but it goes in the same direction: there could be some heterogeneity in dispersal-driven connectivity between local sampling sites at the scale of the island).

That is right, PCu is located in a quite remote area at the far West side of the island, and it is differentiated from the Gulf of Morbihan sites, and also from some sites in the North and South. We now make this point in the Result section lines 502-503: "*Further, they indicate the existence of spatial heterogeneity in dispersal-driven connectivity at the scale of the island.*".

On the other hand, obviously the hypothesis favoured by the authors (i.e. local adaptation: see end of abstract and end of introduction for instance) is also based on interesting observations. And perhaps that all of my points above are refuted simply because there are no loci involved in reproductive isolation (a point that I couldn't quite get clearly from the paper). While I am not sure that these observations are strong enough to conclude, I think that what is important here is the discussion of potential effects, rather than coming to a definitive conclusion.

We now explicitly say in the Introduction on lines 127-128: "*Accordingly, reproductive isolation genes have not been reported so far between mussels in the Kerguelen Islands.*" Though, we would like to draw your attention to the fact that no experimental crosses have been carried on. This claim is based on the observation that our candidate SNPs are mostly freely segregating in the island, whereas they belong to genome regions involved in the interspecific barrier between species in the Northern Hemisphere.

My following point deals with the structure of the paper.

(2.1)

In my view the main topic (and the strongest result) of this paper deals with the role of admixture effects on current population structuring, with possible effects on the detection of barriers to gene flow either due to physical barriers or local adaptation. These important topics are well discussed in the Discussion section, but their presentation in the Introduction could be improved. The Introduction starts with one general paragraph on local adaptation in marine environments, and then almost all of the intro details the *Mytilus* case, except for a few sentences on adaptive introgression near the end. I suggest that a more general presentation of the potential effects of admixture (role of ancestry-informative loci, difference with incomplete lineage sorting) is needed early in the introduction, including effects on signatures of connectivity breaks and links with (or even facilitation of) local adaptation. This would allow the introduction to end up with more formally hypothesis-driven objectives (e.g. no effect of past admixture vs detection of connectivity break vs local adaptation) - this is left to the judgment of the authors.

As you recommended, we extensively modified the Introduction to present our results into a broader evolutionary context and clearly present the different hypotheses, highlighting the "neutral hypothesis" of connectivity break.

(2.2)

Accordingly, the Discussion could also better partition arguments for or against each of these hypotheses (surprisingly, the authors do just that in some instances, but generally in the Results section: e.g. see very good sentences such as the final sentence of the first paragraph of p. 17, and final sentence of first paragraph of p. 18). My point is perhaps best illustrated by the following example: The second-last sentence of the paper reads "Possibly these markers simply better reveals a genome-wide signal of habitat constrained connectivity". This statement contradicts the point apparently preferred throughout the whole article, namely that local adaptation is necessarily involved (e.g. last sentence of the abstract, last sentence of the intro...). I think that the discussion would be clearer if it would expose the arguments for each hypothesis, and then conclude on what the authors think is happening.

Following your suggestions, we modified the Discussion so that the different hypotheses are more clearly exposed, and the presentation of the results related to our study system is simplified. We refer the reviewers directly to the revised manuscript.

Other minor comments

- Briefly explain why different filtering options were chosen for different analyses (end of last paragraph in supplementary methods).

This has now been specified: *"In total, 129,346 SNPs were called. For the genetic network analysis, we used a high-quality dataset (51,878 SNPs retained) to improve the reconstruction of genetic relationships while having a limited loss of information for such genome-wide analysis: genotype quality threshold was set at 30 (error probability of 0.1%) and missing data were not allowed. For the TreeMix and ADMIXTURE analyses, sites with a minor allele frequency below 5% over all populations were excluded (32,162 SNPs retained), as they are uninformative to infer admixture history."*

- In the same paragraph: What is a "maximum a posteriori genotype"?

We corrected this incorrect wording as follows: *"For each variant retained, the genotype with the highest likelihood (i.e. with a null phred-scaled likelihood score) was assigned to each individual assuming Hardy–Weinberg proportions in genotype prior probabilities."*

- Lisibility of figure S5A could be improved (red writings are difficult to see).

This has been changed.

- I could not find any of the supplementary tables.

We apologize for the inconvenience. This issue has been solved in the new bioRxiv version. All Supplementary Tables are included in a single file ("Supporting Information Tables").

- Post-Scriptum: reviewing papers without line numbering is tiresome...

Sorry for that omission in the bioRxiv version. The revised manuscript includes line numbering.

Reviewed by Tatiana Giraud, 2018-01-10 11:23

This study reports genomic analyses of mussel populations for investigating introgression and adaptation. Overall, I found the manuscript interesting, although unclear in many places, with even inappropriate wording in some cases (see below for some examples). There are too many confusing details while it is difficult for a reader not familiar to the system to get a global understanding.

As you recommended, we clarified the vocabulary and simplified the text in many places throughout the manuscript (see below). Thanks for your numerous suggestions. We also provide more general insights to the reader in the Abstract, Introduction and Discussion, especially highlighting how reticulated evolution can affect local genetic structure, and possibly facilitate adaptation.

Other than that, my main concern is that I am not convinced by the interpretation that introgression is more likely than incomplete lineage sorting. These two phenomena are notoriously difficult to disentangle in closely related species as is the case here. Incomplete lineage sorting is actually inferred on P13, rendering even more confusing the interpretation of introgression. If the authors think the methods they used can reliably infer introgression and refute incomplete lineage sorting, they should explicit why and how.

We acknowledge that disentangling between shared ancestry and introgression is challenging in closely-related species. However, in our system, inferences are strengthened by the fact that introgression was not species-wide, but geographically localized with some localities more introgressed than others (and also associated with environmental variation). This is not predicted under incomplete lineage sorting (ILS) alone, because the rate of drift is expected to be the same across populations (assuming similar N_e). For this first reason, we believe that local introgression contributed to the patterns observed in the Kerguelen Islands. Evidence for introgression between *Mytilus* species in our previous papers (e.g., Gosset *et al.* 2013 DOI: 10.1111/jeb.12046, Fraïsse *et al.* 2014 DOI: 10.1534/genetics.114.161380, Fraïsse *et al.* 2016 DOI: 10.1111/mec.13299) relied on similar patterns of local introgression revealed by genome scans of F_{ST} and gene tree reconstruction. We now make this point more explicit (see our reply to your specific suggestions on ILS below).

In this paper, we used both gene tree reconstruction (*Twisst*) and correlations of allele frequencies between populations (*TreeMix*) to demonstrate that the Kerguelen mussels have been introgressed by Northern mussels. *TreeMix* jointly estimates population splits and subsequent admixture events based on a method very similar to the classic *D-statistic* of Patterson *et al.* 2012 (DOI: 10.1534/genetics.112.145037) and *F-statistics* introduced by Reich *et al.* 2009 (DOI: 10.1038/nature08365). These methods have been widely used, especially in the study of human genetics, to robustly test for admixture and estimate ancestry proportions in an admixed population (e.g. Reich *et al.* 2009, Green *et al.* 2010 DOI: 10.1126/science.1188021, Haak *et al.* 2015 DOI: 10.1038/nature14317). As for *TreeMix*, it has been specifically applied to systems with a complex demography (e.g. in humans, Wong *et al.* 2017 DOI: 10.1101/gr.202945.115; in *Arabidopsis thaliana*, Exposito-Alonso *et al.* 2018 DOI: 10.1038/s41559-017-0423-0), because it explores the space of possible admixture graphs and selects the one that best fits the data based on a log-likelihood framework. Thus, this method was particularly suitable for such a complex model system as *Mytilus* mussels. Importantly, in their original paper, Pickrell & Pritchard 2012 (DOI: 10.1371/journal.pgen.1002967) were able to identify with *TreeMix* previously known introgression events in the human history.

We now reinforced these results by conducting a new analysis to explicitly test for the

presence of gene flow during the divergence history of Kerguelen mussels and the three Northern species. We reconstructed their divergence history based on their joint site frequency spectrum with $\partial a \partial i$, and we found that a model of ancient migration with heterogeneity in introgression rates across the genome was the most supported model in all three comparisons. These new results are presented in Table S5, and the associated site frequency spectra and models definition are provided in Text S5 and S6, respectively.

In a complementary way, we computed gene trees and analyzed them with an exploratory method (*Twisst*) that determines how genetic relationships vary across the genome. This tree weighting approach has been successfully applied to *Heliconius* butterflies to evaluate the support of different phylogenies around colour pattern loci (Van Belleghem *et al.* 2017 DOI: 10.1038/s41559-016-0052). Based on this method, we found contrasting genetic relationships between different genomic regions of the *Mytilus* mussels: the Kerguelen mussels were alternatively grouped with each of the three Northern species. We recognize that this pattern is due to a combined effect of introgression and ILS, and we revised the text to make it clearer. In addition, the fact that only 17% of the gene trees showed resolved topologies implies that ILS is preponderant as we stated in the Result section line 458 “A first result of the analysis is therefore a high rate of incomplete lineage sorting”. We improved the representation of the gene trees (Figure 3) by adding observed genealogies in non-recombining blocks (the longest blocks between breaking points estimated with the Difference of Sums of Squares method) for each *Twisst* topology. Moreover, we updated the tree classifications in Table 1 and Table S6.

Finally, we would like to point out that the loci that are currently structured between localities in the Kerguelen Islands are primarily those that are highly differentiated between Northern species, regardless of the mechanisms that explain why they segregate in the island (introgression and/or ILS). However to explain the fine-scale genetic structure currently observed in the island, we need to invoke gene flow, either as opposed to drift (connectivity break hypothesis), or to selection (local adaptation hypothesis).

Please find below more specific suggestions (lines were not numbered so it is not easy to refer to where the comments apply exactly):

Again sorry for that. The revised manuscript includes line numbering.

-Abstract and elsewhere: “proto-*edulis*” should be “proto-*M. edulis*”, but actually this view (as well as the “*M. edulis*-derived” formulation) is biased and incorrect: a species divergence is symmetric and you do not know whether the ancestral species was more *M. edulis*-like or more *M. platensis*-like or even to any of the two.

We now clarified the phylogenetic vocabulary used. We changed the phrasing as proposed by the referee in places where “proto-” and “-derived” occurred, typically with “an ancestor of *M. edulis* and *M. platensis*”.

-Abstract: “Southern lineage haS”.

This has been corrected.

-Abstract and elsewhere: the number of contigs is uninformative without their size. Anyway there are too many details in the abstract; There should be less details and instead a more general conclusion at the end, explaining what this study brings as general insights beyond mussels.

We removed details related to the studied system in the Abstract (including references to the number of contigs), and gave a more general conclusion lines 42-45: “*Our study highlights genetic connectivity of populations is more easily revealed by non-equilibrium secondary introgression clines at a subset of loci, while association with the environment should not be hastily advocated to support adaptation from admixture variation.*”

-Abstract and elsewhere: “ancestry-informative SNPs or markers”: could they be instead SNPs more affected by selection? And/or could this choice of highly differentiated markers bias the results?

In the present context of high-dispersal marine organisms, we deliberately chose to use selected loci that delineate Northern *Mytilus* species in order to infer genetic connectivity at a micro-geographical scale in the Kerguelen Islands. This is because outlier loci can reveal subtle patterns of genetic differentiation in place where neutral markers remain uninformative, as shown by our set of randomly chosen loci. This approach is all the more appealing for marine mussels, because populations have large effective population size and so selection may be more efficient than drift to oppose the homogenizing effect of migration (see Gagnaire *et al.* 2015 for a review, DOI: 10.1111/eva.12288).

As to the selective mechanisms that target the outlier loci (or most probably, loci next to them), they can be diverse including local adaptation, genetic incompatibilities, background selection, etc. It is not clear which one affected our panel of “ancestry - informative loci”, and this is also because reproductive barriers between *Mytilus* species result from the coupling of endogenous and exogenous barriers (Bierne *et al.* 2002 DOI: 10.1111/j.0014-3820.2002.tb01339.x, Bierne *et al.* 2003 DOI: 10.1098/rspb.2003.2404, Simon *et al.* 2017 DOI: 10.1101/237925), which are then hard to tell apart.

We now give more rationale for the usage of “ancestry-informative SNPs” in the Introduction lines 160-163 “*These ancestry-informative loci enabled us to infer genetic structure at a micro-geographical scale in the islands and connect the evolution of the Kerguelen mussels with the history of the Mytilus species in the Northern Hemisphere.*”

-Abstract and elsewhere: “the Kerguelen is a divergent lineage”: the Kerguelen is an island, not a lineage.

We corrected this misphrasing throughout the manuscript by replacing the inaccurate occurrences with “*Kerguelen population*”.

-P2: “evidence that local recruitment occurs in marine environments” does not “highlight the role of high population size”, it only refutes “high levels of dispersal”.

We deleted this sentence.

-P2: taxa cannot be “semi-differentiated” (differentiation is quantitative): they are either “partially-reproductively-

We replaced “semi-differentiated taxa” by partially “*reproductively isolated taxa*”.

-P2: “produce hybrid zones”: replace “produce” by “display” or “show”.

This has been done.

-P2: “encouraged to consider them as different taxonomic entities”: who are “them”? And different from what? This does not seem contradictory to be “related to” *M. edulis* or *M. galloprovincialis*”.

We have now clarified this sentence on lines 97-101: *“In the South, a reevaluation of allozyme data and a review of the results obtained with mtDNA and two nuclear DNA markers (Borsa et al. 2012) encouraged to group Southern mussels in two different taxa, namely M. platensis for those related to M. edulis (the South American and Kerguelen mussels), and M. planulatus for those related to M. galloprovincialis (the Australasian mussels).”*

-P3: replace “unique” by “single”.

This has been done.

-P3-4: there are too many details, dealing with just a few markers, while it is difficult for a reader not familiar with the system to understand the global picture. In particular, it is unclear whether the existence of “reproductive isolation genes” has been shown. One first question of the study could be whether the pattern found using just a few markers holds at the genome-wide level. In addition, it is not clear why incomplete lineage sorting is not considered here, while it seems a more parsimonious hypothesis.

We simplified the description of the model system in the Southern Hemisphere by removing extraneous details in the Introduction. We also make a more direct connection between the aim of our study and what was known based on a few markers on lines 150-154: *“We used published genotyping-by-sequencing (GBS) data of the three Northern species (Fraïsse et al. 2016) and new GBS data of a sample from a single Kerguelen population to reconstruct their genetic relationships, and investigate whether reticulated patterns found with a handful of markers hold genome-wide.”*

Also, we clarified what is known about reproductive isolation genes in the Kerguelen mussels on lines 127-128: *“Accordingly, reproductive isolation genes have not been reported so far between mussels in the Kerguelen Islands.”* Actually, it is not entirely clear whether the genetic structure we observed between localities is due to a connectivity break between two genetic backgrounds (partially isolated by incompatible genes), or to local adaptation only (in such case, we won't expect reproductive isolation genes).

We now explicitly say in the Introduction that both processes (introgression and ILS) have an effect lines 121-123: *“This micro-geographical variation in allele frequency suggests that admixture with Northern mussels contributed to the pattern observed at Glu-5, although shared ancestry certainly affects a large part of the genome in these closely-related species.”*

-P4 and elsewhere: *Macrocystis* is mentioned in several places in relationship with local adaptation, but it is unclear what is meant here: competitive interactions? And why looking at only this organism presence regarding local adaptation? Aren't they any other important competitor, predators, parasites...?

The genetic-environmental associations that we investigated here were constrained by the fact that we relied on a previous study (Gérard et al. 2015) in which the different environmental factors were measured. This previous study put a lot of effort into characterizing habitats, including presence / absence of *Macrocystis* kelp. This macroalgae is a keystone species in the coastal ecosystems of the Antarctic. It forms kelp forest which serves as substrata and refuge for many molluscs species, including *Mytilus* (Adami & Gordillo 1999, DOI: 10.3989/scimar.1999.63s1183). Moreover, kelp beds were generally associated with other environmental factors in the Kerguelen: they were especially found in open coasts with exposed rocky shores, whereas they were absent from protected areas with flat sandy bottoms and low-salinity waters. So it is hard to tell apart the role of each of

these factors in the genetic-environmental association that we observe, though the RDA analyses point toward a primary effect of *Macrocystis* kelps. Moreover, it is interesting to note that such contrasting habitat were already described as being associated with different genetic backgrounds in another context: *M. edulis* (high salinity) vs. *M. trossulus* (low salinity) in the Baltic sea hybrid zone, *M. edulis* (sheltered habitat) vs. *M. galloprovincialis* (exposed habitat) in France and British Isles. This was also true in the North American Atlantic hybrid zone, but with an opposite pattern: *M. edulis* were preferably found within bays and estuaries while *M. trossulus* were observed in oceanic coasts. These reversed genetic-environmental associations could suggest that the coupling of endogenous factors with habitat-related genes have occurred in opposite directions in Europe and America (see Bierne *et al.* 2011 for a review: DOI 10.1111/j.1365-294X.2011.05080.x). In the Kerguelen, the foreign allele is generally associated with low-salinity / sheltered sites / flat sandy bottom / absence of *Macrocystis*, which are typical of the Gulf of Morbihan localities (from Kerguelen, not the Gulf of Morbihan from Brittany). However, RdA shows the opposite pattern: it is characterized by all previous factors while it has the lowest foreign allele frequency across all localities. Thus the role that *Macrocystis* kelps have on the genetic structure of the Kerguelen mussels is not entirely clear, and we now put more emphasis on alternative hypothesis to local adaptation (see also our reply to Thomas Broquet's comments).

-P4: how was introgression distinguished from incomplete lineage sorting in this previous study (Fraisse *et al.* 2016)?

We briefly specify the methodology used on lines 86-89: “Moreover, a recent study based on *Fst* genome scans and small-scale gene genealogies demonstrated that local introgression is widespread, and it is the primary cause of outlying levels of genetic differentiation between conspecific populations (Fraisse *et al.* 2016).”

-P5 and elsewhere: KASpar, GEA, FCT... there are many undefined abbreviations all along the manuscript.

We replaced all occurrences of “GEA” by “genetic-environment associations”. In addition, we now define all abbreviations when they first appear, as follows:

- “KASpar” as “*kompetitive allele specific PCR*”.
- “GBS” as “*genotyping-by-sequencing*”.
- “BAC” as “*bacterial artificial chromosome*”.
- “AIC” as “*Akaike information criterion*”.
- “PCA” as “*principal component analysis*”.

-P5: explain better the “target enrichment sequencing”: enriched in what and how?

We refer the readers to the Supplementary Information, where we now give more details about the target enrichment sequencing in those terms: “BAC sequences (totalling 2 Mb) were obtained from a BAC library constructed based on whole genomic DNA of three *M. edulis* individuals, and composed of 32 tagged clones and a pool of 192 untagged clones. The targeted coding regions were of two types: (i) a random panel of 338 cDNA contigs (1.6 kb on average, totalling 0.5 Mb) from *M. galloprovincialis* produced by RNA sequencing (Romiguier *et al.* 2014); and (ii) a set of 553 publicly released expressed sequence tags (1.0 kb on average, totalling 0.5 Mb) including 262 immunity genes, 133 genes involved in cytonuclear interactions, 30 reproduction-related genes, 20 habitat-related genes, six nucleoporins and 102 housekeeping genes.”

-P5: explain if and how you could disentangle introgression from incomplete lineage sorting.

We now explain in more details why we are confident that we detected introgression on top of ILS on lines 154-157: *“Past introgression events between Northern and Southern mussels were robustly inferred (on top of high rates of incomplete lineage sorting) by testing for admixture with genome-wide allele frequency data and reconstructing gene genealogies at a small chromosomal scale.”*

-P6 and elsewhere: FST is a mathematical index, it should be written with only the F in big capital and ST in subscript capitals.

This has been done.

-P6 and elsewhere: “*trossulus*” should be “*M. trossulus*” and idem for other species names.

This has been changed.

-P6 and 7: “GBS individuals”: replace by “GBS-typed individuals” (GBS is not a trait of individuals).

This has been done.

-P7: Fis should be written with the F in big capital and IS in subscript capitals.

This has been done.

-P8: give the mean contig size and whether we know whether they are unlinked.

We now specify their size and we explicitly point out that they are provided in Table S5 on lines 288-290 *“Neighbour-joining trees of the 395 retained sequences are available in Text S7 and their length are indicated in Table S6 (4.5 kb in average, a minimum length of 1 kb and a maximum length of 25 kb).”*

In addition, we specify information about linkage in the “Genotyping-by-sequencing” section of the Material & Methods lines 185-188: *“The final dataset across the twelve localities consisted of 1269 reference sequences (378 BAC contigs that come from a pool of 224 unlinked clones, and 891 cDNA contigs that correspond to unlinked coding sequences of known-functions or randomly selected) and 129,346 SNPs.”*

-P9: explain better what is an RDA.

We now give a more general introduction to the RDA in the Material & Methods on lines 314-317: *“It performs a multiple linear regression between a matrix of response variables (individual genotypic data) and a matrix of explanatory variables (environmental factors). Notably, the effect of partially confounded explanatory variables can be estimated separately.”*

-P10: *M. platensis* should be introduced from the beginning.

M. platensis (as well as *M. planulatus*) was already introduced in the Abstract and Introduction. But we make this even clearer now in the Introduction on lines 97-101.

-P11: replace “more” by “most”;

This has been changed.

-P11: not sure “edge” is the most appropriate word?

We retained the word “edge”, because this formulation is commonly used in the context of genetic admixture analyses with *TreeMix*.

-P12: not sure “eventually” is the most appropriate word?

Indeed, we removed the word “eventually”.

-P12: “*M. edulis*-derived” is a wrong formulation: the bifurcating node between *M. edulis* and *M. platensis* is symmetric, there is no reason to call the ancestral node “*M. edulis*” (same comment P13).

We changed the formulation by “*a Southern lineage related to M. edulis*”.

-P13: replace “already involved” by “has been previously suggested to be involved”.

This has been replaced.

-P16-17: there are many details that seems rather anecdotal, it is difficult to get an interesting take-home message on the Kerguelen structure results.

We took away some details (e.g. average allele frequencies at the outlier loci, and results on Glu-5’ from previous papers), and we now focus on the description of the spatial genetic pattern (see also our reply to Thomas Broquet’s comments).

-P17: explain also briefly here what is a redundancy analysis.

This has been specified as follows on lines 568-571: “*As such, we performed a redundancy analysis, RDA, (i.e., a multivariate constrained ordination) on the 695 individual genotypes sampled from the 35 sites characterized by different habitats, and estimated the relative contribution of each environmental factor on population genetic structure.*”

-P20: not sure “sample” is the most appropriate word?

We removed this sentence from the revised version.

-P20: “divergence between [mussel populations] in the two hemispheres”.

This has been changed.

-Discussion: it is too long, detailed, too much focused on the model. Try to highlight the main messages better and why your study is of general interest beyond the model species. Avoid citing Tables and Figures in the discussion (all results should have been presented in the result section already).

We substantially modified the Discussion as you suggested, and also by following Thomas Broquet’s comments (see our replies above).