HAL
open science

Evolution of divergence between the river lamprey (Lampetra fluviatilis) and the brook lamprey (L.planeri) inferred by experimental approaches and population genomics

Quentin Rougemont

To cite this version:

Quentin Rougemont. Evolution of divergence between the river lamprey (Lampetra fluviatilis) and the brook lamprey (L.planeri) inferred by experimental approaches and population genomics. Biodiversity and Ecology. Université de Rennes, 2015. English. NNT : 2015REN1S141 . tel-01338038

HAL Id: tel-01338038

https://theses.hal.science/tel-01338038v1

Submitted on 27 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

Ecole doctorale Vie Agro Santé

présentée par

Quentin Rougemont

Préparée à l'Unité Mixte de Recherche 985 ESE
UMR Ecologie et Santé des Ecosystèmes
UFR Science de la Vie et de l'Environnement

Evolution de la divergence entre la lamproie fluviatile (*Lampetra fluviatilis*) et la lamproie de planer (*Lampetra planeri*) inférée par approches expérimentales et de génomique des populations

Thèse soutenue à Rennes le 15/12/2015

devant le jury composé de :

Margaret DOCKER
Professeur Associé University of Manitoba / *rapporteur*

Nicolas BIERNE
Directeur de Recherche CNRS/ *rapporteur*

Louis BERNATCHEZ
Professeur et Chaire de recherche du Canada en Génomique et Conservation des Ressources Aquatiques Université Laval / *examinateur*

Jean-Christophe SIMON
Directeur de Recherche INRA / *examinateur*

Guillaume EVANNO
Chargé de Recherche INRA / *directeur de thèse*

Sophie LAUNEY
Chargé de Recherche INRA / *co-directeur de thèse*

Table of Contents

Acknowledgements

Chapter 1: General introduction **1**

1. Speciation and Reproductive isolation 6
2. Modelling gene flow across space and time 15
3. Studying speciation: selection vs endogenous barriers 21
4. Insight from studies of parallel adaptation and parallel speciation 25
5. Lampreys as a model of speciation research 29
6. Goals of the Thesis 36

Chapter 2: Investigating gene flow and reproductive isolation in lampreys **39**

Article 1: Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys

Chapter 3: Investigating divergence history of European river and brook lamprey **81**

Article 2: Reconstructing the demographic history of divergence between European river and brook lampreys using Approximate Bayesian Computations

Chapter 4: Understanding speciation: moving toward genomics **131**

Article 3: Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and non-parasitic lamprey ecotypes

Chapter 5: Effect of anthropogenic disturbance on population genetic diversity and structure of European brook lamprey **181**

Article 4 Moderate effect of river fragmentation but strong influence of gene flow between ecotypes on the genetic diversity of brook lamprey populations

Chapter 6: Discussion & Perspectives **215**

1. Low levels of reproductive isolation and high viability of F1s at an early developmental stage

2. The importance of the geographical context in studying speciation

3. The complexity of histories of divergence

4. Better characterizing isolated *L. planeri* populations

General Conclusion **230**

Appendices **233**

Appendix 1: Testing selection at linked sites: effects of BGS

Appendix 2: Development of a hybrid linkage map: mapping endogeneous and exogeneous barrier

Appendix 3: Testing outbreeding and heterosis in isolated *L. planeri* populations

References **245**

Scientific Activities **273**

Acknowledgments

Prendre le temps de réfléchir, sur tout, sur rien, semble devenir un luxe dans ce monde gouverné par les notions de rendement, productivisme, utilité et autre valeur ajouté, pourtant sources de déchirement et non du bonheur de l'homme. Durant 3 années, j'ai eu la chance de réfléchir, de me poser des questions que beaucoup considèrent inutile dans cette société productiviste, mais qui permettent de comprendre le fonctionnement du monde qui nous entoure, et qui aujourd'hui me force à rester humble face à l'étendue de mon ignorance. A l'issue de ces 3 ans de thèse, je continuerais encore plus qu'avant à me poser des questions sur tout, sur rien, et à m'égarer dans des digressions scientifiques qui ne peuvent être que bénéfiques pour la culture scientifique, mais aussi car seule la réduction de mon ignorance contribuera à mon épanouissement personnel. En cette fin de thèse, plus que jamais en accord avec Socrate, « je ne sais qu'une seule chose : c'est que je ne sais rien ».

Pour m'avoir donné l'opportunité d'effectuer cette thèse, de l'orienter selon mes propres réflexions, quitte à faire fausse route, je remercie mes encadrants Guillaume et Sophie. Merci à vous pour votre ouverture d'esprit, vos qualités humaines, votre enthousiasme, et vos idées. Merci de m'avoir laissé l'opportunité de partir à l'étranger de rencontrer des personnes qui m'ont permis aujourd'hui de largement enrichir mes connaissances.

Merci à L. Bernatchez, N. Bierne, M. Docker et J-C Simon d'avoir accepter de faire partie de mon Jury.

Merci aux membres de mon comité de thèse : Marie-Agnès Coutellec, P.A. Gagnaire, Eric Petit, Manu Plantegenest et Tony Robinet pour les discussions.

Merci à deux passionnés : Charles Perrier pour les « piqûres de rappel » de génétique des populations, toujours d'actualité à l'aire de la génomique à tout va. Un merci infini à Pierre Alexandre Gagnaire pour m'avoir transmis autant de connaissances sur la spéciation, les balayages sélectifs et autres îlots génomiques. Merci à vous pour vos grandes qualités humaines, j'attends toujours d'aller attraper un bon grand silure avec vous….

Merci à J. Goudet, Samuel Neuenschwander et C. Roux pour m'avoir accueilli à l'UNIL. Merci en particulier à Camille pour m'avoir brièvement enseigné les bases de la coalescence de l'ABC.

Merci à toutes les personnes du laboratoire ESE et U3E….par ou commnencer ?

Merci à Dominique Huteau, Adrien Oger, Victoria Dolo, Anne-Laure Besnard pour votre aide infinie sur le terrain et au laboratoire. J'imagine bien qu'à l'heure actuelle les lamproies vous auront définitivement marquées (parasitées) l'esprit. Sans vous, bien des choses n'auraient jamais pu aboutir et les jeux de données n'auraient jamais atteint une telle envergure !!! Merci particulier à Adrien d'avoir réussi à me supporter en toutes conditions….Merci aux personnes de l'U3E pour leur aide dans les (très) grandes opérations de croisements, tentatives multiples d'élevage, de maintien de lamproies de planer, fluviatiles et même de lamproie marines... et pour le nourrissage de nos bébés lamproies. Un merci particulier à Maïra et Julien pour votre aide. Merci aussi à toutes les autres personnes de l'U3E qui ont quasiment tous contribué à nourrir ou entretenir les aquariums : Yoann, Antoine, Bernard, Cédric et bien sur Didier Azam et Frédéric Merchand pour avoir permis que toutes ces opérations puissent avoir lieu.

Merci à toutes les personnes ayant contribuées à l'échantillonnage de lamproies à travers la France : Merci en particulier aux Benjamin(s) : Bulle, Hérodet, Jacquot, Dufour de la fédération de l'Ain pour leur grand enthousiasme et leur engouement/entêtement à pêcher des lamproies, mobiliser des gens d'autres fédérations et avoir fourni un effort gigantesque malgré l'absence de connaissances préalable sur la bête...J'ai aussi été heureux de pouvoir vous sensibiliser un peu à l'intérêt de la génétique! Merci aussi à toutes les autres personnes des fédérations de pêches et AAPPMA de Bretagne, Normandie, Nord, Loire Atlantique :

Sur l'Odon : Benjamin Dufour et Yannick Salaville de la FDPPMA du Calvados. Sur le Montafilan, la Rance, le Léguer et le Saint Emilion : Hubert Catroux de la FDPPMA Côtes d'Armor. Sur la Risle : Germain Sanson et Victor Zunigas de la FDPPMA de l'Eure. Sur la Tamoute et l'Illet : Richard Pèlerin de la FDPPMA l'Ille-et-Vilaine. Sur le Cens : Vincent Mouren et Maxime Lesimple de la FDPPMA Loire Atlantique Sur l'Aa, la Hem, la Liane et le Wimereux : Benoît Rigault de la FDPPMA du Pas-de-Calais. Sur la Bethune et la Bresle : Geoffroy Garot de l'association Seinormigr, Pascal Domalain et Jean Louis Fagard de l'ONEMA, Jean Marcel le Boucher de l'AAPPMA de Dieppe, l'AAPPMA le Pêcheur Brayon, et l'AAPPMA de la Truite Brayonne. Sur l'ensemble du département du Morbihan : Nicolas Jeanneau de l'Unité Expérimentale U3E, Anne-Laure Caudal de la FDPPMA du Morbihan ainsi que les AAPPMA suivantes : André Robbe de l'AAPPMA la Gaule Alréenne, l'AAPPMA Brochet Basse Vilaine ; l'AAPPMA la truite Questembergoise; l'AAPPMA de Malestroit et Yannick Perraud sur la Loire. Merci aussi à Mathieur Romain, pour les lamproies Alsaciennes ! Enfin, pour terminer cet échantillonnage, merci à Bill Beaumont et Rasmus Lauridsen pour m'avoir fait découvrir les lamproies anglaises, Merci à Phil McGinnity et son enthousiasme pour nous fournir des échantillons de lamproies à la Guinness !! Merci aussi à l'association Migrateur Rhône Méditerranée pour les échantillons de lamproies marines méditerranéennes et merci à P. McCormick pour l'envoie récent de lamproies marines américaines que nous ne manquerons pas d'analyser prochainement.

Merci à Gervaise Février, Marie Thérèse Delaroche pour leur bonne humeur, efficacité considérable et grande capacité à gérer mes changements de trains de dernière minute…. ! Merci à Gilles Maubert (et à Linux!) Merci aussi à Gilles Lassalle pour la bonne humeur constante, l'aide bio-informatique, la patience avec « ce satané logiciel de cartho »…..et la transmission de connaissances !

Merci à tous les anciens collègues de bureau qui ont du me supporter durant les 3 dernières années : Valérie Lopez, Alice Beaudoin, François Martignac. Merci à tous les autres collègues de l'UMR ESE pour les discussions diverses et leurs bonnes humeur : Yannick Bayonna, Anne Treguier, Guillaume F, Marco Limnée, Marion, Isa, Stephane, Yann Echasserieau, Clarrisse, Aurélie, Lucie, Pierre-Lou.

Merci enfin à mes amis, en particulier Adrien et Anne-So ainsi qu'à Mon Poulet Emilien et à ton petit bout qui grandi bien vite…. ! Merci à vous….!! Merci à Benjamin J. malgré la distance...j'attends toujours un thon ! Merci aux personnes qui ont partagés un moment de vie avec moi. Merci aux faux Suisses Arnaud et Prout, merci pour les soirées, pour avoir retourné Genève, pour le Gay Tour International-Français et tous ces bons moments ! Merci le black métal, le stoner, le grindcore et le brutal death metal !

Merci à mes parents pour m'avoir soutenu dans mes projets, et ce depuis toujours.

Chapter 1

General introduction

Introduction

The current diversity of life on earth is the result of millions of years of evolution and past speciation events. Understanding how diversity arises is at the heart of evolutionary biology and has both fundamental and applied consequences. Understanding species diversification therefore requires comprehension of the process of speciation (Felsenstein 1981). To do so, one needs at first to define what a species is. Here comes the trouble as the definition of a species is unclear and gathers more than twenty different concepts (Mayden 1997; Coyne & Orr 2004; Hausdorf 2011). The main difficulty with this number of species concepts is that each one conveys a different definition of what speciation is (Hausdorf 2011). The most widely adopted definition, the biological species concept (Mayr 1942), defines species as "groups of actually or potentially interbreeding individuals reproductively isolated from other such groups". This simple definition allows the study of evolutionary processes leading to the establishment of reproductive isolation (RI) between taxa. It is particularly useful to study the appearance of prezygotic and postzygotic barriers to gene flow. However, hybridization and introgression are commonplace in nature (Mallet 2005). Furthermore, there is accumulating evidence for adaptive introgression (i.e. the introgression of advantageous alleles from one genome to the other through hybridization) in plants (Rieseberg 2009; Arnold & Martin 2009) and animals (Hedrick 2013), including humans (Hawks & Cochran 2005; Racimo *et al.* 2015). These phenomena should not exist under the strict definition of the biological species concept (Coyne & Orr 2004). The "genic species concept" (Wu 2001) on the contrary allows for barrier semi-permeability (Harrisson, 1986) so that most of the genome may freely be exchanged when species co-occur in sympatric areas, while key genomic regions remain impermeable to gene flow (Wu 2001; Harrison & Larson 2014).

Hybrid zones are particularly well suited to the study of speciation (Hewitt 2011) and the effect of semi-permeability and heterogeneity along the genome (Barton & Hewitt 1985). In this context, when significant genetic differentiation is observed in spite of gene flow in sympatry, it is particularly important to determine if differentiation was truly initiated without geographic barriers or whether it was initiated in allopatry and followed by secondary contact. Both theory and empirical evidence overwhelmingly favor scenarios of allopatric divergence and admixture (Turelli *et al.* 2001; Coyne & Orr 2004; Harrison & Larson 2014). Studies of speciation rates demonstrated that speciation only reaches completion over very long time scales, typically millions of years (Hedges *et al.* 2015). However a growing body of literature (e.g. Rundle & Nosil 2005; Schluter 2009; Nosil 2012) suggests that speciation can proceed *i)* quickly (hundreds to thousands of years) *ii*) in the presence of gene flow and *iii)* due mainly to the action of natural selection. Without a solid theoretical framework and explicit tests of hypotheses that do not rely on selection it is complicated to draw conclusions about these new findings (Lynch 2007; Nei *et al.* 2010; Hughes 2012; Sorrells & Johnson 2015). Fortunately, models of speciation have been developed (Coyne & Orr 2004; Gavrilets 2014) and greatly help to understand the conditions necessary for speciation to occur.

As noted by Abbott *et al.* (2013) (Fig 1) speciation is a multi-level process "unfolding through space and time" and the genetic makeup of contemporary species is influenced by alternative cycles of isolation and connection (Hewitt 1996). Taking these multiple components into account can be difficult, but combinations of experimental and modelling approaches together with population genetics could help understan how speciation proceeds through space and time. Understanding how biodiversity arises also has applied consequences for the future. Indeed, we may have entered into the sixth mass extinction on Earth (Barnosky *et al.* 2011), which is largely man-driven (Vitousek *et al.* 1997; Palumbi 2001). Thus, if man aims at preserving biodiversity -at least for the sake of its own survival- understanding how speciation proceeds may be very helpful. For instance, different species concepts will have different implications for species conservation (Frankham *et al.* 2012). Similarly, understanding the effect of human perturbations on the genetic structure and functioning of populations can help preserve these populations.

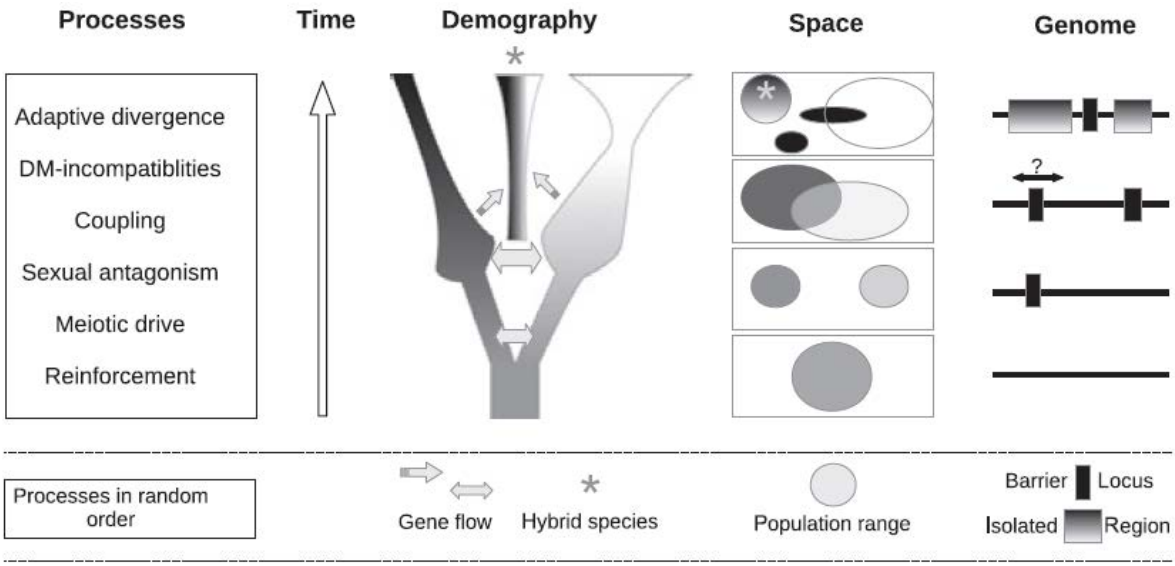


Figure 1: **"Speciation is a multi-level process unfolding through space and time**" (From Abbott et al. 2013). This figure summarizes the main steps that can lead to reproductive isolation over space and time. Over the course of time populations will undergo different demographic events combined with periods of gene flow and periods of strict geographic isolation. Geographic isolation can lead populations to adapt to divergent environments and start to accumulate barrier loci (e.g. DM incompatibilities (see 1.2.2.1) that can either collapse (barrier breakdown) or couple to enhance reproductive isolation so that at a certain point most of the genome becomes differentiated.

It is within this conceptual framework that I have studied the speciation process in the European river lamprey (*Lampetra fluviatilis)* and brook lamprey (*Lampetra planeri*), two species for which relatively little is known in terms of evolutionary biology and divergence history. The major goal was to characterize the level of reproductive isolation and the magnitude of the genetic barriers to gene flow between these closely related species. These inferences were jointly performed with investigations of the historical processes that have generated the observed level of reproductive isolation. Finally I also investigated how natural fragmentation of riverine habitat, enhanced by current human induced perturbations, can impact the level of population genetic diversity and structure and eventually lead to modification of population genetic evolutionary trajectories.

In the following sections of this introduction, I will first present the different mechanisms of speciation, the underlying theoretical genetic models and their limits. In a second part I will explain the importance of the geographical and historical conditions under which speciation arises. The third part presents some new methodological advances in the genomic era and discusses their limits. The last section presents the model species and what makes it a relevant model to study the process of speciation.

1. Speciation and Reproductive isolation

1.1 Overcoming the homogenizing effect of gene flow

Under the biological species concept (BSC), the most basic question to address is to determine which reproductive barriers arise first in space and time and what are the most important barriers that maintain species isolation (Coyne & Orr 2004). The most influential parameter on the probability of establishment of a genetic barrier is simply the effect of the level of gene flow between the diverging populations. Accordingly, recombination breaks up associations between adaptive alleles (Felsenstein 1981) and impedes divergence between nascent species.

The central question is thus to determine the quantity of gene flow that can pass freely between two diverging pools without stopping them from diverging. Mayr (1942; 1963) proposed a classification based on a biogeographical scale (the species distribution range) that indirectly takes into account the intensity of gene flow between diverging pools. Allopatry defines the situation when populations are in disjoint geographical areas and the gene flow m is equal to 0. When populations have completely overlapping ranges they occur in sympatry and $m = 0.5$. When populations have partially overlapping ranges, they are considered parapatric and m lies in between the extreme values of 0 and 0.5. From this classification, it is obvious that speciation will not proceed in the same manner in each situation. In allopatry, when populations are separated by mountains or oceans, no other factor but time is required for populations to eventually become reproductively isolated (Turelli *et al.* 2001). They will undergo the action of genetic drift and randomly accumulate mutations. They can also undergo divergent selection that will largely accelerate the process (Gavrilets 2003). All these processes drive the accumulation of genetic incompatibilities through time (see section 1.2.2.1). The model of allopatry is theoretically well supported and several examples exist in the literature (see examples in Coyne & Orr 2004). In sympatry, conditions for divergence to occur appear rather complicated because even with strong selection on many loci recombination will impede divergence (Felsenstein 1981). In parapatry, the conditions for the evolution of reproductive isolation (RI) are highly conditioned by the level of gene flow (Gavrilets 2003; Bank *et al.* 2012).

It is important to note that this classification of the geographical modes of speciation in discrete categories has been criticized (Butlin *et al.* 2008; Fitzpatrick *et al.* 2008) because it does not reflect the continuous nature of the speciation process. According to Butlin *et al.* (2008) long distance migration between allopatric populations can almost always occur even at very low frequency and complete sympatry almost never occurs because the distribution of habitat between coexisting populations is often patchy. As a result, speciation would always occur in parapatry (Butlin *et al.* 2008). However, it is clear from theory that the geographical context will determine if RI can occur and at which pace. The importance of this context was recognized again recently (Marie Curie SPECIATION Network *et al.* 2012). We will see in chapters 2 and 4 that the geographical context is important in determining the level of gene flow in our focal species. However, since species spatial ranges undergo contractions and extensions (Hewitt 2004); the

current distribution of species may have little to do with the initial conditions of divergence. Hence, to understand under which conditions divergence was initiated one has to reconstruct the underlying history (see section 2.1). But first, an important step under the BSC is to quantify the strength of reproductive barriers that can act to maintain differentiation between populations (Coyne & Orr 2004).

1.2 The barriers to gene flow

A barrier can be endogenous (independent of the environment) or exogenous (dependent on the environment). Barriers that occur before the formation of the zygote are called pre-zygotic barriers, those that occur after are called post-zygotic barriers (Dobzhansky 1937) (Fig. 2). Most post-zygotic barriers involve genetic mechanisms whereas pre-zygotic barriers involve genetic mechanisms but also ecological factors.

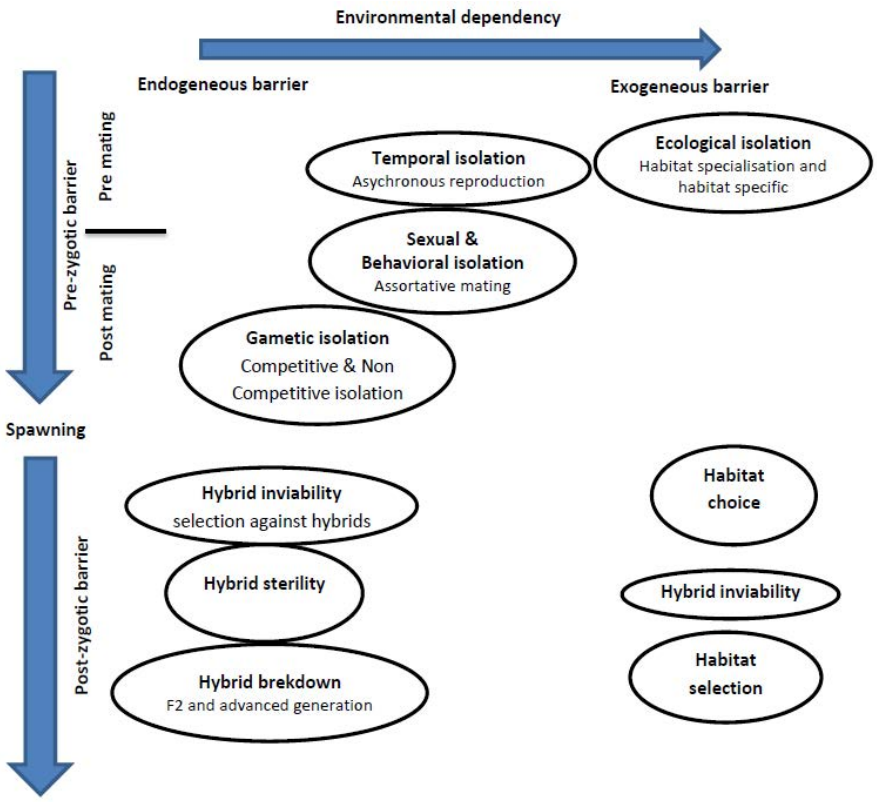


Figure 2: **Classification of the main mechanisms of reproductive isolation according to their environmental dependency and their stage of occurrence during the life cycle**. Taken from Ravigné *et al*. 2007 and modified according to Coyne & Orr (2004) and Nosil *et al.* (2005).

1.2.1 Prezygotic barriers

A first type of barrier called ***habitat isolation*** arises when species colonize two different ecological niches hence reducing their probability of mating together. This type of isolation involves spatial separation of the taxa but differs from a geographic barrier as it is based on genetic differences between taxa, different habitat choice and survival ability in these habitats (Rundle & Nosil 2005). Geographic isolation on

the contrary is linked to historical processes such as mountain formation, glaciations or continental drift (Coyne & Orr 2004). A classic example of habitat isolation is the coexistence of the *Rhagoletis pomonella,* phytophagous insect that inhabits apples and hawthorn, and *Rhagoletis mendax* whose mating and oviposition is restricted to blueberry. Hybrids of the two species have never been found in the wild (Feder & Bush 1989). It is possible that this mutualism between insects and hosts plants is genetically based and efficiently maintains RI (Coyne & Orr 2004).

Temporal isolation is a type of prezygotic barrier that occurs when two species mate at different time periods, reducing their hybridization probability. An interesting example is the anadromous pink salmon *Oncorchynchus gorbuscha* that has a two year life cycle and exists as two genetically isolated populations (odd-year and even-year). Allelic frequencies are generally similar between either distantly spaced odd-year or even-year populations. On the contrary, allelic frequency differences occurred between odd- and even-year populations occupying the same streams suggesting that these two alternate-year broodlines are reproductively isolated (Aspinwall 1974). It is likely that even and odd year broodlines have allopatric origins (Aspinwall 1974; Brykov *et al.* 1996). To better understand the effect of this barrier crosses between broodlines and measurements of hybrid fitness would be required.

Behavioral and mating isolation can also act as prezygotic barriers to gene flow. They include species differences reducing mate attraction and heterospecific mating (Coyne & Orr 2004). One classical example is the coexistence of *Pieris occidentalis* and *P. protodice* sympatric butterflies. In this system, males of *P. occidentalis* have much darker forewings than males of *P. protodice* while female wing patterns do not differ. Field observations have shown that females of *P. occidentalis* mate with conspecific males but reject nearly all heterospecific males. When male *P. protodice* wings were experimentally darkened they were more easily accepted by female *P. occidentalis* (Coyne & Orr 2004). The genetics of behavioral isolation has been studied in the *Drosophila* system and generally supports the view of a polygenic adaptation (Wu *et al.* 1995). In addition, behavioral isolation can evolve rapidly due to sexual selection (West-Eberhard 1983; Wu *et al.* 1995; Ting *et al.* 2001).

Mechanical isolation is defined as the inhibition of fertilization between species due to incompatibilities of their reproductive structures in relation to sexual selection (Coyne & Orr 2004; Arnqvist 1998; Arnqvist *et al.* 2000). A classic example involves pollinator isolation due to structural incompatibilities in plants. Similar examples are found in animals as in *Drosophila* where the evolution of male genitalia in the *D. melanogaster* group is proportional to divergence time (Lachaise *et al.* 1988). Other examples of mechanical isolation include size assortative mating as observed in water striders (Arnqvist *et al.* 1996) and stickleback (McKinnon *et al.* 2004). It seems that most genital differences have a polygenic basis (Coyne & Orr 2004; Eberhard 2010; Masly & Masly 2011) and most of the interspecific variance would be additive (Liu *et al.* 1996). This view is supported by studies in *Drosophila* (Liu *et al.* 1996; Zeng *et al.* 2000; Masly *et al.* 2011; McNeil *et al.* 2011; LeVasseur-Viens *et al.* 2014) and *Carabus* (Sasabe *et al.* 2007, 2010).

Post-mating gametic isolation includes all barriers acting between copulation/pollination and fertilization (Coyne & Orr 2004). These barriers are divided in non-competitive forms (without the presence of conspecific or heterospecific male gametes) and competitive forms (with conspecific or heterospecific male gametes). For instance studies in *Drosophila* have shown that *D. sechelia* males mated with *D. simulans* females had a very low quantity of sperm transferred and when *D. simulans* males were present few hybrids were produced due to noncompetitive and competitive isolating mechanisms (Price *et al.* 2001). Other non-competitive barriers include gamete inviability or failure of gametes to produce fertilization (see Coyne & Orr, 2004). Competitive isolation involves conspecific sperm precedence (CSP) in animals and conspecific pollen precedence (CPP) in plants. CSP can be defined as the greater fertilization success of conspecific *versus* heterospecific sperm in conspecific crosses (Howard 1999). Several studies have shown that CSP can contribute to gene flow reduction and form an efficient barrier between species including fishes, insects, marine invertebrates and corals (Howard 1998; Howard *et al.* 2002; Chang & Noor 2004; Geyer & Palumbi 2005; Ludlow & Magurran 2006; Fogarty *et al.* 2012; Manier *et al.* 2013; Yeates *et al.* 2013) although some other studies did not find such barriers in other taxa (Larson *et al.* 2012; Williams & Mendelson 2014).

In conclusion, understanding the respective roles of these barriers is critical to explain the maintenance of RI. However, it appears fundamental but challenging to reconstruct the temporal order of appearance of these barriers as currently observed reproductive barriers may not be those involved in the initiation of RI (Coyne & Orr 2004). Some barriers may also be a consequence of speciation and not the initial cause generating RI. We will see that most of them may contribute to reproductive isolation in our focal lamprey species pair.

1.2.2 Postzygotic barriers to gene flow

Various postzygotic barriers to gene flow can arise after zygote formation: (i) hybrid lethality, (ii) hybrid sterility and (iii) reduced viability and/or fitness of advanced hybrid generations. Under the BSC, RI is complete when the fitness of every F1 is zero. Selection against hybrids can be endogenous (independent of the environment) when hybrids show reduced fertility or viability. Alternatively it can be exogenous when different phenotypes (or genotypes) are favoured in different environments. The genetic basis of postzygotic isolation has been extensively studied (Coyne & Orr 1998, 2004) and is presented below.

1.2.2.1 Endogenous barriers

Several models exist to explain the evolution of endogenous barriers but few have received empirical support (Coyne & Orr, 2004). The best supported is the Dobzhansky-Muller (DM) model of hybrid incompatibility. Models of polyploid speciation, hybrid underdominance, and extensions of the DM model have been developed (Nei & Nozawa 2011) but will not be presented here.

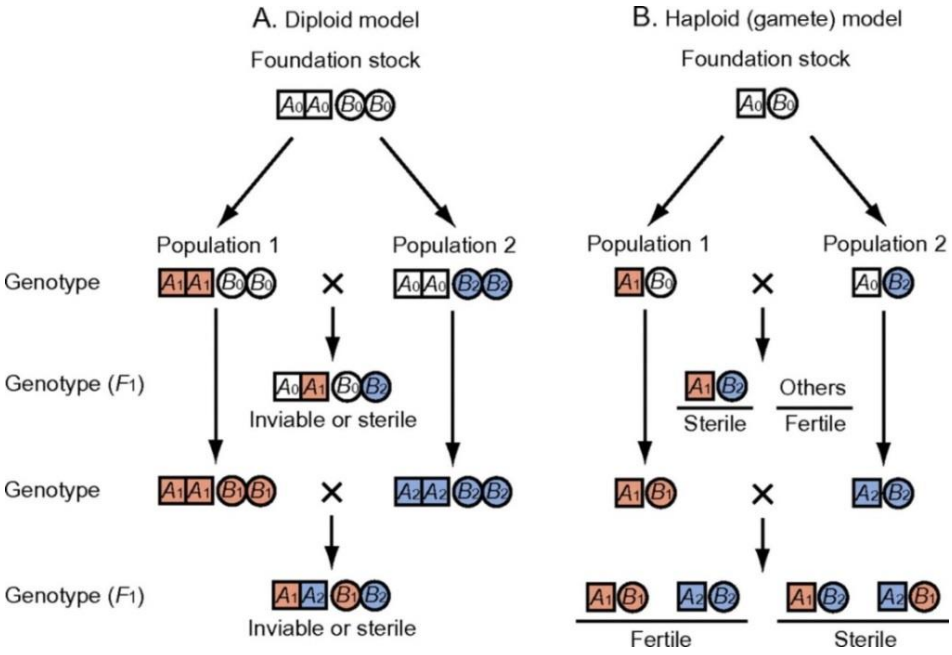


Figure 3: **Model of Dobzhansky-Muller incompatibility for the evolution of postzygotic reproductive isolation in diploid (A) and haploid organisms (B).** A new mutation appears in population 1 and 2 derived from the ancestral population. These populations evolved independently (allopatry) and the mutation can become fixed by drift and/or by the action of natural selection. These fixed alleles may become incompatible between the two genetic backgrounds and are revealed upon secondary contact. Compatibility alleles here are dominant. Note that the ancestral genotype can be $A_1A_1B_1B_1$ and unchanged in population 1 but changed to $A_2A_2B_2B_2$ in the population 2. (Taken from Nei & Nozawa, 2011).

The DM model of hybrid incompatibility (Dobzhansky 1937; Muller 1942) requires the negative epistatic interaction of two or more biallelic loci (A and B) in hybrid individuals (Fig. 3). These loci are initially neutral (or advantageous) in their respective parental background. The DM model does not require crossing of an adaptive valley or display of a deleterious mutation (Orr 1995). Let's consider an ancestral population containing A and B loci and its genotype $A_0A_0B_0B_0$. This population now diverges in two allopatric populations. In population 1, allele A_0 mutates to allele A_1, advantageous or neutral, and this mutant becomes fixed by genetic drift or selection. In population 2 allele B_0 mutates to allele B_2 that also becomes fixed. If the populations hybridize (e.g. upon secondary contact) the two alleles will be combined in the hybrids. It becomes possible then that the combination of A_1 and B_2 reduces the fitness of hybrids as compared to parental populations. In this model, mutations accumulated during population divergence interact negatively through epistatic interactions to produce hybrids of low fitness. Note that the ancestral genotype could have been $A_1A_1B_1B_1$ maintained in one population and changed to $A_2A_2B_2B_2$ in population 2. When taking into account a dominance effect between alleles hybrid breakdown can be observed in F1 (Turelli & Orr 2000). However, in diploids, hybrid depression/breakdown is generally observed more frequently in advanced hybrid generations (F2s and backcrosses) than in F1 (Edmands 1999). Recombination breaks down co-adapted gene complexes in F2s so that hybrids become homozygote for parental alleles at each locus. If the A_1 and B_2 alleles are recessive and localized on autosomes, then hybrid depression will take place only in homozygotes ($A_1A_1B_2B_2$) (F2s). If alleles are carried out by sexual chromosomes, then DMI will be stronger in the heterogametic sex (Haldane's Rule. Haldane 1922).

A central question is the **average fixation time** of A_1 and B_2 in allopatry for DMI to occur. If a neutral mutation occurs in a population, its fixation is conditioned only on genetic drift so that the waiting time (in generations) until A_0 is replaced by A_1, will be $\frac{1}{\mu} + 2N\ generations$ with μ the mutation rate and *N* the effective population size (Nei & Nozawa 2011). This time can thus be particularly long under pure drift and when the effective population size is large. If the mutation is advantageous however, the allele will invade the population more rapidly than under the sole effect of drift. When the population size is small, the waiting time until replacement is approximately the same as the time of appearance of the new mutation (Li & Nei 1977).

Mathematical models (Orr 1995; Turelli & Orr 1995; Orr & Turelli 2001) suggest that populations will diverge through a series of weak and independent mutations that should accumulate exactly with the square of time. With simple DMI involving two loci, this accumulation is quadratic, an observation called the **"snowball effect"**. Some studies have found support for this hypothesis in *Drosophila* species (Matute *et al.* 2010), *Solanum* (Moyle & Nakazato 2009) and more recently in hybrid males of *Mus* (Wang *et al.* 2015). Another mathematical model attempted to extend the Orr's model (1995). It uses gene networks for the evolution of DMI but does not predict a snowball effect (Palmer & Feldman 2009).

In addition, models suggested that **complex DMI** should be common. Complex DMI involve negative epistatic interaction at more than two loci. Similarly to the single-gene *versus* two-gene speciation case, where the path to speciation requires passing through the unfit genotype in the single-gene case, several paths exist in the two-gene case to obtain two species. The reasoning is the same as with DMI, the more loci are involved, the more paths exist to produce two divergent lineages (Coyne & Orr 2004). Empirical studies have often shown that multiple DMI contribute to male sterility in *Drosopihla* (Tao *et al.* 2003a; Masly & Presgraves 2007). However, one must clearly distinguish whether incompatibilities represent many evolutionarily independent two-locus DMI caused by distinct mechanisms or a small number of multilocus DMI with the same mechanistic basis (Maheshwari & Barbash 2011). Here also, evidence for complex DMI is supported by experiments in *Drosophila* (Davis & Wu 1996; Orr & Irving 2001; Tao *et al.* 2003b; Coyne & Orr 2004). Finally, DMI have been identified in animals and plants diverging in allopatry (reviewed in Lowry *et al.* 2008; Presgraves 2010) but one can ask what is their ability to maintain separate taxa in other geographical contexts?

Model of evolution of postzygotic isolation in parapatry

The simple DM model with two bi-allelic loci has been used by Bank *et al.* (2012) in a parapatric case under an island continent model. They showed that the maximum rate of gene flow is limited by exogenous selection. When gene flow occurs and DMI are neutral, they cannot be maintained. On the other hand when selection occurs, DMI can evolve either by selection against immigrants or by selection against hybrids. In the first case, this involves exogenous selection on one locus involved in the DMI. In the second case, this involves endogenous selection due to incompatibility of the two parental genetic backgrounds. The authors also showed that the genetic architecture maximizing gene flow supported by a

DMI was not the same in the two cases. In selection against immigrants, tightly linked DMI of any strength are favoured. Selection against hybrids favours the evolution of strong unlinked DMIs. In addition, if selection act against hybrids and the environment is homogenous, the order of mutations is important (mutation order speciation).

An older model was that of Navarro & Barton (2003a) who proposed that chromosomal rearrangements favour the evolution of RI between species, in the presence of gene flow when DM incompatibilities accumulate within these rearrangements. In line with their model, the authors have shown that differentiation between human and chimpanzee was twice as high in genes mapped in rearranged chromosomes compared to collinear chromosomes (Navarro & Barton 2003b). However, their results (not the model) were largely controversial, as one study did not find such differences (Zhang *et al.* 2004) and technical criticism of the approach of Navarro & Barton was made by others (e.g. Lu *et al.* 2003). Finally, one of the conclusions about the debate following these studies was that finding evidence for or against parapatric speciation "remains a fascinating but elusive goal" (Lu *et al.* 2003; Navarro *et al.* 2003).

Model of evolution in sympatry

Evolution of RI in sympatry requires individuals to adapt to divergent habitats within the same area and the underlying genetic polymorphisms to be different in each habitat. The evolution of RI through divergent ecological selection also requires the linkage disequilibrium between genes involved in postzygotic isolation and those involved in prezygotic isolation. The main difficulty is that recombination will break up the coupling between these postzygotic and prezygotic isolating factors (genes) at each generation (Felsenstein 1981). To understand how coupling can be maintained, Felsenstein (1981) proposed two models: a one allele and a 2 allele model. In the one allele model, a new allele becomes fixed in all populations and allows individuals to recognize each other and to mate preferentially with the same adapted genotype. In this model, homogamy is controlled by a single locus. In the 2 allele model, two alleles (A_1 and A_2) have to be fixed in the populations. These alleles will also favor homogamy so that individuals fixed for A_1 will mate together and those fixed for A_2 will mate together. In this model, divergence is possible only if the 2 alleles are associated differentially to habitat choice, which requires strong linkage disequilibrium. The second model is generally considered as more realistic (Felsenstein 1981) but some authors considered the first model more reasonable (Kirkpatrick & Ravigné 2002). Indeed, recombination does not play a homogenizing role in this case, so that RI can evolve more easily. Under the 2 allele model another way to establish associations between prezygotic and postzygotic factors is through pleiotropic interactions between alleles of habitat choice and of local adaptation (Rice 1984; Doebeli 1996). Such pleiotropic traits were called "magic traits'" (Servedio *et al.* 2011). One example of this kind of interaction may be the threespine stickleback (Nagel & Schluter 1998). However few other convincing examples exist in nature (Servedio & Noor 2003).

There are many other quantitative models of sympatric speciation. They usually rely on the variance of some quantitative traits that will promote variance in resource use through competition

between individuals within a sympatric population (Dieckmann & Doebeli 1999). In this model, frequency dependent selection and disruptive selection favor extreme individuals, consuming unexploited resources. In these conditions, assortative mating may lead to RI between ecologically diverging subpopulations. In the model, if assortative mating depends on a trait unlinked to resource use, genetic drift is necessary to break linkage equilibrium between the assortative trait and the trait for resource use. This model was criticized since during resource use, competition for extreme resources arises (e.g. competition for small and big seeds) leaving more intermediate resources and then favoring intermediate individuals.

Finally although theory predicts that sympatric speciation may occur, model assumptions are numerous and their empirical validity remains contentious (Bolnick & Fitzpatrick 2007). Today, few convincing examples of sympatric speciation exist as the magnitude of RI between putatively incipient species is generally low (e.g. in *Rhagoletis* (Feder *et al.* 2005) or *Tinema* (Soria-Carrasco *et al.* 2014)) suggesting that local adaptation, rather than speciation, is at play, or more simply because the null hypothesis of build-up of RI in allopatry has not been tested or could not be rejected. For instance, in *Rhagoletis* it is likely that the divergence was initiated in allopatry, with one race having diverged in North America and the other in Mexico, the latter having accumulated chromosomal rearrangements that have been maintained upon secondary contact in the United States.

1.2.2.2 The genes of post-zygotic isolation

The identification of genes involved in RI, their number, effect size, pleiotropic and epistatic effects is a central question in speciation (Orr 2005). Population genetic theory predicts that many alleles of small effects should lead to adaptation of populations in different environments (Fisher 1930), a view defended by others (Pritchard *et al.* 2010; Hancock *et al.* 2010; Pritchard & Di Rienzo 2010; Rockman 2012). Regarding speciation, several studies have suggested that a large number of loci were involved. Studies of hybrid zones for instance suggested that 50 loci explained the maintenance of RI between *Bombina bombina* and *B. variegata* while a minimum of 150 loci could be involved in the isolation of races of *Podisma pedestris.* More recent studies have shown that many genes with moderate effects were involved in hybrid breakdown of *Arabidopsis* complex (Bomblies & Weigel 2010). In the lake whitefish complex *Coregonus clupeaformis*, it seems that a polygenic basis is also involved (Rogers & Bernatchez 2007; Bernatchez *et al.* 2010; Gagnaire *et al.* 2013a). Finally, reviews demonstrated a substantial variability in the number of genes implied in RI (Allen Orr 2001; Coyne & Orr 2004). As stated by Coyne & Orr (2004) it is probable that many genes contribute to the total post-zygotic isolation. However, the number of genes that initiate RI may be much lower (Seehausen *et al.* 2014). Indeed according to Seehausen *et al.* (2014) when taking into account standing genetic variation, gene flow or changing environment, a few number of genes of large effects may be sufficient. In that sense a simulation study has showed that a concentrated architecture containing fewer, larger, and more tightly linked divergent alleles favored adaptation under migration-selection-drift balance (Yeaman & Whitlock 2011; Yeaman & Otto 2011). Thus, finding QTLs of large effect composed of many tightly linked alleles of smaller effect can be facilitated in these conditions. However, adaptation is not speciation, and it is not clear whether such settings will facilitate divergence.

1.2.2.3 Exogenous barriers

Exogenous barriers imply that hybrids from differentially adapted populations have a smaller fitness than parental populations in their respective environments. For example the sympatric butterflies *Heliconius melpomene* and *H. cnydo* in Central America are closely related and display differences in wing colour. One is black, red and yellow coloured while the other is black and white. When hybrids are formed in the wild they have a lower fitness than the parental species (Jiggins *et al.* 2001). Another example is that of the benthic/limnetic ecotypes of *G. aculeatus* in British Columbia (Hatfield & Schluter 1999; Rundle *et al.* 2000; Rundle 2002). The two groups differ in their habitat use: the limnetic group eats plankton in open water and displays a smaller body with a narrow jaw gape whereas the benthic morph feeds on invertebrates in the littoral zone and has a larger body with a wide jaw gape. F1 crosses obtained in laboratory conditions were fit while they performed poorly in the wild environment.

1.2.2.4 Coupling of barriers to gene flow

Coupling of endogenous barriers to gene flow is a well-known process in the hybrid zone literature (Barton 1983; Barton & Hewitt 1985; Kruuk *et al.* 1999). It is well established that when selection affects many loci, coupling will occur between these loci and the strength of the barriers to gene flow will be strongest (dependent on selection strength, recombination and the number of loci (Barton 1983)). One recent paper extended Felsenstein's (1981) two allele model which focused only on prezygotic isolation, to any number of incompatibilities and showed how coupling of any kind of barrier, either prezygotic or postzygotic, can emerge to maintain strong RI within a single population (Barton & de Cara 2009). In their model the authors showed that coupling of multiple (single-locus or multi-locus) incompatibilities increase mean fitness in case of positive epistasis. In addition, they showed that single locus incompatibilities involved in assortative mating can be coupled with loci reducing hybrid fitness, hence contributing to the theory of reinforcement (Servedio & Noor 2003).

Another question is the coupling between *exogenous* barrier and *endogenous* barrier (Moore & Price, 1993). Theoretical developments in hybrid zones have already shown that endogenous barriers were preponderant and formed "tension zones[1]", independent of the environment (Barton & Hewitt 1985) and this was also applicable for multi-locus barriers (Kruuk *et al.* 1999). It was later shown that these tension zones were often trapped by environmental barriers (Barton 1979; Barton & Hewitt 1985), meaning that DMI and signatures of ecological adaptation may be confounded. A recent study has investigated in detail how such coupling can result in genetic-environmental-associations (GEA) especially in heterogeneous environments (Bierne *et al.* 2011).

2 Modelling gene flow across space and time

[1] Tension zones: cline maintained by a balance between random dispersal and selection against hybrids (Barton et Hewitt, 1985)

Understanding which barriers reduce gene flow and promote speciation is crucial to our understanding of the speciation process, especially under the BSC. However, within the last two decades, the development of the coalescent theory, together with new methods to model gene flow have allowed us to further our understanding of how speciation proceeds through time and space, and also across the genome. Understanding the history of speciation across space and time is fundamental to draw general inferences about the geographical settings under which speciation can happen and determine the quantity of gene flow necessary to impede divergence. In addition, a number of studies focus on detecting positive selection at the genome scale. However, disentangling the effects of selection from those of other factors, including historical demographic processes and other non-selective forces that act together to leave their footprints along the genome, remains a challenging task.

2.1 The history of speciation

Studying the historical processes that have shaped the current distribution of life on earth lies at the heart of phylogeography (Avise, 2009; Hewitt 2011). Phylogeography has greatly improved our understanding of the evolutionary processes involved in species diversification through the study of expansion-fragmentation and colonization events on patterns of genetic differentiation (Hewitt 1996, 2011; Taberlet *et al.* 1998). Understanding the consequences of the climatic oscillations during ice ages on the spatial distribution of genetic diversity on earth was a crucial step (Hewitt 1996, 2004). Throughout historical times, populations have undergone spatial range expansions and contractions during which they were connected or separated so that gene flow has varied. For instance, the Pleistocene glaciations, which lasted approximately two million years and terminated around 10 000 years ago with a period of global warming (Hewitt 1996), were suggested to be one of the most important historical events involved in shaping the large-scale population structure and genetic differentiation of contemporary species (Bernatchez & Wilson 1998; Taberlet *et al.* 1998). During Pleistocene glaciations, ice sheets restricted populations to southern glacial refuges (Hewitt 1996, 1999). Following glacial retreats, the recolonization of suitable northern areas frequently involved a few founder individuals, leading to a reduction in genetic diversity at the front of colonization (Hewitt 1996). Differentiated populations then had large opportunities to interbreed in areas of secondary contact zones (or hybrid zones). These hybrid zones formed “windows on evolution” (Hewitt 2011) and have provided great insights into the evolutionary process of speciation (Barton & Hewitt 1985). Phylogeographic studies have allowed us to trace the movement of species from refugial areas into previously glaciated regions and to identify the main glacial refugia and sutures zones from Europe in Italy, the Balkans, and the Iberian Peninsula, where several species were known to occur (Taberlet *et al.* 1998; Hewitt 2000, 2004). However, finer investigations of refugial areas depicted a more complex story and some studies have identified the English Channel area as a potential refugium for some species (eg. Finnegan *et al.* 2013). Overall, reconstructing the history of species divergence implies understanding in which areas and for how long populations and species were connected by gene flow or were separated. Phylogeographic methods do not allow such inference but new methods based on the

coalescent theory allow us to to draw more complex and realistic inferences (Beaumont *et al.* 2002; Beaumont 2010; Csilléry *et al.* 2010).

In addition, new genomic data gathered across a wide range of species allowed addressing with greater accuracy a basic question in speciation research: what is the average time to speciation (TTS)? A recent study found that the average TTS in plants and animals was approximately 2 million years (Hedges *et al.* 2015). Based on these results the authors suggested that the accumulation of genetic incompatibilities proceeded mainly by neutral processes and that adaptive changes were almost decoupled from the speciation process. The authors subsequently suggested that described species separated by only tens of thousands of years are not real species. Similarly, species divergence in hybrid zones was estimated to be at least 1 million years, and was likely underestimated due to hybridization (Barton & Hewitt 1989). In addition, estimations of the age of adaptive alleles in some species (Colosimo *et al.* 2005; Brawand *et al.* 2014) suggested that they predated glacial retreats. These estimates are in contrast with the current literature on ecological speciation (Nosil 2012) and speciation with gene flow (Smadja & Butlin 2011) which suggests that natural selection can rapidly promote the evolution of RI. This can be explained by the fact that speciation rates are in fact decoupled from RI (Rabosky 2013; Rabosky & Matute 2013). To reconcile these contradictory results, Rosenblum *et al.* (2012) introduced the notion of ephemeral species for groups of populations which may finally never reach the status of isolated species. Finally, to disentangle the role of gene flow, ecology and historical processes on species formation, more complex scenarios of divergence must be contrasted. These considerations are fundamental for our focal lamprey species pair, for which the level of reproductive isolation and divergence time are not well known.

2.2 New methods to infer the history of speciation

The development of coalescent models based on gene genealogy allows a better estimation of population genetic parameters such as effective population size, migration rate and timing of divergence between populations (Pinho & Hey 2010). However, a first prerequisite before estimating demographic parameters is to obtain a null realistic model of divergence for the studied populations. Fitting phylogenetic tree-based models allows drawing explicit inferences about history. However, these models assume a simple bifurcating tree with no subsequent gene flow, which may be incorrect when populations are connected by gene flow (Edwards 2009). Solutions for this issue were proposed by Pickrell & Pritchard (2012) and Gautier & Vitalis (2013). The first model is an extension of Cavalli-Sforza and Edward approaches that estimates allelic frequencies based on a multivariate Gaussian model (Pickrell & Pritchard, 2012). Migration allows for population split and mixture among multiple populations and is represented as edges along a graph instead of a tree. However, divergence estimates based on a Gaussian model may be reliable only for recent divergence times. The method of Gauthier & Vitalis (2013) relies on a diffusion process (forward in time) in contrast to most phylogenetic approaches and also allows handling several populations at a time. However, the method of Pickrell & Pritchard (2012) does not allow explicit comparisons of

alternative scenarios and works better for full genome sequences with outgroup data. Overall the two methods do not allow estimating parameters such as effective population size or migration rates.

Coalescent methods such as the one developed by Nielsen & Wakeley (2001) and Hey (2010) allow estimating demographic parameters, but assume either continuous gene flow (i.e. the Isolation with Migration model) or no gene flow. In addition, such models use a full likelihood approach, which is computationally intensive. Other recently developed methods used Hidden Markov Model (HMM) and full genome data to reconstruct the history of divergence. However, these methods are often limited to a few genomes or populations (eg. PSMC, Li & Durbin 2011). Methods based on admixture fraction (Liang & Nielsen 2014b) or admixture tract length (eg. Harris & Nielsen 2013; Liang & Nielsen 2014a; Sedghifar *et al.* 2015) are currently under development in human populations and should be soon applicable to natural populations, as full genome data become available. On the other hand, new methods such as Approximate Bayesian Computation (ABC) bypass the need to compute likelihoods (Csilléry *et al.* 2010) by making use of simulated data under a given historical scenario. The method simply compares a set of observed summary statistics with a set of simulated statistics under different scenarios (Beaumont *et al.* 2002; Beaumont 2010). These methods have been widely developed in recent years and used to infer the history of species or population divergence (Ross-Ibarra *et al.* 2008, 2009; Duvaux *et al.* 2011; Pettengill & Moeller 2012; Roux *et al.* 2013, 2014). ABC methods work as follows: first a set of different models are simulated based on prior distributions of a set of parameters corresponding to the tested model. Typically a set of around one million datasets (i) are computed under each model. Then for each simulation a set of summary statistics S(i) are computed and compared to the observed set of summary statistics (So) based on a distance, such as the Euclidian distance. When the distance is below a tolerance threshold (δ), the parameter value is accepted. Parameter values can then be adjusted by local linear regressions, logistic regressions or neural network layers, giving more weight to the simulated summary statistics that are closest to the observed data and allowing a posterior distribution then to be drawn. Ultimately, model selection is performed based on posterior probabilities and Bayes factor.

Other methods use information from the site frequency spectrum (SFS, the number of derived alleles within a population), which constitutes a full summary of the data and implies no information loss (Gutenkunst *et al.* 2009). When a pair of populations is compared under a given divergence scenario, then the Joint Site Frequency Spectrum (JSFS) is used. In general, data from an outgroup species is needed to polarize the JSFS (i.e. to determine proportions of shared ancestral alleles versus proportions of shared derived alleles). JSFS can be simulated under a given demographic scenario using diffusion[2] approximation to the one-locus, two-allele, Fisher-Wright Model (Gutenkunst *et al.* 2009). The method of Gutenkunst *et al.* (2009) is primarily designed to work on few populations (one, two or three) with stable allele frequency changes at each generation and assumes independence of the polymorphisms, otherwise, the likelihood becomes a composite likelihood and bootstrap methods must be used to validate the accuracy of estimates

[2] Diffusion process: A continuous-time stochastic process that tracks a quantity that changes continuously in time and whose future depends only on the present state.

(Gutenkunst *et al.* 2009). The method has been improved, to compare more complex divergence scenarios, taking into account the heterogeneity of migration rates along the genome and by improving the exploration of the likelihood landscape through the use of simulated annealing methods in addition to the Broyden-Fletcher-Goldfarb-Shanno (BGFS) (Tine *et al.* 2014). These methods, combined with genomic data, allow exploring the heterogeneity of migration rates along the genome in line with the view of semi-permeable barriers to gene flow (Harrison, 1989, Wu 2001; Harrison 2012; Harrison & Larson 2014) (**box 1)**.

Box 1. The Semi Permeable nature of barrier to gene flow

So far we have focused on Mayr's BSC which is a "whole genome concept". The varying degree of RI observed across the genome has led to the emergence of the genic view of speciation (Fig 4). This view is directly linked to the fact that reproductive barriers are generally semipermeable due to variations in recombination rates and selection (Harrison, 1986). As a consequence the migration rate is heterogeneous across the genome, with some regions of strong differentiation harboring barrier loci and linked neutral alleles for which introgression into the genome of the foreign population is delayed in proportion to their linkage. Revealing which barriers are impermeable to gene flow is best performed in hybrid zones where only loci involved in RI will resist the homogenizing effect of gene flow. One important question to address will be whether the hybrid zone arises from primary differentiation (where two populations diverge in the face of gene flow) or from secondary contact between formerly isolated populations. 37% of the 106 hybrids zones in Barton & Hewitt (1985) are likely secondary contact zones. More recently the term "genomic island of speciation" was proposed (Turner *et al.* 2005) and a theory of divergent hitchhiking during speciation in the face of gene flow was developed (Feder & Nosil 2010; Feder *et al.* 2012c; Nosil & Feder 2012). However, the role of islands of differentiation during divergence is currently debated (see section 3.).

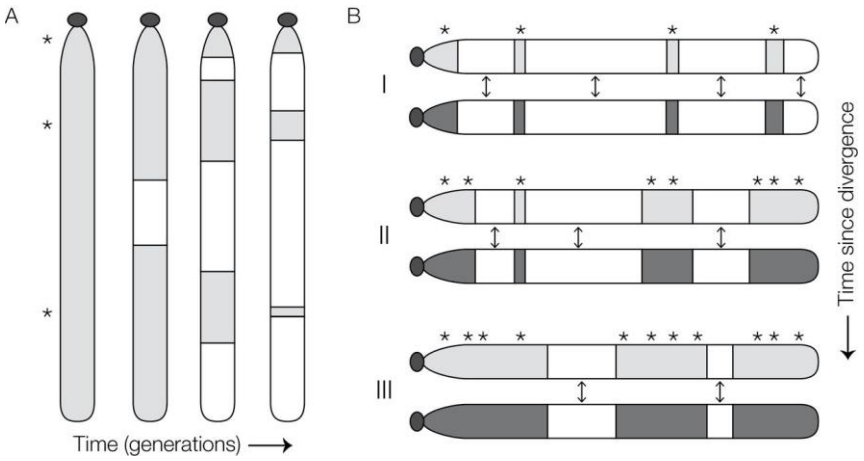


Figure 4: **Illustrations of the semipermeable nature of species boundaries**. (A) Gene flow following secondary contact as depicted by Barton and Hewitt (1981). The vertical bars represent a chromosome with 3 loci contributing to reproductive barriers (indicated by *). Immediately after contact, linkage disequilibrium along the chromosome will be high, but over time recombination breaks down associations among loci. Barrier genes or genes under divergent selection will remain differentiated (light grey regions), whereas alleles at loci that are neutral (white regions) will be exchanged between species. Many generations of recombination in hybrid zones allow fine-scale mapping of genes contributing to reproductive isolation and estimates of the strength of selection on individual alleles. (**B**) The idea that the genomes of diverging lineages become less permeable over time, as shown by Wu (2001). Each pair of horizontal bars represents chromosomes of 2 diverging lineages. Very recently diverged species (pair I) may have few genes contributing to reproductive isolation (indicated by *). These regions will remain differentiated (light and dark grey

regions), whereas gene exchanges can occur in other parts of the genome (white regions). With increasing genetic divergence (chromosome pairs II and III), more loci contribute to reproductive barriers, thus restricting gene flow on a larger proportion of the genome (from Harrison & Larsson 2014).

2.3 Heterogeneous genomic divergence

Recent genome wide studies have documented heterogeneous genomic divergence, corroborating the idea of barrier permeability. Some studies revealed a few regions of large size (Turner *et al.* 2005; Jones *et al.* 2012; Via *et al.* 2012; Martin *et al.* 2013) whereas others have identified multiple regions of smaller size spread across the genome (Michel *et al.* 2010; Renaut *et al.* 2013; Burri *et al.* 2015). These observations have led to the development of a verbal theory of "divergence hitchhiking" facilitating divergence with gene-flow (Via & West 2008; Via 2012). This theory states that divergent selection and non-random mating will reduce recombination in the face of gene flow and generate large islands of differentiation. The four steps of the model are shown in Fig 5 from (Feder *et al.* 2012a).

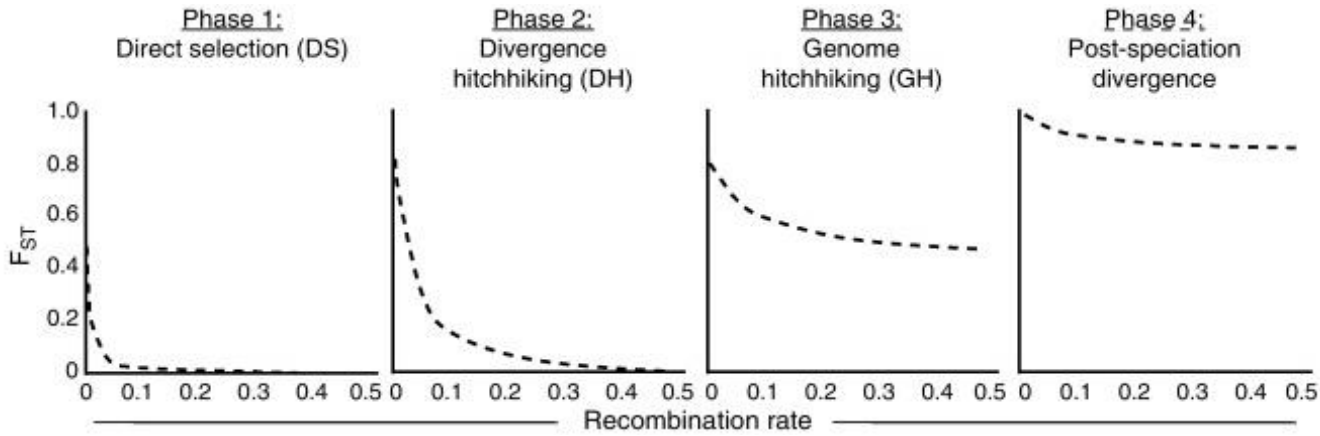


Figure 5: **The four potential phases of speciation-with-gene-flow involving differences in the relative importance of DS, DH, and GH**. Plots depict the general expected relationship of divergence (FST) for a neutral site at varying recombination rates ranging from *r* = 0 cM (completely linked) to *r* = 0.5 (unlinked) to a divergently selected locus, as speciation proceeds through the four phases. Taken from Feder *et al.* (2012).

However, the formalization of the model and simulation studies have shown that conditions for such genomic islands to be generated were restricted (Feder & Nosil 2010; Feder *et al.* 2012c; Flaxman *et al.* 2014). In addition the historical demography has rarely been investigated in empirical studies (Harrison & Larson 2014) and one cannot exclude that similar patterns would arise from secondary contacts. Most importantly, the role of genomic islands in speciation has been questioned for several reasons (Noor & Bennett 2009; Turner & Hahn 2010). First, most empirical studies have relied on relative measures of divergence (mainly F_{ST}) to identify genomic islands. However F_{ST} is known to be dependent upon both divergence between and variations within populations so that inflated F_{ST} may arise because of reduced nucleotide diversity (Charlesworth 1998; Charlesworth & Campos 2014) or other processes such as reduced recombination rates (Noor & Bennett 2009). Accordingly, recent, empirical studies have shown that genomic islands often occur in regions of low recombination in stickleback (Roesti *et al.* 2013), sunflower (Renaut *et al.* 2013) and *Ficedula* flycatchers (Burri *et al.* 2015). Second, in a recent review Cruickshank & Hahn (2014) proposed the use of estimators of absolute divergence, such as the D_{XY}, jointly with measures

of F_{ST} .The D_{XY} (Nei & Li 1979) measures the average number of pairwise differences between sequences from two populations but excludes all comparisons between sequences within populations (so it is not affected by current levels of within population diversity). $D_{XY} = \sum_{ij} x_i y_j d_{ij}$ with X and Y the two populations and d_{ij} the number of nucleotide differences between the ith haplotype from X and j^{th} haplotype from Y. If genomic islands are indeed caused by barrier permeability to gene flow, then both F_{ST} and D_{XY} should be higher than the neutral background. Alternatively, if D_{XY} in islands is similar to the measure in neutral regions, then the authors suggest that this is not due to variable levels of ongoing gene flow but rather to variations in local genomic architecture and suggest a model of postspeciation selection at linked sites (Cruickshank & Hahn 2014). Applying their model to a series of well described organisms i.e. *Mus* (Geraldes *et al.* 2011), *Ficedula* (Ellegren *et al.* 2012), *Anopheles* (Turner *et al.* 2005) *Oryctolagus* (Carneiro *et al.* 2010) and *Heliconius* (Nadeau *et al.* 2012) they showed that these species were not experiencing gene flow. As illustrated in Fig 6. this prediction was also recently validated in *Ficedula* flycatchers (Burri *et al.* 2015).

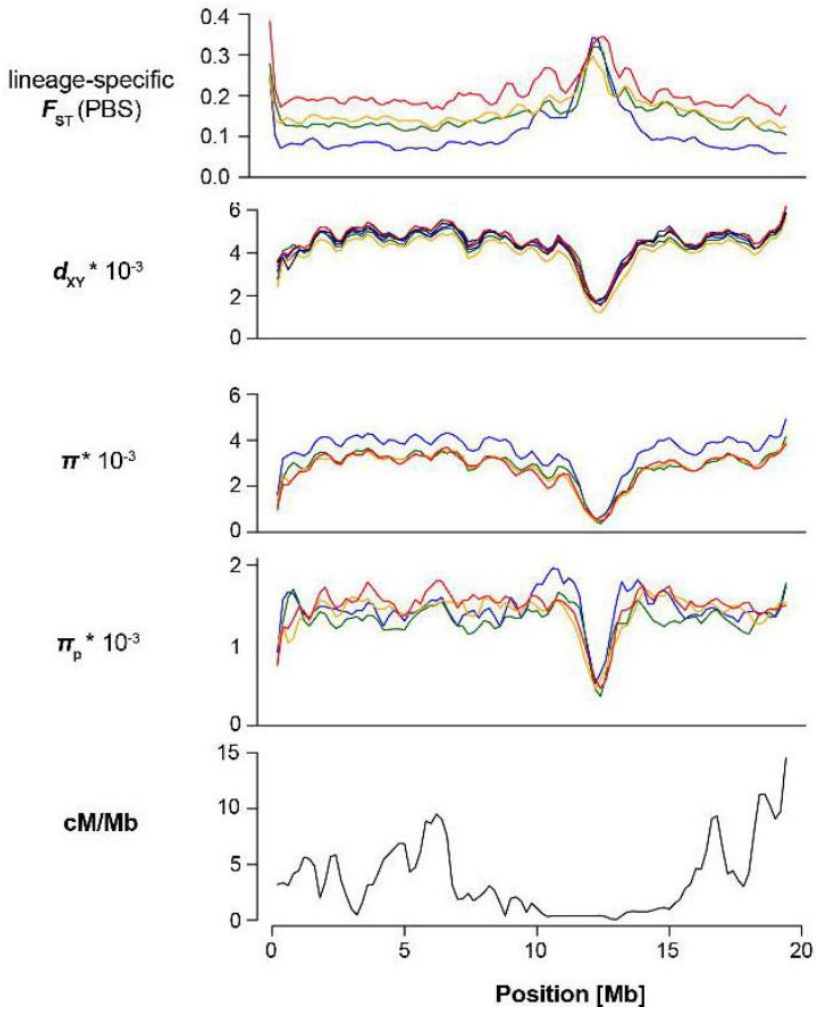


Figure 6: **Example of population genomic parameters along a chromosome from different flycatcher species** (Burri *et al*. 2015). Blue: collared; green: pied; orange: Atlas; red: semicollared. Color codes for dxy: green, collared-pied; light blue, collared-Atlas; blue, collared-semicollared; orange, pied-Atlas; red, pied-semicollared; black, Atlas-semicollared. Differentiation islands clearly occurred in regions of low recombination. Dxy is reduced in differentiation islands showing the reduction of Ne *via* Hill-Robertson Interference.

3 Studying speciation: selection *vs* endogenous barriers

With the recent advent of Next Generation Sequencing technologies (Mardis 2008) which allow for genome-wide studies in non-model organisms, several new questions can now be addressed in speciation research, notably the respective roles of ecological divergence and endogenous barriers in genomic divergence. In this section, I will first recall some of the footprints of selection left across the genome and how to detect them with genome scans, and then discuss some current debates in speciation genomics.

3.1 Genetic hitchhiking, hard and soft sweeps

When a single new adaptive mutation occurs with a selective advantage (s), its frequency quickly rises until fixation. Individuals carrying the mutation will be favored by selection generating a selective sweep, the so-called "hard"[3] sweep. This allele frequency shift leads to a similar shift in allele frequency of loci in close proximity to the selected locus, a process called genetic hitchhiking (Smith & Haigh 1974). Under a strong selective sweep, a local loss in genetic diversity occurs in the neighborhood of the hitchhiker alleles, so that almost only one haplotype remains. The selective sweep size will be enhanced by selection that will increase the frequency of all alleles in the same neighborhood, but it will be eroded by recombination and this erosion will increase as a function of time (Kim & Stephan 2002). Hitchhiking is supposed to be efficient regardless of effective population size (Gillespie 2000, 2001). In the classical model of local hitchhiking, genetic differentiation decreases as a function of the distance from the hitchhiker locus (Fig 7a,b)(Charlesworth *et al.* 1997a). This corresponds to the case where a mutation appears favorable in its deme but unfavorable in another deme. This model implies a strong genetic differentiation between populations at the (selected) hitchhiker locus and other loci in its neighboring environment. In a second model, a mutation can be favorable globally, that is to say in two structured populations, and its classical signature implies two peaks of differentiation on each side of the selected locus (Fig7 c, d). This model of global hitchhiking in structured populations implies that the intensity of the sweep will be smoother than in the first model (details in Bierne 2010). Such patterns may lead to false inferences of genome scan data in search of footprints of ecological selection, as will be discussed below.

[3] There is currently a debate on the role of "hard sweeps" (adaptation from a single beneficial mutation) *versus* "soft sweeps" (adaptation from multiple beneficial mutation due to mutation or migration) (Hermisson & Pennings 2005). Some argue that the probability of a soft sweep to occur is very low (Jensen 2014) while others have proposed that adaptation from soft sweeps should be easier (Messer & Petrov 2013). Empirical evidence suggests that distinguishing a "soft sweep" from a "hard sweep" is difficult (e.g. Schrider *et al.* 2015). In addition, most tools were initially designed to identify "hard sweep" patterns. Investigations on this debate are clearly beyond the scope of this thesis, however one should bear in mind that this is *another* confounding factor for which further investigations are certainly required in the future. To add to complexity, partial sweeps may also occur: either the beneficial mutation is still increasing in frequency and has not reached fixation (the sweep is "in action") or the initially beneficial sweep started to increase in frequency and then lost its selective advantage so began to drift. A potential alternative scenario to the many partial sweeps inferred from common methods (e.g. iHs) is that they could be false positive located in the "shoulders" (Schrider et al. 2015) of a true hard sweep that is fixed (P.A. Gagnaire, personal communication).

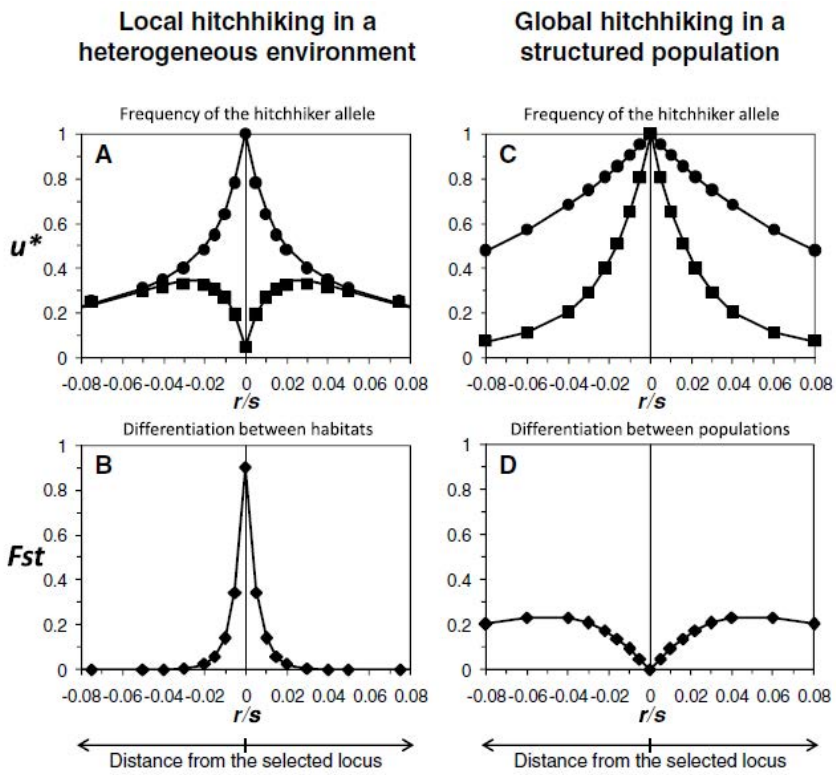


Figure 7: **The distinct signatures of local hitchhiking in a heterogeneous environment and global hitchhiking in a structured population**. Distance (x-axis) is represented as the ratio between the recombination rate and the selection coefficient (r/s). (A) Local hitchhiking: if an advantageous mutation appears in habitat 1 (circles), it rises in frequency (u*) sweeping its genomic background (20 000 generations simulated). Circles depict frequency in habitat 2. (B) Local hitchhiking producing the peak of genetic differentiation at the hitchhiker allele between the two habitats (filled diamonds) after 20,000 generations. Differentiation decreases as a function of the distance from the site. (C) Global hitchhiking: frequency of an initially rare neutral allele that hitchhiked with an unconditionally favorable mutation, in the population in which the favorable mutation originated and in a population reached by the favorable mutation by migration. (D) Global hitchhiking: genetic differentiation between the two populations. Two domes of differentiation are observed on each side of the advantageous mutations together with the signature of a selective sweep of various intensity. The intensity of the sweep decreases as a function of LD with the selected locus. Taken from Bierne (2010).

3.2 Genome scans of adaptation

A series of techniques to detect the classical signature of a hard sweep have been developed. They rely on the SFS (Tajima 1989; Kim & Stephan 2002), linkage disequilibrium or most commonly on patterns of decay of homozygosity along a haplotype as a function of recombination (e.g. Sabeti *et al.* 2002; Voight *et al.* 2006 ; Gautier & Vitalis 2012; Ferrer-Admetlla *et al.* 2014). Another class of methods relies on patterns of genomic differentiation (F_{ST}) across a large number of markers and search for those with a stronger differentiation ("outliers") as compared to neutral expectations. Such "genome scans" may be useful to find genomic regions contributing to RI between populations. This method was first proposed by Lewontin & Krakauer (1973) (LK test) and tests the heterogeneity of F_{ST} between loci. The central assumption is that if a few markers depart from a neutral distribution obtained under an island model, then these loci may be under disruptive selection. The LK test was criticized since it assumed no influence of the structure between populations on the variance of F_{ST}, and did not take into account the variance in

mutation rates or in population size (Nei & Maruyama 1975; Robertson 1975; Meirmans 2012; De Mita *et al.* 2013). The method was later shown to be robust to different population structures and was then implemented under the infinite island model assuming equal sizes and equal migration rates between islands (Beaumont & Nichols 1996). Vitalis *et al.* (2001) then developed a method assuming an instantaneous split between population pairs of constant size without gene flow. Major improvement were then made by Foll & Gaggiotti (2008) who implemented a Bayesian approach to solve the among deme allele frequencies correlation problem. More specifically, assuming an island model, subpopulation allele frequencies can be correlated through a common migrant gene pool (common ancestor), from which they will differ through the use of a multinomial Dirichlet distribution. Other methods were then developed to take into account more complex demographic scenarios or the hierarchical structure occurring among populations (Bonhomme *et al.* 2010; Günther & Coop 2013; Vitalis *et al.* 2014). Another class of methods aiming at detecting selection using correlations of ecological/environmental data with genetic datasets are also increasingly developed (Coop *et al.* 2010; Frichot *et al.* 2013; Duforet-Frebourg *et al.* 2015).The two last methods are based on principal components or 'latent factors' to correct for population structure.

Genome scans generally display high rates of false positives due to several factors (reviewed in Bierne *et al.* 2011). First, populations are often spatially structured, and not correcting for the higher variance in F_{ST} is known to lead to false inferences (Robertson 1975; Bonhomme *et al.* 2010). Bonhomme *et al.* (2010) introduced an extension of the LK test that corrects for genealogical relationship in structured populations. Similarly Günther & Coop (2013) standardized the F_{ST} by the covariance among populations. A recent review has shown that these two methods were the less biased under a pattern of isolation by distance or range expansion, whereas methods such as Bayescan (Foll & Gaggiotti 2008) and FDIST2 (Beaumont & Nichols 1996) displayed higher rates of false positives (Lotterhos & Whitlock 2014). Indeed, a second potential great concern for the focal species of this thesis is the high rate of false positives that occurs in fractal networks such as hydrographic network. A high rate of outliers is generally observed in these systems, where departures from the one or two dimensional stepping stone model can be very strong (Fourcade *et al.* 2013). Fragmented populations in these landscapes can show strong departures from demographic equilibrium, with founding events and fixations events (e.g. Perrier *et al.* 2013) potentially leading to false inferences. A third potential bias is that of allele surfing, which is expected during population expansion, when a few founder individuals colonize a new habitat. A neutral mutation arising at the front could quickly spread in frequency, generating a pattern similar to that observed under hitchhiking (Klopfstein *et al.* 2006). Fourth, background selection against deleterious mutations can reduce effective population size in subpopulations hence contribute to increase genetic differentiation (Charlesworth *et al.* 1997a). Ultimately, the coupling of endogenous barriers (see 1.2.3) (Bierne *et al.* 2011) creates tension zones expected to coincide with exogenous barriers through time and to be trapped at environmental boundaries (Barton & Hewitt 1985) forming Genetic Environmental Association (GEA). These GEAs may be easily and spuriously detected by genome scans.

Overall, to circumvent these problems, one possible solution is to jointly infer demography and selection (Li *et al.* 2012; Singh *et al.* 2013; Bank *et al.* 2014). Inferring a neutral demographic model allows drawing subsequent inferences about the proportion of loci departing from the neutral envelope of that model (cf Chapter 4). In addition, comparisons between replicate pairs of populations should be performed to test the extent of parallelism (cf 3.4) and verify the robustness of those inferences. Ideally, inferences should be performed in hybrid zones where only reproductive barriers will resist the homogenizing effects of gene flow. Hybrid zones best highlight which genomic regions are affected by selection and which are affected by neutral processes. This is very important in species with small effective population sizes or species in fragmented habitats where founder events can play a key role in the spatial distribution of genetic diversity. Finally, several of the restrictions made above may apply to our focal species. In particular, the distribution of brook lampreys along fragmented habitats, together with the possibility of founder events may have profound impacts on genome scans. Thus, studying river and brook lampreys in sympatry in "hybrid zones" may be a better setting to perform adequate genome scans.

3.3 Polygenic adaptation

Most common traits in humans (and other model species) are known to have a polygenic basis (Hindorff *et al.* 2009; Yang *et al.* 2010). The response to selection at any locus is usually modest, displaying subtle and coordinated allele frequency changes going in the same direction (Le Corre & Kremer 2003, 2012; Kremer & Le Corre 2012). Such a pattern is not restricted to man but is also becoming increasingly recognized in other organisms (eg. *Arabidopsis thaliana* (Atwell *et al.* 2010)) and Pritchard & Di Rienzo (2010) suggested that this may be the case for many traits. Footprints of polygenic adaptation at the level of individual loci is likely to go undetected without a great amount of annotation of the sites that are the target of selection (Pritchard *et al.* 2010). In these cases, the traditional sweep framework and genome scan methods relying on F_{ST} are likely to be inadequate (Le Corre & Kremer 2012). These authors showed first that when neutral differentiation is already high, local adaptation does not necessarily further increase differentiation and thus adaptation may be undetected. Second, when increased covariance of allelic effects (under the infinitesimal model) is involved to reach adaptation, then genome scans will be inefficient. This is a possible scenario under high gene flow and recent selection involving a high number of loci. Le Corre & Kremer (2012) proposed to shift to a quantitative framework, combining Q_{ST}-F_{ST} analyses, together with the integration of the covariance of allele frequencies and the covariance of allelic effects. Another interesting approach was that of Renaut *et al.* (2011) who combined phenotypic, transcriptomic and functional analysis to identify the possible genes involved in speciation in the *Coregonus clupeaformis* complex. Recently Berg & Coop (2014) developed a model based on Genome Wide Assocation Study (GWAS) data to detect polygenic adaptation. They used allele frequency data to estimate the mean additive genetic value of a phenotype in any number of populations. They then developed a neutral model accounting for drift, population history and population structure between populations. From this model, they developed methods to detect 1) genetic-environmental correlations, 2) over-dispersion of genetic values among populations based on the Q_{ST}-F_{ST} generalization and 3) individual populations that contribute

to this signal. To conclude, it is clear that genome scans are just a first preliminary step that needs to be combined with other approaches to detect more subtle signals. Then functional analyses should be performed to dissect in detail the nature and effects of candidate loci on phenotypes.

The current literature is full of examples of genome-wide analysis reporting on average 5-10% of the loci as “outliers” and attributing this result to local adaptation (review in Nosil *et al.* 2009a). However, the probability of finding loci or genes that are direct targets of selection is known to be very low hence many false positives are probably detected by genome scans (Tiffin & Ross-Ibarra 2014). Nevertheless, when done carefully, genome scans can provide great results. For instance, Fournier-level *et al.* (2011) used common garden experiments combined with a powerful GWAS analysis. They screened diverse ecotypes from different regions of Europe genotyped at 213,248 SNPs to test for climate adaptation in *Arabidopsis thaliana* and found 0.002% of SNPs putatively involved in local adaptation to climate. More recently, Barson *et al.* (2015) combined GWAS, genome scans and genome resequencing in *Salmo salar* to identify the genetic basis of age at maturity They identified one large effect locus (vestigial like family member 3, VGLL3) explaining nearly 40% of phenotypic variation in this trait.

4 Insight from studies of parallel adaptation and parallel speciation

A promising way to understand the origin of species and their historical mode of divergence is to study independent replicate pairs of populations. The independent evolution of the same phenotypic trait in independent populations is called parallel evolution and it is suggested that this is “strongly due to the action of natural selection” because genetic drift is unlikely to result in such concerted patterns in different places (Schluter & Nagel 1995; Johannesson 2001). A classical scenario of parallel evolution at the phenotypic level is provided in Figure 8. When a trait that induces RI evolves independently in different populations, Schluter & Nagel (1995) proposed that it shows a case of parallel speciation. In general, the best examples demonstrated a strong role for size assortative mating.

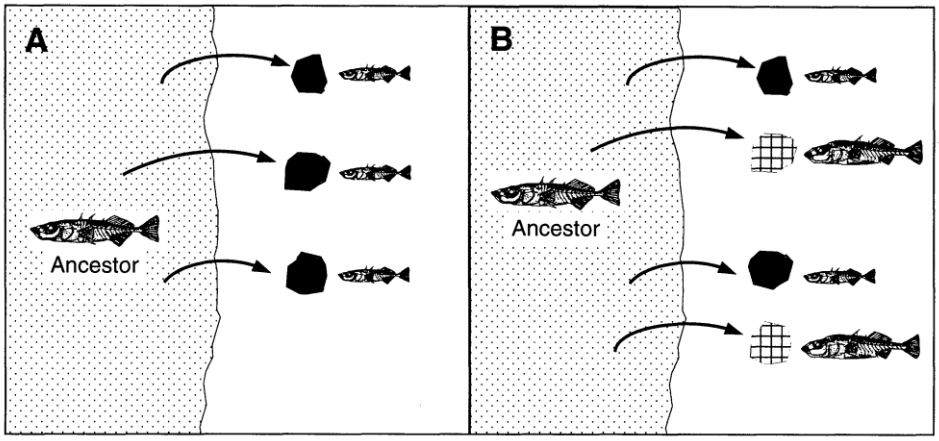


Figure 8: **Example scenarios for parallel evolution and parallel speciation.** Shaded areas are geographical ranges; shading type indicates environment type. Arrows indicate establishment of new derived populations from a phenotypically uniform ancestral species/population. (A) Colonization of replicate new environments leads to repeated evolution of small body size (and of mate preferences for small size). (B) Colonization of two environments causes repeated divergence in body size (and of mate preferences for body size) between daughter populations inhabiting different environments. (Extracted from Schluter & Nagel, 1995).

The best studied systems of parallel speciation are *Gasterosteus aculeatus* (Rundle *et al.* 2000), *Tinema* (Nosil *et al.* 2002), *Litorina* (Butlin *et al.* 2014a) *Coregonus clupeaformis* (Bernatchez *et al.* 2010) and cichlid fishes complexes (Elmer *et al.* 2014). However, parallel divergence at the phenotypic level may have little to do with genetic parallelism (Elmer & Meyer 2011)). The examples of parallel evolution at the molecular level involve parallel adaptation, but not necessarily parallel speciation. These include the repeated evolution of resistance to insecticides within insect species (Ffrench-Constant *et al.* 2000), the resistance of malaria to antimalarial drugs (Pearce *et al.* 2009) and the evolution of pigmentation within vertebrate species such as cavefish (Gross *et al.* 2009) or mice (Hoekstra *et al.* 2006; Steiner *et al.* 2007). Each of these examples involved independent mutations that have led to the same or functionally equivalent adaptive phenotypes (Ralph & Coop 2010). More generally, there are three genetic sources of adaptation to explain parallel phenotypic divergence. Either a mutation occurs *de novo* (as in the evolution of malaria resistance), or mutations can segregate into the standing variation (as in the evolution of pigmentation in mice (Steiner *et al.* 2007). Alternatively, variation can arise through gene flow, either by a connecting population, i.e. the transporter hypothesis Schluter & Conte (2009) Bierne *et al.* (2013) or by interspecific hybridization (e.g. (Abbott *et al.* 2013). Given the current popularity of the ecological speciation framework, it is important to distinguish the sources of adaptation. For instance, the origin of adaptation in marine-freshwater populations of *Gasterosteus aculeatus* is generally attributed to recent (<10 000 years old) and independent ecological adaptation of freshwater populations from the marine population due to standing variation, but this is still unclear. Below I present this system and associated literature in more detail, as it is conceptually similar to the lamprey species pair system studied in this thesis.

In stickleback, completely plated marine populations evolved repeatedly towards low-plated freshwater populations, an often cited example of parallel and independent evolution (Colossima et al. 2005). The lateral plate formation is mainly controlled by one major gene with pleiotropic effects: Ectodysplasin (*Eda*) (Colosimo *et al.* 2005). Recently Roesti *et al.* (2014) performed simulations to investigate the origin of adaptation in marine-freshwater populations of *Gasterosteus aculeatus.* They simulated the replicated colonization of a freshwater habitat (where the derived population occurs) by a source population adapted to the marine habitat and analyzed the pattern observed at three strong candidate loci for local adaptation. Their simulations showed a peak of F_{ST} between source and derived populations due to a barrier to gene flow. When comparing freshwater populations, they obtained the peak-valley-peak signature (cf section 3.1.1) which is somewhat similar to the pattern described above when a selective sweep arises in a structured population (Bierne 2010). The authors hypothesized that the freshwater populations adapted in parallel to freshwater, using alleles introgressed from the marine population. In a review Welch & Jiggins (2014) proposed four scenarios (Fig 9) to explain the results from Roesti *et al.* (2014) and all of these were already more or less presented in Bierne *et al.* (2013).

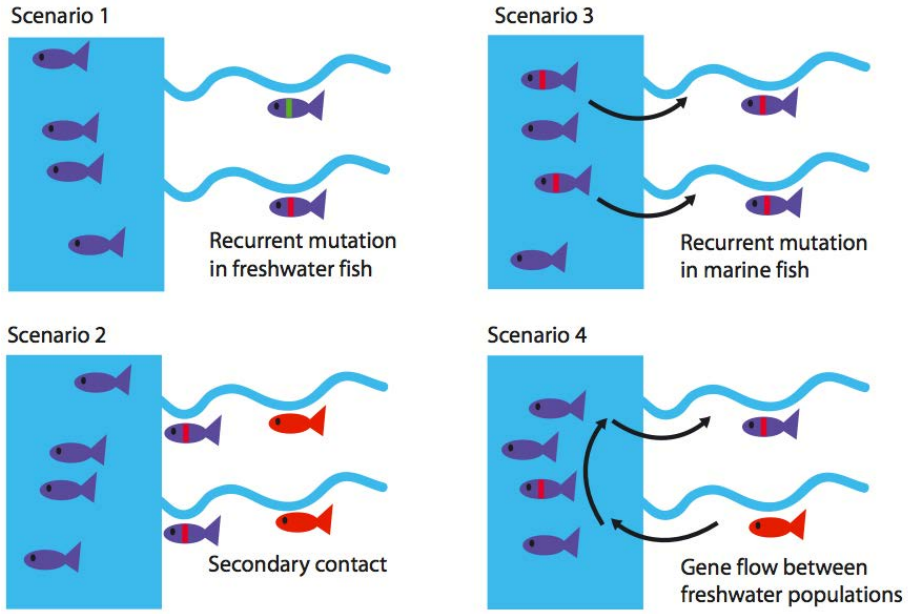


Figure 9: **Some of the possible scenarios underlying adaptation to the freshwater environment**. (1) Recurrent mutations occur independently in different freshwater populations. (2) Freshwater alleles were retained in multiple refugia, which then came into secondary contact in the marine population. Over time admixture may lead to considerable homogenization of genomes between the two habitats apart from regions involved in local adaptation. (3) Freshwater alleles arose by recurrent mutations in the marine habitat. These alleles, which may be identical by descent, are then introduced to the novel freshwater habitats. (4) Freshwater alleles were maintained in a refugium and introduced into novel freshwater habitats *via* the marine habitat. Although both (3) and (4) involve adaptation from standing variation in the marine habitat, the source of that variation is fundamentally different in the two cases. From Welsh & Jiggins 2014.

The first scenario corresponds to parallel adaptation from independent *de novo* mutations, a scenario that is rather unlikely in that case. The second scenario is secondary contact involving segregation of freshwater alleles in multiple refugia. The authors proposed that their data do not fit this scenario because the peak-

valley-peak between freshwater populations was unlikely. The third and fourth scenarios are adaptations from shared variation and are the most difficult to discriminate; they involved the transport of freshwater allele at very low frequency in the marine habitat to colonize other freshwater habitats. They differed in that they imply recurrent mutations (third scenario) (i.e. adaptation from standing variation) or recurrent migration (fourth scenario) requiring a population preadapted to freshwater (transporter hypothesis). In this last scenario adaptive alleles can be very old, whereas this is not the case in the third scenario. Finally Roesti *et al.* (2014) suggested that their data matched a scenario of divergence with gene flow that would be similar to the fourth scenario (Welch & Jiggins 2014). However, since scenario 4 can involve secondary contacts, the question of primary versus secondary differentiation remains the same. Finally, one should bear in mind that many populations remain polymorphic for lateral plate number (McCairns & Bernatchez 2008; 2012; Berner *et al.* 2012). In addition, there are also anadromous populations that may significantly contribute to gene flow between marine and freshwater populations (Raeymaekers *et al.* 2014). As pointed out by Raeymaekers *et al.* (2014) there is a bias in the literature with some widely cited studies "*investigating sharp contrast divergence between completely plated marine and low-plated resident freshwater populations*" (eg. Colosimo *et al.* 2005; Schluter & Conte 2009; Jones *et al.* 2012) mostly from Northern Europe and the Pacific Coast of North America, while vast regions in Western and Central Europe and the Atlantic Coast of North America show weaker contrasts (Raeymaekers *et al.* 2007; McCairns & Bernatchez 2008). In their study, Raeymaekers *et al.* (2014) investigated the effect of gene flow and selection on the dynamics of the *Eda* locus in natural populations that were polymorphic for lateral plate numbers. Their analysis clearly showed a lack of correlation between differentiation at *Eda* and habitat characteristics. They also performed a meta-analysis and showed that signatures of selection at *Eda* were rather weak compared to widely cited studies. Their results further suggested that introgression between marine and freshwater populations is possible (corresponding to the transporter hypothesis). Overall, their study casts some doubts onto the role of gene flow under "ecological" selection in maintaining local adaptation of populations. A more recent study of stickleback inhabiting different environments performed one of the first demographic analyses using genome wide data (Ferchaud & Hansen 2015). The authors showed that most freshwater populations have undergone bottlenecks, a result that may impact genome wide patterns of divergence and seems to have been largely overlooked so far. Finally, it is worth pointing out that sympatric species pairs of benthic-limnetic stickleback probably evolved from a double colonization event from marine stickleback (i.e. lakes were colonized twice) resulting in the speciation of the benthic morph and then the limnetic ecotype following the second colonization event (Schluter & McPhail 1992). This period of isolation may thus explain some of the divergence between ecotypes and is undoubtedly a confounding factor in studies related to the ecological process of speciation.

In the case of *Littorina* periwinkle ecotypes coexisting in adjacent coastal habitats, alternative models of divergence have been tested recently (Butlin *et al.* 2014a) and the best supported model was that of parallel divergence in the face of continuous gene flow in different locations. Alternative scenarios are possible because the authors used mainly neutral markers that are expected to converge to the same equilibrium after secondary contact to that observed under primary differentiation (see Chapter 3 and

details in Bierne et al. 2013). In addition, excluding long distance gene flow remains difficult. Further investigations yielded puzzling results since levels of outlier sharing was very low at either a large or small geographical scale (Westram *et al.* 2014; Ravinet *et al.* 2015). These results would suggest the independent evolution of the populations through *de novo* mutations rather than from standing variation. In the *Coregonus* whitefish limnetic-benthic pairs, it was hypothesized that divergence was initiated in allopatry and that the two glacial lineages then came into secondary contact in different lakes (Bernatchez *et al.* 1996, 2010). The divergence between ecotypes probably involved a polygenic basis (Bernatchez *et al.* 2010; Gagnaire *et al.* 2013a,b). Despite the common history of divergence, only partial parallelism was found with genome scans (Gagnaire *et al.* 2013a). This result could be explained by the independent erosion of the weakest barriers to gene flow in different lakes. A polygenic basis was also found recently by linkage mapping of traits involved in body shape (Laporte *et al.* 2015).

5 Lampreys as a model of speciation research

5.1 General presentation

Lampreys (Petromyzontiformes) form, together with hagfishes, the cyclostomes that are the only surviving representatives of jawless vertebrates (agnathans) (Hardisty & Potter, 1971). They have diverged from the gnathosthome (jawed vertebrate) lineage more than 500 milions years ago (Kuraku & Kuratani 2006; Hedges *et al.* 2015) and represent the most basal position in the vertebrate lineage. This position explains why lampreys are often used to understand the origin and evolution of vertebrate genomes and body plans (reviewed in Shimeld & Donoghue 2012; Green & Bronner 2014). However, lampreys are not an easy study system because they are difficult to maintain in experimental settings and larvae have a long larval period (> 4 years) and are very difficult to rear (Kuratani *et al.* 2002). These limitations may explain why so little is known about the evolutionary relationships among lamprey species and their underlying mechanisms of divergence.

42 species separated into three families are currently recognized in lampreys. The family Petromyzontidae lives in the Nothern Hemisphere and the Geotriidae and Mordaciidae are found in the Southern Hemisphere. 18 lamprey species display a parasitic (mostly hematophagous) lifestyle. Among them nine are anadromous and the others remain entirely freshwater resident (e.g. the three *Ichthyomyzon* species, three *Entosphenus* species, *Tetrapleurodon spadiceus, Eudontomyzon danfordi*). The 24 remaining species display a nonparasitic lifestyle, do not feed at the adult stage and are strictly freshwater resident (Potter & Potter 1971; Docker 2009). Most lampreys (seven out of ten genera) occur as “paired” species (Zanandrea 1959; Docker 2009) with one species displaying a parasitic strategy and the other, closely related, and supposedly derived from the parasitic one, being nonparasitic. These paired species therefore share some similarities with the anadromous/marine and freshwater resident stickleback *Gasterosteus aculeatus*, a classical model of speciation research and thus make lampreys a potentially interesting model to study the processes of speciation. In Western Europe, lampreys are represented by

the parasitic and anadromous sea lamprey *Petromyzon marinus,* the parasitic and anadromous river lamprey *Lampetra fluviatilis* and its paired species, the brook lamprey *Lampetra planeri* which is non parasitic and freshwater resident (Fig 10 and Fig 11). The two latter have split from the sea lamprey around 10-30 Mya ago (Docker *et al.* 1999; Kuraku & Kuratani 2006). These three species are widespread in Western Europe, where the brook lamprey can be found all along the river networks whereas the river lamprey is often restricted to downstream areas due to anthropogenic barriers to upstream migration (dams, weirs…). The sea lamprey is also present in North America, especially in the Great Lakes where it has become invasive. Both *Lampetra* species reproduce in freshwater and can be found in areas of sympatry where they sometimes spawn together (Huggins & Thompson 1970; Lasne *et al.* 2010). The larvae (called ammocoetes) then spend 3 to 7 years buried in river bed sediments (Fig. 11) (Maitland 1980). At this stage, larvae of both taxa are morphologically indistinguishable. It is only after metamorphosis that both species can be morphologically distinguished. They can be best recognized at the adult stage due to an important size difference: *L. planeri* measures 10-15 cm while *L. fluviatilis* adults are about 25-30 cm long. This size difference is found in all paired species, since parasitic lampreys feed during the adult stage and thus display a larger body size than the non parasitic taxa which stop feeding after metamorphosis-(Potter & Potter 1971; Vladykov & Kott 1979; Potter 1980).

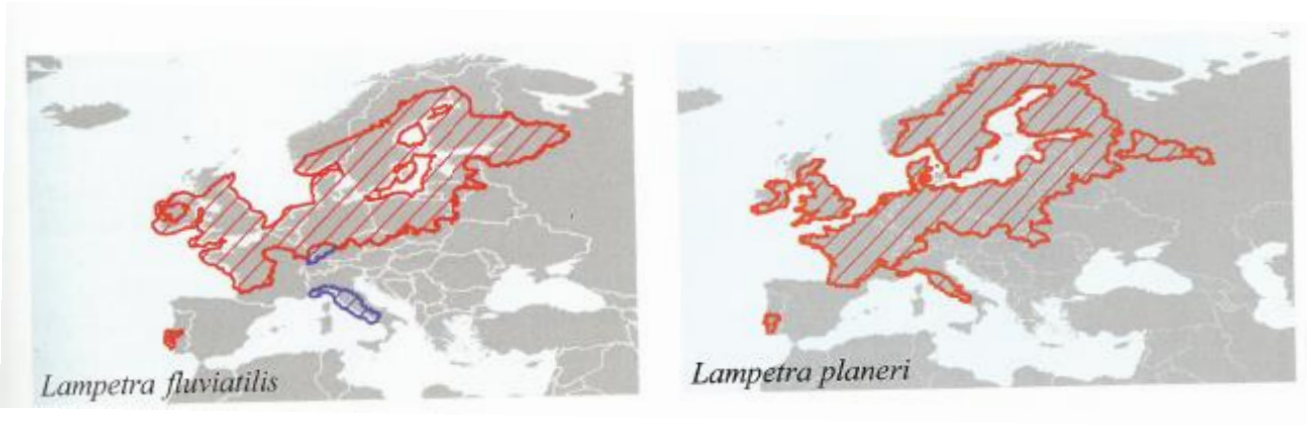


Figure 10: **European distribution of the genus Lampetra.** Blue outline represents areas in which the species is now extirpated. Extracted from Kottelat, 2007.

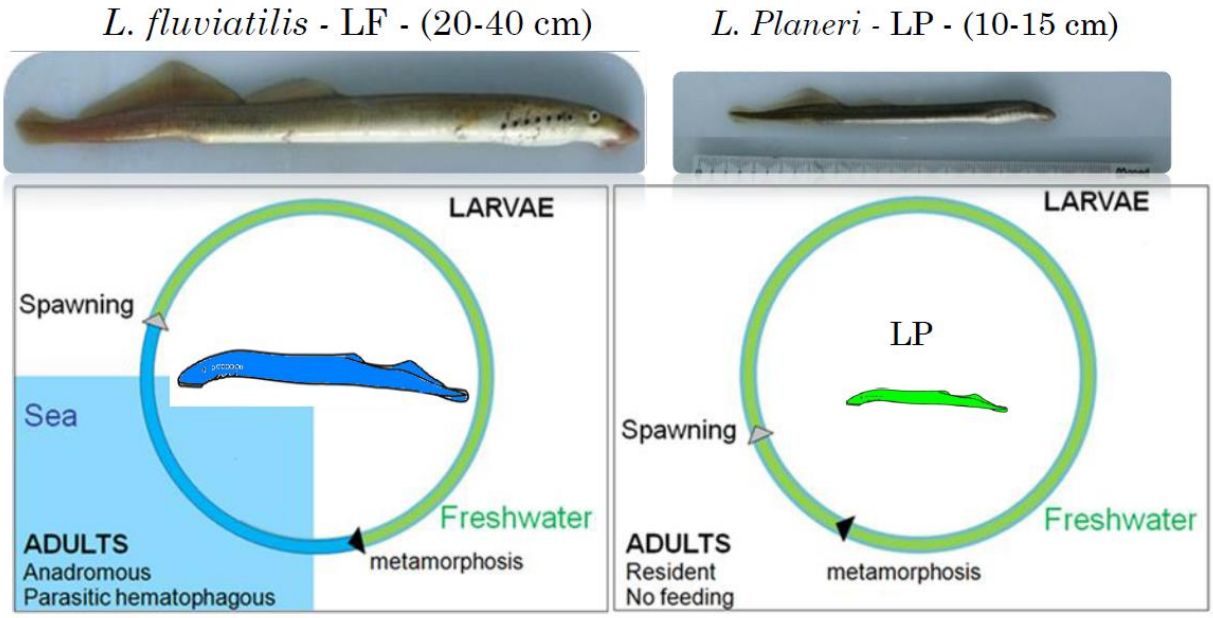


Figure 11: **Life cycle of the parasitic anadromous *L. fluviatilis* (left); of the nonparasitic resident *L. planeri* (right).**

During the larval period, it is likely that ammocoetes move passively downstream (Maitland 1980), especially during important flood events. Studies in *Petromyzon marinus* have shown that age-0 siblings tended to be found up to 0.9 km from each other 3 months after emergence and age-1 larvae had dispersed more than 1 km downstream (Derosier et al. 207 in Dawson et al. 2015). Rapid downstream disperal is also suggested in pouched lamprey *Geotria australis* larvae (Kelso and Todd 1993 in Dawson et al. 2015). Some studies suggest that larval lampreys are also capable of short upstream movement but, given the relatively poor swimming capacity of larvae, upstream movements are more limited (Dawson et al. 2015). The metamorphosis occurs from September to November when ammocoetes reach the macrophthalmia stage and develop functional eyes and an oral disc which in the parasitic *L. fluviatilis* has sharpened teeth for feeding while these teeth are blunt in brook lampreys (Hardisty 1944; Maitland 1980; Docker 2009). During metamorphosis, brook lampreys begin sexual maturation, with females developing oocytes (Huggins & Thompson 1970, Rougemont, personal observation); they reproduce in the following months and subsequently die. Upon metamorphosis, river lampreys migrate to sea where they become parasitic for one or two years (Taverny & Élie 2010). The transformation of river lampreys when going from freshwater to seawater is quite similar to the smoltification process in anadromous salmonids (the suite of morphological and physiological changes involved with adaptation to sea-water) (Reis-Santos *et al.* 2008; Stefansson & al. 2008). They have different host species including herring, sea trout, twaite and Allis shad, smelt and sprat (Taverny & Élie 2010). Anadromous lampreys then migrate back into rivers for spawning but they may not reproduce in their natal river: there is probably no 'homing' (Maitland 1980, Waldman *et al.* 2008). In *Petromyzon marinus*, the 'choice' of a river for spawning could be linked to the presence of conspecifics as migrating adults may follow hormones (pheromones) released by ammocoetes in rivers

(Sorensen *et al.* 2005; Johnson *et al.* 2009). This pheromone-induced behavior seems to occur also in *L. fluviatilis* (Gaudron & Lucas, 2006). This behavior can explain why the 'homing' is supposed to be modest in lampreys (Waldman *et al.* 2008), hence potentially resulting in high levels of gene flow among rivers (Spice *et al.* 2012; Hess *et al.* 2013, Bracken *et al.* 2015).

5.2 Lamprey paired species: one or two species?

The taxonomic status of paired species is often controversial. Some pairs represent two good species and others different ecotypes from a single species with various degrees of divergence (reviewed in Docker 2009). Most paired species are phylogenetically closely related (Fig. 12) and it is usually assumed that nonparasitic species derived from their parasitic counterparts (Zanandrea 1959; Vladykov & Kott 1979; Docker 2009).

The taxonomy of lampreys is generally based on morphological criteria such as body proportions, dentition and the number of trunk myomeres (Potter & Potter 1971; Vladykov & Kott 1979; Potter 1980), which have remained contentious so far (Potter *et al.* 2015. Furthermore, apart from differences in adult body size, the closely related paired species possess few diagnostic morphological differences. Other characters (e.g., the presence of chloride cells in the gills, a key element for osmoregulation in marine water of adult parasitic lamprey) may also fail to distinguish among recently diverged paired species. The nonparasitic *Lampetra planeri* and *L. richardsoni*, for example, still develop chloride cells following metamorphosis, despite residency entirely in fresh water (Bartels *et al.* 2011, 2015). The presence of such cells together with the low differentiation of the dentition in *L. planeri* and *L. richardsoni* from their respective ancestors *L. fluviatilis* and *L. ayresii* was interpreted as evidence of a very recent divergence (Bartels *et al.* 2011, 2015). However, chloride cells are also present in *Lethenteron appendix,* which is distinctly allopatric from its parasitic ancestor *Lethenteron camtschaticum*; this species pair display fixed difference in their mtDNA (Docker et al. 1999) suggesting that divergence is not as recent as with the above two species pairs. To date, the only nonparasitic species in which chloride cells are absent is a very old species (Bartels et *al.* 2015). Chloride cells seem to take a long time for freshwater resident lamprey to lose so that their retention is not a good evidence of a very recent divergence. These results clearly highlight the high uncertainty associated with the use of morphological criteria to determine the evolutionary relationships between paired species. Using molecular data this difficulty remains and is summarized in Fig 12 showing that the phylogeny in some species pairs is not well resolved (Docker 2009). Current barcoding analyses have also not provided enough resolution to discriminate paired species (April *et al.* 2011; Knebelsberger *et al.* 2014). For instance, 13 out of the 27 North American lamprey species could not be distinguished using DNA barcoding. The 13 species were separated into 5 groups, all containing at least one parasitic and one non-parasitic taxon (April *et al.* 2011). This is particularly the case between *Lethenteron camtschaticum* and *L. kessleri* where few differences were found based on mtDNA analysis (Yamazaki *et al.* 2006; Okada *et al.* 2010). Similarly, mtDNA, RFLP and microsatellite analyses of the parasitic silver lamprey (*Ichthyomyzon unicuspis*) and nonparasitic northern brook lamprey *(I. fossor*) indicated two mtDNA lineages that were

shared between the two species as well as strong levels of gene flow in areas of sympatry (Docker *et al.* 2012). Some nonparasitic species, however, appear to have diverged from a parasitic ancestor a long time ago. One such example is the nonparasitic *Lampetra aepyptera,* which has been separated from its (now extinct) ancestor for at least 2 million years (Docker *et al.* 1999; Bartels *et al.* 2015). *Lampetra aepyptera* is not paired with any parasitic species and is morphologically and genetically distinct.

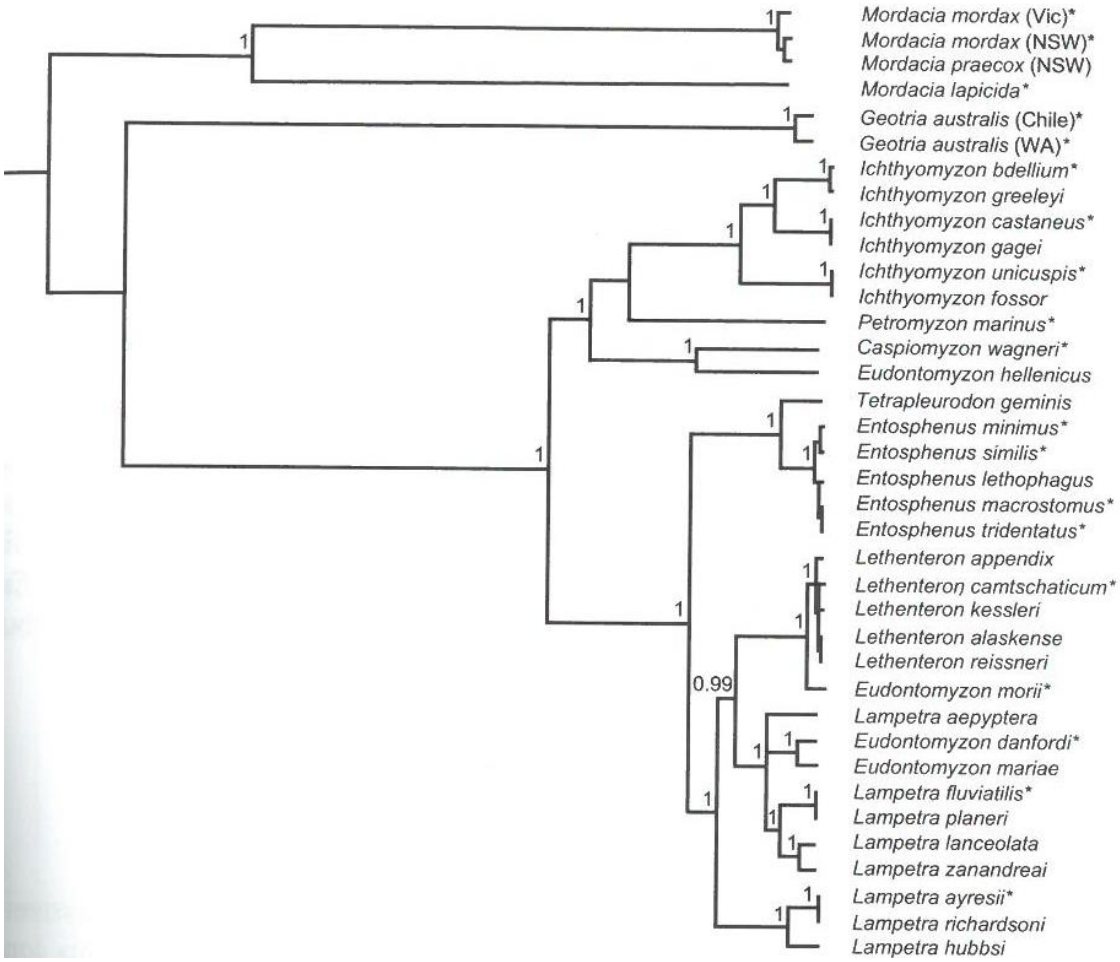


Figure 12 **Phylogenetic relationships between parastic and non parasitic species of the three lamprey families**. Dervied from cytochrome b sequences data using Bayesian analyses. Asterisks designate parasitic species. Data derive from *Lang* et al. *(2009)* with additional data for *Mordacia mordax* from New South Wales, Australia (NSW). VIC: Victoria, WA: Western Australia. Bayesian posterior probabilities are given when values are higher than 0.95. Taken from Potter *et al*. 2015.

The divergence between the European river lamprey and brook lamprey may have occurred since the last glacial retreat 10 000 years ago (Espanhol *et al.* 2007; Blank *et al.* 2008). A first study found a low level of divergence between both species based on allozymes (Schreiber & Engelhorn 1998). mtDNA analysis revealed the existence of three distinct lineages: one was restricted to the Iberian Peninsula and the two other lineages were widespread among the studied populations (Espanhol *et al.* 2007). In addition, the Iberian Peninsula displayed more genetic diversity than other populations suggesting ancient glacial refugia. More recently, further investigations in the Iberian Peninsula performed by Mateus *et al.* (2011)

revealed several differentiated clades of *L. planeri* but did not allow drawing further inferences about the taxonomic relationships between *L. fluviatilis* and *L. planeri*. A recent investigation using microsatellite markers found that contemporary gene flow was ongoing between landlocked *L. fluviatilis* and *L. planeri* in areas of sympatry around Loch Lomond in Scotland (Bracken *et al.* 2015). The first genome-wide study on *L. planeri* and *L. fluviatilis* found a strong divergence between both species (F_{ST}= 0.37) based on RAD-sequencing data from a single putatively sympatric pair located at the southern limit of the *L. fluviatilis* range (Mateus *et al.* 2013). The authors found 166 fixed markers between species and subsequently identified a list of genes putatively involved in their divergence. They also concluded that they provided "the first genetic evidence for the taxonomic validity of the two European lamprey species *L. planeri* and *L. fluviatilis*". Unfortunately, there are several shortcomings in this study. First, it is based on a single population pair (n = 37 individuals) located at the southern limit of the species' range, which prevents any generalization of results at a wider scale. Second, *L. fluviatilis* has almost disappeared in this area and has probably a very low population size (Mateus *et al.* 2012), which can strongly alter the genetic diversity due to genetic drift. In addition, given the temporal heterogeneity of the sampling and the large size of the Sorraria watershed, it is probable that individuals were sampled in different (potentially disconnected) sections of the river since in such river systems, anadromous lampreys are typically captured by nets in estuarine areas, while resident individuals are more likely to be found in upstream sections of smaller size. We will see in chapters 2 and 4 that the spatial design of sampling has strong implications on downstream conclusions about levels of gene flow between *L. planeri* and *L. fluviatilis*. Overall, all the studies performed so far on the genetic divergence between river and brook lampreys suffered from limitations in the sampling design (no replicate pairs in sympatry) and no study investigated the demographic history of population pairs in this system, which is extremely important to understand the processes underlying the genomic divergence (see section 2 above).

5.3 Level of reproductive isolation between *L. fluviatilis* & *L. planeri* inferred from experimental approaches

Investigations about the level of pre- and post-zygotic isolation and overall strength of reproductive barriers between lamprey paired species have been limited due to the difficulty (if not impossibility) of rearing larvae for more than a few months. The most important isolating barrier that is widely recognized is the size difference between parasitic and nonparasitic taxa (Beamish & Neville 1992). Beamish & Neville (1992) suggested complete size assortative mating when size difference was higher than 25% between North American river and brook lampreys, *L. ayresii* and *L. richardsoni*. However, size assortative mating is an exogenous prezygotic barrier, which does not necessarily imply that genetic incompatibilities (e.g. DMI) will accumulate if the two species co-occur in sympatry. In particular, during breeding of European river and brook lampreys, intraspecific communal spawning is frequently observed (Huggins & Thompson, 1971, Lasne *et al.* 2010). However, this communal spawning can also be interspecific (Huggins & Thompson 1970;

Lasne *et al.* 2010). In this situation, sneaker[4] males of brook lamprey have been observed (Hume et al. 2013) and this sneaking strategy may generate some gene flow that will break up any association between assortative mating traits and other traits that could putatively play a role in RI. The few experimental measurements made so far have shown that it was possible to artificially cross the two species and to obtain viable hybrid larvae (Hume et *al.* 2013). In their study however, Hume *et al.* (2013) did not separate the effect of mortality *before* or *after* fertilization of the eggs.

Finally, lamprey paired species share many similarities with other pairs of diadromous species, especially, the marine-freshwater stickleback system. Like the stickleback, lampreys were proposed a few years ago as a model to study the process of "ecological speciation" (Salewski 2003). While my thesis was originally embedded within this framework, which remains contentious and largely lacks theoretical support, I decided to investigate additional scenarios of divergence. An integrative approach combining experiments, population genetics, genomics, linkage mapping and simulations was developed as a starting point that may improve our understanding of how speciation operates in this system.

5.4 Possible threats on fragmented populations

L. planeri, in opposition to *L. fluviatilis,* is widespread in many rivers that are fragmented by anthropogenic barriers to migration (dam, weirs…). Approximately 1 million dams are fragmenting riverine habitats across the world (Nilsson *et al.* 2005). Habitat fragmentation divides the populations in smaller subsets and may also cause significant losses to the spawning and nursery habitat of both brook and river lampreys (Renaud, 1997). For instance, in the Iberian Peninsula, 80% of *P. marinus* and *L. fluviatilis* spawning habitat is now unreachable due to the construction of dams in the lower part of rivers (Mateus *et al.* 2012). Since the brook lamprey is supposed to display a reduced migratory behavior at the adult stage (Malmqvist 1980; Mateus *et al.* 2011), populations can be highly isolated in upstream reaches and be subject to founder events and losses of diversity through genetic drift (Brook *et al.* 2002; Perrier *et al.* 2013). Recent investigations by Bracken *et al.* (2015) suggest that anthropogenic factors may play a role in reducing the genetic diversity of *L. planeri* populations in the United Kingdom, but further analyses are required to quantify the effect of fragmentation on the distribution of diversity within and among populations. Understanding this effect may also be very useful for the conservation of both *Lampetra* species. Other factors such as pollution, over-exploitation, and habitat degradation (Dudgeon et al. 2006) are among the major threats that man imposed on biodiversity and this includes lampreys. Indeed, their long larval phase makes lampreys particularly vulnerable to pollution events (Moyle *et al.* 2009). It is even possible that some populations have been extirpated because of strong pollution (Mateus *et al.* 2012). The river lamprey, as most anadromous species (e.g. salmon, shad) has undergone a strong decline due to habitat fragmentation and pollution. Its low migratory ability as compared to other anadromous species may render this species even more vulnerable (Lucas *et al.* 2009; Foulds & Lucas 2013). In spite of this strong decline for over 40 years (Taverny & Elli 2005) no specific conservation efforts have been planned in

[4] Male of the non migratory species (or early maturing male in other species than lampreys) that used a sneaking strategy to fertilize eggs of females during spawning with a (larger) conspecific male.

Europe or in France. The river lamprey is nevertheless considered vulnerable in France on the IUCN red list, whereas the brook lamprey is listed as 'least concern'.

6 Goals of the thesis

The major goal of my thesis was to understand the process of divergence at play between the European river lampreys *L. planeri* and *L. fluviatilis.* This question was addressed *via* a multidisciplinary approach including experimental tests of reproductive isolation, investigations of gene flow in natural populations based on microsatellite and genome wide data and simulations to investigate the demographic history of divergence. A second goal related to conservation issues was to characterize the effect of natural or human-induced river fragmentation on the genetic integrity of *L. planeri* populations. We used a landscape genetics approach to test the effect of distance, barriers to migration as well as admixture between ecotypes on the spatial distribution of genetic diversity among *L. planeri* populations. To address these two main issues a multi-step approach was developed, which will be presented in the next chapters in the form of scientific papers:

In Chapter 2; the level of RI between *L. planeri* and *L. fluviatilis* was quantified both experimentally with controlled crosses and in the wild with measures of gene flow based on microsatellite data. I also characterized the extent of genetic structure among and within *L. planeri* and *L. fluviatilis* populations in French coastal rivers. Given the moderate levels of reproductive isolation and variable levels of gene flow, I suggested that *L. fluviatilis* and *L. planeri* may form partially reproductively isolated ecotypes. I identified replicated population pairs connected by high levels of gene flow as the most relevant populations to study the historical process of speciation.

In Chapter 3, these most relevant populations are used to identify the most likely evolutionary scenario of ecotypic divergence using Approximate Bayesian Computation on microsatellite data. Given the limited power of these markers, the scenarios of secondary contacts versus ongoing migration could not be disentangled, which highlighted the necessity of a genome-wide approach.

In Chapter 4, I used RAD-seq to determine the historical divergence of *L. fluviatilis* and *L. planeri* by a diffusion approximation approach and to investigate whether a signature of parallel genomic divergence was found among replicate population pairs. I identified markers putatively involved in the divergence of lamprey ecotypes and validated the hypothesis of partial parallel genomic divergence possibly stemming from a common divergence history and generating genomic islands (heterogeneous differentiation).

In Chapter 5, the goal was to quantify the effect of river fragmentation on the genetic integrity of *L. planeri* populations. Furthermore, taking advantage of our sampling of the two ecotypes in sympatry, parapatry and allopatry we investigated the potential benefit of recurrent introgression from *L. fluviatilis* in maintaining the genetic diversity of *L. planeri* populations. A moderate effect of barriers to migration was

found on level of genetic diversity in *L. planeri* populations. A greater influence of the distance to the source and admixture with *L. fluviatilis* was revealed.

I close this manuscript with a discussion synthesizing the main results and proposing new ideas to answer unresolved questions or bypass some model assumptions that were used during my PhD.

Chapter 2

Investigating gene flow and reproductive isolation between European river and brook lampreys

Article 1
Low reproductive isolation and highly variable levels of gene flow reveal limited progress toward speciation between European river and brook lampreys

Rougemont Q, Gaigher A, Lasne E, Côte J, Coke M, Besnard AL, Launey S, Evanno G

In press in Journal of Evolutionary Biology DOI: 10.1111/jeb.12750

Low reproductive isolation and highly variable levels of gene flow reveal limited progress toward speciation between European river and brook lampreys

Quentin Rougemont[1,2]*, Arnaud Gaigher[1,2,3]*, Emilien Lasne[4,5], Jessica Côte[1,2], Maïra Coke[6], Anne-Laure Besnard[1,2], Sophie Launey[1,2], Guillaume Evanno[1,2]

[1]INRA, UMR 985 Ecologie et Santé des Ecosystèmes, 35042 Rennes, France

[2]Agrocampus Ouest, UMR ESE, 65 rue de Saint-Brieuc, 35042 Rennes, France

[3]Laboratory for Conservation Biology, Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015, Switzerland

[4]Muséum National d'Histoire Naturelle, CRESCO, 35800 Dinard, France

[5]INRA, UMR CARRTEL, 74203 Thonon-les-Bains, France

[6]INRA, Unité Expérimentale d'Ecologie et d'Ecotoxicologie Aquatique, 35042 Rennes, France

*Both authors contributed equally to this work

Corresponding author: guillaume.evanno@rennes.inra.fr

Phone: +33 2 23 48 54 45, Fax: +33 2 23 48 54 40

Running title: Speciation between lamprey ecotypes

Abstract

Ecologically based divergent selection is a factor that could drive reproductive isolation even in the presence of gene flow. Population pairs arrayed along a continuum of divergence provide a good opportunity to address this issue. Here we used a combination of mating trials, experimental crosses and population genetics analyses to investigate the evolution of reproductive isolation between two closely related species of lampreys with distinct life histories. We used microsatellite markers to genotype over 1000 individuals of the migratory parasitic river lamprey (*Lampetra fluviatilis*) and freshwater-resident nonparasitic brook lamprey (*L. planeri*) distributed in 10 sympatric and parapatric population pairs in France. Mating trials, parentage analyses and artificial fertilizations demonstrated a low level of reproductive isolation between species even though size assortative mating may contribute to isolation. Most parapatric population pairs were strongly differentiated due to the joint effects of geographic distance and barriers to migration. In contrast, we found variable levels of gene flow between sympatric populations ranging from panmixia to moderate differentiation, which indicates a gradient of divergence with some population pairs that may correspond to alternative morphs or ecotypes of a single species and others that remain partially isolated. Ecologically-based divergent selection may explain these variable levels of divergence among sympatric population pairs but incomplete genome swamping following secondary contact could have also played a role. Overall, this study illustrates how highly differentiated phenotypes can be maintained despite high levels of gene flow that limit the progress toward speciation.

Keywords: genetic structure, *Lampetra*, life history strategy, parapatry, reproductive barrier, sympatry.

Introduction

Understanding the process of speciation, i.e. the evolution of reproductive isolation, is a central issue in evolutionary biology. Reproductive barriers among populations can be due to genetic incompatibilities that cause intrinsic reproductive isolation and / or divergent selection that produces extrinsic reproductive isolation (Coyne & Orr, 2004; Seehausen *et al.*, 2014). In allopatry, the cumulative effects of selection (including sexual selection), genetic drift and mutation can lead to speciation. Alternatively, sympatric speciation has been regarded as less likely since gene flow between nascent species will contribute, *via* recombination, to continuously break down associations between alleles linked to divergent adaptation (Felsenstein, 1981; Kirkpatrick & Ravigné, 2002). In parapatry, reproductive isolation can be maintained only under restricted values of gene flow (Bank *et al.*, 2012). The spatial context of speciation could thus greatly influence the evolution of reproductive isolation by constraining or facilitating gene flow (Sobel *et al.* 2010; Marie Curie SPECIATION Network *et al.* 2012)Studies under these different geographical settings (sympatry, allopatry and parapatry) have provided evidence for the role of natural selection in promoting speciation (Jiggins *et al.* 2001; Nosil *et al.* 2002; McKinnon *et al.* 2004; Barluenga *et al.* 2006; Langerhans *et al.* 2007; Soria-Carrasco *et al.* 2014) and have shown variable levels of divergence from panmixia to complete reproductive isolation resulting in a divergence continuum (Nosil *et al.* 2009b; Hendry 2009; Præbel *et al.* 2013). It remains challenging to discriminate the respective roles of gene flow, mutation and population size relative to the action of natural selection along this continuum (Barrett & Hoekstra 2011a). The relative importance of these factors is usually assessed by studying replicate pairs of taxa or populations either in sympatry (Nosil *et al.* 2009a; Gagnaire *et al.* 2013b; Powell *et al.* 2013) or parapatry (Berner *et al.* 2009; Kaeuffer *et al.* 2012; Roesti *et al.* 2012a) but rarely in both situations simultaneously. In most cases, results have been interpreted as evidence of recent and independent parallel divergence due to ongoing (ecological) selection and the role of demographic history has been usually overlooked (Bierne *et al.,* 2013). However, it can be particularly difficult to disentangle the role of different past demographic events that can leave similar signatures in the genetic makeup of present-day populations (e.g. Hewitt, 1996, 2011). For instance, it is often challenging to distinguish between primary divergence in sympatry *versus* a secondary contact following differentiation in allopatry because neutral markers often used for demographic inference may converge to the same equilibrium under both scenarios (Endler 1977; Barton & Hewitt 1985; Bierne *et al.,* 2013). In addition, population divergence after primary or secondary contact does not always lead to complete reproductive isolation (sensu Mayr, 1947) and to the formation of a new species (Mallet 2008; Hendry 2009; Nosil 2012; Elias *et al.* 2012). As a result, it is important to combine experimental approaches analyzing reproductive barriers with inferences of gene flow in replicated population pairs to measure the level of reproductive isolation between putative species (e.g. Dey *et al.* 2012; Sobel & Streisfeld 2015). Here we used such an integrative approach by focusing on both parapatric

and sympatric population pairs in an emerging model species that displays two extremely different life history strategies.

Lampreys are ancient jawless vertebrates (agnathans) in which most genera include 'paired' species (Zanandrea, 1959) with divergent life histories, which represent putative cases of ecological speciation (Salewski 2003). These paired species reproduce in freshwater but have extremely different feeding strategies at the adult stage: some taxa are parasitic (hematophagous) and migratory (either migrating to sea, i.e. anadromous, or migrating entirely within fresh water) whereas others are nonparasitic and freshwater resident. Larvae from paired parasitic and nonparasitic taxa are morphologically indistinguishable but adults can be distinguished mainly by the larger body size of parasitic taxon (Hardisty & Potter, 1971; Vladykov & Kott 1979; Potter 1980). Paired species are phylogenetically closely related and it is usually assumed that nonparasitic species derived from their parasitic counterparts (Zanandrea, 1959; Vladykov & Kott, 1979; Docker, 2009).

In Western Europe, the nonparasitic brook lamprey (*Lampetra planeri*, Bloch 1784) and the parasitic river lamprey (*L. fluviatilis*, Linnaeus 1758) display a low to moderate genetic differentiation as measured with allozymes (Schreiber & Engelhorn, 1998), mitochondrial DNA and microsatellite markers (Espanhol *et al.*, 2007; Blank *et al.*, 2008; Mateus *et al.*, 2011; Bracken *et al.,* 2015). In addition, the high viability of F1 hybrid larvae (Hume *et al.*, 2013a), communal spawning of both species on the same nest (Huggins & Thompson, 1970; Lasne *et al.*, 2010) and sneaking behavior of males towards spawning females from the other species have been observed (Hume *et al.*, 2013b). These results led to the hypothesis that brook and river lampreys may represent alternative life-histories strategies (or ecotypes) within a single species (Beamish 1987; Yamazaki *et al.* 2006; April *et al.* 2011; Docker *et al.* 2012; Knebelsberger *et al.* 2015). Alternatively, it was argued that the divergence between the two species may be very recent (Docker *et al.*, 1999; Salewski, 2003; Espanhol *et al.*, 2007; Okada *et al.*, 2010). However, Mateus *et al*. (2013) found a strong genome-wide divergence between *L. fluviatilis* and *L. planeri* sampled in a single river and concluded on the taxonomic validity of the two species. In addition, the different size of adults of both species has been hypothesized to induce size assortative mating leading to reproductive isolation (Beamish & Neville 1992). Nevertheless, this hypothesis has never been thoroughly investigated by testing whether the sneaking behavior of *L. planeri* males can lead to the fertilization of *L. fluviatilis* eggs and the production of viable hybrids (Hume *et al.* 2013b). The various population genetics studies led so far have also rarely distinguished the situations of sympatry and parapatry and it remains unclear whether the moderate to strong levels of genetic differentiation observed between both species within the same river was due to reproductive isolation in sympatry, isolation by distance or anthropogenic barriers (e.g. dams or weirs) in parapatry (Schreiber & Engelhorn, 1998; Mateus *et al*., 2013; Bracken *et al*. 2015).

In this study, we used an integrative approach combining experimental measures of reproductive isolation and estimates of gene flow in replicated population pairs to better understand the evolution of divergence between *L. fluviatilis* and *L. planeri*. Hereafter we use the term 'species' as it is the current taxonomic status of these lampreys but we acknowledge that other terms like 'ecotypes' or 'forms' may also be appropriate. First, we measured the reproductive success of *L. fluviatilis* and *L. planeri* males under semi-natural conditions where only *L. fluviatilis* females were present in order to test whether size-assortative mating induces a strong prezygotic isolation. Second, we performed *in vitro* fertilizations of *L. fluviatilis* eggs with semen from *L. fluviatilis* and *L. planeri* males to compare the fertilization and hatching rates of eggs from intra- and inter-specific crosses. Third, we performed a large-scale population genetic analysis including five sympatric and five parapatric population pairs to infer the level of gene flow between species and among populations within species. We hypothesized that if the level of reproductive isolation between *L. fluviatilis* and *L. planeri* is high, a strong level of pre- and post-zygotic isolation will be observed in our experiments as well as a low level of gene flow among sympatric populations. Alternatively, if *L. fluviatilis* and *L. planeri* were ecotypes of a single species at a very early stage of divergence or lineages subject to a secondary contact after a period of allopatric divergence, we expected a low reproductive isolation combined with high levels of gene flow in natural populations.

Materials and methods

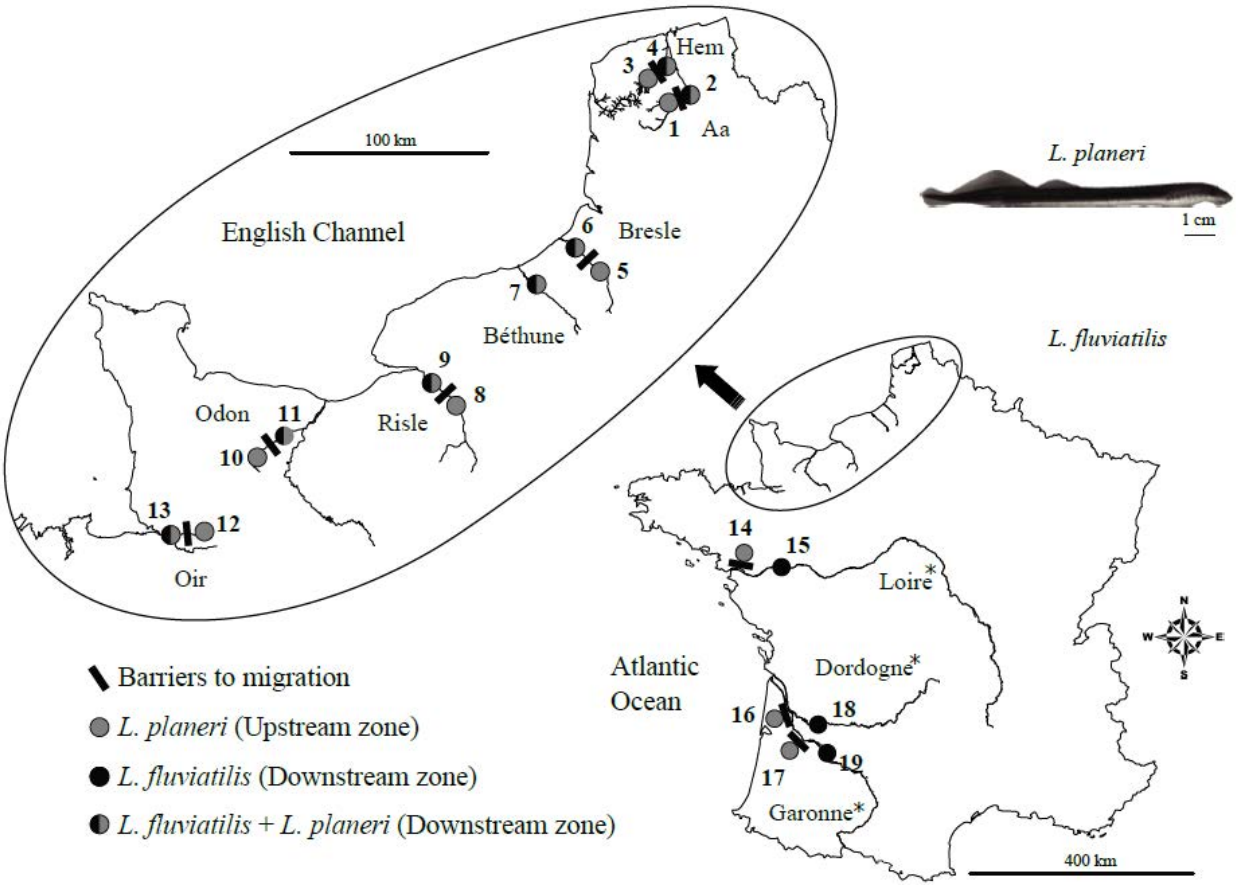


Figure 1: Map showing the sampling sites in France (numbers match those given in table 1).

*The Loire, Dordogne and Garonne populations are composed of upstream populations of *L. planeri* sampled in smaller streams located on the watersheds: the Cens, the Jalles de Tiquetorte and the Saucats rivers respectively.

Reproductive isolation: reproductive success under semi-natural conditions

We quantified the reproductive success of *L. fluviatilis* and *L. planeri* males under semi-natural conditions where only *L. fluviatilis* females were present. We aimed at testing whether size assortative mating may prevent any mating between *L. planeri* males and *L. fluviatilis* females. Four *L. planeri* males, two *L. fluviatilis* males and two *L. fluviatilis* females were caught by electrofishing in March 2013 on the downstream part of the Oir River (Fig. 1). We used a greater number of *L. planeri* males to compensate for their samller size. We did not use *L. planeri* females, because we were especially interested in testing whether smaller males could fertilize females of greater size as a previous study (Beamish & Neville 1992) suggested that mating was not possible when size differences are greater than 20%. Individuals were kept together in a 300 liter tank with 3-5 cm of fine gravel (0.5 - 1.5 cm diameter) in a recirculated water system. Temperature was set at 12±1°C with a 12:12 photoperiod. After spawning of both females, 129 larvae as well as tissue samples from each adult were collected and genotyped using 13 microsatellite markers (Gaigher *et al.*, 2013). Parentage analyses were performed with CERVUS 3.0 (Kalinowski *et al.* 2007) using the trio logarithm of the odd score, a 95% confidence level and allowing either no or up to two mismatches between putative parents and offspring.

Reproductive isolation: *artificial fertilization*

We performed *in-vitro* fertilizations of *L. fluviatilis* eggs with sperm from both species. Six *L. fluviatilis* females were crossed with four males of each species in a full factorial design producing 48 sib groups. Eight males and three females were captured by electrofishing in the Oir River (Fig. 1) whereas three other females were collected in the Loire River by a professional fisherman. We used females from two genetically differentiated populations (Oir and Loire) to discriminate the effects of outbreeding among populations and reproductive isolation between species (Waser & Price 1985, 1994; Schierup & Christiannsen 1996). We used an experimental design similar to that of Rodríguez-Muñoz & Tregenza (2009), presented in detail as supporting information. The fertilization success for each sib group was estimated three hours after fertilization based on the presence of a perivitelline space in the eggs (Ciereszko *et al.* 2000). We then measured the hatching rate at the individual level in microplates using only successfully fertilized eggs to avoid confounding dead and non-fertilized eggs. The average time to hatch was approximately 288°C-days (240-324°C-days).

Generalized linear mixed models (GLMM) with a binomial error family were used to test the influence of the cross type (within species ♀*Lf* × ♂*Lf* *versus* between species ♀*Lf* × ♂*Lp*) and maternal population (Oir *versus* Loire) on fertilization success and hatching rate. Cross type, population and cross type × population were considered as fixed effects. Sire, dam, sire × dam and microplates (only for hatching

rate) were treated as random effects. To test the significance of each factor on the response variable, we compared models including or not the focal variable using likelihood ratio tests (LRT) based on a χ^2 distribution (Zuur *et al.* 2009). Differences among populations were also investigated for each cross type separately to account for significant cross type × population interactions. Statistical analyses were performed with the lme4 package (Bates *et al.* 2014) in the R software (R Development Core Team 2011). Experiments were approved by the Ethics Committee in Animal Experiment of Rennes (file number: R-2012-EG-02).

Sampling for population genetic analyses

A total of 1023 lampreys were sampled in 19 sites spread over 13 rivers in France during the spawning period in 2010, 2011 and 2014 (Table 1, Fig. 1). Individuals were anesthetized with benzocaine, measured to the nearest millimeter and a fin clip was collected on each specimen and preserved in 95% EtOH. We sampled both species in sympatry simultaneously on the same spawning ground in the Aa, Béthune and Oir Rivers. Two other sites were also considered as sympatric: the Bresle River, where the two species were captured 8 km apart (with no anthropogenic barrier in between) and the Hem River where *L. planeri* individuals were sampled above a dam occasionally passable for *L. fluviatilis,* depending on water level. We also sampled *L. planeri* in five parapatric upstream sites inaccessible to *L. fluviatilis* due to dams or weirs: Risle, Odon, Cens, Saucats and Jalle de Tiquetorte Rivers. Such upstream sites inaccessible to *L. fluviatilis* were also sampled in the Aa, Hem, Bresle and Oir Rivers (i.e. the sympatric sites) in order to quantify the within-river genetic variability of *L. planeri*. In addition, two *L. planeri* and one *L. planeri* individuals were captured in sympatry in the Odon and Risle River respectively.

Molecular and statistical analyses

Genomic DNA was extracted from fin clips using a modified Chelex protocol (Estoup *et al.*, 1996). Genotyping was performed with 13 microsatellite markers specifically developed for *L. planeri* and *L. fluviatilis* (Gaigher *et al.*, 2013). Allelic richness (*Ar*), observed heterozygosity (*Ho*), expected heterozygosity (*He*) and inbreeding coefficient (*Fis*) for each population were calculated using FSTAT 2.9.3 (Goudet, 2001). The number of private alleles (*Pa*) was estimated with GENALEX 6.5 (Peakall & Smouse 2012). Exact tests implemented in GENEPOP 4.1.0 (Rousset 2008) were used to test the Hardy-Weinberg Equilibrium (HWE) and the linkage disequilibrium. The Bonferroni correction was used to adjust the significance level for multiple tests (Rice 1989; $\alpha = 0.05$). Potential differences of expected heterozygosity (*He*) and allelic richness (*Ar*) between species were investigated using the permutation test implemented in FSTAT (15 000 permutations). Differences of *He* and *Ar* between *L. fluviatilis* and *L. planeri* from the same river and the same year were further tested with Wilcoxon paired signed-rank tests (using values per locus) in R. F_{ST} among populations within species was estimated by θ (Weir & Cockerham, 1984) and pairwise values were tested using 17 000 permutations and the Bonferroni correction in FSTAT. An Analysis of Molecular Variance

(AMOVA) was performed with ARLEQUIN 3.5.1.2 (Excoffier & Lischer 2010) to quantify the hierarchical distribution of genetic variability between the two species, among populations within each species and within populations. The significance of variance components was tested with 15 000 permutations.

Genetic structure and gene flow

The genetic structure was analyzed with the Bayesian clustering approach implemented in STRUCTURE 2.3.3 (Pritchard *et al.*, 2000). The number of genetic clusters (*k*) varied from 1 to 12 and 20 independent replicates per *k* value were performed. Markov Chain Monte Carlo simulations (MCMC) were based on 250 000 burn-in followed by 500 000 iterations using an admixture model with correlated allele frequencies (Falush *et al.*, 2003). The most likely number of clusters was determined with the estimated log likelihood Ln Pr(X|K) and the ΔK method (Evanno *et al.*, 2005) using STRUCTURE HARVESTER (Earl & vonHoldt, 2012). Plots were drawn using DISTRUCT 1.1 (Rosenberg 2004) by averaging individual membership values over the 20 runs for the best *k*. A global analysis using the full dataset of 1023 individuals was performed and then the level of divergence was investigated within population pairs and in each species separately.

Ongoing gene flow between sympatric and parapatric pairs of *L. planeri* and *L. fluviatilis* was estimated using BAYESASS 1.3 (Wilson & Rannala, 2003). Delta values for migration rates, inbreeding coefficient and allele frequencies were optimized to obtain acceptance rates between 40 and 60% of the total number of iterations to ensure proper chain mixing. The program was run with a burn in of 2 000 000 followed by 7 000 000 iterations with 5 runs initiated with random seed numbers.

To test for a pattern of isolation by distance (IBD), the correlation between pairwise geographic and genetic distances measured by $F_{ST}/(1-F_{ST})$ (Rousset, 1997) was tested using a Mantel test in R with 10 000 permutations. Geographic distances between each sampling site along coastline and within rivers were computed using ArcGIS 9.3.

Results

Reproductive success under semi-natural conditions

The 129 larvae were all successfully assigned with a 95% confidence level and up to two loci mismatches to a pair of parents and 73 out of 129 were assigned with no loci mismatches between parents and offspring (Fig. 2). From these 73 larvae, 59 (81%) were assigned as pure *L. fluviatilis* and 14 as hybrids (19%). Each *L. planeri* male produced two to four offspring while the two *L. fluviatilis* males sired 35 and 24 offspring, respectively. The two *L. fluviatilis* females produced 47 and 26 offspring, respectively. Size of the 2 *L. fluviatilis* females was 206 and 211 mm respectively. The size of each male is provided in Figure 2.

Artificial fertilization: fertilization success and hatching rates

Fertilization rates (± s.e.) of eggs from both homospecific (♀*Lf* × ♂*Lf*) and heterospecific (♀*Lf* × ♂*Lp*) crosses were extremely high: 95.5% ± 3.2 and 95.8% ± 3.8, respectively. We found a significant interaction between cross type and maternal population on fertilization success (Table S1). However, considering each cross type separately, no population effect was found for both hetero- and homospecific crosses (LRT χ^2 = 3.28, P = 0.07; LRT χ^2 = 1.50, P = 0.22). Hatching rates were also extremely high: all eggs from homospecific crosses successfully hatched (100%) and only 5 out of 563 eggs (99.1% ± 1.74) from heterospecific crosses failed to hatch. Given the lack of variance in homospecific crosses, we could not statistically test for an effect of cross type on hatching rate. Considering heterospecific crosses, no population effect was found (LRT χ^2 = 0, P = 1).

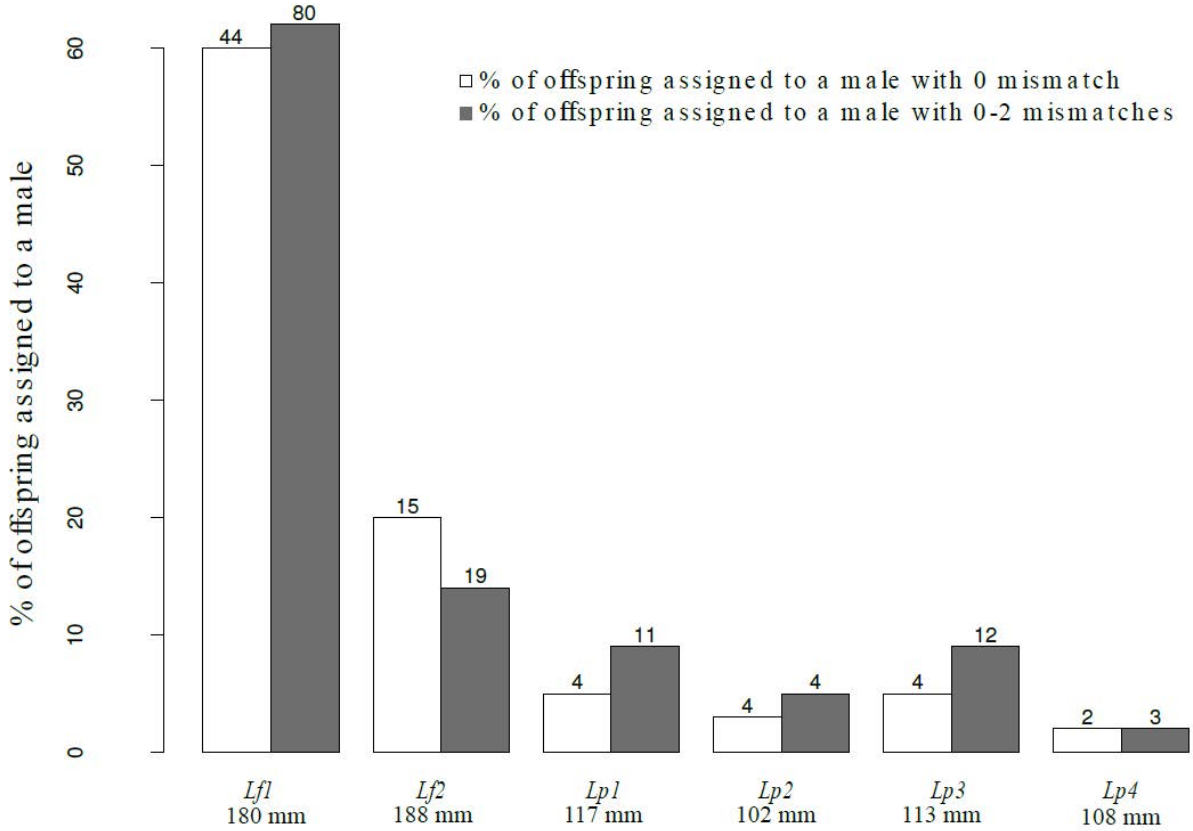


Figure 2: Reproductive successes of four *L. planeri* and two *L. fluviatilis* males under semi-natural conditions after spawning with two *L. fluviatilis* females. White bars represent the percentage of offspring assigned to a male at the 95% confidence level when no mismatch was allowed in assignment tests (73 individuals assigned). Grey bars represent the percentage of offspring assigned to a male at the 95% confidence level when up to 2 mismatches were allowed (129 individuals assigned). Numbers on top of each bar depict the absolute number of offspring assigned.

Within-species and population genetic diversity

A total of 1023 individuals were successfully genotyped at 13 loci. The number of alleles per locus varied from 4 to 12, for an average of 5.8 (Table S2). No linkage disequilibrium was observed between pairs of loci after Bonferroni correction. Tests on deviation from HWE showed a significant excess of heterozygotes only for the *Lf* Dor D 2010 population after correction (Table 1). Average allelic richness

(based on 14 samples) and average expected heterozygosity were significantly higher (15 000 permutations, $P < 0.01$) in *L. fluviatilis* (3.270 and 0.501 respectively) than in *L. planeri* (2.871 and 0.444) (Table1). *L. planeri* sampled in upstream and downstream parts of the Aa, Hem, Bresle and Oir Rivers showed no difference of genetic diversity (two-sided paired Wilcoxon test, $P > 0.05$, Table S3) except for expected heterozygosity in Aa 2014 and allelic richness in Oir 2014, which were both significantly higher downstream (Table S3). In most rivers, the allelic richness was significantly higher in *L. fluviatilis* than *L. planeri* (two-sided paired Wilcoxon test, $P < 0.05$ and Tables 1 and S3), except in sympatric populations and in the Risle River. The same trend was observed for expected heterozygosity (Table S3).

Table 1 Genetic diversity estimates of *L. fluviatilis* (*Lf*, n = 523) and *L. planeri* (*Lp*, n = 500) populations based on 13 microsatellite loci for each site and year. Site numbers refer to Figure 1, N = Number of individuals, *Ar* = Allelic richness (based on resampling of 14 individuals), *An* = Number of alleles (averaged by loci), *Ho* = Observed heterozygosity, *He* = Expected heterozygosity, *Fis* = Inbreeding coefficient (*significant deviation from the Hardy-Weinberg Equilibrium) and (S), U, and D refer to sympatric, upstream and downstream sites, respectively. Rivers (abbreviation) : Béthune (Bet) - Bresle (Bre) - Risle (Ris) - Odon (Odon) - Oir (Oir) - Loire (Loi) - Cens (Cens) - Dordogne (Dor) - Garonne (Gar) - Jalle de Tiquetorte (Jal) - Saucats (Sau). Jalle de Tiquetorte is a tributary of the Gironde estuary that is common to Garonne and Dordogne Rivers. The Bresle, Risle and Jalle de Tiquetorte *L. planeri* samples from 2011 include 1, 2 and 5 ammocoetes individuals, respectively. The *L. planeri* Risle U, Hem U and Aa U samples from 2014 are composed only of ammocoetes. The Odon *L. fluviatilis* samples include 7 juveniles (smolts). All other samples include adult individuals only.

	Site	N	Mean Size (mm)	*Ar*	*An*	*Ho*	*He*	*Fis*
Populations								
Lf Aa D 2014	2 (S)	34	310.5	3.511	3.846	0.504	0.514	0.019
Lf Hem D 2014	4 (S)	30	300.1	3.313	3.769	0.497	0.504	0.013
Lf Bre D 2010	6 (S)	38	335.9	3.415	3.692	0.467	0.480	-0.005
Lf Bre D 2011	6 (S)	41	334.6	3.252	3.615	0.478	0.484	-0.004
Lf Bet D 2014	7 (S)	17	306.8	3.761	3.538	0.475	0.514	0.078
Lf Ris D 2010	9	19	NA	3.555	3.769	0.526	0.517	-0.018
Lf Ris D 2011	9	40	320.4	3.212	3.615	0.478	0.500	0.041
Lf Ris D 2014	9	35	315.9	3.769	3.769	0.515	0.503	0.024
Lf Odo D 2011	11	32	316	3.207	3.692	0.502	0.486	-0.035
Lf Oir D 2010	13 (S)	34	222	3.399	3.846	0.554	0.542	-0.048
Lf Oir D 2011	13 (S)	40	229.4	3.188	3.692	0.506	0.505	-0.009
Lf Oir D 2014	13 (S)	30	222.9	3.462	3.462	0.505	0.52	-0.03
Lf Loi D 2010	15	32	290.8	3.089	3.308	0.468	0.461	-0.016
Lf Loi D 2011	15	32	NA	3.223	3.38	0.483	0.507	0.047
Lf Dor D 2010	18	39	250.7	3.068	3.308	0.527	0.486	-0.110*
Lf Dor D 2011	18	15	291.9	3.138	3.154	0.543	0.500	-0.09
Lf Gar D 2011	19	15	258.3	3.210	3.231	0.547	0.542	-0.016
Lp Aa D 2014	2 (S)	30	129.9	2.921	3.231	0.563	0.528	-0.068
Lp Aa U 2014	1	39	129	3.091	3.462	0.492	0.522	0.059
Lp Hem D 2014	4 (S)	39	155	3.738	3.462	0.469	0.477	0.017
Lp Hem U 2014	3	26	126	2.996	2.923	0.504	0.471	-0.071
Lp Bre D 2011	6 (S)	21	132.4	2.763	2.923	0.349	0.347	-0.005
Lp Bre U 2011	5	28	133.1	2.690	2.923	0.342	0.345	0.009
Lp Bet D 2014	7 (S)	17	144	3.791	3.385	0.482	0.472	-0.023

Lp Ris D 2011	9 (S)	1	147	NA	NA	NA	NA	NA
Lp Ris U 2011	8	16	143.3	3.323	3.385	0.335	0.335	-0.114
Lp Ris U 2014	8	28	136.5	3.770	3.769	0.466	0.435	0.066
Lp Odo D 2011	11 (S)	2	117.5	NA	NA	NA	NA	NA
Lp Odo U 2011	10	33	124.7	2.257	2.385	0.343	0.329	-0.043
Lp Oir D 2010	13 (S)	34	112	3.114	3.385	0.503	0.511	-0.043
Lp Oir D 2011	13 (S)	17	125.2	2.980	3.077	0.516	0.481	-0.075
Lp Oir D 2014	13 (S)	23	129.1	3.154	3.154	0.492	0.534	-0.087
Lp Oir U 2011	12	35	114.6	2.929	3.154	0.481	0.495	0.027
Lp Oir U 2014	12	31	122.8	3.000	3.000	0.458	0.466	-0.018
Lp Cen U 2011	14	33	163.3	2.248	2.462	0.257	0.262	0.008
Lp Jal U 2011	16	17	117.9	2.699	2.769	0.443	0.457	0.031
Lp Sau U 2011	17	30	109	2.459	2.538	0.405	0.400	-0.012
Species (mean)								
L. fluviatilis			287.9	3.270	3.570	0.502	0.501	-0.002
L. planeri			131.2	2.871	3.077	0.446	0.444	-0.005

Genetic structure

Pairwise F_{ST} values ranged from 0 to 0.324 (Tables 2, 3 and S4), with an overall F_{ST} of 0.082 (99% IC = 0.065-0.106). The AMOVA revealed that the percentage of variance among populations of the same species (6.25%) was much higher than between species (1.55%) and the largest part of variance (92.20%) was found within populations. F_{ST} among *L. fluviatilis* populations was significant but much smaller than among *L. planeri* populations: 0.022 and 0.134, respectively (15 000 permutations, $P < 0.001$). *L. planeri* populations sampled upstream and downstream of barriers in the same river (Aa, Hem, Bresle and Oir Rivers) were not significantly differentiated ($F_{ST\text{-}Aa} = 0.006$; $F_{ST\text{-}Hem} = 0.008$; $F_{ST\text{-}Bresle} = 0.005$) except in the Oir River ($F_{ST\text{-}Oir\text{-}2011} = 0.031$, $F_{ST\text{-}Oir\text{-}2014} = 0.020$).

We observed contrasting levels of population differentiation between *L. fluviatilis* and *L. planeri* depending on rivers (Table 2 and Table 3). The sympatric population pair in the Oir (2014) was not significantly differentiated ($F_{ST} = 0.008$) whereas a moderate structuration was observed in the Aa, Hem, Oir (2010 and 2011), Béthune and Bresle (2011) Rivers (F_{ST} = 0.080; 0.081; 0.048; 0.032; 0.028 and 0.074, respectively, Table 2). F_{ST} was generally higher in parapatry with population pairs from the Odon, Loire-Cens and Garonne-Saucats Rivers being the most differentiated (Table 3). The parapatric population pair from the Risle River was an exception with a low F_{ST} of 0.028 and 0.036 in 2011 and 2014 respectively (Table 3). Overall, F_{ST} between *L. planeri* and *L. fluviatilis* from the same river system was always smaller than the mean F_{ST} between the *L. planeri* population from this river and all other *L. planeri* populations (Fig. 3). Similarly, the mean F_{ST} between a given *L. planeri* population and all other *L. fluviatilis* populations was always smaller than among *L. planeri* populations (Fig. 3).

A positive correlation between genetic and geographic distances was found in the global data set ($r_{Spearman}$ = 0.27, P = 0.004) but when each species was considered separately the pattern of isolation by distance was stronger in *L. fluviatilis* ($r_{Spearman}$ = 0.79, P < 0.001) than in *L. planeri* ($r_{Spearman}$ = 0.40, P = 0.005).

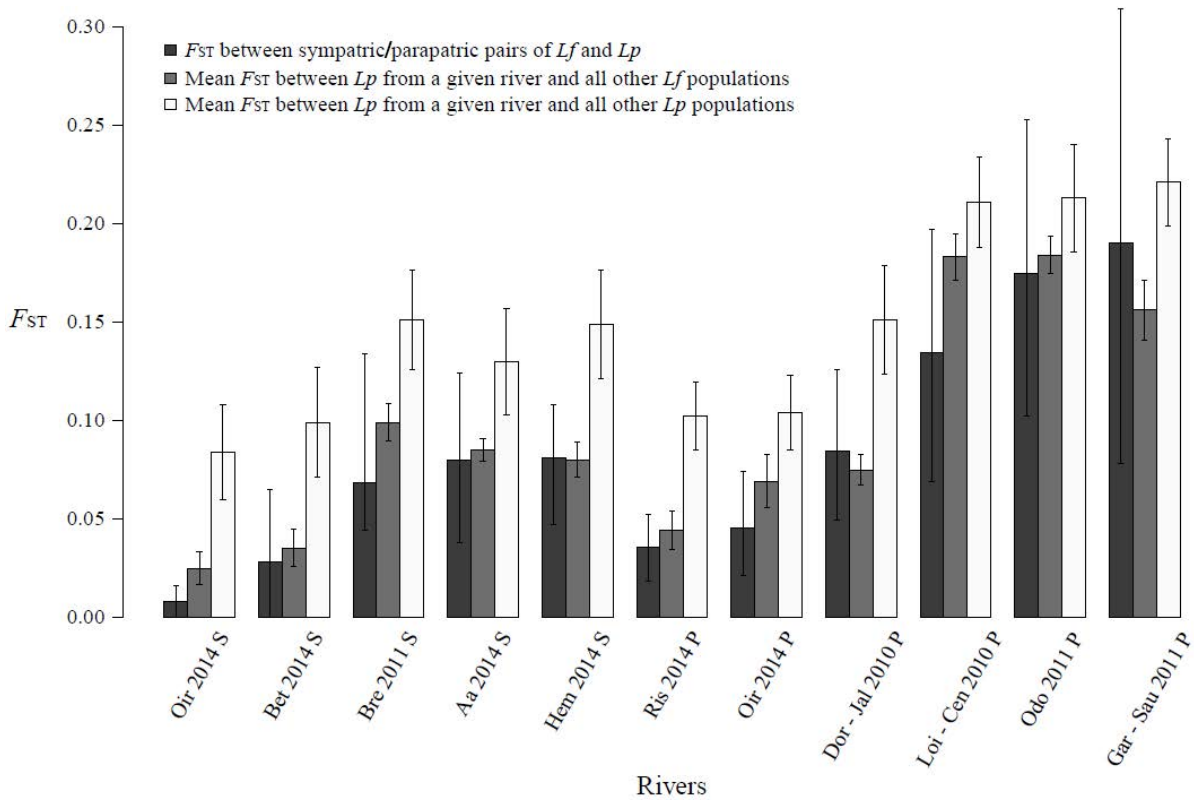


Figure 3: Comparison of F_{ST} values between sympatric (S) and parapatric (P) pairs of river and brook lampreys (dark-grey bars with their 95% confidence intervals), brook lampreys from a given river and all other brook lamprey populations surveyed (grey bars ± s.e.) and brook lampreys from a given river and all other river lamprey populations (white bars ± s.e.).

Bayesian clustering analyses

Results from STRUCTURE for each population pair illustrated a continuum of differentiation from apparent panmixia in some rivers to strong differentiation in other cases (Fig. 4). In the Oir, Béthune (sympatry) and Risle (parapatry) Rivers all individuals were assigned to both clusters, hence suggesting that *L. planeri* and *L. fluviatilis* formed a single population. However, in the three other sympatric situations (Aa, Hem and Bresle) where the differentiation was higher, two clusters were observed but some individuals were assigned to the cluster of the other species. In the parapatric Loire-Cens, Dordogne-Garonne-Saucats and Dordogne-Garonne-Jalles de Tiquetorte systems both species clustered in two groups and very few individuals were assigned to the cluster of the other species. Interestingly, the few *L. planeri* individuals (n =

3) captured in sympatry on the Risle and Odon Rivers were mostly assigned to the cluster of *L. fluviatilis* (Fig. 4 and Fig. S1a, b, c).

Results from STRUCTURE based on the 1023 individuals showed the highest likelihood for k = 8 and 10 (Table S5a). The best Δk values were observed at k = 3, 6 and 8 (Table S5a). We also investigated k = 2 to illustrate the level of admixture between both species (Fig. S1a). At this clustering level, all individuals had mixed membership proportions between the two species. When considering k = 8 (Fig. S1b), we found two widely admixed clusters for *L. fluviatilis*: the first one included samples from the Atlantic coast and the second one those from the English Channel area. The six other clusters included *L. planeri* samples: four clusters corresponded each to one river (the Aa, Hem, Jalles and upstream part of the Oir River), the fifth cluster gathered samples from the Odon and Garonne-Saucats Rivers and the sixth cluster included samples from the Bresle and Loire-Cens Rivers. *L. planeri* samples from the Oir, Risle and Béthune were strongly admixed with *L. fluviatili*s populations. Analyzing species separately confirmed the existence of two clusters for *L. fluviatilis* (Table S5b and Fig. S1d). For *L. planeri*, the most probable number of cluster was k = 9 and each population formed a distinct cluster except the Béthune and Risle, which clustered together (Table S5c and Fig. S1e).

Estimates of recent migration rates within the different population pairs obtained with BAYESASS revealed an asymmetric pattern when considering the whole dataset with a significant tendency for a higher gene flow from *L. planeri* to *L. fluviatilis* (two-sided paired permutations test, T = 78, P = 0.015) (Table S6). Interestingly, this pattern was driven by asymmetric migration occurring mainly in parapatric situations (T = 45, P = 0.006) whereas there was no significant difference in the direction of migration in sympatry (T = 3, P = 0.343).

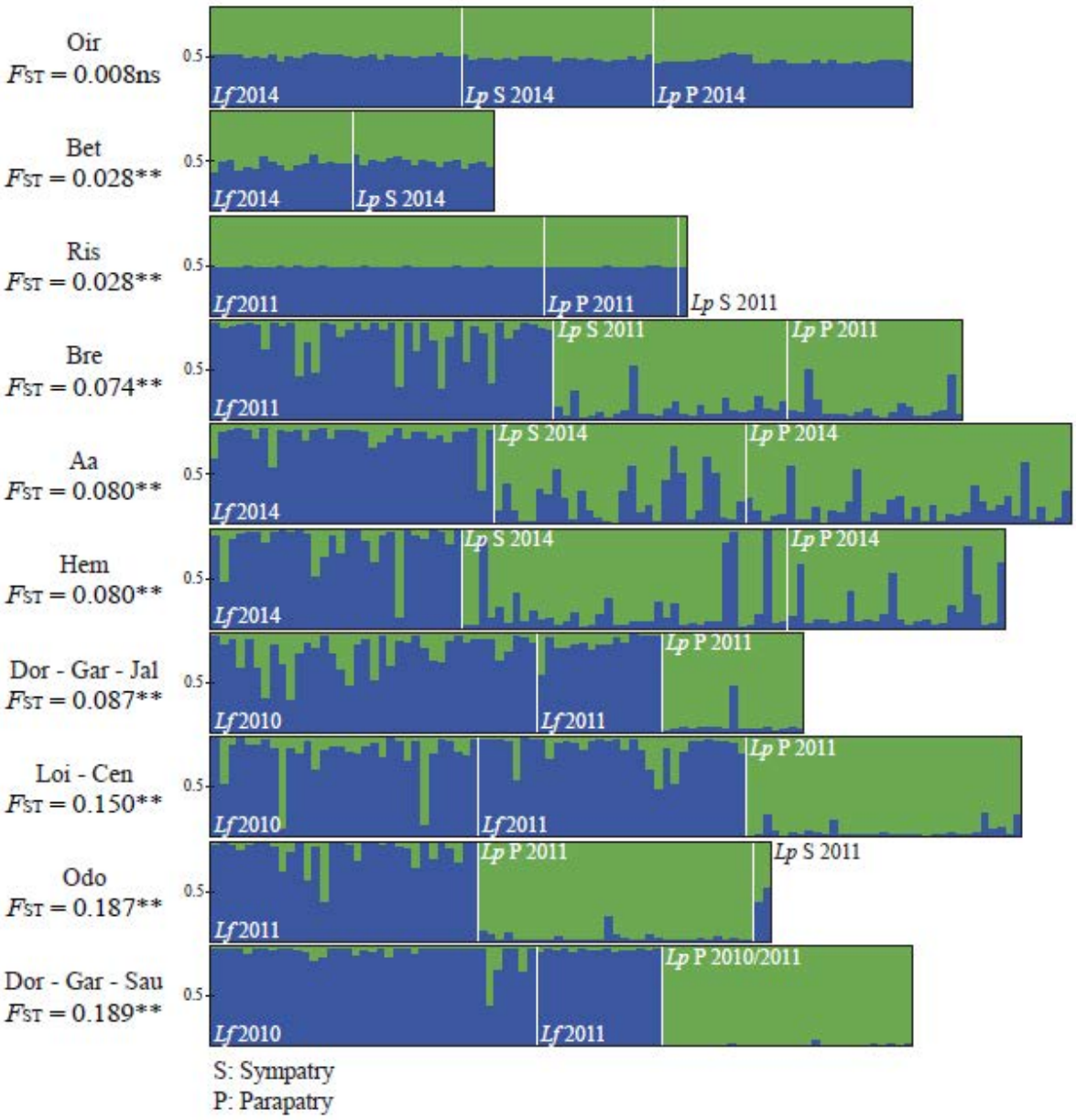


Figure 4: Bayesian analysis performed with STRUCTURE for each sympatric (S) or parapatric (P) pair of river (*Lf*) and brook (*Lp*) lamprey populations.

Table 2: Pairwise F_{ST} values for sympatric populations (non-significant values are grey coloured and negative values were set to zero).

Populations	*Lf* Aa 2014	*Lf* Hem 2014	*Lf* Bre 2011	*Lf* Bet 2014	*Lf* Oir 2010	*Lf* Oir 2011	*Lf* Oir 2014	*Lp* Aa 2014	*Lp* Hem 2014	*Lp* Bre 2011	*Lp* Bet 2014	*Lp* Oir 2010	*Lp* Oir 2011
Lf Hem D 2014	0												
Lf Bre D 2011	0	0											
Lf Bet D 2014	0.002	0.001	0.003										
Lf Oir D 2010	0.013	0.023	0.021	0.024									
Lf Oir D 2011	0.001	0.001	0.002	0.017	0.018								
Lf Oir D 2014	0	0	0	0	0.018	0							
Lp Aa D 2014	0.080	0.083	0.092	0.076	0.074	0.102	0.083						
Lp Hem D 2014	0.083	0.081	0.082	0.094	0.087	0.074	0.062	0.112					
Lp Bre D 2011	0.081	0.092	0.074	0.111	0.089	0.077	0.087	0.185	0.187				
Lp Bet D 2014	0.022	0.025	0.018	0.028	0.034	0.015	0.027	0.110	0.093	0.080			
Lp Oir D 2010	0.080	0.087	0.080	0.086	0.048	0.073	0.071	0.091	0.12	0.115	0.072		
Lp Oir D 2011	0.030	0.041	0.025	0.027	0.033	0.032	0.029	0.078	0.122	0.090	0.034	0.032	
Lp Oir D 2014	0.013	0.019	0.013	0.018	0.019	0.013	0.008	0.082	0.075	0.072	0.031	0.035	0.019

1 **Table 3:** Pairwise F_{ST} values for parapatric populations (non-significant values are grey coloured and negative values were set to zero).

Populations	*Lf* Aa 2014	*Lf* Hem 2014	*Lf* Bre 2011	*Lf* Ris 2011	*Lf* Ris 2014	*Lf* Odo 2011	*Lf* Oir 2011	*Lf* Oir 2014	*Lf* Loi 2010	*Lf* Loi 2011	*Lf* Dor 2010	*Lf* Dor 2011	*Lf* Gar 2011	*Lp* Aa 2014	*Lp* Hem 2014	*Lp* Bre 2011	*Lp* Ris 2011	*Lp* Ris 2014	*Lp* Odo 2011	*Lp* Oir 2011	*Lp* Oir 2014	*Lp* Cen 2011	*Lp* Jal 2011
Lf Hem D	0																						
Lf Bre D	0	0																					
Lf Ris D	0	0	0																				
Lf Ris D	0	0	0.00	0																			
Lf Odo D	0	0	0.00	0	0																		
Lf Oir D	0.01	0.00	0	0.00	0.00	0.00																	
Lf Oir D	0	0	0	0	0	0.00	0																
Lf Loi D	0.02	0.03	0.01	0.02	0.03	0.03	0.02	0.02															
Lf Loi D	0.02	0.04	0.02	0.02	0.03	0.03	0.03	0.03	0.00														
Lf Dor D	0.05	0.06	0.05	0.05	0.05	0.06	0.06	0.06	0.03	0.01													
Lf Dor D	0.08	0.10	0.10	0.09	0.08	0.09	0.09	0.09	0.07	0.03	0.01												
Lf Gar D	0.03	0.05	0.04	0.03	0.03	0.04	0.04	0.04	0.03	0.00	0.01	0.02											
Lp Aa U	0.09	0.09	0.09	0.10	0.09	0.10	0.11	0.09	0.09	0.08	0.11	0.12	0.08										
Lp Hem U	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.07	0.08	0.11	0.12	0.09	0.08									
Lp Bre U	0.08	0.10	0.07	0.09	0.09	0.11	0.08	0.09	0.05	0.08	0.09	0.13	0.13	0.18	0.17								
Lp Ris U	0.02	0.03	0.03	0.02	0.02	0.03	0.02	0.02	0.03	0.04	0.07	0.10	0.04	0.11	0.09	0.08							
Lp Ris U	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.04	0.04	0.07	0.10	0.04	0.12	0.09	0.09	0						
Lp Odo U	0.16	0.18	0.17	0.16	0.17	0.18	0.17	0.17	0.17	0.16	0.21	0.22	0.17	0.20	0.22	0.23	0.13	0.15					
Lp Oir U	0.09	0.09	0.09	0.10	0.10	0.11	0.08	0.08	0.08	0.08	0.12	0.14	0.10	0.09	0.09	0.14	0.11	0.11	0.18				
Lp Oir U	0.05	0.05	0.04	0.05	0.06	0.07	0.05	0.04	0.05	0.07	0.10	0.15	0.09	0.08	0.08	0.10	0.07	0.08	0.16	0.01			
Lp Cen U	0.16	0.18	0.15	0.18	0.18	0.19	0.16	0.17	0.13	0.15	0.17	0.20	0.20	0.24	0.24	0.16	0.17	0.16	0.27	0.22	0.20		
Lp Jal U	0.05	0.07	0.05	0.06	0.07	0.06	0.06	0.06	0.06	0.07	0.08	0.12	0.10	0.14	0.13	0.10	0.10	0.09	0.27	0.14	0.11	0.22	
Lp Sau U	0.12	0.11	0.12	0.11	0.12	0.13	0.13	0.13	0.16	0.15	0.19	0.22	0.18	0.18	0.19	0.27	0.16	0.14	0.19	0.18	0.15	0.32	0.21

2

3

Discussion

The main aim of our study was to investigate the level of reproductive isolation between two lamprey species by combining experimental measurements of reproductive barriers and analyses of gene flow between sympatric and parapatric population pairs. Our experiments demonstrated that brook lamprey males could reproduce with river lamprey females under semi-natural conditions despite the important size difference between species. Results from artificial fertilizations further supported a low level of reproductive isolation. Population genetic analyses of replicated pairs revealed a continuum of gene flow between species with a pattern of panmixia in some sympatric populations, a moderate differentiation in some other sympatric sites and a strongly reduced gene flow between populations separated by anthropogenic barriers. This gradient of divergence suggests ongoing gene flow in certain sympatric population pairs and some degree of reproductive isolation in other sites. However, secondary contacts after a period of allopatry could also explain this variable degree of genetic differentiation among sympatric sites. In addition, anthropogenic barriers strongly restrict the level of gene flow between species and may thus ultimately promote the evolution of reproductive isolation.

Using artificial fertilizations, we found that *L. planeri* and *L. fluviatilis* males have the same capacity to fertilize eggs of *L. fluviatilis* females. Hatching survival of larvae was also high (nearly 100%) and identical regardless of the male's species. Using a genetically distinct female population (F_{ST} = 0.055), we found that hatching rates were as strong as with crosses using females from the same population, further suggesting a low postzygotic isolation and no outbreeding depression. Recently, Hume *et al.* (2013a) also observed viable hybrids using a similar experimental approach but they obtained much lower values of survival potentially because their experimental design did not allow a distinction of unfertilized and dead embryos. Viable hybrid crosses have been obtained between different pairs of lamprey species but they have never been raised up to the adult stage due to the difficulty of rearing juvenile lampreys (Weissenberg, 1925; Piavis *et al.* 1970; Beamish & Neville, 1992; Hume *et al.*, 2013a). As a result, the fitness of hybrids has never been thoroughly assessed and was only limited to the F1 generation, which prevents an accurate assessment of intrinsic postzygotic barriers. Indeed, genetic incompatibilities are generally best revealed in F2s or backcrosses while heterosis is expected in the F1 generation (e.g. Edmands 1999; Wiley *et al.* 2009). For instance, Bierne *et al.* (2002, 2006) found a pattern of heterosis in F1 crosses of *Mytillus edulis* and *M. galloprovincialis* while F2s were selected against at the larval stage. As a consequence, further studies are needed to better understand the potential mechanisms of postzygotic isolation in lampreys.

Even when postzygotic isolation is low, premating barriers can contribute to reproductive isolation (e.g. Sobel & Streisfeld, 2015). Our results based on mating trials under semi-natural conditions showed for the first time that *L. planeri* males were able to fertilize *L. fluviatilis* females despite the important size difference between species. This interbreeding produced viable hybrid larvae, which suggested a low level of premating isolation and confirmed the low postzygotic isolation at an early developmental stage observed with *in vitro* fertilizations. Size-assortative mating has been suggested to promote divergence and partial or complete reproductive isolation in many taxa including seahorses (Jones *et al.* 2003), sticklebacks (McKinnon *et al.* 2004) or water striders (Han *et al.* 2010). A similar process has been suggested to induce reproductive isolation in lampreys when size differences are greater than 20 % (Beamish & Neville, 1992), a hypothesis that was not confirmed by our results. However, in our experiment the reproductive success of *L. planeri* males was much lower than that of *L. fluviatilis* males and their size differences were far greater than 20 % (Fig. 2) hence some degree of size-assortative mating may occur. *L. planeri* males may also adopt a sneaking strategy, in which they would fertilize some eggs of a *L. fluviatilis* female despite its tendency to mate with larger conspecific males (Hume et *al.,* 2013b). This tactic is widespread in many fish species and may thus limit the evolution of prezygotic isolation in species pairs of lampreys (Gross, 1984; Gage *et al.* 1995; Fleming 1996). However, in the absence of *L. planeri* females in our experiment, *L. planeri* males may have been somehow "forced" to mate with interspecific females, which may have led to an underestimation of the strength of prezygotic barriers. Further experiments including males and females from both species are thus required to produce quantitative estimates of prezygotic isolation in this system.

The low reproductive isolation measured in our experiments on individuals from a single river (Oir) was mirrored by high levels of gene flow in this sympatric site. However, by studying a total of 10 population pairs we found a gradient of increasing differentiation with some sympatric pairs forming a genetically homogeneous population, some others being significantly differentiated and parapatric pairs displaying a high level of divergence. Such a gradient of divergence across multiple pairs in sympatry suggests variable levels of reproductive isolation within a single species complex, which has been observed in relatively few systems (e.g. Gagnaire *et al.*, 2013b; Powell *et al.*, 2013) and emphasizes the interest of using lampreys as a model in speciation studies.

Sympatric pairs in the Oir and Béthune Rivers were not (or weakly) genetically differentiated demonstrating that gene flow can be high between resident and migratory lampreys as suggested in earlier studies (Schreiber & Engelhorn, 1998; Espanhol *et al.*, 2007; Blank *et al.*, 2008; Bracken *et al.* 2015). Similarly, a low differentiation has been observed between ecotypes of resident and migratory rainbow trout *Onchorynchus mykiss* (e.g. Docker & Heath 2003) and is also well documented in

brown trout *Salmo trutta* (e.g. Hindar *et al.,* 1991; Cross *et al.* 1992; Pettersson *et al.* 2001; Charles *et al.* 2005). This suggests that *L. planeri* and *L. fluviatilis* may also represent two ecotypes of a single species. However, we found other sympatric situations (Aa, Hem and Bresle Rivers) where the two species were moderately but significantly differentiated. Accordingly, Mateus *et al.* (2013) found a strong differentiation (F_{ST} = 0.37) in a population pair sampled in the same river system in Portugal. These pairs may be a step further along the divergence continuum, which suggests that disruptive selection and other isolating factors may act in these systems. For instance, some temporal isolation during the spawning season and patchiness of breeding habitat may contribute to ecological divergence between the two species. Analogously, temporal and spatial differences in spawning were linked to genetic differentiation between ecotypes or sub-populations of various salmonid species (Deiner *et al.* 2007; Pearse *et al.* 2009). In addition, the magnitude of size differences between species seems to vary among rivers (Table 1) and this factor may contribute to variations of reproductive isolation as predicted by theory (Bolnick 2011) and observed in other taxa (Arnqvist *et al.* 1996; McKinnon *et al.* 2004; Martin 2013)). Our experiments showed a low premating isolation in lampreys from the Oir River where the size difference and the genetic differentiation are low between species. Similar experiments in other sympatric sites would thus be required to better understand the role of size assortative mating in the evolution of reproductive isolation. Besides, the significant genetic differentiation observed in sympatric situations may also reflect some genetic barriers to gene flow (Wu, 2001; Tuner *et al.,* 2005). However, if they exist, these barriers may not be distributed over large portions of the genome and may not efficiently counteract gene flow since genetic differentiation as well as overall reproductive isolation were low (Barton & Bengtsson 1986; Wu, 2001).

The continuum of genetic differentiation observed in sympatric sites could arise from two different historical scenarios of divergence: (1) ecologically based speciation with gene flow or (2) differential introgression following a secondary contact after a period of allopatric divergence. If the populations have diverged in allopatry for a period of time too short to allow complete reproductive isolation it is also possible that secondary contacts have occurred at different times in different areas so that in some cases (e.g. Oir River) a single panmictic population is currently found while in other situations the genome swamping between *L. planeri* and *L. fluviatilis* is still incomplete. Such scenarios of secondary contacts have been suggested in many taxa including Cameroon crater lake cichlids (Martin *et al.* 2015), whitefish (Gagnaire *et al*., 2013b), voles (Beysard & Heckel 2014) and may have also played a crucial role in the evolution of reproductive isolation in the apple maggot, which is considered as a classical model of sympatric speciation (Feder *et al.* 2003). Nevertheless, cases of ongoing gene flow in sympatry between closely related species have been often interpreted as evidences for ecological speciation whereas the hypothesis of admixture following a secondary

contact (or the one of local adaptation) was either not considered or could not be definitively rejected (Via 2001; Michel *et al.* 2010; Hohenlohe *et al.* 2012; Kautt *et al.* 2012). These different scenarios of divergence are difficult to disentangle but new modelling approaches may help tackle this issue as described in several recent studies (Duvaux *et al.* 2011; Roux *et al.,* 2013; Roux *et al.,* 2014; Butlin *et al.,* 2014; Tine *et al.,* 2015).

In contrast to sympatric situations, we observed high levels of differentiation in most parapatric sites as expected under the joint effects of isolation by distance and anthropogenic barriers to migration. In these situations migration was reduced and asymmetric from *L. planer*i (upstream) to *L. fluviatili*s (Table S6), highlighting the low migratory ability of *L. fluviatili*s in the presence of obstacles (Russon *et al.* 2011; Foulds & Lucas 2013; Bracken *et al.* 2015). Combined negative effects of distance and barriers on gene flow have also been observed in several fish species (Thrower *et al.* 2004; Raeymaekers *et al.* 2008; Gomez-Uchida *et al.* 2009) and more generally in many taxa (see Templeton *et al.* 2001; Fahrig 2003). Our results thus highlight the importance of untangling the effects of habitat fragmentation inducing restricted gene flow from ecological divergence between habitats when studying speciation. Inferences about the speciation process might be obscured by the effect of barriers to migration and isolation by distance. Sampling should thus be carefully designed to clearly distinguish sympatric and parapatric sites even at a within-river scale. Ultimately, habitat fragmentation could promote the evolution of reproductive isolation and lead to founder-induced speciation but it seems more likely to induce local extinctions (Templeton 1980, 2008).

Anthropogenic barriers to migration did not have the same impact on patterns of genetic diversity and differentiation of *L. planer*i and *L. fluviatili*s populations. We found high levels of genetic structure combined with clustering at the river level in *L. planeri* suggesting that each resident population tends to evolve as an independent evolutionary unit due to low dispersal ability and isolation by anthropogenic barriers in upstream reaches. Similar results were obtained in earlier studies based on allozyme or mtDNA in northern Europe and in the Iberian Peninsula (Schreiber & Engelhorn, 1998; Pereira *et al.*, 2010) and also recently based on microsatellite data in United Kingdom (Bracken *et al.,* 2015). In contrast, populations of the migratory *L. fluviatilis* were weakly differentiated and structured at the regional level. The genetic diversity (both allelic richness and expected heterozygosity) of *L. planeri* populations was also lower than the one of *L. fluviatilis* populations indicating an important role of genetic drift in isolated brook lamprey populations. A similar pattern has been observed in several salmonid species between freshwater resident and anadromous populations (Gomez-Uchida *et al.*, 2009; Perrier *et al.*, 2013). Finally, the fact that isolation by distance was lower in *L. planeri* than in *L. fluviatilis* further highlights the impact of

anthropogenic barriers on the genetic differentiation among *L. planeri* populations (see also Bracken *et al.*, 2015).

To conclude, our results suggest that *L. fluviatilis* and *L. planeri* may form partially reproductively isolated ecotypes. The variable levels of gene flow among sympatric sites show that different pairs of populations are either at different stages of divergence along the speciation continuum or at different levels of fusion following secondary contacts. In the first hypothesis, size assortative mating and selection could act together to maintain phenotypic differences in the face of gene flow. The relative strength of these factors may vary among rivers, resulting in variable progress toward sympatric speciation or even stalled speciation (Bolnick, 2011). In the second hypothesis, similar patterns of varying levels of divergence would arise by secondary contacts (Bierne *et al.,* 2013). Combining genome-wide analyses and modeling of complex historical processes may help untangling these different scenarios. Ultimately, experimental approaches testing the long-term fitness of F1s and later generation hybrids will allow deeper investigations of the mechanisms of postzygotic reproductive isolation in this system. Common garden experiments may also allow clarify the relative roles of phenotypic plasticity and genetic factors in the emergence of parasitic and nonparasitic life histories.

Acknowledgements

We thank F. Marchand, J. Tremblay, A. Oger, V. Dolo, Y. Salaville, R. Lemasquerier, V. Lauronce, C. Taverny, B. Rigault C. Rigaud, Y. Perraud, C. Perrier, G. Sanson, J.-L. Fagard and P. Domalain (ONEMA) who helped us collect the samples. We thank L. Benestan and J.- S. Moore for valuable comments on the manuscript as well as D. R. Matute and two anonymous referees. This study was funded by the European Regional Development Fund (Transnational program Interreg IV, Atlantic Aquatic Resource Conservation Project).

Data deposited at Dryad: doi: 10.5061/dryad.5qv85

Supporting information

Detailed protocol for artificial fertilizations

Each progenitor was anesthetized with benzocaïne and gametes were then collected by manual stripping. Eggs from each female were distributed into eight Petri dishes. We added 20 to 30 µl of milt in each Petri dish, which were then half-filled with dechlorated water. The fertilization success for each sib group was estimated three hours after fertilization based on the presence of a perivitelline space in the eggs (Ciereszko *et al.*, 2000).

A total of 24 fertilized eggs from each sib group were then individually distributed in 24 wells plates filled with 2ml of chemically standardized water reconstituted according to OECD guidelines (OECD 1992). Eggs were then incubated in a climate chamber at a constant temperature of 12°C +/- 1°C to measure hatching rate.

References

OECD. 1992. OECD guideline for the testing of chemicals 203 (fish acute toxicity test). Annex 2.

Figure S1: Bayesian individual clustering results with STRUCTURE: *k* = 2, 8 and 10 for the two species analysed together (a, b and c, respectively) and *k* = 2 and 9 for *L. fluviatilis* (d) and *L. planeri* (e) respectively.

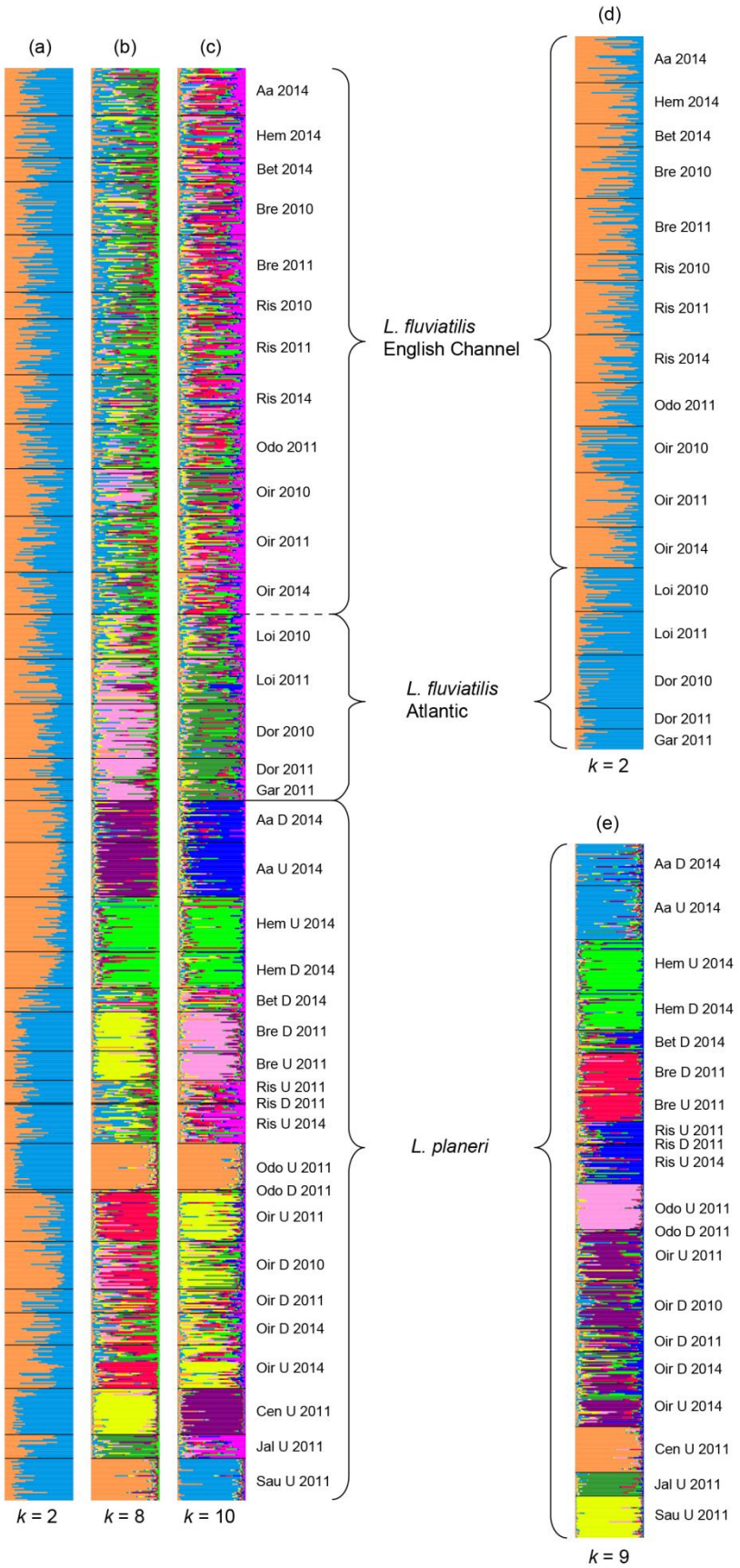


Table S1: Results of a Generalized linear mixed model (GLMM) testing the effect of cross type (homo- *versus* heterospecific) and maternal population (Loire *versus* Oir) on fertilization success of *L. fluviatilis* eggs.

		d.f.	χ^2	p-value	Estimates (± s.e.)
Fertilization success	Cross type	1	0.16	ns	-0.11 ± 0.12
	Pop.	1	2.45	ns	0.51 ± 0.45
	Cross type × Pop.	1	4.31	**0.04**	0.52 ± 0.24

Table S2: Genetic diversity estimates for each locus in each population.

		LP-03	LP-06	LP-43	LP-09	LP-18	LP-22	LP-27	LP-28	LP-30	LP-37	LP-39	LP-45	LP-46
Lf Aa D 2014	Na	3	2	7	3	2	4	6	3	4	3	4	4	5
	Ar	2.357	1.859	4.849	2.565	1.697	3.762	4.206	2.143	3.097	2.917	3.423	2.817	3.511
	Ho	0.4412	0.1765	0.1471	0.6176	0.7576	0.3824	0.7059	0.7059	0.5882	0.1765	0.5588	0.6176	0.6765
	He	0.4091	0.2107	0.1383	0.6119	0.7846	0.3863	0.6787	0.7392	0.6234	0.2998	0.5931	0.5571	0.6457
	Fis	-0.08	0.165	0.035	0.01	-0.065	-0.041	0.046	0.415	-0.009	0.057	0.059	-0.111	-0.048
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Hem D 2014	Na	2	3	9	4	2	4	5	2	3	3	4	4	4
	Ar	1.993	1.982	5.928	2.871	1.813	3.798	4.073	1.861	2.665	2.556	3.157	3.274	3.313
	Ho	0.4	0.2	0.2	0.6333	0.8276	0.4333	0.6	0.8	0.5333	0.1667	0.5	0.6333	0.5333
	He	0.3977	0.1859	0.1831	0.5554	0.8167	0.4492	0.6898	0.7107	0.552	0.2096	0.6266	0.6028	0.5655
	Fis	-0.006	-0.077	-0.014	0.036	-0.094	0.132	-0.128	0.208	-0.143	0.034	0.205	-0.052	0.058
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Bet D 2014	Na	3	2	8	4	2	4	4	2	2	3	4	3	5
	Ar	2.356	1.971	5.681	2.822	1.971	3.901	3.741	1.971	2	2.943	3.353	2.803	3.761
	Ho	0.3529	0.2353	0.3529	0.5882	0.6875	0.4118	0.6471	0.4375	0.6471	0.2353	0.5294	0.5882	0.4706
	He	0.3084	0.2995	0.2995	0.4991	0.8065	0.508	0.7041	0.631	0.5971	0.2995	0.6132	0.4938	0.6275
	Fis	-0.15	0.22	0.152	0.194	-0.185	0.083	0.314	0.22	-0.185	-0.086	0.14	-0.199	0.256
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Bre D 2010	Na	2	2	2	4	5	3	6	5	4	2	3	4	3
	Ar	2	1.933	1.998	4	4.933	3	5.8	4.933	3.933	1.933	3	3.931	3
	Ho	0.333	0.067	0.133	0.357	0.733	0.533	0.6	0.933	0.667	0.067	0.467	0.6	0.6
	He	0.286	0.067	0.129	0.544	0.679	0.536	0.69	0.776	0.652	0.067	0.519	0.605	0.514
	Fis	-0.167	0	-0.037	0.343	-0.081	0.004	0.131	-0.202	-0.022	0	0.101	0.008	-0.167
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Bre D 2011	Na	4	2	2	3	7	3	5	4	3	2	3	4	5
	Ar	2.91	1.982	1.971	2.999	5.556	2.951	4.331	3.973	2.982	1.971	3	3.745	3.904
	Ho	0.463	0.22	0.195	0.585	0.625	0.415	0.692	0.625	0.537	0.195	0.634	0.585	0.61
	He	0.388	0.198	0.178	0.605	0.716	0.401	0.661	0.668	0.6	0.178	0.626	0.527	0.609
	Fis	-0.193	-0.111	-0.096	0.032	0.127	-0.035	-0.047	0.064	0.105	-0.096	-0.013	-0.112	-0.002
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	yes	no
Lf Ris D 2010	Na	3	2	2	4	8	3	6	4	4	2	4	3	4

	Ar	2.736	1.986	1.936	3.723	7.48	2.737	5.41	3.999	3.737	2	3.737	2.997	3.737
	Ho	0.316	0.158	0.105	0.632	0.737	0.368	0.684	0.737	0.789	0.211	0.632	0.737	0.737
	He	0.279	0.149	0.102	0.607	0.781	0.482	0.713	0.734	0.652	0.275	0.674	0.595	0.675
	Fis	-0.131	-0.059	-0.029	-0.041	0.056	0.236	0.041	-0.004	-0.211	0.234	0.063	-0.238	-0.091
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Ris D 2011	Na	3	2	3	4	8	3	4	5	3	2	4	3	3
	Ar	2.359	1.99	2.282	3.349	6.71	2.892	3.992	4.331	2.999	1.932	3.349	2.58	2.997
	Ho	0.436	0.2	0.125	0.625	0.7	0.35	0.65	0.725	0.65	0.15	0.575	0.45	0.6
	He	0.449	0.222	0.164	0.638	0.772	0.445	0.694	0.696	0.621	0.14	0.592	0.506	0.562
	Fis	0.03	0.098	0.238	0.02	0.094	0.213	0.063	-0.042	-0.046	-0.068	0.028	0.11	-0.068
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Ris D 2014	Na	3	2	9	3	3	4	5	3	3	3	4	3	4
	Ar	2.197	1.752	5.722	2.346	2.148	3.566	4.153	2.007	2.806	2.849	3.155	2.806	3.147
	Ho	0.4	0.1143	0.1429	0.7429	0.8	0.3429	0.6571	0.8	0.6286	0.1429	0.6176	0.6	0.5429
	He	0.446	0.159	0.3118	0.5909	0.8033	0.4008	0.6658	0.7068	0.5996	0.2083	0.6264	0.576	0.5959
	Fis	0.104	0.284	0.004	0.146	0.545	0.013	-0.134	0.317	-0.262	-0.049	0.014	-0.042	0.09
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Odo D 2011	Na	2	3	2	4	8	4	5	4	3	3	4	3	3
	Ar	2	2.424	1.997	3.434	6.453	2.875	4.266	3.997	2.998	2.125	3.685	2.437	2.998
	Ho	0.484	0.188	0.219	0.625	0.625	0.438	0.75	0.688	0.719	0.094	0.531	0.625	0.563
	He	0.403	0.226	0.246	0.632	0.742	0.395	0.684	0.717	0.61	0.092	0.543	0.501	0.532
	Fis	-0.2	0.171	0.111	0.011	0.158	-0.109	-0.096	0.041	-0.179	-0.022	0.022	-0.249	-0.058
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Oir D 2010	Na	3	3	2	3	9	3	4	4	3	3	5	3	4
	Ar	2.842	2.894	2	3	6.767	2.657	3.988	3.998	2.944	2.393	4.627	2.658	3.42
	Ho	0.613	0.281	0.545	0.697	0.824	0.294	0.706	0.727	0.485	0.235	0.618	0.588	0.667
	He	0.53	0.303	0.469	0.585	0.793	0.343	0.652	0.732	0.571	0.213	0.596	0.533	0.623
	Fis	-0.156	0.073	-0.164	-0.191	-0.039	0.143	-0.083	0.007	0.152	-0.102	-0.036	-0.103	-0.071
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Oir D 2011	Na	4	2	2	4	7	3	4	5	3	2	4	3	4
	Ar	2.93	1.99	1.984	3.349	5.443	2.725	3.915	4.348	2.932	2	3.592	2.892	3.344
	Ho	0.475	0.15	0.225	0.725	0.575	0.325	0.513	0.825	0.6	0.4	0.667	0.5	0.625
	He	0.456	0.222	0.202	0.611	0.684	0.303	0.571	0.735	0.571	0.379	0.677	0.546	0.589

	Fis	-0.041	0.326	-0.114	-0.187	0.16	-0.074	0.102	-0.123	-0.052	-0.054	0.015	0.085	-0.062
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Oir D 2014	Na	2	2	7	4	3	4	4	2	3	3	3	3	5
	Ar	1.995	1.861	4.654	2.783	1.997	3.352	3.837	1.946	2.838	2.68	2.981	2.413	3.443
	Ho	0.4333	0.2333	0.1154	0.8214	0.7667	0.3667	0.5333	0.8276	0.5862	0.2	0.5862	0.6333	0.6552
	He	0.413	0.2096	0.1802	0.5844	0.7469	0.4898	0.6164	0.7241	0.5693	0.2825	0.6358	0.4842	0.6316
	Fis	-0.05	-0.115	-0.027	0.255	0.364	0.137	-0.146	0.296	-0.416	-0.03	0.079	-0.315	-0.038
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Loi D 2010	Na	2	3	2	3	5	3	5	4	4	2	4	3	3
	Ar	2	2.431	2	2.951	4.313	2.818	4.409	3.95	3.937	1.703	3.822	2.829	3
	Ho	0.375	0.281	0.438	0.563	0.633	0.2	0.563	0.813	0.625	0.065	0.531	0.344	0.656
	He	0.308	0.249	0.38	0.486	0.644	0.244	0.59	0.691	0.658	0.063	0.573	0.477	0.625
	Fis	-0.216	-0.13	-0.151	-0.156	0.017	0.181	0.046	-0.176	0.051	-0.017	0.072	0.279	-0.05
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Loi D 2011	Na	2	3	2	3	5	3	5	4	4	2	5	3	3
	Ar	2	2.987	2	3	4.636	2.925	4.63	3.974	3.684	1.688	4.463	2.907	3
	Ho	0.406	0.469	0.375	0.563	0.594	0.313	0.563	0.781	0.563	0.063	0.375	0.531	0.688
	He	0.449	0.524	0.381	0.598	0.676	0.304	0.54	0.717	0.627	0.061	0.486	0.558	0.67
	Fis	0.094	0.106	0.016	0.059	0.122	-0.028	-0.042	-0.089	0.102	-0.016	0.228	0.048	-0.026
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Dor D 2010	Na	2	3	2	3	5	3	5	4	4	2	5	2	3
	Ar	2	2.998	2	2.903	4.826	2.661	4.531	3.9	3.348	1.359	4.354	2	3
	Ho	0.757	0.538	0.769	0.368	0.692	0.211	0.622	0.846	0.711	0.026	0.333	0.436	0.658
	He	0.503	0.531	0.487	0.356	0.736	0.196	0.663	0.711	0.581	0.026	0.417	0.398	0.674
	Fis	-0.504	-0.014	-0.581	-0.035	0.06	-0.074	0.062	-0.191	-0.223	0	0.201	-0.095	0.024
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Dor D 2011	Na	2	3	2	3	6	3	4	4	3	2	3	3	3
	Ar	2	3	2	2.933	5.933	2.998	4	3.998	2.998	2	2.931	2.998	3
	Ho	0.867	0.533	0.533	0.333	0.8	0.2	0.533	0.867	0.6	0.2	0.2	0.733	0.667
	He	0.495	0.55	0.495	0.343	0.824	0.305	0.607	0.676	0.569	0.186	0.19	0.564	0.681
	Fis	-0.75	0.03	-0.077	0.028	0.029	0.344	0.122	-0.282	-0.054	-0.077	-0.05	-0.3	0.021
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lf Gar D 2011	Na	2	3	2	3	5	2	4	4	4	2	5	3	3

	Ar	2	2.998	2	3	5	2	3.933	4	3.933	2	4.931	2.933	3
	Ho	0.786	0.4	0.6	0.643	0.667	0.267	0.733	0.733	0.6	0.071	0.533	0.4	0.733
	He	0.505	0.548	0.476	0.599	0.793	0.238	0.657	0.707	0.617	0.071	0.607	0.555	0.679
	Fis	-0.554	0.27	-0.26	-0.073	0.159	-0.12	-0.116	-0.037	0.027	0	0.122	0.279	-0.081
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Aa D 2014	Na	2	2	5	3	2	3	5	2	3	4	4	3	4
	Ar	1.986	1.986	3.535	2.415	1.999	2.925	4.167	1.241	2.861	3.806	3.507	2.98	2.921
	Ho	0.4667	0.4667	0.5333	0.8	0.6	0.4	0.7333	0.7333	0.8667	0.0345	0.6333	0.5333	0.5172
	He	0.3638	0.3638	0.4723	0.5949	0.5718	0.5356	0.6198	0.7616	0.7124	0.0345	0.5847	0.6571	0.5862
	Fis	-0.289	-0.289	-0.05	0.256	-0.132	-0.187	0.038	0	-0.353	-0.221	-0.085	0.191	0.119
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Aa U 2014	Na	2	3	6	3	4	3	4	2	3	4	4	3	4
	Ar	1.858	2.323	4.216	2.179	2.353	2.811	3.95	1.452	2.92	3.797	3.354	2.849	3.091
	Ho	0.2368	0.4103	0.4872	0.5897	0.5946	0.4872	0.4615	0.8108	0.7368	0.0769	0.5263	0.4615	0.5128
	He	0.2116	0.4462	0.4466	0.5931	0.7079	0.4952	0.5861	0.7579	0.7172	0.0749	0.5582	0.5991	0.5941
	Fis	-0.213	-0.017	0.529	-0.088	0.06	0.115	-0.137	NA	-0.226	-0.03	0.113	0.204	0.003
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Hem D 2014	Na	2	3	6	3	2	4	4	1	3	4	4	4	5
	Ar	1.999	2.12	3.376	2.179	1.904	3.232	3.86	1	2.993	2.342	3.028	2.892	3.738
	Ho	0.5641	0.3077	0.2308	0.8205	0.2308	0.5641	0.5641	0.8205	0.4103	0	0.5385	0.3846	0.6667
	He	0.4662	0.3027	0.2454	0.671	0.4862	0.5191	0.6364	0.7229	0.3986	0	0.6061	0.4822	0.6687
	Fis	-0.121	0.082	0.162	0.016	-0.092	0.215	-0.071	-0.027	0.006	-0.028	0.058	0.232	0.138
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Hem U 2014	Na	2	3	5	2	3	4	4	1	3	3	3	2	3
	Ar	1.906	2.45	3.327	2	2.279	3.523	3.922	1	2.981	2.239	2.864	1.998	2.996
	Ho	0.2692	0.3077	0.3333	0.6538	0.5	0.6923	0.8077	0.8077	0.4231	0	0.5769	0.3846	0.8
	He	0.2376	0.3989	0.3927	0.6365	0.4879	0.5068	0.6584	0.7443	0.3477	0	0.6041	0.4344	0.6784
	Fis	-0.136	0.232	-0.025	-0.376	0.154	-0.232	-0.087	NA	-0.028	-0.222	0.046	0.117	-0.184
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Bet D 2014	Na	4	2	7	2	2	3	5	2	3	3	3	4	4
	Ar	2.823	1.993	5.171	1.661	1.944	2.356	4.286	1.944	2.895	2.964	2.942	2.795	3.791
	Ho	0.5882	0.3529	0.2941	0.7059	0.5625	0.1176	0.3529	0.8235	0.5882	0.1765	0.4706	0.4706	0.7647
	He	0.5276	0.3708	0.2585	0.6132	0.754	0.1141	0.3084	0.713	0.5847	0.2585	0.5348	0.3939	0.6988

	Fis	-0.119	0.05	0.26	-0.032	-0.143	-0.15	-0.161	0.324	-0.157	-0.006	0.123	-0.202	-0.098
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Bre U 2011	Na	3	1	2	2	6	2	2	6	3	2	3	3	3
	Ar	2.699	1	2	1.882	5.134	1.988	1.944	5.326	3	1.975	2.5	2.754	2.755
	Ho	0.214	0	0.429	0.107	0.679	0.214	0.143	0.786	0.714	0.083	0.357	0.25	0.464
	He	0.2	0	0.381	0.103	0.627	0.194	0.135	0.747	0.663	0.158	0.391	0.335	0.545
	Fis	-0.073	NA	-0.125	-0.038	-0.082	-0.102	-0.059	-0.051	-0.077	0.471	0.086	0.253	0.148
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Bre D 2011	Na	3	2	2	2	6	1	2	5	3	2	3	3	4
	Ar	2.664	1.667	2	1.667	5.594	1	1.894	4.982	3	2	2.998	2.789	3.666
	Ho	0.286	0.048	0.19	0.048	0.762	0	0.095	0.81	0.81	0.2	0.381	0.19	0.714
	He	0.256	0.048	0.319	0.048	0.7	0	0.093	0.751	0.665	0.332	0.487	0.181	0.631
	Fis	-0.116	0	0.403	0	-0.088	NA	-0.026	-0.078	-0.216	0.397	0.218	-0.053	-0.132
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Ris U 2011	Na	4	3	3	3	5	2	4	4	3	2	3	3	5
	Ar	3.875	2.863	2.999	2.875	4.862	2	3.874	3.999	3	1.999	3	3	4.851
	Ho	0.625	0.188	0.375	0.625	0.813	0.375	0.5	0.813	0.813	0.188	0.625	0.563	0.563
	He	0.498	0.179	0.377	0.481	0.621	0.313	0.423	0.702	0.671	0.175	0.656	0.671	0.571
	Fis	-0.255	-0.047	0.006	-0.299	-0.309	-0.2	-0.182	-0.157	-0.211	-0.071	0.048	0.161	0.015
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Ris U 2014	Na	7	3	6	3	3	4	4	2	3	4	3	3	4
	Ar	3.769	2.795	3.386	2.12	1.83	3.047	3.636	1.598	2.659	3.281	2.878	2.934	2.446
	Ho	0.4419	0.3953	0.1628	0.5116	0.4884	0.1628	0.5814	0.5952	0.6667	0.1163	0.5814	0.6279	0.3256
	He	0.5193	0.4462	0.1529	0.49	0.5425	0.3253	0.5352	0.6899	0.6687	0.1108	0.5289	0.6293	0.4134
	Fis	0.151	0.115	0.101	0.503	-0.065	-0.088	0.139	-0.05	-0.045	0.003	-0.101	0.002	0.214
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Odo U 2011	Na	3	2	2	2	4	2	1	3	3	1	3	2	3
	Ar	2.815	1.898	1.424	1.97	3.888	2	1	3	2.999	1	2.424	1.944	2.984
	Ho	0.455	0.121	0.03	0.182	0.606	0.455	0	0.606	0.848	0	0.364	0.152	0.636
	He	0.372	0.116	0.03	0.168	0.634	0.501	0	0.623	0.638	0	0.511	0.142	0.535
	Fis	-0.221	-0.049	0	-0.085	0.045	0.093	NA	0.027	-0.329	NA	0.288	-0.067	-0.189
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Oir U 2011	Na	3	2	2	3	6	2	4	4	3	2	4	2	4

	Ar	2.871	2	2	2.874	4.608	2	3.545	3.958	2.878	2	3.965	2	3.377
	Ho	0.371	0.543	0.343	0.4	0.629	0.457	0.382	0.571	0.543	0.4	0.743	0.429	0.457
	He	0.321	0.5	0.388	0.34	0.637	0.504	0.478	0.647	0.56	0.438	0.676	0.466	0.487
	Fis	-0.157	-0.086	0.117	-0.175	0.013	0.093	0.2	0.116	0.031	0.086	-0.099	0.081	0.061
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Oir D 2010	Na	2	3	2	3	7	3	4	4	4	2	5	2	3
	Ar	2	2.969	2	2.867	5.714	2.907	3.474	3.995	2.875	2	4.683	1.998	2.997
	Ho	0.758	0.552	0.545	0.303	0.688	0.469	0.536	0.636	0.375	0.382	0.667	0.303	0.625
	He	0.502	0.504	0.484	0.272	0.675	0.526	0.539	0.736	0.537	0.381	0.614	0.26	0.529
	Fis	-0.509	-0.094	-0.127	-0.115	-0.018	0.109	0.006	0.135	0.301	-0.005	-0.085	-0.164	-0.182
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Oir D 2011	Na	2	2	2	3	6	3	4	4	3	2	4	2	3
	Ar	2	2	2	3	5.47	2.824	3.823	4	2.824	2	3.973	2	2.824
	Ho	0.353	0.412	0.353	0.706	0.588	0.588	0.529	0.765	0.471	0.294	0.529	0.471	0.647
	He	0.298	0.489	0.298	0.546	0.68	0.439	0.513	0.761	0.509	0.257	0.577	0.368	0.504
	Fis	-0.185	0.158	-0.185	-0.293	0.135	-0.339	-0.032	-0.005	0.076	-0.143	0.083	-0.28	-0.285
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Oir D 2014	Na	2	2	6	3	2	4	5	2	3	3	4	2	3
	Ar	1.999	1.96	4.535	2.304	1.853	3.222	4.151	1.956	2.667	2.884	3.614	1.999	2.979
	Ho	0.5909	0.2609	0.2174	0.6087	0.7	0.5217	0.3636	0.8182	0.6667	0.2381	0.6818	0.5	0.7727
	He	0.4598	0.2937	0.1981	0.4841	0.659	0.5034	0.519	0.7336	0.5889	0.2846	0.5613	0.4598	0.6543
	Fis	-0.294	0.114	-0.064	-0.037	-0.1	0.304	-0.118	0.167	-0.265	-0.136	-0.221	-0.09	-0.186
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Oir U 2014	Na	2	2	6	2	2	3	4	2	3	3	4	3	3
	Ar	1.888	1.996	4.093	2	1.8	2.519	3.783	1.977	2.38	2.8	3.736	2.209	2.65
	Ho	0.2581	0.3226	0.1935	0.3548	0.7	0.6452	0.5161	0.6452	0.6774	0.2903	0.7419	0.3548	0.3548
	He	0.2285	0.4188	0.1777	0.3866	0.665	0.5034	0.4098	0.6531	0.5732	0.3369	0.6838	0.3802	0.5336
	Fis	-0.132	0.233	-0.054	-0.288	-0.091	-0.265	0.012	0.14	0.083	-0.185	-0.087	0.068	0.339
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Cen U 2011	Na	3	1	3	2	5	1	2	3	3	1	2	2	4
	Ar	2.416	1	2.807	1.898	3.869	1	1.974	2.951	2.984	1	2	2	3.322
	Ho	0.273	0	0.242	0.061	0.3	0	0.188	0.406	0.387	0	0.406	0.515	0.576
	He	0.242	0	0.294	0.116	0.301	0	0.172	0.525	0.403	0	0.364	0.46	0.504

	Fis	-0.125	NA	0.176	0.48	0.004	NA	-0.088	0.226	0.039	NA	-0.116	-0.119	-0.143
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Jal U 2011	Na	2	2	2	3	4	2	4	3	4	3	3	2	2
	Ar	2	2	1.824	3	3.824	2	3.82	3	3.973	2.824	2.824	2	2
	Ho	0.588	0.118	0.059	0.529	0.647	0.176	0.706	0.588	0.706	0.353	0.412	0.294	0.588
	He	0.511	0.305	0.059	0.551	0.618	0.342	0.618	0.686	0.697	0.384	0.346	0.338	0.496
	Fis	-0.151	0.614	0	0.04	-0.048	0.484	-0.143	0.142	-0.013	0.081	-0.191	0.13	-0.185
	Nul allele	no	no	no	no	no	no	no	no	no	no	no	no	no
Lp Sau U 2011	Na	3	1	1	3	3	3	4	2	3	2	2	3	3
	Ar	2.995	1	1	2.964	3	2.982	3.467	2	2.72	1.855	2	2.992	2.998
	Ho	0.433	0	0	0.367	0.433	0.567	0.833	0.367	0.6	0.1	0.533	0.433	0.6
	He	0.473	0	0	0.345	0.667	0.567	0.683	0.345	0.53	0.097	0.487	0.434	0.575
	Fis	0.084	NA	NA	-0.062	0.35	0.001	-0.221	-0.063	-0.131	-0.036	-0.094	0.001	-0.043
	Nul allele	no	no	no	no	yes	no	no	no	no	no	no	no	no

Table S3: Results of permutations tests comparing Allelic richness (*Ar*) and expected heterozygosity (*He*) between *Lf* and *Lp* in each river system. Significant values are in bold-italic and all tests were performed using paired Wilcoxon tests in R.

Lf	***Lf***	**V**	***P-value***
	Ar		
Aa D 2014	Aa D 2014	63	0.244
Aa D 2014	Aa U 2014	61	0.305
Hem D 2014	Hem D 2014	68	0.127
Hem D 2014	Hem U 2014	69	0.108
Bre D 2011	Bre D 2011	67	0.142
Bre D 2011	Bre U 2011	73	0.057
Bet D 2014	Bet D 2014	51	0.727
Oir D 2011	Oir D 2011	58	0.147
Oir D 2010	Oir D 2010	66	***0.038***
Oir D 2011	Oir U 2011	65	0.046
Oir D 2014	Oir D 2014	55	0.542
Oir D 2014	Oir U 2014	71	0.080
Ris D 2011	Ris U 2011	41	0.787
Ris D 2014	Ris U 2014	60	0.340
Odo D 2011	Odo U 2011	84	***0.005***
Loi D 2011	Cen U 2011	82	***0.008***
Dor D 2011	Jal U 2011	66	***0.037***
Gar D 2011	Jal U 2011	50	***0.025***
Gar D 2011	Sau U 2011	75	***0.043***
	He		
Aa D 2014	Aa D 2014	30	0.294
Aa D 2014	Aa U 2014	52	0.685
Hem D 2014	Hem U 2014	44	0.946
Hem D 2014	Hem D 2014	42	0.839

Bre D 2011	Bre D 2011	66	0.168
Bre D 2011	Bre U 2011	76	***0.033***
Bet D 2014	Bet D 2014	46	1.000
Oir D 2011	Oir D 2011	59	0.376
Oir D 2010	Oir D 2010	55	0.542
Oir D 2011	Oir U 2011	53	0.636
Oir D 2014	Oir D 2014	39	0.685
Oir D 2014	Oir U 2014	55.5	0.507
Ris D 2011	Ris U 2011	44	0.947
Ris D 2014	Ris U 2014	64	0.216
Odo D 2011	Odo U 2011	81	***0.011***
Loi D 2011	Cen U 2011	91	***0.000***
Dor D 2011	Jal U 2011	52	0.685
Gar D 2011	Jal U 2011	67	0.149
Gar D 2011	Sau U 2011	78	***0.021***

***Lp* U**	***Lp* D**	**V**	***P-value***
Ar			
Aa D 2014	Aa U 2014	38	0.636
Hem D 2014	Hem U 2014	48	0.505
Bre D 2011	Bre U 2011	38	0.689
Oir D 2011	Oir U 2011	18	0.636
Oir D 2014	Oir U 2014	75	***0.040***
He			
Aa D 2014	Aa U 2014	73	***0.057***
Hem D 2014	Hem U 2014	19	0.415
Bre D 2011	Bre U 2011	52	0.685
Oir D 2011	Oir U 2011	35	0.497
Oir D 2014	Oir U 2014	51	0.367

Table S4: Pairwise F_{ST} among all populations (non-significant values are grey coloured).

Populations	*Lf* Aa D 2014	*Lf* Hem D 2014	*Lf* Bet D 2014	*Lf* Bre D 2010	*Lf* Bre D 2011	*Lf* Ris D 2010	*Lf* Ris D 2011	*Lf* Ris D 2014	*Lf* Odo D 2011	*Lf* Oir D 2010	*Lf* Oir D 2011	*Lf* Oir D 2014	*Lf* Loi D 2010	*Lf* Loi D 2011	*Lf* Dor D 2010	*Lf* Dor D 2011	*Lf* Gar D 2011	*Lp* Aa D 2014	*Lp* Aa U 2014	*Lp* Hem D 2014	*Lp* Hem U 2014	*Lp* Bet D 2014	*Lp* Bre D 2011	*Lp* Bre U 2011	*Lp* Ris U 2011	*Lp* Ris U 2014	*Lp* Odo U 2011	*Lp* Oir U 2011	*Lp* Oir D 2010	*Lp* Oir D 2011	*Lp* Oir D 2014	*Lp* Oir U 2014	*Lp* Cen U 2011	*Lp* Jal U 2011
Lf Hem D 2014	-0.005																																	
Lf Bet D 2014	0.002	0.001																																
Lf Bre D 2010	0	0.005	0.006																															
Lf Bre D 2011	-0.002	-0.002	0.003	-0.001																														
Lf Ris D 2010	-0.007	-0.004	-0.003	0.006	0.004																													
Lf Ris D 2011	-0.004	-0.001	0.002	0.002	-0.001	0.001																												
Lf Ris D 2014	-0.006	-0.008	-0.001	0.008	0.002	-0.005	-0.004																											
Lf Odo D 2011	-0.002	-0.002	0.008	0.009	0.006	0.004	-0.003	-0.003																										
Lf Oir D 2010	0.013	0.023	0.024	0.024	0.021	0.031	0.013	0.013	0.018																									
Lf Oir D 2011	0.001	0.001	0.017	0.012	0.002	0.006	0.006	0.004	0.009	0.018																								
Lf Oir D 2014	-0.006	-0.003	0	0.004	0	-0.006	-0.004	-0.006	0.001	0.018	-0.001																							
Lf Loi D 2010	0.024	0.034	0.032	0.018	0.015	0.032	0.028	0.031	0.033	0.021	0.025	0.026																						
Lf Loi D 2011	0.026	0.043	0.041	0.026	0.026	0.041	0.027	0.031	0.035	0.01	0.032	0.03	0.008																					
Lf Dor D 2010	0.059	0.068	0.078	0.06	0.056	0.089	0.055	0.058	0.063	0.022	0.065	0.061	0.033	0.014																				
Lf Dor D 2011	0.086	0.102	0.116	0.102	0.101	0.12	0.092	0.086	0.091	0.029	0.094	0.091	0.072	0.032	0.011																			
Lf Gar D 2011	0.037	0.051	0.053	0.043	0.049	0.057	0.037	0.039	0.046	0.005	0.046	0.044	0.036	0.003	0.01	0.02																		
Lp Aa D 2014	0.08	0.083	0.076	0.08	0.092	0.067	0.094	0.079	0.091	0.074	0.102	0.083	0.086	0.075	0.105	0.108	0.072																	
Lp Aa U 2014	0.091	0.098	0.084	0.087	0.099	0.075	0.1	0.093	0.102	0.092	0.112	0.094	0.09	0.082	0.119	0.126	0.082	0.006																

Lp Hem D 2014	0.083	0.081	0.094	0.099	0.082	0.072	0.073	0.076	0.064	0.087	0.074	0.062	0.105	0.091	0.129	0.138	0.105	0.112	0.111															
Lp Hem U 2014	0.076	0.074	0.081	0.081	0.072	0.064	0.072	0.074	0.06	0.076	0.067	0.06	0.078	0.081	0.119	0.129	0.096	0.097	0.089	0.008														
Lp Bet D 2014	0.022	0.025	0.028	0.035	0.018	0.041	0.031	0.028	0.023	0.034	0.015	0.027	0.027	0.033	0.066	0.095	0.058	0.11	0.125	0.093	0.089													
Lp Bre D 2011	0.081	0.092	0.111	0.077	0.068	0.106	0.087	0.089	0.089	0.089	0.077	0.087	0.055	0.088	0.1	0.135	0.132	0.185	0.174	0.187	0.152	0.08												
Lp Bre U 2011	0.087	0.107	0.117	0.088	0.076	0.113	0.099	0.099	0.112	0.097	0.082	0.091	0.056	0.085	0.095	0.138	0.13	0.191	0.18	0.197	0.172	0.084	0.005											
Lp Ris U 2011	0.027	0.031	0.042	0.029	0.03	0.029	0.028	0.025	0.035	0.039	0.021	0.025	0.039	0.041	0.076	0.106	0.042	0.115	0.114	0.111	0.098	0.052	0.077	0.089										
Lp Ris U 2014	0.033	0.036	0.051	0.026	0.035	0.041	0.028	0.036	0.031	0.048	0.028	0.033	0.048	0.042	0.075	0.108	0.047	0.129	0.122	0.111	0.094	0.056	0.08	0.097	-0.003									
Lp Odo U 2011	0.162	0.185	0.223	0.174	0.178	0.167	0.169	0.171	0.187	0.169	0.174	0.173	0.179	0.161	0.21	0.228	0.179	0.22	0.207	0.247	0.226	0.235	0.212	0.239	0.132	0.151								
Lp Oir U 2011	0.094	0.096	0.089	0.095	0.09	0.086	0.104	0.101	0.114	0.093	0.081	0.088	0.083	0.087	0.121	0.142	0.1	0.102	0.097	0.146	0.098	0.095	0.14	0.146	0.114	0.116	0.183							
Lp Oir D 2010	0.08	0.087	0.086	0.088	0.08	0.097	0.082	0.084	0.091	0.048	0.073	0.071	0.055	0.052	0.039	0.047	0.043	0.091	0.092	0.12	0.096	0.072	0.115	0.108	0.103	0.104	0.219	0.047						
Lp Oir D 2011	0.03	0.041	0.027	0.02	0.025	0.041	0.037	0.043	0.053	0.033	0.032	0.029	0.016	0.023	0.054	0.093	0.034	0.078	0.079	0.122	0.088	0.034	0.09	0.086	0.057	0.052	0.193	0.031	0.032					
Lp Oir D 2014	0.013	0.019	0.018	0.022	0.013	0.019	0.009	0.019	0.019	0.019	0.013	0.008	0.02	0.022	0.048	0.08	0.033	0.082	0.078	0.075	0.062	0.031	0.072	0.081	0.033	0.035	0.15	0.058	0.035	0.019				
Lp Oir U 2014	0.054	0.056	0.046	0.052	0.045	0.046	0.059	0.061	0.073	0.074	0.052	0.045	0.052	0.071	0.107	0.15	0.096	0.094	0.083	0.125	0.087	0.069	0.094	0.102	0.079	0.081	0.164	0.013	0.052	0.014	0.02			
Lp Cen U 2011	0.161	0.182	0.234	0.145	0.159	0.204	0.187	0.183	0.19	0.166	0.163	0.173	0.136	0.155	0.173	0.203	0.203	0.219	0.243	0.276	0.246	0.175	0.156	0.164	0.173	0.161	0.273	0.222	0.208	0.152	0.207	0.204		
Lp Jal U 2011	0.053	0.073	0.087	0.073	0.059	0.077	0.061	0.071	0.066	0.07	0.065	0.066	0.064	0.07	0.086	0.12	0.102	0.143	0.142	0.136	0.136	0.058	0.117	0.108	0.103	0.094	0.275	0.149	0.091	0.079	0.073	0.113	0.22	
Lp Sau U 2011	0.124	0.113	0.143	0.122	0.126	0.121	0.117	0.12	0.137	0.143	0.138	0.135	0.167	0.157	0.192	0.228	0.189	0.185	0.186	0.215	0.196	0.215	0.24	0.275	0.162	0.149	0.194	0.181	0.211	0.181	0.139	0.156	0.324	0.21

Table S5: Results of STRUCTURE analysis for each dataset: (a) full dataset, (b) *L. fluviatilis* only and (c) *L. planeri* only. Best *k* and Δ*k* values are in bold.

(a)

k	Replicate	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-25615.3133	0.0352	NA	NA	NA
2	20	-25385.4267	4.0411	229.886667	34.620000	8.566928
3	**20**	**-25120.9200**	**3.7126**	**264.506667**	**82.513333**	**22.225421**
4	20	-24938.9267	28.7455	181.993333	63.606667	2.212749
5	20	-24693.3267	9.1600	245.600000	18.573333	2.027647
6	**20**	**-24466.3000**	**5.3195**	**227.026667**	**186.376667**	**35.036462**
7	20	-24425.6500	11.3011	40.650000	216.333333	19.142722
8	**20**	**-24168.6667**	**8.0090**	**256.983333**	**301.263333**	**37.615679**
9	20	-24212.9467	54.6664	-44.280000	110.811282	2.027045
10	**20**	**-24146.4154**	**56.8238**	**66.531282**	**250.635897**	**4.410754**
11	20	-24330.5200	431.2098	-184.104615	72.157949	0.167338
12	20	-24442.4667	389.5193	-111.946667	459.993333	1.180926
13	20	-25014.4067	302.4341	-571.940000	428.646667	1.417323
14	20	-25157.7000	386.5390	-143.293333	NA	NA

(b)

k	Replicate	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-12861.2267	0.0458	NA	NA	NA
2	**20**	**-12732.6067**	**2.5647**	**128.620000**	**314.826667**	**122.752647**
3	20	-12918.8133	54.0977	-186.206667	476.553333	8.809126
4	20	-13581.5733	210.5316	-662.760000	298.886667	1.419676
5	20	-13945.4467	128.4933	-363.873333	95.186667	0.740791
6	20	-14404.5067	332.8450	-459.060000	317.206667	0.953016
7	20	-14546.3600	242.4818	-141.853333	35.733333	0.147365
8	20	-14723.9467	243.8772	-177.586667	42.626667	0.174787
9	20	-14944.1600	396.4927	-220.213333	124.566667	0.314171
10	20	-15039.8067	260.0662	-95.646667	NA	NA

(c)

k	Replicate	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-12478.7667	0.0488	NA	NA	NA
2	**20**	**-12041.1267**	**0.6486**	**437.640000**	**187.593333**	**289.233338**
3	20	-11791.0800	7.0161	250.046667	37.193333	5.301122
4	20	-11503.8400	4.4946	287.240000	68.473333	15.234686
5	20	-11285.0733	2.7886	218.766667	158.206667	56.733035
6	20	-11224.5133	31.5086	60.560000	12.853333	0.407931
7	20	-11151.1000	55.0251	73.413333	66.313333	1.205148
8	20	-11011.3733	61.4216	139.726667	7.501212	0.122127
9	**20**	**-10864.1455**	**2.5426**	**147.227879**	**309.122424**	**121.578115**
10	20	-11026.0400	61.6090	-161.894545	196.727879	3.193169
11	20	-10991.2067	40.4605	34.833333	223.246667	5.517642
12	20	-11179.6200	54.9826	-188.413333	NA	NA

Table S6: Estimates of ongoing migration rates (*m*) obtained with BayesAss (Wilson and Ranalla 2003) from river lampreys to brook lampreys (*m* in *Lp* from *Lf*) and from brook lampreys to river lampreys (*m* in *Lf* from *Lp*). Parapatric populations are indicated (P) and other estimates correspond to sympatric populations. River abbreviations match those given in table 1.

Rivers	*m* in *Lp* from *Lf* (95% CI)	*m* in *Lf* from *Lp* (95% CI)
Aa 2014	0.076 (0 – 0.17)	0.025 (0 – 0.070)
Aa 2014 (P)	0.016 (0 – 0.046)	0.020 (0 – 0.054)
Hem 2014	0.028 (0 – 0.07)	0.060 (0 – 0.13)
Hem 2014 (P)	0.023 (0 – 0.066)	0.076 (0.033 – 0.011)
Bre 2011	0.022 (0 – 0.060)	0.040 (0 – 0.111)
Bre 2011 (P)	0.015 (0 – 0.038)	0.070 (0 – 0.18)
Bet 2014	0.136 (0 – 0.37)	0.245 (0.07–0.42)
Ris 2011 (P)	0.281 (0.206 – 0.355)	0.014 (0 – 0.042)
Ris 2014 (P)	0.025 (0 – 0.088)	0.163 (0 – 0.306)
Odo 2011 (P)	0.011 (0 – 0.033)	0.051 (0 – 0.12)
Oir 2010	0.138 (0.064 – 0.211)	0.050 (0 – 0.121)
Oir 2011	0.290 (0.19 – 0.41)	0.075 (0 – 0.19)
Oir 2011 (P)	0.023 (0 – 0.061)	0.077 (0.038 – 0.116)
Oir 2014	0.290 (0.25 – 0.33)	0.030 (0 – 0.06)
Oir 2014 (P)	0.045 (0 – 0.11)	0.260 (0.18 – 0.41)
Cen – Loi 2011 (P)	0.013 (0 – 0.037)	0.051 (0 – 0.11)
Jal – Dor 2011 (P)	0.021 (0 – 0.589)	0.066 (0 – 0.171)
Sau – Gar 2011 (P)	0.011 (0 – 0.033)	0.059 (0 – 0.13)

Chapter 3:

Investigating divergence history of European river and brook lamprey

Article 2: Reconstructing the demographic history of divergence between European river and brook lampreys using Approximate Bayesian Computations

Quentin Rougemont, Camille Roux, Samuel Neuenschwander , Jérôme Goudet , Sophie Launey[2], Guillaume Evanno

In prep for Peer J

Reconstructing the demographic history of divergence between European river and brook lampreys using Approximate Bayesian Computations

Quentin Rougemont[1,2], Camille Roux[3], Samuel Neuenschwander[3] , Jérôme Goudet[3] , Sophie Launey[1,2], Guillaume Evanno[1,2]

[1]INRA, UMR 985 Ecologie et Santé des Ecosystèmes, 35042 Rennes, France

[2]Agrocampus Ouest, UMR ESE, 65 rue de Saint-Brieuc, 35042 Rennes, France

[3]Department of Ecology and Evolution, University of Lausanne, CH-1015, Switzerland

Abstract

Inferring the history of isolation and gene flow during species divergence is a central question in evolutionary biology. The European river lamprey (*Lampetra fluviatilis*) and brook lamprey (*L. planeri*) show low reproductive isolation despite highly distinct life histories, the former being parasitic-anadromous and the latter non-parasitic and freshwater resident. Here we analyzed six replicated population pairs and attempted to reconstruct their history of divergence using an Approximate Bayesian Computation framework combined with a Random Forest model on 13 microsatellite loci. Scenarios of divergence with recent isolation are outcompeted by scenarios proposing ongoing gene flow. The estimation of demographic parameters under the Secondary Contact model (SC) indicates a time of secondary contact close to the time of speciation, explaining why the support of SC over Isolation-and-Migration (IM) is poor. In case of an ancient secondary contact, the historical signal of divergence is simply lost and neutral markers converge to the same equilibrium as under the less parameterized model allowing ongoing gene flow. Our results imply that models of secondary contacts should be systematically compared to models of divergence with gene flow and given the difficulty to discriminate among these models we suggest that genome-wide data are needed to adequately reconstruct divergence history.

Introduction

Understanding the spatio-temporal conditions favouring species emergence is a fundamental question in evolutionary biology. One long standing controversy concerns the geographical setting promoting species divergence (Butlin *et al.* 2008; Fitzpatrick *et al.* 2008). Theory predicts that the accumulation of genetic incompatibilities is rather straightforward under allopatric conditions without gene-flow (Turelli *et al.* 2001; Coyne & Orr 2004; Barton & de Cara 2009). In contrast, speciation with gene flow theoretically requires (*i)* strong divergent selection and non-random mating, *(ii)* high genetic variance and (*iii)* non-random association of traits under disruptive selection and those involved in assortative mating (Dieckmann & Doebeli 1999; Gavrilets 2003, 2014; Coyne & Orr 2004). Importantly, the current geographical distribution of contemporary species may not reflect the initial conditions of divergence as most species may have undergone alternative phases of separation and contact over historical periods (Hewitt 1996, 2011; Bierne *et al.* 2011). As a result, reconstructing the history of demographic events that have shaped the genetic architecture of present-day populations is of primary importance to understand how speciation occurred and infer the role of gene flow during divergence. The accuracy of this reconstruction will depend on an adequate statistical method for demographic inferences, but also on the relevance of the sampling scheme.

Simulation-based methods are helpful for inferences although the tested models are always simplification of the real - and usually unknown - demographic history of the populations studied (Wakeley 2008). For instance, several studies using full likelihood approaches implemented in the IM and IMa programs (Hey & Nielsen 2004, 2007; Hey 2010) have compared Isolation-and-Migration (IM) models against a model of strict isolation (SI) and revealed a widespread effect of gene flow during divergence (Pinho *et al.* 2008; Niemiller *et al.* 2008; Strasburg & Rieseberg 2008). However, the method makes a number of simplifying assumptions (Strasburg & Rieseberg 2010, 2011) and does not allow for reconstruction of complex scenarios with several parameters, due to computation burden or intractable likelihood computation. Thus the complexity of demographic events may have been missed (Nielsen & Wakeley 2001; Hey 2010). For instance, most of these studies failed to distinguish between primary *versus* secondary differentiation (i.e. allopatric divergence followed by secondary contact), hence no general conclusion about the ubiquity of either mechanism during speciation can be drawn yet. Recent advances in coalescent theory (Wakeley 2008) and Bayesian methods (Tavare *et al.* 1997; Beaumont *et al.* 2002; Beaumont 2010) now allow for explicit tests of alternatives and complex models of divergence. In particular, Approximate Bayesian Computation (ABC) bypasses the need to compute full likelihoods, as this is not possible or is computationally too intensive for complex models with many parameters and large datasets (Beaumont *et al.* 2002). ABC

has been used with success to test alternative models of divergence in various taxa and has provided useful information on the level of interspecific introgression and complexity of demographic history underlying population divergence (Fagundes *et al.* 2007; Duvaux *et al.* 2011; Roux *et al.* 2013, 2014; Nadachowska-Brzyska *et al.* 2013; Nater *et al.* 2015).

Several studies have focussed on single population pairs to infer demography, so that drawing more general conclusions may be complicated. On the other hand, studies of replicated pairs of diverging natural populations have proven very useful to understand the genetic mechanisms of divergence and speciation and have shown that populations can independently evolve similar reproductive barriers in the face of ongoing gene-flow (e.g.Schluter & McPhail 1993; Nosil *et al.* 2002; Colosimo *et al.* 2005; Johannesson 2010). Their results were generally interpreted as evidence for parallel adaptation of diverging populations due to the action of recent natural selection. However, alternative scenarios of divergence including secondary contact after periods of allopatry have rarely been investigated (Bierne et al., 2013; Welch & Jiggins, 2014; Butlin et al., 2014).

Lampreys are jawless vertebrates (agnathans) thought to have diverged from the gnathostomes lineage (jawed vertebrates) approximately 590 million years ago (Hedges et al., 2015). At least half of the approximately 40 species of lampreys around the world occur as 'paired' species which often overlap in geographic distribution but which show strong divergence in adult life history strategies. One member of each pair is migratory (migrating to sea or downstream to lakes or large rivers) and becomes parasitic-hematophagous, while the other member (the so-called brook lampreys) are non-migratory (i.e., are entirely freshwater resident, remaining within their natal stream) and non-parasitic (i.e., not feeding at all after metamorphosis) (Docker, 2009). Despite a large number of evolutionary and developmental studies in lampreys (Heimberg *et al.* 2010; Shimeld & Donoghue 2012; Smith *et al.* 2013; Lagadec *et al.* 2015), there is a high uncertainty about the taxonomic relationships among lamprey paired species (but see Docker, 2009). For instance, the European river lamprey (*Lampetra fluviatilis*) and brook lamprey (*L. planeri*) display marked morphological differences at the adult stage, with adults of the anadromous and parasitic river lamprey on average 2.2 times longer than resident and non-parasitic brook lampreys. While adults of both species have been found on the same spawning ground (Lasne et al. 2010), this size difference likely forms the most important prezygotic barrier to gene-flow (Beamish & Neville, 1992, Rougemont et al. 2015). However, the genetic differentiation between these two taxa is usually low when measured either with allozymes (Schreiber & Engelhorn, 1998), mtDNA (Espanhol *et al.* 2007; Blank *et al.* 2008) or microsatellites markers (Bracken *et al,* 2015, Rougemont *et al.* 2015) and these species have also been hypothesized to be different ecotypes of a single species (Docker, 2009). Only one study based on restriction site-associated-DNA sequencing of a single population pair reported a

strong differentiation between *L. planeri* and *L. fluviatilis* (Mateus *et al.* 2013). Currently, there has been only one large scale phylogeographic study using mtDNA to investigate demographic history among *Lampetra* (Espanhol *et al.* 2007). The authors found a very low level of divergence, that was hypothesized to result from ongoing gene flow or very recent divergence following postglacial dispersion. However, it is known that widespread mtDNA introgression among sympatric taxa can easily obscure their taxonomic relationship (Shaw 2002). More recently Bracken *et al*. (2015) drew similar conclusions of recent divergence following postglacial dispersion. They concluded that founder event fuelled the evolution of diversity and may have promoted speciation. However, phylogeographic approaches do not allow contrasting alternative scenarios of divergence and do not address gene flow following divergence. As a consequence, relatively little is known so far about the history of divergence within lampreys, and most conclusions have been related to recent postglacial divergence (Espanhol *et al.* 2007; Bracken *et al.* 2015) or linked to ecological processes (Salewski 2003). Overall, few studies have used a wide number of pairs of river and brook lamprey connected by gene flow and realistic scenarios of demographic history have never been modelled.

Recently Rougemont *et al.* (2015) studied ten pairs of sympatric and parapatric populations of *L. fluviatilis* and *L. planeri* and found varying levels of genetic differentiation ranging from very low differentiation (F_{ST} = 0.008) to moderate levels of gene flow (F_{ST} = 0.189) depending on population pair. They concluded that these two ”species” may actually represent partially reproductively isolated ecotypes, a statement that was consistent with the low degree of reproductive isolation measured in experimental crosses (Hume et al. 2013; Rougemont et al. 2015). However, this pattern of low genetic differentiation observed among population pairs can be explained by two opposite hypotheses: (*i)* ongoing gene flow reduces differentiation even between ancient gene-pools; (*ii)* species have recently diverged in allopatry, which did not allow the accumulation of different alleles, including endogenic barriers. Here we used an ABC approach on genetic data obtained in multiple population pairs of *Lampetra* to test whether one of these two competing scenarios is shared across population pairs, or whether different pairs show contrasted scenarios of divergence.

Material and Methods

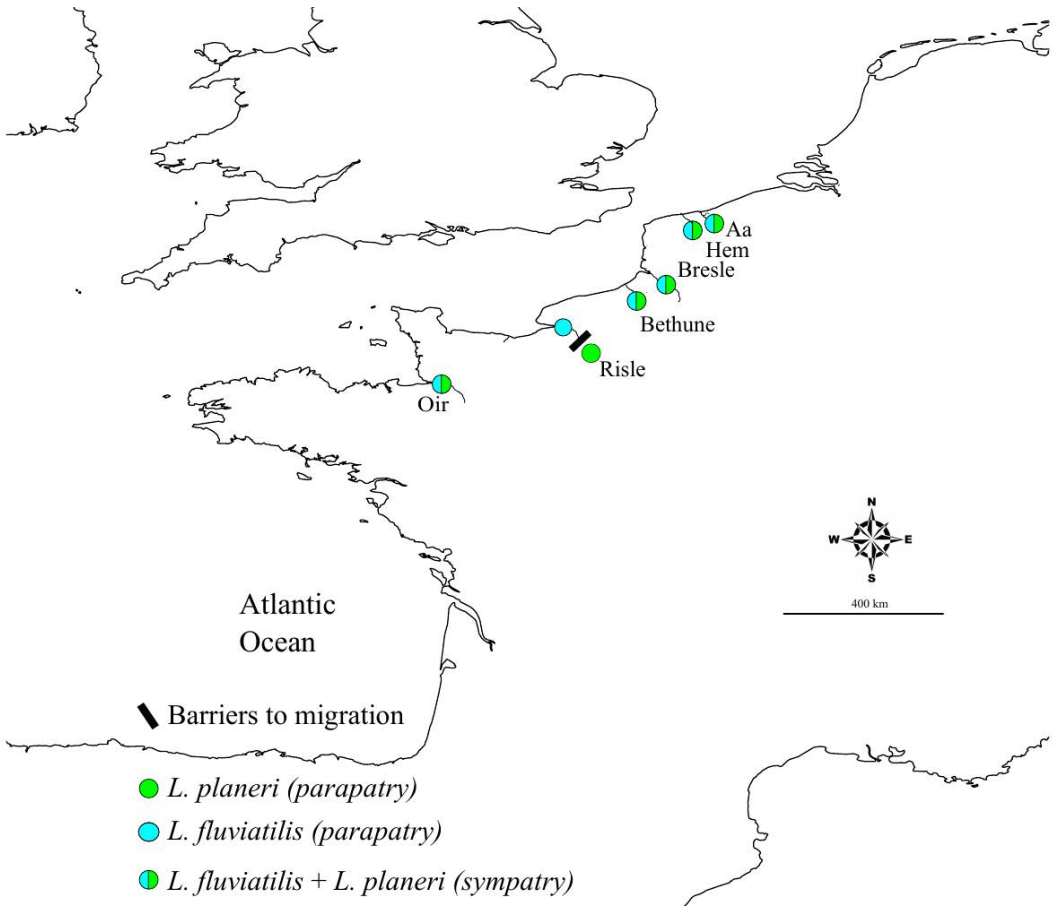


Figure 1: **Map of sampling sites across the channel area**.

Sampling and genotyping

L. fluviatilis and *L. planeri* samples were collected from 2010 to 2014 in 6 population pairs from northern France (data from Rougemont *et al.* 2015). Three pairs were collected in sympatry (Aa, Bethune and Oir Rivers). Two pairs are not strictly sympatric as on one river (Hem) there is a small obstacle between both samples, while in the second case (Bresle), populations were located 8 km apart, on the same stream section (Bresle). The last pair is a parapatric pair showing a moderate F_{ST} value similar to what is observed in sympatric populations (Risle) (see Rougemont *et al.* 2015). We chose the weakly differentiated pairs as they were less likely to deviate from demographic equilibrium than most parapatric pairs from the Rougemont *et al.* study. In parapatry, populations of *L. planeri* were generally highly geographically isolated in upper parts of the stream, subject to genetic drift, with no opportunity for gene flow with *L. fluviatilis*, hence we hypothesised that these populations were probably less appropriate to investigate the speciation process than the most connected pairs. The sampling included temporal replicates on the Oir (2010, 2011 and 2014), Bresle

(2011 and 2014), and Risle rivers (2011 and 2014). A set of 13 microsatellites was used to genotype a total of 727 individuals following the protocol described in (Gaigher *et al.* 2013).

Summary statistics

Given the lack of genetic differentiation between samples collected in different years in the same river, they were merged together. Similarly, we pooled brook lamprey individuals sampled in upstream and downstream areas on the Aa and Hem river as they displayed no significant genetic differentiation (Table S1). To obtain samples of river lampreys of similar size we also pooled individuals from the Aa and Hem river together as our previous genetic analysis indicated a single panmictic population of river lamprey across the English Channel area (Rougemont et al. 2015). Summary statistics were then computed for each pooled sample. For summary statistics used for comparison between simulated and observed datasets, we computed the average and standard deviation values of: the number of alleles (A), Allelic richness (*Ar*), observed and expected heterozygosity (*Ho* and *He*), allele size in base pairs, the Garza-Williamson index (GW,Garza & Williamson 2001), G_{ST} , $\delta\mu^2$ (Goldstein *et al.* 1995), Weir and Cockerham (1984) estimation of F_{ST}. All statistics were computed using R scripts (R Core Team, 2015) available upon request to the authors.

Testing alternatives demographic scenario

ABC coalescent simulations

For each population pair we used an Approximate Bayesian Computation (Beaumont *et al.* 2002; Csilléry *et al.* 2010)) framework to statistically compare five alternative models of demographic history (Fig. 2): (1) the two studied populations derive from a single panmictic gene pool (PAN); (2) the strict isolation model (SI) between sister populations; (3) isolation with migration model (IM); (4) a model allowing ancient migration but recent isolation (AM); (5) a model of secondary contact after past isolation (SC). The PAN model assumes a single panmictic population without size change. The SI model assumes a strict and instantaneous split of the ancestral population into two daughter populations with constant size and no subsequent gene-flow. The IM model assumes continuous gene-flow between daughter populations after the initial split at constant rate over generations. The AM model assumes gene-flow between the two diverging populations at the first generations following the split of the ancestral population.The SC model describes the split of an ancestral population in two isolated daughter populations; the two evolving lineages then experience gene flow through a secondary contact starting T_{SC} generations ago. For IM, AM and SC models, migration rates were scaled by $M = 4.N_0.m$, with M_1 the migration rate from *L. fluviatilis* to *L. planeri* and M_2 the

migration rate from *L. planeri* to *L. fluviatilis* and *m* is the fraction of the population made of migrants from the other population at each generation.

Coalescent simulations were performed using the ms software (Hudson, 2002) assuming an infinite-sites model of mutation, in which most parameters are scaled by the effective population size of an arbitrarily chosen reference population (N_{ref}) with impact on conclusions drawn by the ABC analysis. Each model was also characterized by a scale effective population size ϑ: $\vartheta_A / \vartheta_{Ref}$, $\vartheta_{lf} / \vartheta_{Ref}$, $\vartheta_{lp} / \vartheta_{Ref}$ where $\vartheta_{Ref} = 4N_{Ref}\mu$, μ represents the mutation rate per generation. Patterns of genetic diversity suggested that river lamprey display a greater *Ne* than populations of brook lamprey (Rougemont *et al.* 2015). Thus, ϑ_{lf} was sampled on the interval 0-3 and ϑ_{lp} in the interval 0-max(ϑ_{lf}). ϑ_{Ref} was set to 1 (i.e. we assumed N_{Ref} = 1,000 and μ=2.5e^{-4}). The panmictic model was only characterized by the unique effective mutation rate ϑ which was also modelled on the interval 0-3. All models (except PAN) also incorporated the scaled time of divergence, $\tau_{split}/4N_{Ref}$, where τ_{split} is the time measured in number of generations and drawn from a uniform distribution in the interval 0-25. The two parameters τ_{iso} (AM model) and τ_{sc} (SC model) were computed from uniform distributions defined on the interval 0-τ_{split}. Since the genetics and ecology of lampreys is poorly known, we chose large uninformative prior distributions for all parameters to include commonly used parameters from the literature (Pinho & Hey, 2010) after exploring different combinations of priors following Cornuet *et al.* (2010) (Table 1). Binary simulated data from ms were converted into microsatellite data using a stepwise mutation model (SMM). Probability of changes of the repeat number in each mutation event was modelled by a geometrical parameter α distributed following a uniform prior distribution sampled on the interval 0 – 0.5. All computations were run in R and took into account differences in sample size for each of the thirteen loci. Summary statistics were computed from the transformed microsatellite data. One million simulations composed of the thirteen microsatellite loci were computed under each demographic model. All R code used for ABC computation is available from the authors.

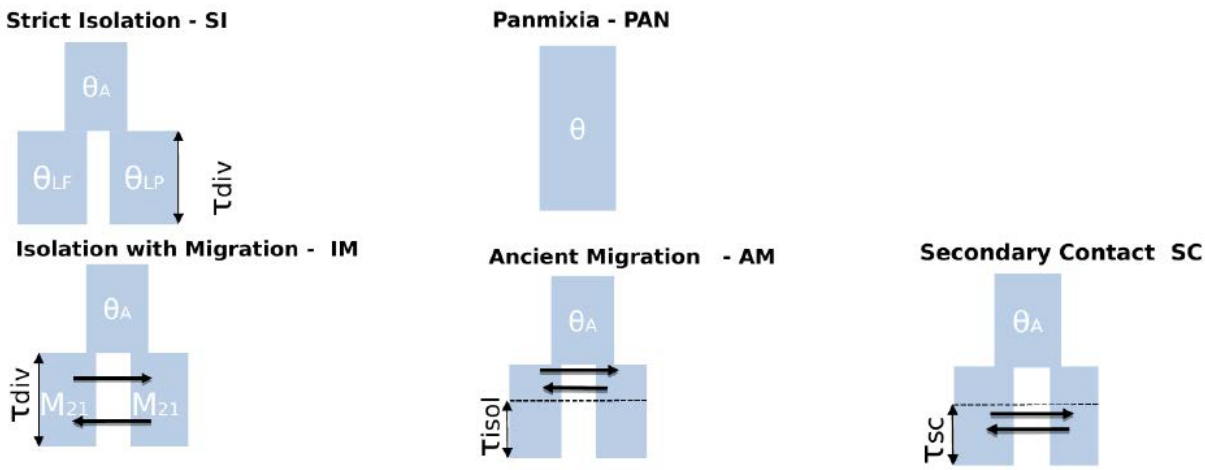


Figure 2: **Different scenarios of divergence between *L. planeri* and *L. fluviatilis.*** (A) five models with different parameters are tested and compared. Two null models: Stict Isolation (SI) and Panmixia (PAN). Three models of migration: isolation with constant migration (IM), ancient migration (AM) and secondary contact (SC). The following parameters are shared by all models: τ_{div}: number of generation since divergence time. θ_A, θ_{Lf} , θ_{Lp} : effective population size of the ancestral population, of *L. fluviatilis* and *L. planeri* respectively. τ_{isol} is the number of generations since the two ecotypes have stopped exchanging genes. τ_{sc} is the number of generations since the two ecotypes have entered into a secondary contact after a period of isolation. M_{12} and M_{21} represent the migration rates expressed in 4.*Nm* units per generation with *m* the proportion of population made of migrants from the other populations

Table 1: Prior for all models. θ_A, θ1, θ2 = effective mutation rate for t he ancestral, river lamprey and brook lamprey populations respectively. M_1, M_2, M_{Anc} = *Effective migration rate* for the ancestral, river lamprey and brook lamprey populations respectively. τ = divergence time, τ_{isol}, τ_{sc} divergence time under the ancient migration model and time of secondary contact respectively. SI: strict isolation, IM: isolation with migration, AM: ancient migration, PAN: Panmixia SC: secondary contact model.

Parameters	***Models***	***Prior***
$\theta_A = 4N_{Anc}\mu$	SI, IM, AM, SC	Uniform [0-3]
$\theta_1 = 4N_1\mu$	SI, IM, AM, SC, PAN	Uniform [0-3]
$\theta_2 = 4N_2\mu$	SI, IM, AM, SC	Uniform [0- (θ_1)]
$M_1 = M_2 = 4N_1\, m$	IM, SC	Uniform [0-20]
$M_{ANC} = 4N_1\, m$	AM	Uniform [0-20]
$\tau = 4N_1 t$	SI, IM, AM, SC	Uniform [0-25]
$\tau_{isol} = 4N_1 t$ $\tau_{sc} = 4N_1 t$	AM SC	Uniform [0- τ] Uniform [0- τ]

Model Selection

ABC approach

We evaluated the posterior probabilities of each demographic model using an ABC framework implemented in the abc package in R (Csilléry *et al.* 2012). We compared all models simultaneously by computing posterior probabilities using a feed forward neural network based on a nonlinear conditional heteroscedastic regression in which the model is considered as an additional parameter to be inferred. This procedure allows taking into account correlations of summary statistics and distortion hence reducing the problem of curse of dimensionality (Blum & Francois 2010). In the rejection step, we retained the 0.02% simulations closest to the observed summary statistics, which were subsequently weighted by an Epanechnikov kernel that peaks when $S_{obs} = S_{sim}$. The regression step was performed using 50 neural networks and 15 hidden layers.

ABC cross-validation

We performed model checking to compute the robustness of the inferred model using pseudo-observed simulated datasets (PODS). We randomly selected 1,000 PODS from one million simulations computed under each simulated model. We used the same ABC selection procedure as above to compute the probability that the best model was indeed the best model given the posterior probability computed from the observed dataset: we kept the 0.02% simulated closest simulations, weighted them with an Epanechnikov kernel in the rejection step and performed the regression with 50 neural networks and 15 hidden layers. We then computed the robustness of each scenario: we computed the type I error rate that corresponds to the risk of excluding the previously inferred scenario when it is the true scenario and the type II error rate that corresponds to the risk of selecting the previously inferred scenario when it is false.

Random Forest model selection and cross-validations

In parallel to our ABC based model selection and cross-validation procedure we explored the ability of a Random-Forest algorithm (Breiman 2001) to discriminate the different models and to estimate which summary statistics were the most informative. Random Forest (RF) is a machine-learning algorithm whose use has recently been advocated for model choice in ABC inference to circumvent curse of dimensionality problems and those linked to the choice of summary statistics (Pudlo *et al.* 2014). This approach is a non-parametric classification algorithm that uses bootstrapped decision trees to perform classification using a set (p) of defined predictor variables (here the summary statistics). Multiple (i.e. hundred to thousand) decision trees are grown and merged together and the ensemble makes the forest (Breiman 2001). Simulations that are not used in tree

building at each bootstrap (the so called out-of-bag simulations OOB) are then used to compute the OOB error rate, which provides a direct method for cross-validation (Breiman 2001, Cutler *et al.* 2007). This method allows reducing the dimensionality of the data (Cutler *et al.* 2007) but also estimating the relative importance of variables (here the summary statistics) through rankings. Variable importance is measured by random permutations of the specified variable in OOB observations and new predictions are then obtained and compared to the original OOB data (Cutler *et al.* 2007). One particularly attracting feature of random-forest is its insensitivity to strong correlations and high noise within data (Pudlo et al. 2014).

We first constructed 6 random forests (one by river) using the randomForestSRC package in R allowing for parallelization and fast computations (Iswharan & Kogalur 2007, 2015; Ishwaran *et al.* 2008). We grew 1,000 trees on subsets of 50,000 simulated dataset (5%) that were used as a training set. Prior analysis using different numbers of trees and training set sizes indicated that the OOB errors reach stationarity using between 500-1,000 trees (see also Fig 4), so we did not grow a bigger forest that would have required extensive computations. All summary statistics were included to get an estimation of the importance of each variable. This allowed us to estimate the OOB error rate for each comparison, which is similar to a prior error rate in ABC inference (Pudlo et al. 2014). Ultimately our forest was used as a prediction tool to compute the probability that our observed data belongs to one of the 5 alternatives models.

Parameter Estimation and cross-validation

Parameter estimation was performed for the best models using nonlinear regressions. We first used a logit transformation of the parameters on the 2,000 best replicate simulations providing the smallest Euclidian distance δ (Csilléry *et al.* 2012). We then jointly estimated parameters' posterior probability using the neural network procedure implemented in the abc package. We obtained the best model by weighted nonlinear regressions of the parameters on the summary statistics using 50 feed-forward neural networks and 15 hidden layers. We performed posterior predictive checks for cross-validation in an attempt to check the ability of our parameter estimates to generate data summary statistics close to the observed summary statistics. For each best model, we selected 10,000 posterior samples obtained after parameter estimation (from the abc package) and simulated 10,000 new datasets by using again ms and custom R scripts. We then again plotted the distance between our observed original values and our new simulations and computed the p-value for each statistic.

Results

Population diversity and divergence

A total of 6 populations pairs (727 individuals) were analysed using 13 microsatellite markers. As already observed (Rougemont et al. 2015), the genetic diversity of river lamprey, as measured by the averaged allelic richness was significantly greater than that of brook lamprey (Ar_{Lf} =3.43, Ar_{Lp} = 3.116, 15,000 permutations, *P = 0.0010*, Table 2). On the contrary there was no significant difference in expected heterozygosity between river lamprey (He_{Lf} = 0.507) and brook lamprey (He_{Lp} =0.46) (15,000 permutations, *P* =0.208). Global population genetic differentiation between river and brook lamprey was F_{ST} = 0.061 (99%IC = 0.044-0.079) and ranged from 0 to 0.192. Genetic differentiation among river lamprey populations was significantly lower (F_{ST} = 0.002) than among brook lamprey populations (F_{ST} = 0.109) (15,000 permutations, *P*= 0.003). No river lamprey populations differed significantly from the others whereas all brook lamprey populations were significantly differentiated from one another (Table S1). Pairwise comparison within rivers revealed significant differentiation between river and brook lamprey populations in all cases except for the Bethune River. This differentiation varied between rivers and ranged from 0.028 (Oir and Bethune river) to 0.091 on the Bresle River.

Table 2: Estimates of populations genetic parameters for each pairs of river and brook lamprey populations. N = number of individuals used for ABC analysis Ar= Allelic richness, He= expected heterozygosity, GW= Garza-Williamson Index. *For the ABC inference, individuals of river lamprey from the AA and Hem (F_{ST}=0) river were pooled together to obtain a sample size similar to the one of brook lampreys. $^{\delta}$ Brook lamprey samples from the AA and Hem rivers are composed of upstream and downstream samples from Rougemont et al. (2015) study.

Pop	N *Lf*	N *Lp*	F_{ST}	Ar *Lf*	Ar *Lp*	He *Lf*	He *Lp*	GW *Lf*	GW *Lp*	Delta μ^2
OIR	104	74	0.028	4.45	3.61	0.52	0.508	0.525	0.622	0.204
BET	14	14	0.028	3.51	3.36	0.516	0.471	0.452	0.464	0.507
RIS	75	75	0.033	3.84	3.92	0.503	0.472	0.497	0.421	0.842
HEM	30*	65$^{\delta}$	0.077	4.21	3.53	0.504	0.477	0.406	0.487	1.633
AA	34*	69$^{\delta}$	0.084	4.21	3.76	0.514	0.522	0.406	0.505	0.915
BRE	93	80	0.091	4.14	4.91	0.49	0.49	0.466	0.263	34.37

Model comparisons

The classical ABC model-choice and random forest approaches generally yielded similar results as detailed in Table 3. In all population pairs, the model of strict isolation (SI) and of ancient migration followed by a period of strict isolation (AM) were clearly rejected. In two population pairs (Aa and Bresle), the best supported model by both the ABC and RF approaches was the SC model. In the Bethune, the best supported model by both methods was the IM model. In two cases (Hem and Risle), none of the methods was able to accurately discriminate between the two scenarios (SC and IM model). Finally, in the Oir river the two methods gave incongruent results with the model of panmixia (PAN) being the best supported model under the ABC framework, while the RF failed to distinguish between the IM and SC models.

Table 3: ABC classification (posterior probability), Random-Forest (RF) prediction and robustness (ABC only) of each model of speciation in each river.

	MODEL									
	SI		IM		AM		SC		PAN	
RIVER	ABC	RF	ABC	RF	ABC	RF	ABC	RF	ABC	RF
AA	0	0	0.3	0.39	0.01	0.06	**0.69**	**0.54**	0	0
BET	0	0	**0.45**	**0.57**	0	0.02	**0.46**	0.35	0.08	0.06
BRE	0.01	0.02	0.12	0.24	0.14	0.12	**0.73**	**0.62**	0	0
HEM	0	0	0.42	**0.53**	0.01	0.05	**0.57**	0.42	0	0
RIS	0	0	0.46	**0.47**	0	0	**0.54**	**0.52**	0	0
OIR	0	0	0.15	**0.46**	0	0.02	0.14	**0.47**	**0.71**	0.05
Average	0.00	0.00	0.32	0.44	0.03	0.05	0.52	0.49	0.13	0.02
Robustness	0.82	-	0.25	-	0.6	-	0.46	-	0.1	-

Robustness and misclassifications errors from ABC analysis and random-forest

We checked whether our model comparison analysis was reliable and fitted well to the data by using pseudo observed datasets (PODS) and running the same ABC procedure as for our observed data. We found that robustness of our analysis was high for both the SI and PAN model in all population pairs (Table 3). However, the accuracy of the IM and SC models was always low and highly unreliable based on the classical ABC model choice procedure (Table 3). Since our simulations were mainly the same (the sole difference was the number of individual genotyped at each locus in each dataset) and model checks were similar between population pairs, we therefore present only robustness results for one population pair (The Aa River, Table 3, Robustness). We secondly tested if a random-forest

approach could help confirm the robustness of the rejected models and distinguish between the IM and SC models. The RF results confirmed that the models of strict isolation, ancient migration and panmixia were classified with a high accuracy (Table 4, Figure 3). The overall error rate (28.79%) hides very different accuracies depending on the models. The OOB error (averaged over the population pairs) was still high between the AM and SI models (24.2%, see also table 4). However both the ABC and RF analyses (Table 3) clearly showed that these two models were never supported by our data so the classification error between these two models was a minor concern. On the contrary average OOB errors in pairwise analyses of IM *versus* SC model were as high as 45% demonstrating that it was generally not possible to correctly classify our simulated data in their correct categories (see details in table 4). The estimation of variable importance (example in Figure 3) indicated that the most informative variables were systematically the mean and variance of G_{ST} and of Delta mu^2, generally followed by the estimators of allelic richness and expected heterozygosity in each population and globally (Figure 4 and Table S2).

Table 4: Random Forests out-of-bag confusion matrix and classification error. Data based on 6 random forests, each composed of 1,000 trees based on a trained set of 50,000 simulated predictor variables (summary statistics). The response variable is the demographic model. Proportions of correctly classified demographic models are in bold. The two grey italic values represent models with high error rates. Simulation between rivers differed only by the number of individual loci simulated and produced very similar values that were subsequently averaged over each demographic model.

	Predicted Model (Averaged over each river)					Averaged OOB error rate
Observed	AM	I	IM	PAN	SC	
AM	**78.0%**	16.7%	2.4%	0.0%	2.9%	21.99%
I	25.2%	**73.6%**	0.6%	0.0%	0.6%	26.38%
IM	1.6%	0.2%	**57.2%**	0.8%	*40.1%*	42.76%
PAN	0.0%	0.0%	0.2%	**99.7%**	0.1%	0.30%
SC	2.1%	0.3%	*43.6%*	0.6%	**53.3%**	47.12%

Parameters estimation from the best models

We estimated the parameters in each population pair for both the IM and SC models that we failed to accurately distinguish and for the panmictic model in the Oir population pair. Results of parameter estimation are presented in Table 4. The accuracy of posterior parameter estimation varied greatly among populations pairs, with the Aa, Hem and Risle presenting accurate parameter estimation under both the IM and SC model. On the contrary posterior parameter estimates on the Bethune population pairs were almost flat and not different from the prior. In general thetas

estimated from the IM model were more accurate than under the SC model. Under both models we generally observed a reduction of thetas in both river and brook lampreys as compared to their ancestral population. Under the IM model the respective median effective mutation rate (θ_{Lf} and θ_{Lp}) of the river lamprey and brook lamprey were on average 1.67 and 4.71 times smaller than the ancestral population.

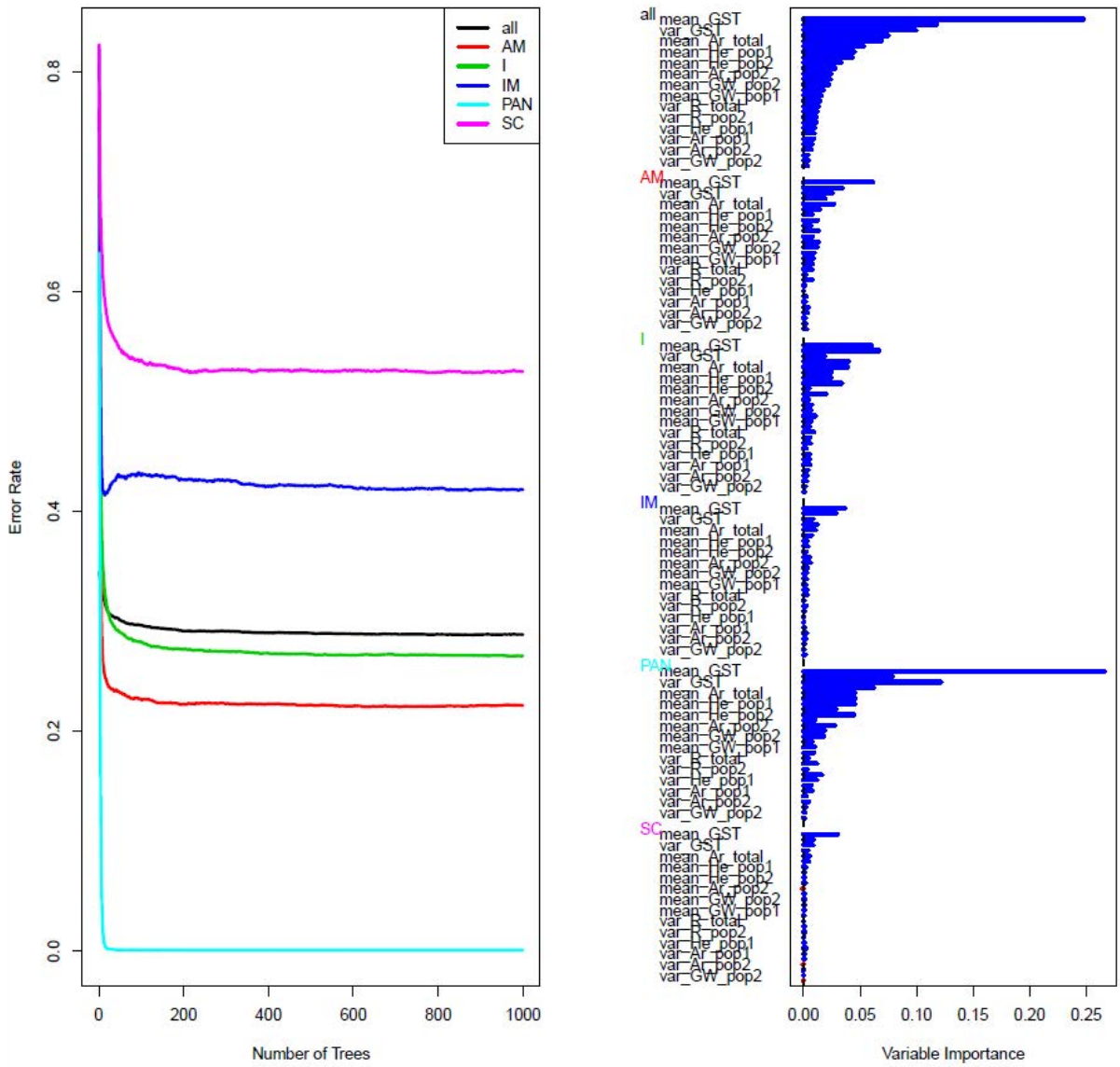


Figure 3: **Curves of out-of-bag errors rates and Estimation of variable importance for the Aa river.** Data based on one random forest, each composed of 1,000 trees obtained from a trained set of 50,000 simulated predictor variables (summary statistics). The response variable is the demographic model. Estimation for the 6 remaining rivers yielded similar results and are presented in table S2 and Figure S1

Under the SC model, the (averaged) median effective mutation rate was 1.11 and 2.26 times smaller in river lamprey and brook lamprey respectively than their ancestral populations (Table 4). Similarly, θ_{Lf} was always larger than θ_{Lp}, reflecting the probable lower effective population size of the latter. Estimates of thetas were on average 2.82 times larger in river lampreys than in brook lampreys under the IM model. Under the SC model estimates of thetas were 2.03 times larger in river lampreys than in brook lampreys. Under both the IM and SC models we also consistently noted an asymmetric migration, with a tendency towards higher migration from river lamprey to brook lamprey: averaged median *4*Ne*m* =16.45 and 16.86 under IM and SC respectively *versus* averaged *4*Ne*m =3.48* and 6.64 under IM and SC respectively (Table 5). However, in one case out of 5 the reverse tendency was observed under the IM model (Bethune River), and in two cases under the SC model (Risle and Hem). Estimates of divergence time and timing of secondary contact (SC only) yielded variable results and were not always accurately estimated (Table 5). Estimates from the Aa and Hem population pairs were the most accurate under both scenarios. Overall estimates of separation times were highly congruent under the SC model but revealed that populations would have come into secondary contact for a long period with the averaged time of secondary contact representing one third of the time since divergence. Finally, simulations of the Oir population pair were summarized by a single parameter (the effective mutation rate of a panmictic population) that was estimated with high accuracy (median = 0.57 95HPD:0.56-0.59]).

Posterior predictive checks

We performed posterior predictive checks in order to assess the ability of the models to accurately reproduce summary statistics close to our observed statistics based on 10,000 simulated datasets and computed the robustness of our inference. Under both the IM and SC models we consistently found some statistics that differed significantly from our observed data. These were the variance of expected heterozygosity ($p<0.05$) and the variance of allelic number ($p<0.05$). In some cases the mean Garza-Williamson index was not accurately reproduced (Table S4). Similarly, in three cases, the variance in F_{ST} did not yield accurate results (p=0) (Table S4). Under the panmictic model (Oir population pairs) we found that the variance in Allelic Richness and the Garza Williamson index were not accurately reproduced by our data (p=0, Table S4).

Table 5: Estimates of demographic parameters under the model of ongoing migration (IM) and secondary contact (SC) in each river.

Ne = effective population size. Lf = *lampetra fluviatilis*, Lp = *Lampetra planeri*.

		Ne Lf		**Ne Lp**		**Ne Anc**		**migration Lf to Lp (=4N12 m12)**		**migration Lp to Lf = 4 N1 m21**		**Split time**		**Time of Secondary Contact**	
RIVER	model	median	95HPD	median	95HPD	median	95HPD	median	95HPD	median	95HPD	median	95HPD	median	95HPD
AA	SC	1550	[900-2580]	480	[220-890]	1720	[1570-1960]	17.2	[8.67-30.44]	1.6	[0.90-6.17]	222200	[200000-261200]	124600	[92000-150400]
	IM	1310	[980-1880]	230	[290-510]	2380	[2260-2470]	19.6	[14.41-28.69]	2.5	[2.50-6.9]	272000	[247000-298600]		
BET	SC	1930	[1010-2800]	870	[500-1380]	1040	[220-2190]	28.9	[7.86-50.57]	4.9	[0.58-25.35]	316400	[132800-448400]	140800	[64000-261000]
	IM	1650	[840-2780]	950	[410-1870]	1880	[990-2620]	18.2	[4.86-44.92]	12.8	[2.89-17.35]	185600	[73800-320600]		
BRE	SC	1350	[800-2270]	800	[300-1680]	2100	[930-2760]	18.2	[5.08-41.0189]	6.3	[0.84-27.70]	280600	[140000-442400]	19000	[5400-71300]
	IM	1420	[760-2580]	340	[150-830]	1440	[750-2260]	13.1	[3.05-41.85]	0.6	[0.11-11.37]	274000	[161800-417200]		
HEM	SC	840	[770-1840]	360	[340-1080]	1690	[1680-2640]	7.1	[6.37-27.89]	11.1	[4.60-19.73]	244200	[240400-430000]	99600	[101200-163600]
	IM	1190	[930-1770]	190	[130-330]	2540	[2460-2600]	21.6	[16.42-32.57]	1.4	[0.76-5.29]	320800	[304400-332200]		
RIS	SC	1230	[860-2100]	710	[370-1150]	1090	[820-1440]	16.2	[10.53-30.41]	9.7	[6-20.07]	161400	[115600-253800]	60800	[24200-111600]
	IM	700	[440-1590]	500	[420-670]	1000	[890-1160]	8.6	[4.30-23.94]	4.6	[3.32-4.12]	123000	[107200-135200]		
average	SC	1380	[868-2318]	644	[346-1236]	1528	[1044-2198]	16.9	[7.69-35.96]	6.6	[2.61-19.16]	244960	[165760-367160]	88960	[57360-151580]
	IM	1254	[790-2120]	442	[280-842]	1848	[1470-2222]	16.4	[8.21-34.38]	5.2	[2.45-5.13]	235080	[178840-300760]		
OIR	PAN	Ne :	2094 [2091-2095]												

Discussion

Our goal was to test whether we could discriminate alternative scenarios of divergence between river and brook lampreys using a set of microsatellite markers and an ABC approach. We were able to reject the model of strict isolation and of ancient migration. In one case, the model of panmixia received the best support, whereas in other population pairs it was not possible to discriminate divergence with ongoing gene flow from a model of allopatric divergence followed by secondary contact.

Difficulty in distinguishing between ongoing migration and secondary contact

In spite of the availability of large amounts of genetic data and computer resources, few studies have explicitly tested alternative models of speciation (e.g. Ross-Ibarra *et al.* 2009; Duvaux et *al.* 2011; Roux *et al.* 2013, 2014; Butlin *et al.* 2014). While populations may diverge (and eventually become reproductively isolated) under various demographic scenarios, our results indicate that distinguishing between primary differentiation (divergence with gene-flow) *versus* allopatric divergence followed by secondary contact remains difficult when using genetic data from a limited number of neutral markers, even with advanced computational tools. Indeed, ABC as well as RF cross-validation clearly showed that the two models were wrongly classified almost half the time. The SC model tended to display a greater proportion of simulations wrongly classified into the IM model, a result that can be explained by the greater complexity of this model that displays one supplementary parameter and is inherently more difficult to infer. On the contrary, even though the OOB error rate was still high in the IM model, it tends to display fewer simulations wrongly classified in the SC model. Given the inherent difficulty of properly classifying the SC model even when it is true, our support for this model in some cases may suggest that it could be the true model under which lampreys have diverged.

Inability to distinguish between a scenario of isolation with migration and secondary contact is in accordance with theoretical expectations from Bierne *et al.* (2013). Using a simple modelling approach, these authours showed how genetic environmental associations at neutral markers such as microsatellites can quickly be lost after secondary contacts and then reach migration/drift equilibrium together with a pattern of isolation by distance, which corresponds to the populations in our study (see Rougemont *et al.* 2015). In particular, they applied their model to the well-studied freshwater/marine stickleback (e.g. Colosimo *et al.* 2005; Hohenlohe *et al.* 2010, 2012) which shares several characteristics with our lamprey system, such as the existence of a single nearly panmictic marine population and small and almost independent freshwater populations. The application to the stickleback model showed that introgression proceeded independently between the different streams and was strongly asymmetric from the migratory to the resident populations, which is exactly the pattern we observed here (Table 4) under both the isolation with migration model (migration 4.2 times greater from river lamprey to brook lamprey) and the secondary contact model (migration 2.6 times greater).

The failure to reject panmixia in the Oir River can also be explained in the light of the conclusions of Bierne *et al.* (2013). It could be attributable to the low genetic divergence observed (F_{ST} = 0.028) especially

given the small number of markers we used, but this pattern of near panmixia can also be attributed to a stronger introgression in this system than in all other investigated streams. In this particular case the mean size of river lampreys (225 mm, *n=134*) was much smaller than the size observed in other rivers (mean =303 mm, *n* = 389). Assuming that size difference is the most important cause of reproductive isolation (Beamish & Neville, 1992) a smaller size difference may facilitate mating of the two ecotypes and subsequent genome swamping. In both cases, inferences from this system based on neutral markers are necessarily difficult as this pattern may be explained by strong gene flow after an isolation period as well as by a very early stage of ongoing divergence.

Demographic parameter estimations and new insights on lamprey history

We assumed a mutation rate of $2.5e^{-4}$, that is somewhat similar to what is observed in several fishes species (Shimoda *et al.* 1999; Steinberg *et al.* 2002; Yue *et al.* 2006) and other vertebrates (e.g. Nance *et al.* 2011). However, this mutation rate remains a rough estimate that was necessary for scaling, together with N*ref* that was set to 1,000 based on prior knowledge of possible population size in lampreys. Every parameter estimate discussed below should thus be considered cautiously. Estimates of current effective population size confirmed the strong dissymmetry in effective population size between river lamprey (averaged mode = 1342, 95% CI = 868-2318 under SC, 1150 and 95% CI=790-2120 under IM) and brook lamprey (averaged mode = 662, 95% CI=346-1236, 408 and 95% CI= 280-842 under SC and IM respectively). Importantly, confidence intervals were rather large and overlapping between the two ecotypes. It is possible that we underestimated population size of brook lampreys in areas where they are strongly connected with river lampreys. Our analysis also suggested a population size reduction in brook lamprey as compared to its ancestral population (averaged mode = 1498, 95% CI = 1044-2198 under SC, 1922 and 95% CI=1470-2221 under IM). Similarly, since these estimates overlap with those from river lamprey, the evidence for a size reduction remains thin given that the two sister populations appeared connected.

Our estimates of timing of divergence provided similar estimates under both IM and SC. They suggested that the two ecotypes may have separated around 228,000 years ago (95% CI:178,000-307,000 years ago) under the IM model and around 259,000 years ago (95% CI:165,000-367,160 years ago) under the secondary contact model (assuming a generation time of 5 years, Hardisty & Potter, 1971). Such estimates are rather similar to what was observed in *Dicentrarchus labrax* (Tine *et al*. 2014) but differ drastically from mtDNA estimates of divergence time available so far in lamprey (Espanhol *et al.* 2007; Bracken *et al.* 2015). This discrepancy might be explained by the different type of molecular data used in each study, but also by the rough estimates of molecular evolution and microsatellite mutation rate. Importantly, under the SC model, the secondary contact would have started around 85,000 years ago, representing one third of the total divergence time in lamprey. This result is rather older than the hypothesis of postglacial colonization of river by resident ecotypes as generally assumed in European fishes

(Bernatchez & Wilson 1998; Aldenhoven *et al.* 2010). Such an ancient secondary contact implies that the genetic signature of historical geographic isolation carried by neutral markers may have been lost. In these conditions, neutral markers can converge to the same state than the one observed under primary differentiation (Barton & Hewitt 1985; Charlesworth *et al.* 1997; Bierne *et al.* 2013). The SC model implies the accumulation of some Dobzhansky-Muller DMI in allopatry when the two ecotypes started to diverge. While both theory (Orr 1995) and empirical evidence (Matute et *al.* 2010) predict that DMI should accumulate faster than linearly in time, our results suggest a limited amount of isolation. In this case the amount of time was certainly not enough to allow for sufficient DMI to occur and to develop strong barriers to gene flow. This would likely explain the low differentiation observed for mtDNA (Espanhol *et al.* 2007 Blank *et al.* 2008; Bracken *et al.* 2015) and is fully compatible with the observation of viable F1 (Hume *et al.* 2013; Rougemont *et al.* 2015).

Our results also suggest that the RF model provides a valuable complement to the standard ABC model comparison (Robert *et al.* 2011; Pudlo *et al.* 2014; Marin *et al.* 2014). The two methods provided similar outcomes in terms of model choice and subsequent cross-validation except in one case (Oir River). The ability to distinguish alternatives between SC and IM was low in both cases. In line with Pudlo *et al.* (2014) we find that the RF approach possesses a series of advantages over the ABC approach such as 1) fast model choice procedure with simultaneous cross validation through OOB computations, 2) estimation of variable importance (i.e. of the summary statistics), 3) considerable reduction of computational time. Estimating variable importance can be particularly interesting when a large set of variables are used without prior knowledge about the pertinence of the summary statistics used. Choice of summary statistics is an important process in ABC methodology (Csilléry *et al.* 2010) for which relatively few tools are available. RF may provide such an objective tool that may be complementary to conventional ABC model choice and cross validation procedures. Note, however, that the neural network method provided in the abc packages performed very well and provided similar results to the RF model.To the extent of our knowledge, we provided the first study which empirically combines ABC and RF for model choice and cross-validation.

Conclusion and perspectives

Our study sheds new light on the demographic process that has shaped the current genetic makeup of population pairs of European river and brook lampreys. In particular, we were able to reject a scenario of divergence in strict isolation and a scenario of ancient sympatric divergence. The scenario of panmixia was also supported only once and it is thus unlikely to be a generalizable scenario across the species range. In addition it illustrates the necessity of explicitly exploring alternative models of divergence before concluding on the prevalence of rapid parallel speciation (Bierne *et al.* 2013). This study also illustrates how combining new modelling approaches can help improve our understanding of the complex process of speciation. However, it was not possible to firmly discriminate the SC or IM models but it is likely that

distinguishing between these alternative scenarios is complicated in cases of ancient secondary contacts, especially when investigations are performed with a limited number of neutral markers. Finally combining modelling approach (e.g. SFS, Tine *et al.* 2014) with a higher number of markers and allowing for heterogeneous migration rate among loci (e.g. Roux *et al.* 2013, 2014), variation of migration rate in time and variation of effective population size along the genome, may allow fine-tuning of our demographic investigations and provide great insight into the prevalence of secondary contact versus speciation with continuous gene-flow in nature.

Table S1: Pairwise F_{ST} values. Non-significant values are in italics and grey-colored.

	LF Aa	LF Hem	LF Beth	LF Bre	LF Ris	LF Oir	LP Aa	LP Hem	LP Bet	LP Bre	LP Ris
LF Hem	*0.000*										
LF Beth	*0.002*	*0.001*									
LF Bre	*0.000*	*0.000*	*0.003*								
LF Ris	*0.000*	*0.000*	*0.001*	*0.003*							
LF Oir	*0.000*	*0.003*	*0.011*	*0.006*	*0.003*						
LP Aa	0.084	0.090	0.079	0.090	0.088	0.090					
LP Hem	0.079	0.077	0.088	0.083	0.071	0.066	0.102				
LP Bet	*0.022*	*0.025*	*0.028*	0.027	0.032	*0.020*	0.115	0.090			
LP Bre	0.099	0.116	0.135	0.091	0.102	0.087	0.192	0.189	0.088		
LP Ris	0.033	0.035	0.050	0.034	0.033	0.030	0.121	0.103	0.055	0.097	
LP Oir	0.040	0.048	0.044	0.044	0.046	0.028	0.075	0.087	0.040	0.087	0.063

Table S2: Confidence measure in model selection inferred from 1000 pseudo observed dataset

River	Focal scenario	Type I error	Type II error				
			SI	IM	AM	SC	PAN
AA	I	0.435	-	0.071	0.478	0.077	0.194
	IM	0.404	0.048	-	0.078	0.685	0.427
	AM	0.522	0.352	0.071	-	0.049	0
	SC	0.811	0.035	0.071	0.051	-	0.049
	PAN	1	0	0.071	0	0	-
BET	I	0.989	-	0.007	0.012	0.005	0
	IM	0.326	0.027	-	0.049	0.763	0.975
	AM	0.988	0.573	0.007	-	0.106	0
	SC	0.897	0.387	0.007	0.298	-	0.106
	PAN	0.975	0.002	0.007	0	0.023	-
BRE	I	0.73	-	0.098	0.109	0.087	0
	IM	0.698	0.086	-	0.186	0.253	0.015
	AM	0.891	0.107	0.098	-	0.058	0
	SC	0.4	0.537	0.098	0.614	-	0.058
	PAN	0.741	0	0.098	0	0.002	-
HEM	I	0.632	-	0.107	0.17	0.076	0
	IM	0.744	0.151	-	0.354	0.337	0.941
	AM	0.83	0.37	0.107	-	0.051	0
	SC	0.464	0.111	0.107	0.294	-	0.051
	PAN	1	0	0.107	0	0	-
OIR	I	0.885	-	0.02	0.058	0.003	0
	IM	0.742	0.236	-	0.263	0.226	0.062
	AM	0.942	0.491	0.02	-	0.0119	0
	SC	0.925	0.119	0.02	0.169	-	0.019
	PAN	0.16	0.039	0.02	0.04	0.677	-
RIS	I	0.876	-	0.041	0.115	0.051	0
	IM	0.868	0.2	-	0.328	0.151	0.01
	AM	0.885	0.647	0.041	-	0.026	0
	SC	0.698	0.029	0.041	0.095	-	0.026
	PAN	0.011	0	0.016	0.041	0.047	-

Table S3: Confidence measure in model selection inferred from 1000 pseudo observed dataset by pairs for the strict isolation (SI) model and panmictic (PAN) model

		Type I error						Type II error					
Focal model		**AA**	**BET**	**BRE**	**HEM**	**OIR**	**RIS**	**AA**	**BET**	**BRE**	**HEM**	**OIR**	**RIS**
SI versus	IM	0	0	0	0	0	0	0	0.092	0.031	0.087	0.037	0.066
	AM	0.337	0.949	0.078	0.283	0.745	0.98	0.609	0.103	0.881	0.7	0.214	0.046
	SC	0	0	0	0	0	0	0	0.068	0.029	0.073	0.023	0.05
	PAN	0	0.001	0	0	0.01	0.004	0	0	0	0	0.106	0

		Type I error						Type II error					
Focal model		**AA**	**BET**	**BRE**	**HEM**	**OIR**	**RIS**	**AA**	**BET**	**BRE**	**HEM**	**OIR**	**RIS**
PAN versus	**IM**	0	0	0	0.002	0	0	0.005	0.023	0	0	0.009	0.027
	AM	0	0	0	0.001	0	0	0.001	0.01	0	0	0.095	0.022
	SC	0	0	0	0	0	0	0	0	0	0	0	0
	SI	0	0	0	0	0	0	0	0.001	0	0	0.01	0.004

Table S4: Estimation of variable importance from random-forest analyses. Estimation is performed pairwise for computational tractability. For each pairwise comparison, the three most discriminatory variables are in bold.

	MODEL											
Summary Stat	IvsAM	IvsIM	IvsSC	IvsPAN	AmvsIM	Scvs AM	AmvsPAN	SCvsIM	ImvsPAN	PANvsIM	ScvsPAN	Average
mean He pop1*	**76.64**	31.58	30.98	18.37	35.76	38.31	19.66	24.23	37.28	37.28	38.94	32.62
var He pop1	55.06	26.87	25.24	12.90	24.56	27.62	10.53	19.00	27.76	27.76	27.45	23.35
mean He pop2*	45.32	27.99	29.70	19.03	38.88	40.61	21.76	18.55	46.75	46.75	48.50	37.16
var He pop2	37.46	21.60	22.84	13.88	33.64	35.13	13.10	17.17	29.06	29.06	28.48	25.33
mean He total	58.42	**48.69**	**48.79**	17.88	**67.61**	**66.55**	24.14	25.90	42.59	42.59	45.80	41.26
var He total	42.27	35.04	32.25	16.30	32.66	36.05	12.12	20.94	28.28	28.28	28.18	25.64
mean Ar pop1	**90.18**	27.56	28.04	13.33	35.17	36.22	13.05	25.68	30.79	30.79	26.85	27.23
var Ar pop1	49.78	23.27	22.09	10.40	36.39	38.03	14.59	19.10	25.36	25.36	25.48	24.66
mean Ar pop2	57.54	25.95	25.37	13.16	34.15	31.60	15.00	28.02	22.31	22.31	22.68	23.65
var Ar pop2	46.81	25.68	22.56	10.06	36.83	35.89	8.23	23.40	16.03	16.03	18.05	19.61
mean Ar total	56.41	38.12	44.20	14.32	43.63	52.08	16.48	23.71	27.01	27.01	24.98	28.54
var Ar total	61.87	21.91	22.81	11.07	31.53	35.19	11.97	19.85	23.42	23.42	25.35	23.20
mean R pop1	47.75	30.25	30.34	8.61	37.93	39.47	10.77	21.22	21.85	21.85	20.81	22.66
var R pop1	57.89	41.89	42.03	7.08	43.97	47.00	9.64	27.19	18.67	18.67	18.13	23.22
mean R pop2	59.87	27.06	32.17	11.53	26.88	32.14	12.87	27.23	26.38	26.38	23.35	24.72
var R pop2	69.44	36.76	46.21	9.39	39.87	46.10	9.37	23.71	21.94	21.94	18.44	23.58
mean R total	44.62	27.08	28.98	11.01	32.01	35.34	11.42	17.09	22.67	22.67	22.07	21.88
var R total	46.63	21.70	21.47	7.86	30.13	29.35	9.54	21.86	19.34	19.34	16.84	19.38
var GW pop1	**94.41**	39.08	43.83	11.70	37.76	37.70	11.88	22.05	28.71	28.71	28.28	26.22
mean GW pop1	50.64	25.46	20.65	9.07	37.06	38.73	6.98	12.90	13.26	13.26	16.00	16.85
var GW pop2	64.23	31.60	37.27	12.00	31.62	37.62	12.35	25.52	26.56	26.56	25.91	25.75
mean GW pop2	43.42	26.60	21.40	7.83	40.62	38.83	7.77	8.55	15.02	15.02	14.02	16.54
var GW total	59.44	31.61	31.51	9.88	35.88	38.19	10.89	20.88	26.82	26.82	27.07	25.11
mean GW total	39.53	24.50	22.06	6.14	38.91	34.47	5.46	18.77	12.46	12.46	13.00	16.10
mean GST	65.74	**62.09**	**66.67**	**34.69**	**80.61**	**85.93**	**57.28**	**34.11**	**84.59**	**84.59**	**84.41**	**71.82**
var GST	37.25	37.52	38.70	27.29	51.50	50.54	40.93	**30.75**	**104.78**	**104.78**	**107.81**	**73.27**
mean deltamu^2	71.55	**74.59**	**72.38**	**52.09**	**76.51**	**76.05**	**61.36**	**33.14**	**103.26**	**103.26**	**98.59**	**79.28**
var deltamu^2	60.48	43.46	44.62	**41.85**	47.22	48.70	**45.17**	23.59	57.10	57.10	60.18	48.64

*Pop 1 = *Lampetra fluviatilis* population *Pop 2 = *Lampetra planeri* population

Table S5: **Robustness of summary statistics computation under the two best models (IM and SC).** Estimation based on 10 00 dataset drawn from the posterior distribution of the parameters. Summary statistics that deviate significantly from the observed distribution are in bold.

	IM					SC					PAN
Statistics	AA	BET	BRE	HEM	RIS	AA	BET	BRE	HEM	RIS	OIR
mean_He_pop1	0.9915	0.9912	0.997	0.9888	1	1	1	1	0.9485	0.9994	1
var_He_pop1	**4.00E-004**	**2.00E-004**	**0.002**	**4.00E-004**	**0**	**0**	**1.00E-004**	**0**	**0.0089**	**0**	1
mean_He_pop2	0.8836	0.9962	0.962	0.8489	1	0.948	1	1	0.8033	0.9859	1
var_He_pop2	**0.0471**	**1.00E-004**	**0.01**	**0.0156**	**6.00E-004**	**0.033**	**0**	**0**	**0.0451**	**0.0316**	1
mean_He_total	0.9539	0.9984	0.999	0.9752	1	0.997	1	1	0.9111	0.9995	1
var_He_total	**0.0011**	**0**	**0.002**	**1.00E-004**	**0**	**0.01**	**0**	**0**	**0.011**	**0**	1
mean_A_pop1	0.9909	0.9912	0.999	0.9889	1	1	1	1	0.9491	0.9994	1
var_A_pop1	**4.00E-004**	**5.00E-004**	**0.001**	**4.00E-004**	**0**	**0**	**1.00E-004**	**0**	**0.0097**	**0**	1
mean_A_pop2	0.887	0.9962	0.967	0.8542	1	0.962	1	1	0.8079	0.9864	1
var_A_pop2	**0.0384**	**1.00E-004**	**0.01**	**0.0132**	**7.00E-004**	**0.031**	**0**	**0**	**0.0405**	**0.0342**	1
mean_A_total	0.9549	0.9984	0.999	0.9765	1	0.997	1	1	0.9128	0.9995	1
var_A_total	**0.0011**	**0**	**0.002**	**0**	**0**	**0.01**	**0**	**0**	**0.0103**	**0**	1
mean_Ar_pop1	0.9928	0.9906	1	0.9978	1	1	1	1	0.9776	0.9999	1
var_Ar_pop1	0.2832	0.3585	0.435	0.2961	0.4841	0.598	0.4826	0.732	0.2514	0.6358	**0**
mean_Ar_pop2	0.9075	0.9962	0.934	0.9931	1	0.999	1	1	0.8948	0.99	1
var_Ar_pop2	0.6802	0.4688	**0.045**	0.3868	0.5767	0.882	0.4959	0.352	0.4541	0.6036	**0**
mean_Ar_total	0.9909	0.9999	0.98	0.9982	1	1	1	0.999	0.9849	0.9991	1
var_Ar_total	0.3833	0.4709	0.109	0.4236	0.6422	0.629	0.569	0.376	0.3807	0.6889	**0**
mean_V_pop1	0.9926	0.9906	1	0.9978	1	1	1	1	0.9767	0.9999	1
var_V_pop1	0.2701	0.3257	0.439	0.2796	0.4769	0.586	0.4457	0.736	0.2421	0.6298	0
mean_V_pop2	0.9057	0.996	0.928	0.9924	1	0.999	1	1	0.8929	0.9895	1
var_V_pop2	0.6692	0.4395	**0.044**	0.3845	0.5762	0.874	0.458	0.341	0.4504	0.6036	**0**
mean_V_total	0.9909	0.9999	0.979	0.9981	1	1	1	0.999	0.9845	0.9987	1
var_V_total	0.3747	0.4431	0.109	0.4195	0.638	0.621	0.5385	0.377	0.3764	0.6837	**0**

mean_R_pop1	0.9785	0.9916	1	0.9916	1	0.999	1	1	0.9508	0.9999	1
var_R_pop1	0.8185	0.9522	0.989	0.8252	0.9995	0.916	0.9818	1	0.7275	0.9991	1
mean_R_pop2	0.974	0.9997	0.164	0.9702	0.9999	1	1	0.713	0.8687	0.9832	1
var_R_pop2	0.9956	0.9867	**0.006**	0.6688	0.962	1	0.9954	0.249	0.5827	0.9346	1
mean_R_total	0.9874	1	0.441	0.9884	1	1	1	0.848	0.9425	0.9988	**0**
var_R_total	0.8553	0.9729	0.108	0.8635	0.9782	0.934	0.9892	0.635	0.787	0.967	**0**
mean_GW_pop1	0.4724	**0.0458**	0.075	0.5732	**4.00E-004**	0.513	**0.0192**	**0.009**	0.6205	**0.003**	**0**
var_GW_pop1	0.783	0.6298	0.954	0.8314	0.6711	0.83	0.6003	0.921	0.8376	0.7077	**0**
mean_GW_pop2	**0.0285**	**0.0242**	0.99	**0.0197**	0.2775	**0.002**	**0.0055**	0.901	0.0849	0.2945	**0**
var_GW_pop2	0.9567	0.5846	0.182	**0**	0.462	0.967	0.5361	0.216	**0.012**	0.4693	1
mean_GW_total	0.4408	**0.0326**	1	0.7969	0.0809	0.456	**0.0119**	0.975	0.8305	0.0972	1
var_GW_total	0.7364	0.6681	0.907	0.9359	0.2853	0.794	0.6458	0.852	0.9416	0.2726	**0**
mean_GST	0.2031	0.3586	0.546	0.5067	0.7748	0.191	0.2718	0.015	0.3924	0.6114	**0**
var_GST	0.0845	**0.0245**	0.76	0.2878	0.2154	0.046	**0.0045**	**0.091**	0.3036	0.2239	1
mean_deltamu2	0.9328	0.9917	0.117	0.8329	0.979	0.976	0.9969	0.181	0.7456	0.9464	1
var_deltamu2	0.9387	0.9883	0.17	0.6931	0.9705	0.979	0.9944	0.243	0.6096	0.9378	1

Figure S1: Curves of out-of-bag errors rates and Estimation of variable importance for each population pairs.

Data based on one random forest, each composed of 1,000 trees obtained from a trained set of 50,000 simulated predictor variables (summary statistics). The response variable is the demographic model.

Each page corresponds to a different river:
(A) AA

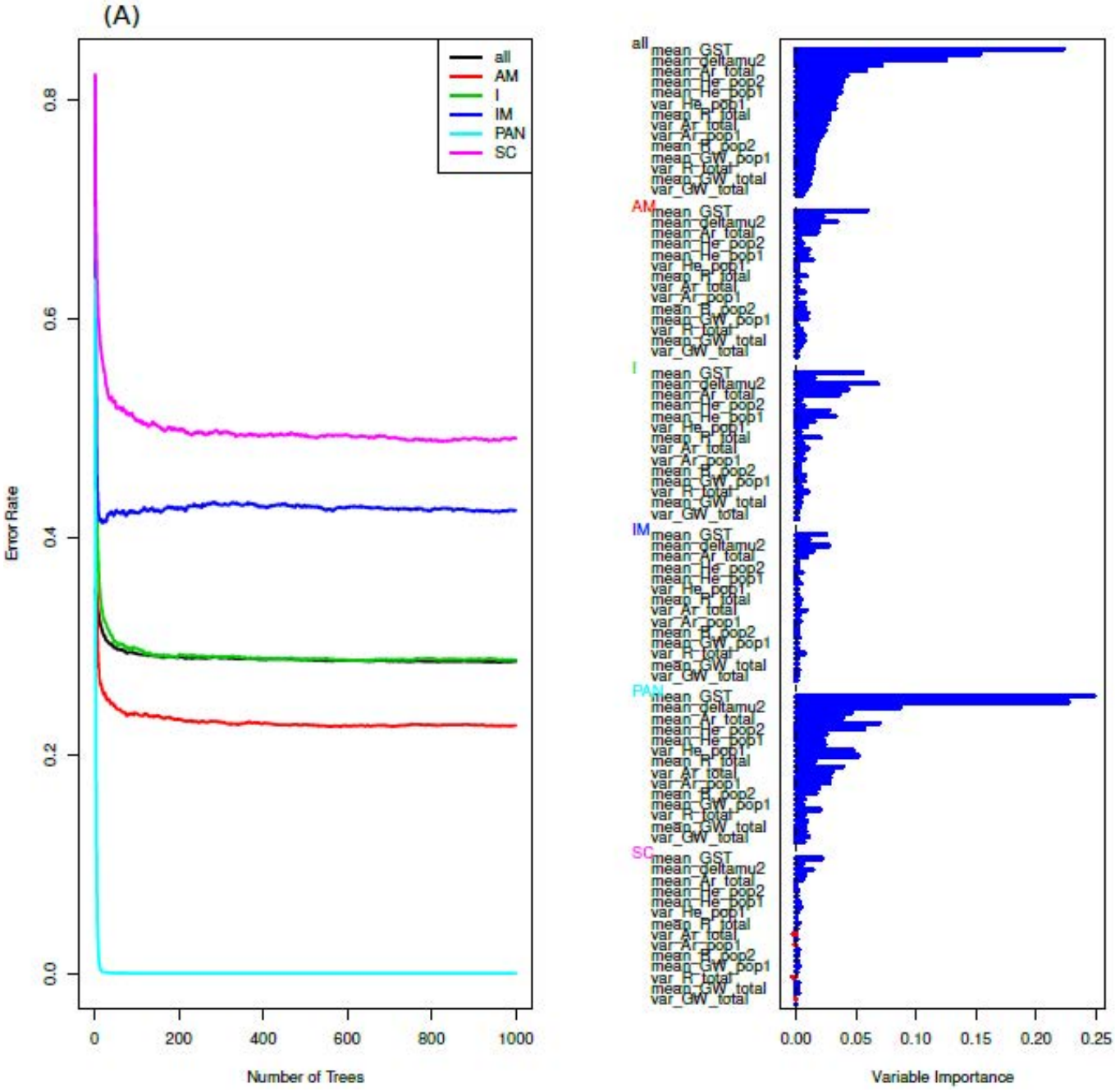
(B) BET

(C) BRE

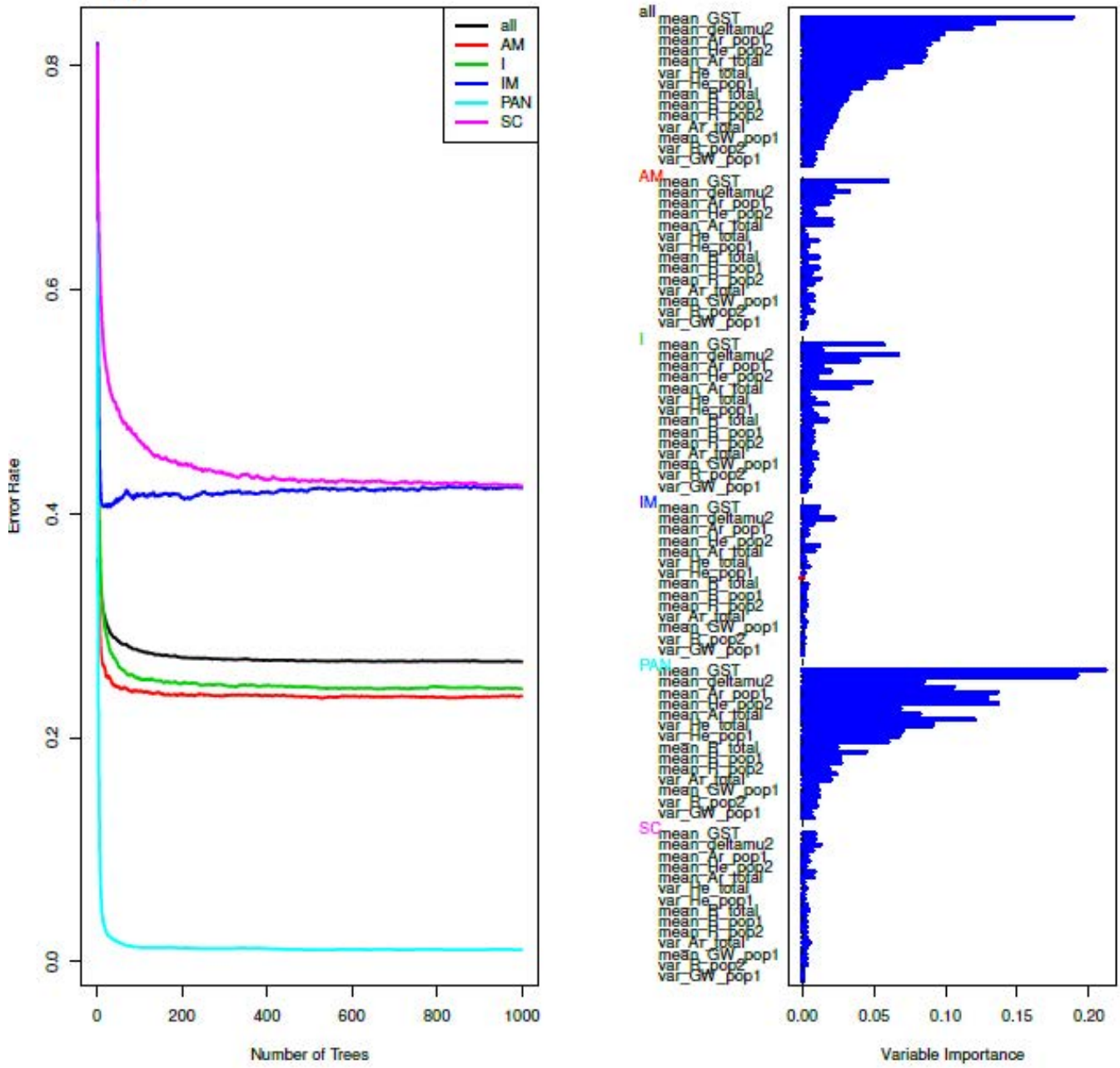
(D) HEM

(E) OIR

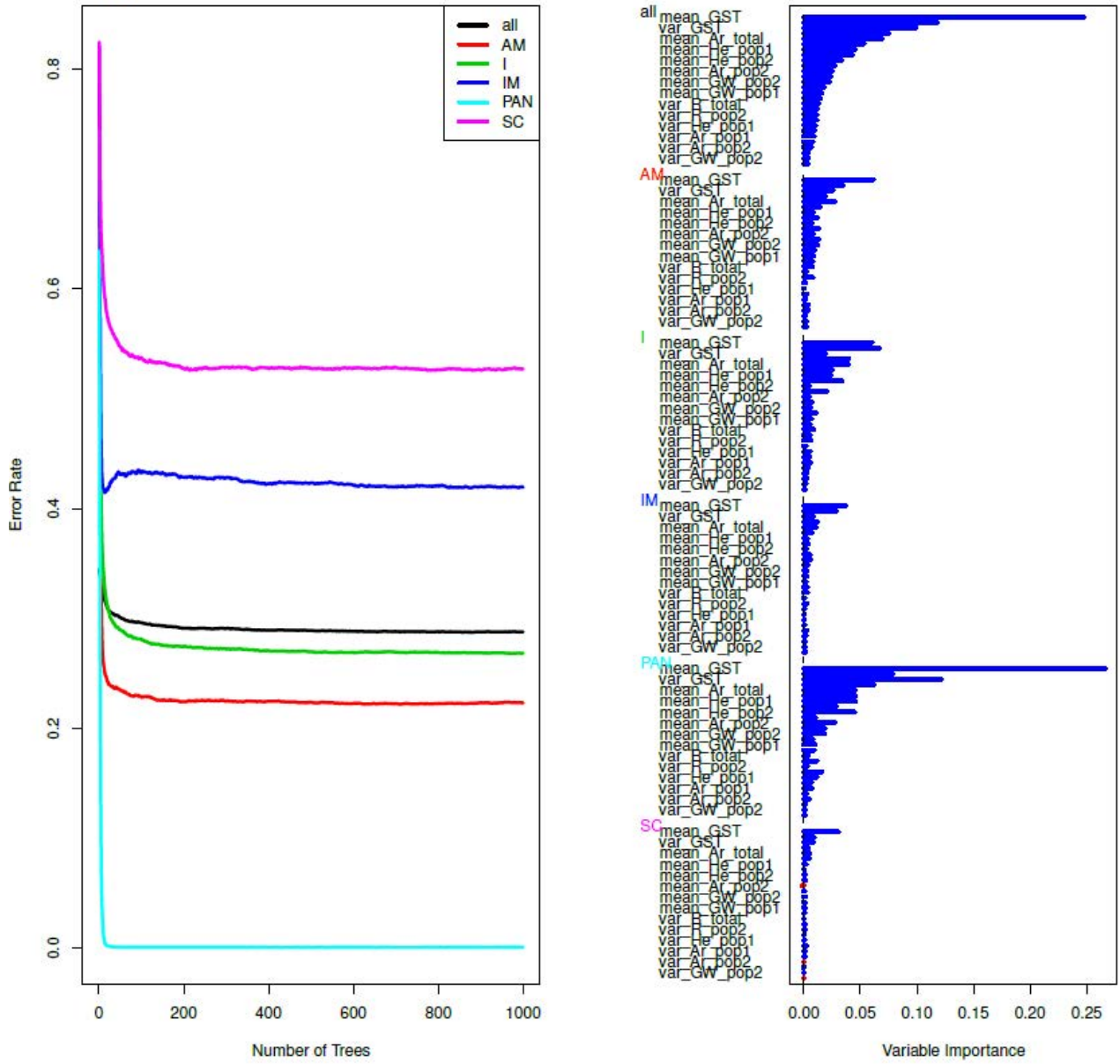
(F) RIS


(A)
all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees
mean_GST
mean_deltamu2
mean_Ar_total
mean_He_pop2
mean_He_pop1
var_He_pop1
mean_R_total
var_Ar_total
var_Ar_pop1
mean_R_pop2
mean_GW_pop1
var_R_total
mean_GW_total
var_GW_total
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance

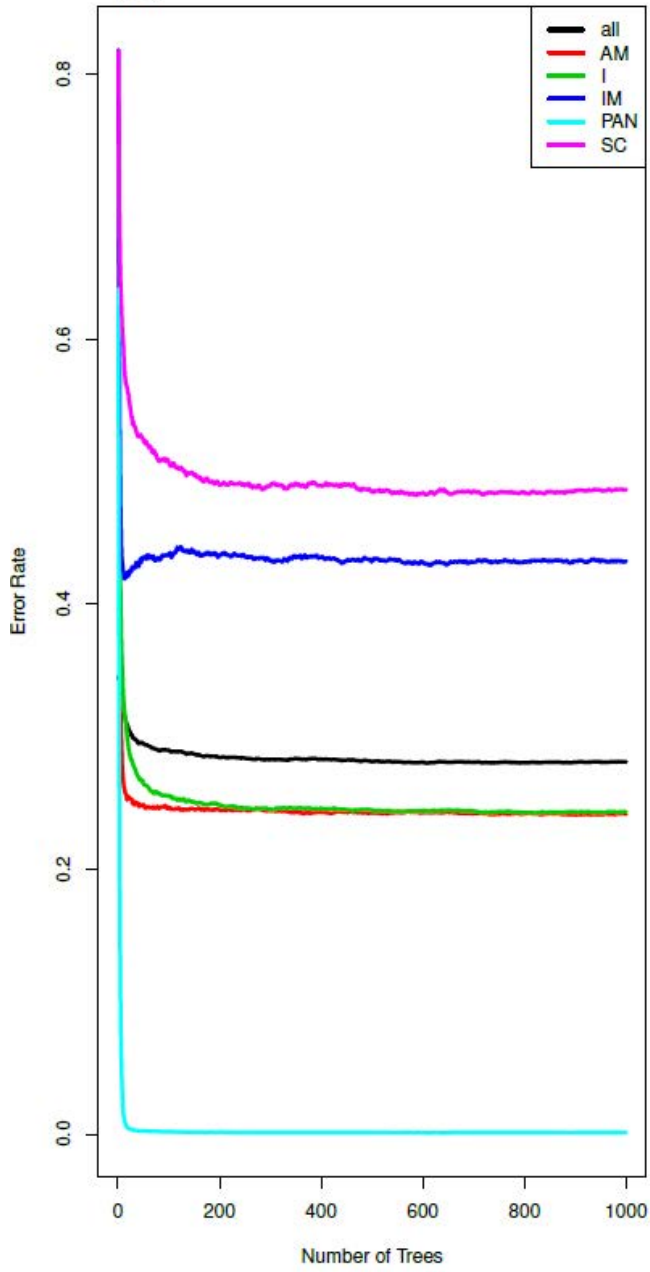
(B)

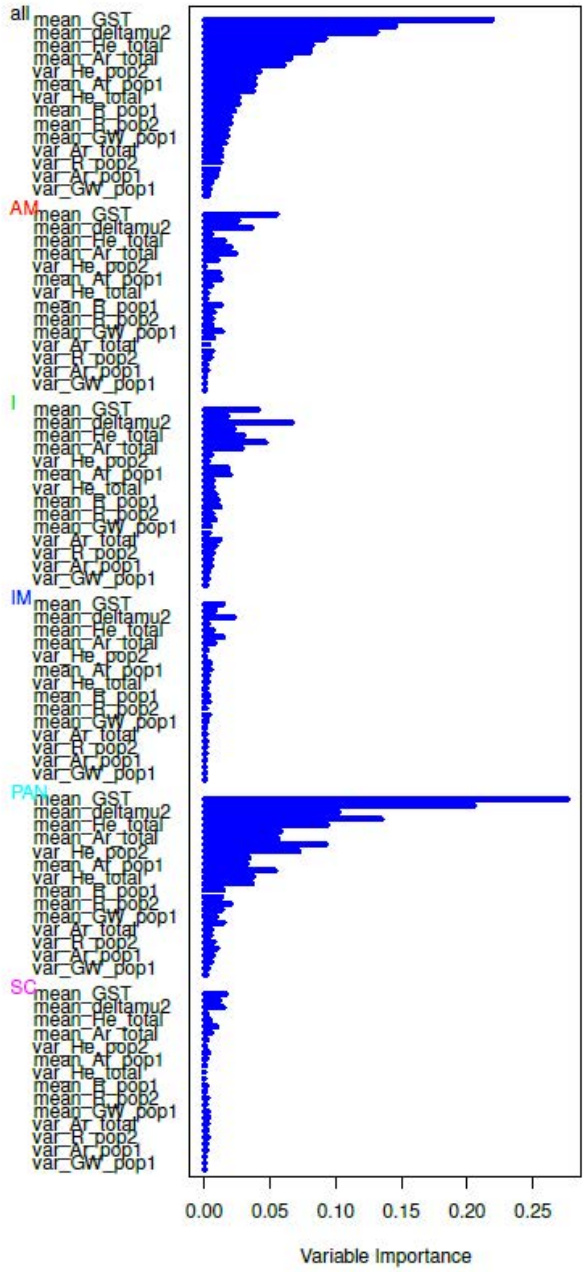

all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees
mean_GST
mean_deltamu2
mean_Ar_pop1
mean_He_pop2
mean_Ar_total
var_He_total
var_He_pop1
mean_B_total
mean_B_pop1
mean_H_pop2
var_Ar_total
mean_GW_pop1
var_B_pop2
var_GW_pop1
0.00
0.05
0.10
0.15
0.20
Variable Importance

(C)

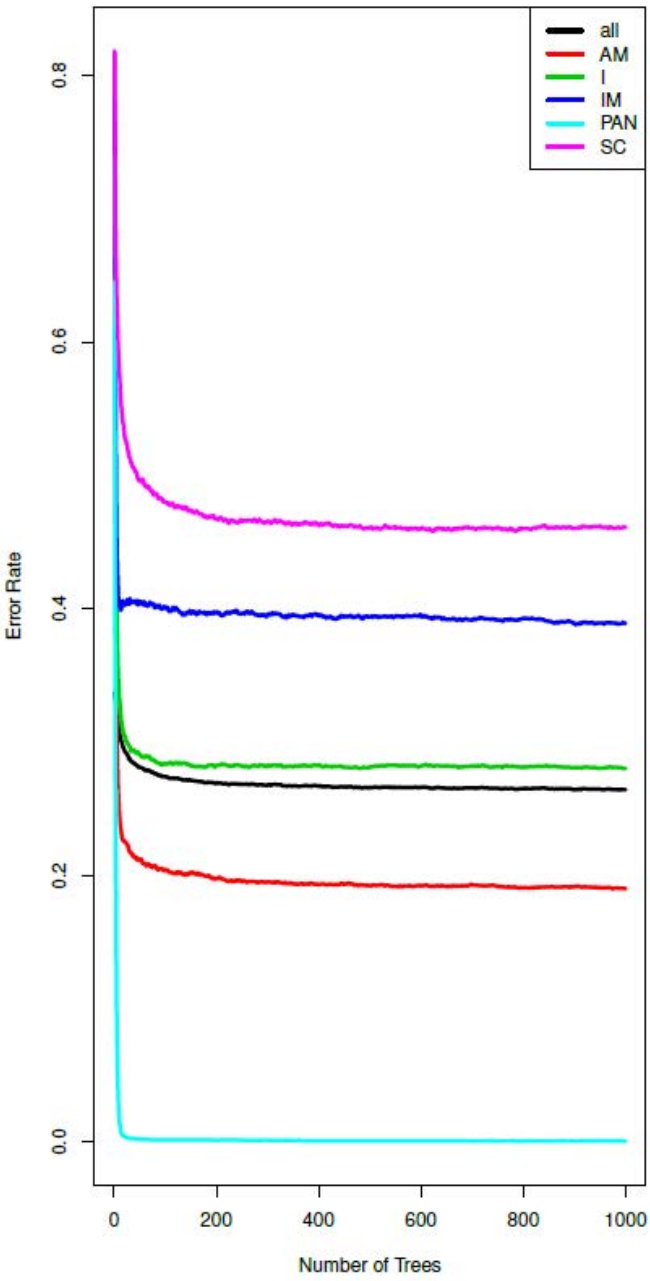

all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees
mean_GST
var_GST
mean_Ar_total
mean_He_pop1
mean_He_pop2
mean_Ar_pop2
mean_GW_pop2
mean_GW_pop1
var_R_total
var_R_pop2
var_He_pop1
var_Ar_pop1
var_Ar_pop2
var_GW_pop2
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance

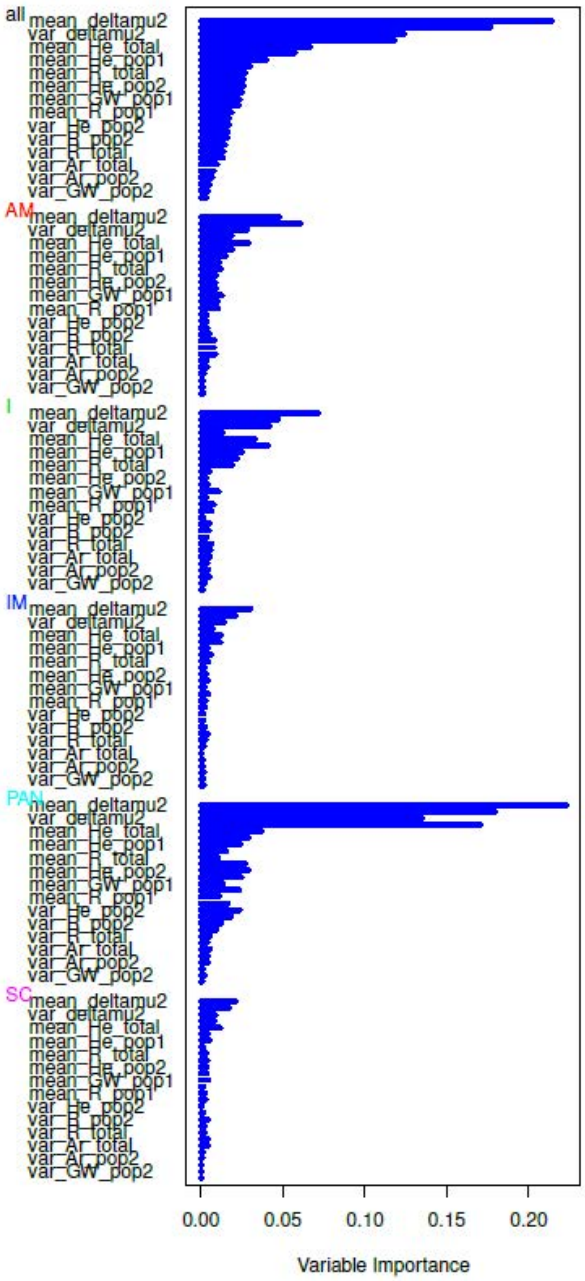
(D)


all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees


all
AM
I
IM
PAN
SC
mean_GST
mean_deltamu2
mean_He_total
mean_Ar_total
var_He_pop2
mean_Ar_pop1
var_He_total
mean_B_pop1
mean_B_pop2
mean_GW_pop1
var_Ar_total
var_B_pop2
var_Ar_pop1
var_GW_pop1
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance

(E)


all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees


all
AM
I
IM
PAN
SC
mean_deltamu2
var_deltamu2
mean_He_total
mean_He_pop1
mean_H_total
mean_He_pop2
mean_GW_pop1
mean_H_pop1
var_He_pop2
var_H_pop2
var_H_total
var_Ar_total
var_Ar_pop2
var_GW_pop2
0.00
0.05
0.10
0.15
0.20
Variable Importance

(F)

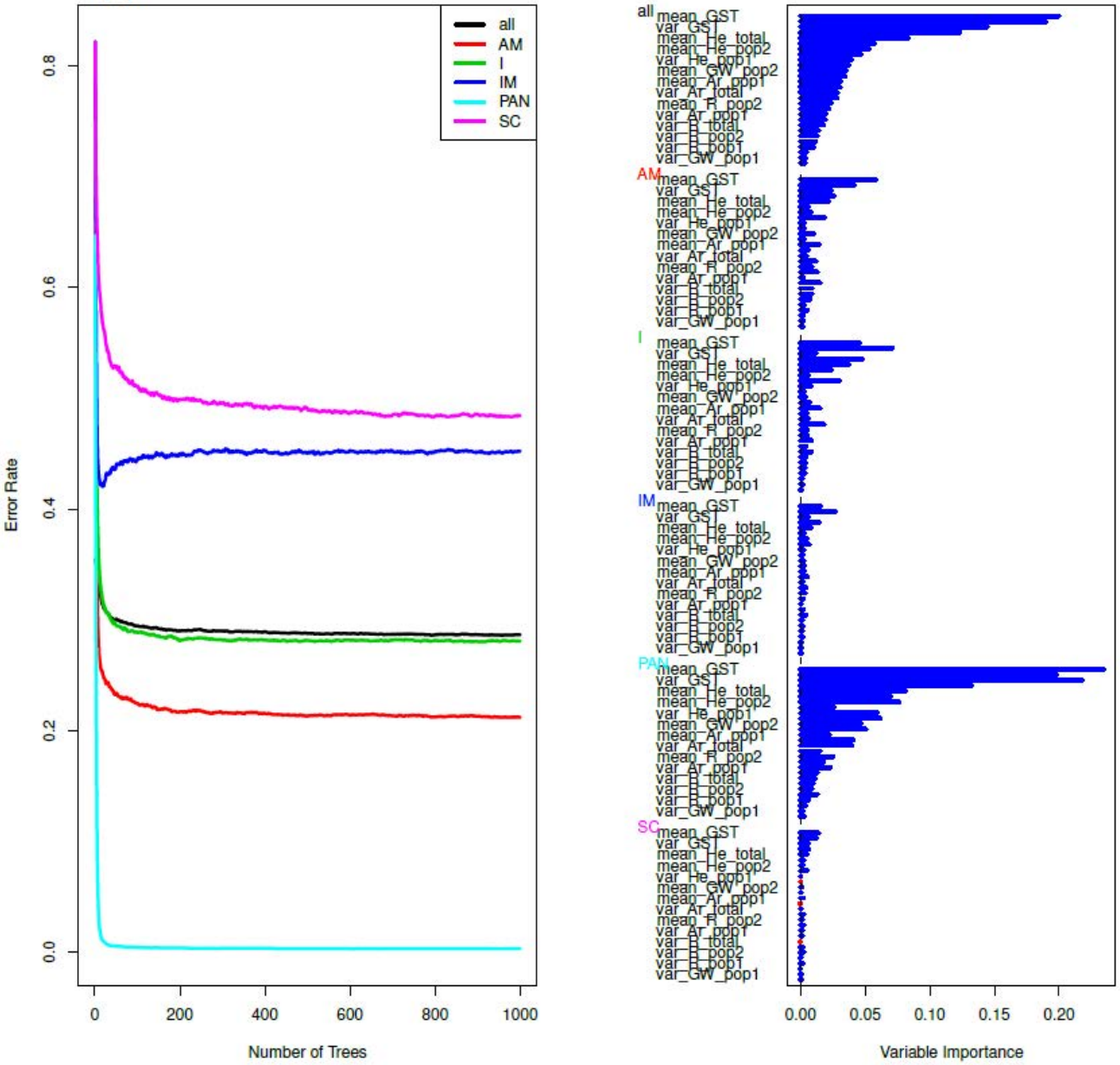

all
AM
I
IM
PAN
SC
Error Rate
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Number of Trees
mean_GST
var_GST
mean_He_total
mean_He_pop2
var_He_pop1
mean_GW_pop2
mean_Ar_pop1
var_Ar_total
mean_R_pop2
var_Ar_pop1
var_R_total
var_R_pop2
var_R_pop1
var_GW_pop1
0.00
0.05
0.10
0.15
0.20
Variable Importance

Figure S2: Distribution of summary statistics obtained from the posterior predictive check.

Statistics obtained after 10 000 simulations based on posterior distributions.
Each page corresponds to a different river and a different model

(A) AA – IM

(B) BET – IM

(C) BRE – IM

(D) HEM – IM

(E) RIS – IM

(F) AA – SC

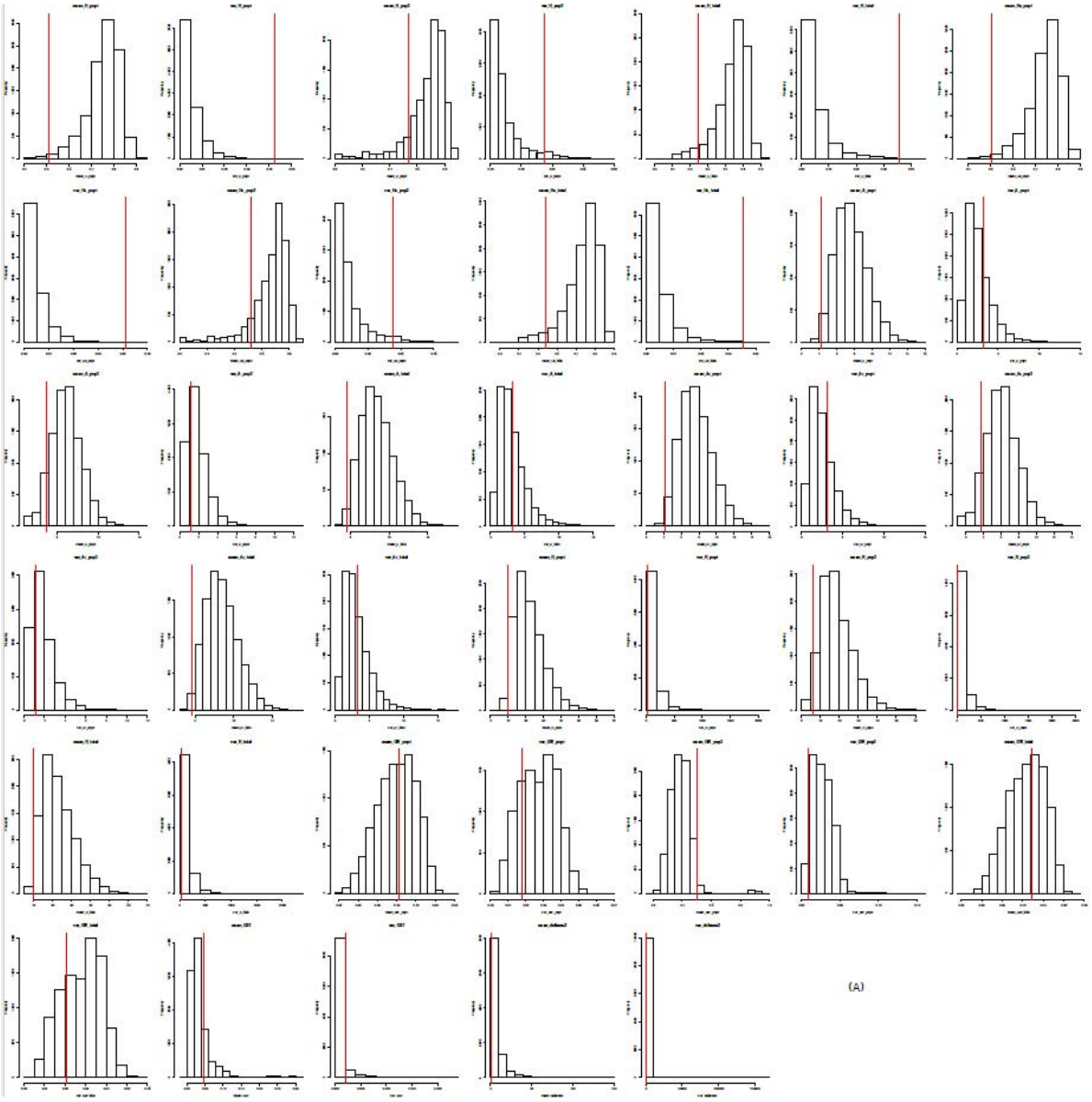
(G) BET – SC

(H) BRE – SC

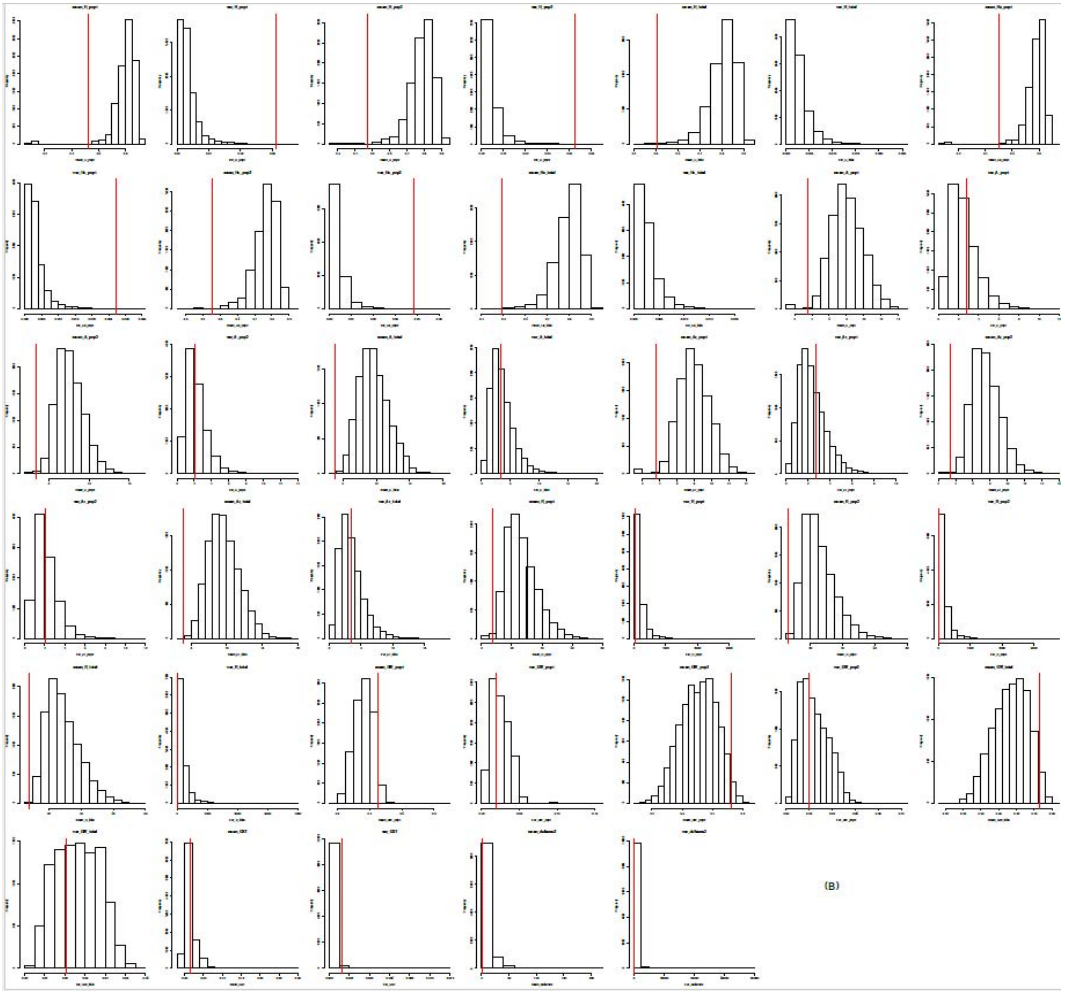
(I) HEM – SC

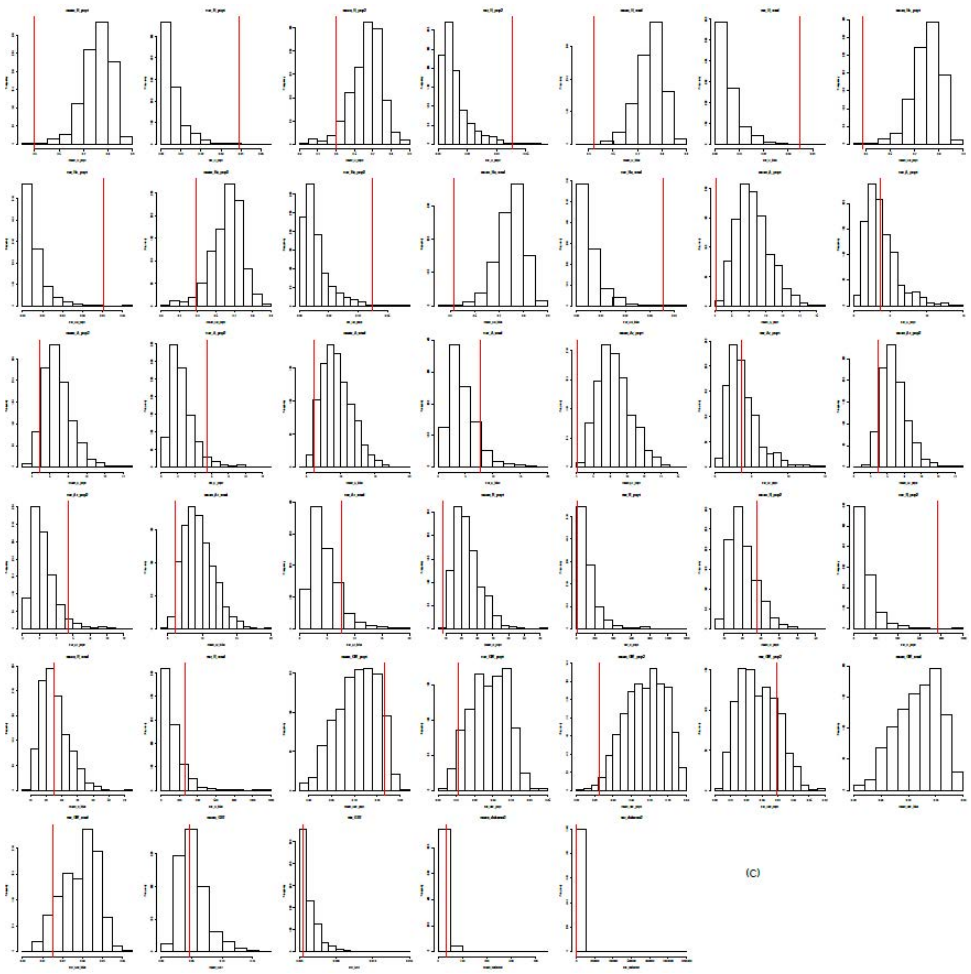
(J) RIS – SC

(K) OIR – PAN

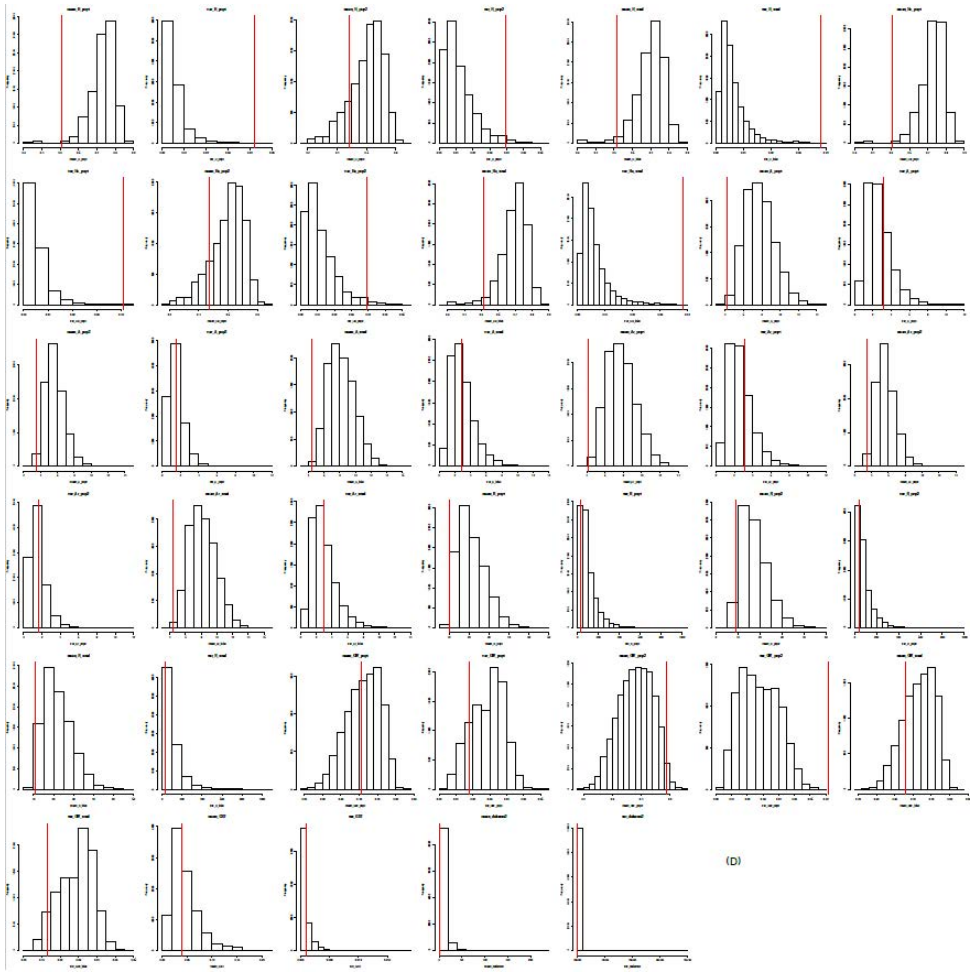


(A)

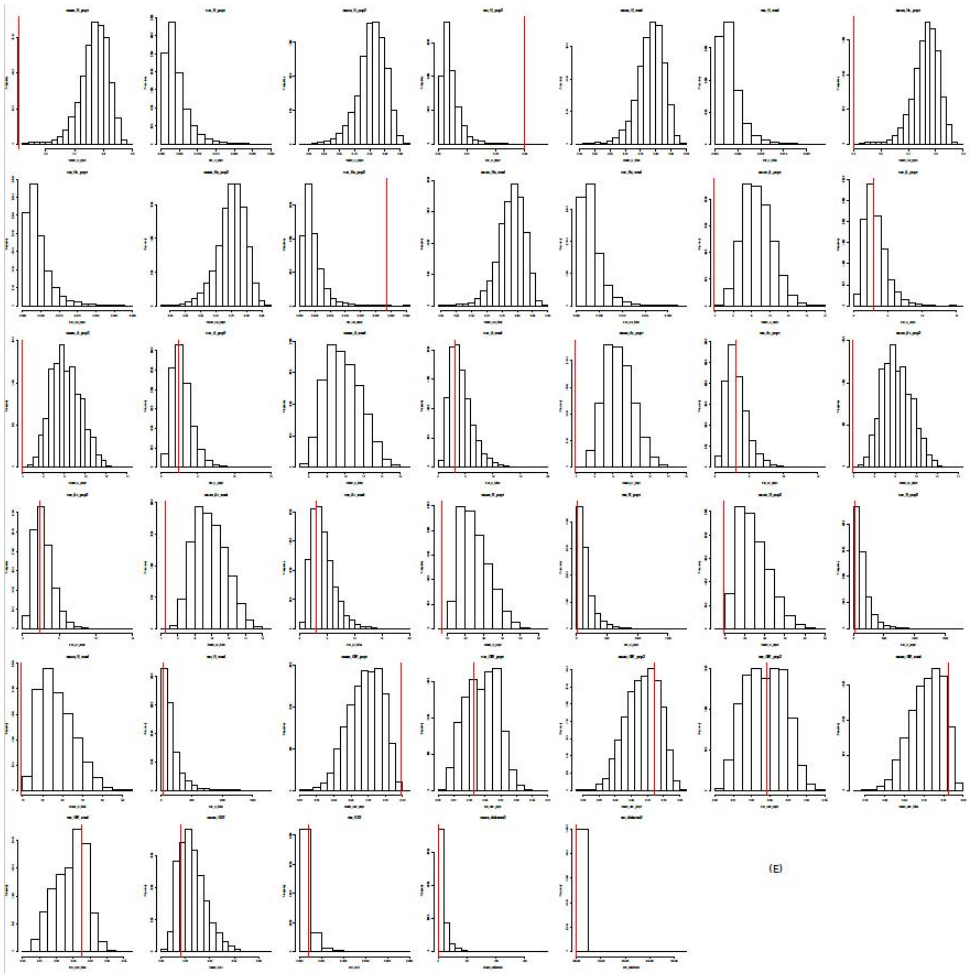

(B)



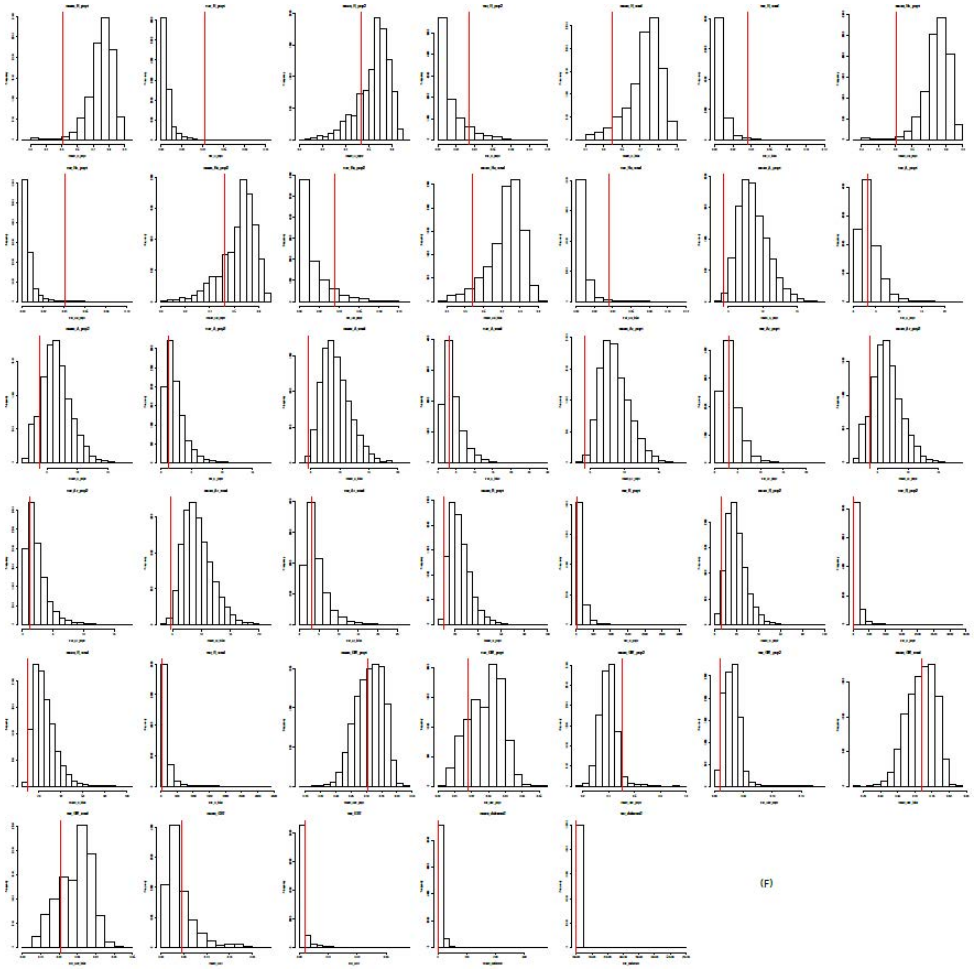
(c)



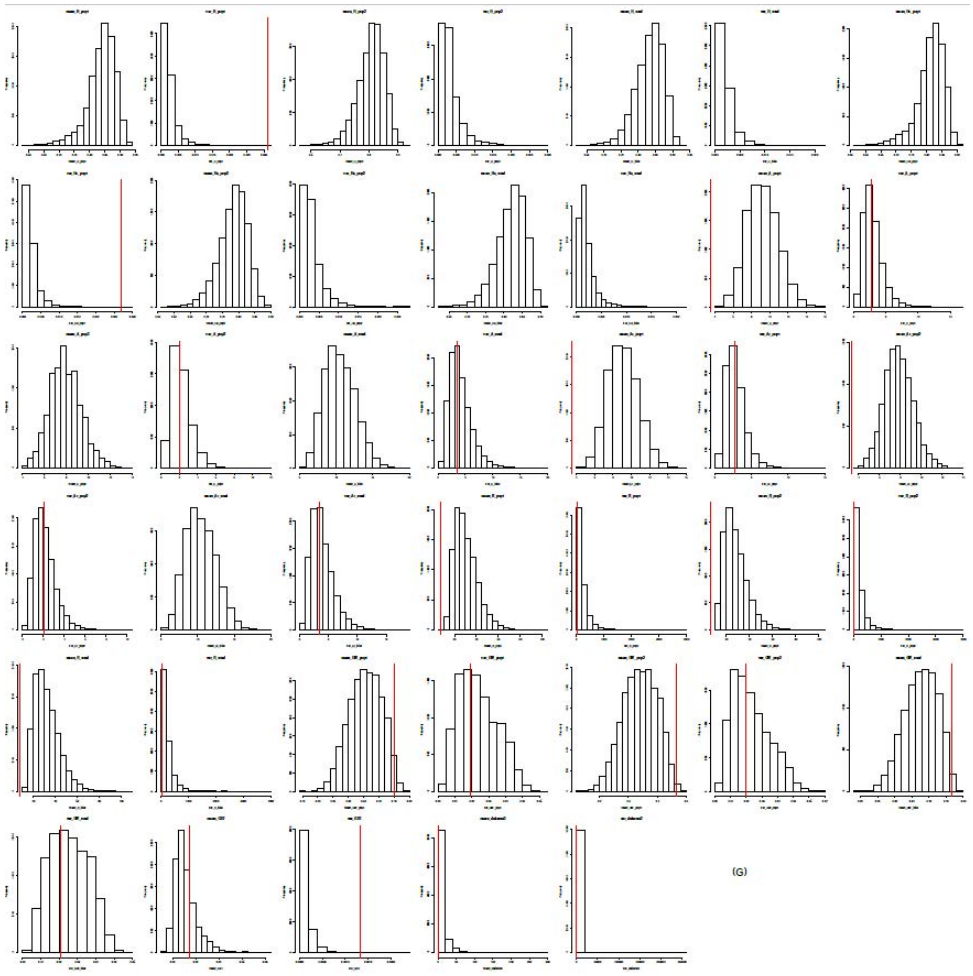
(D)



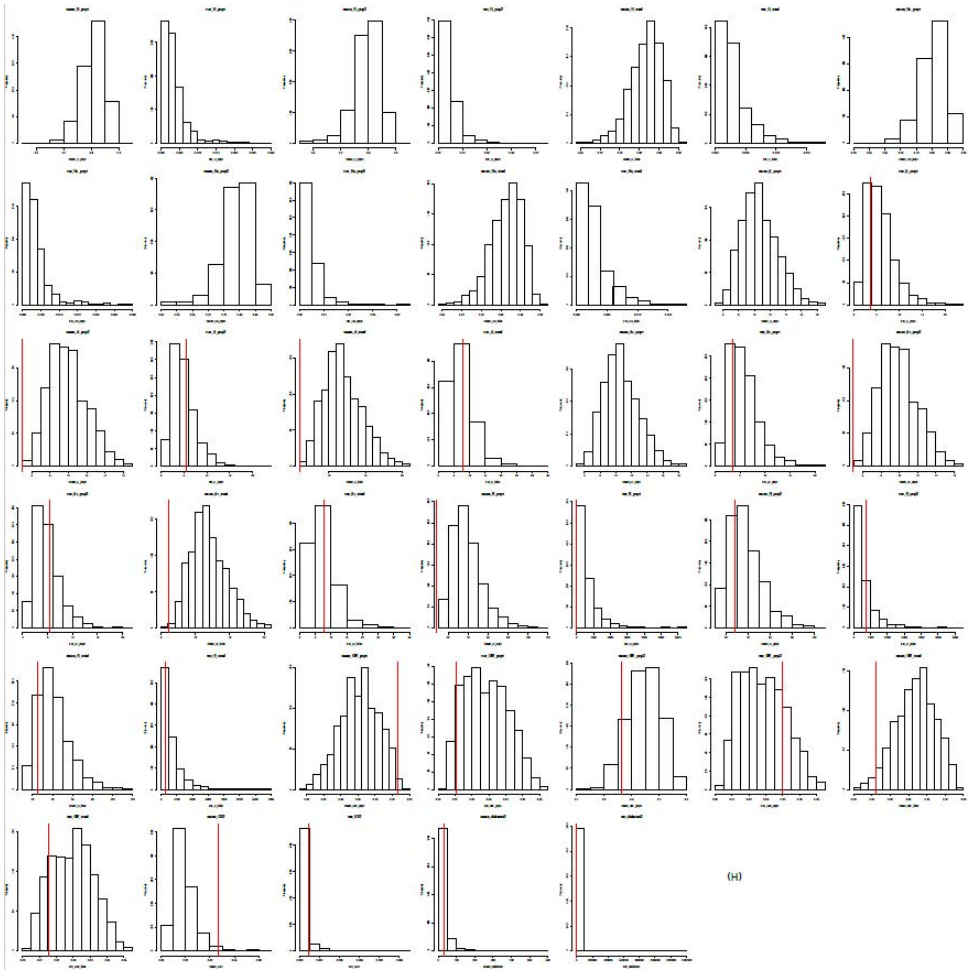
(E)



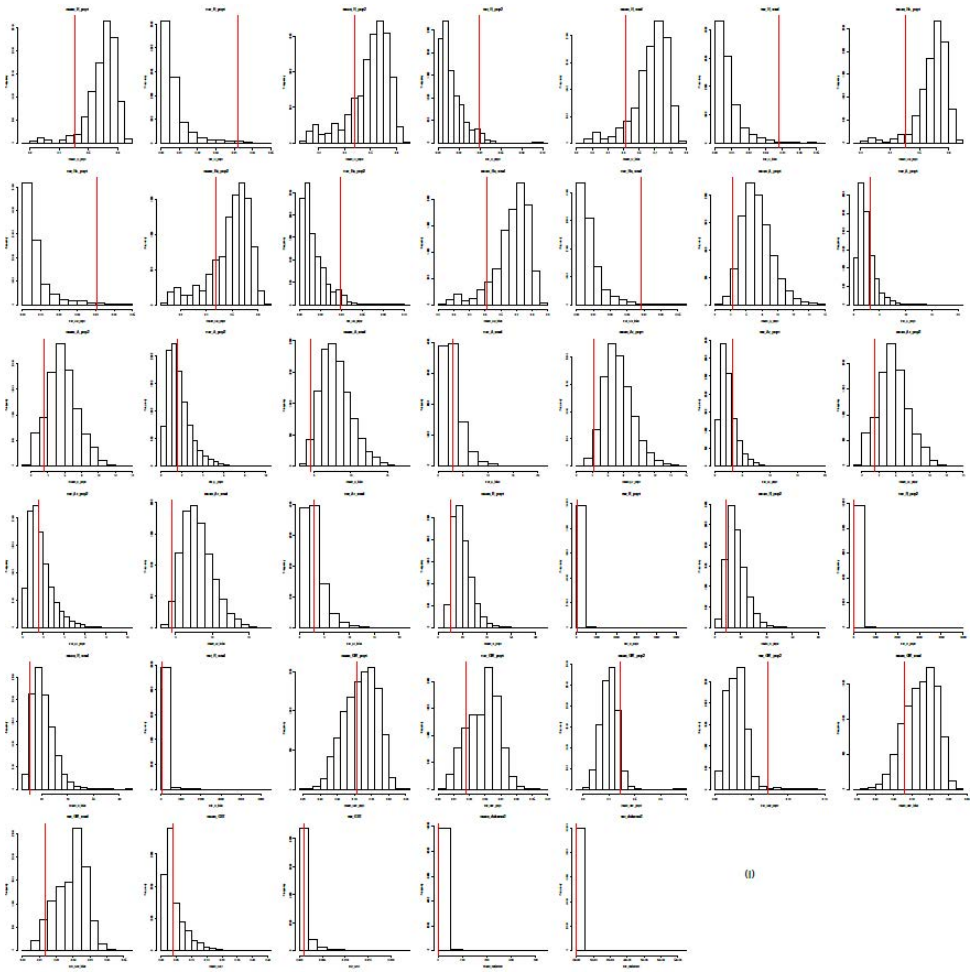
(F)



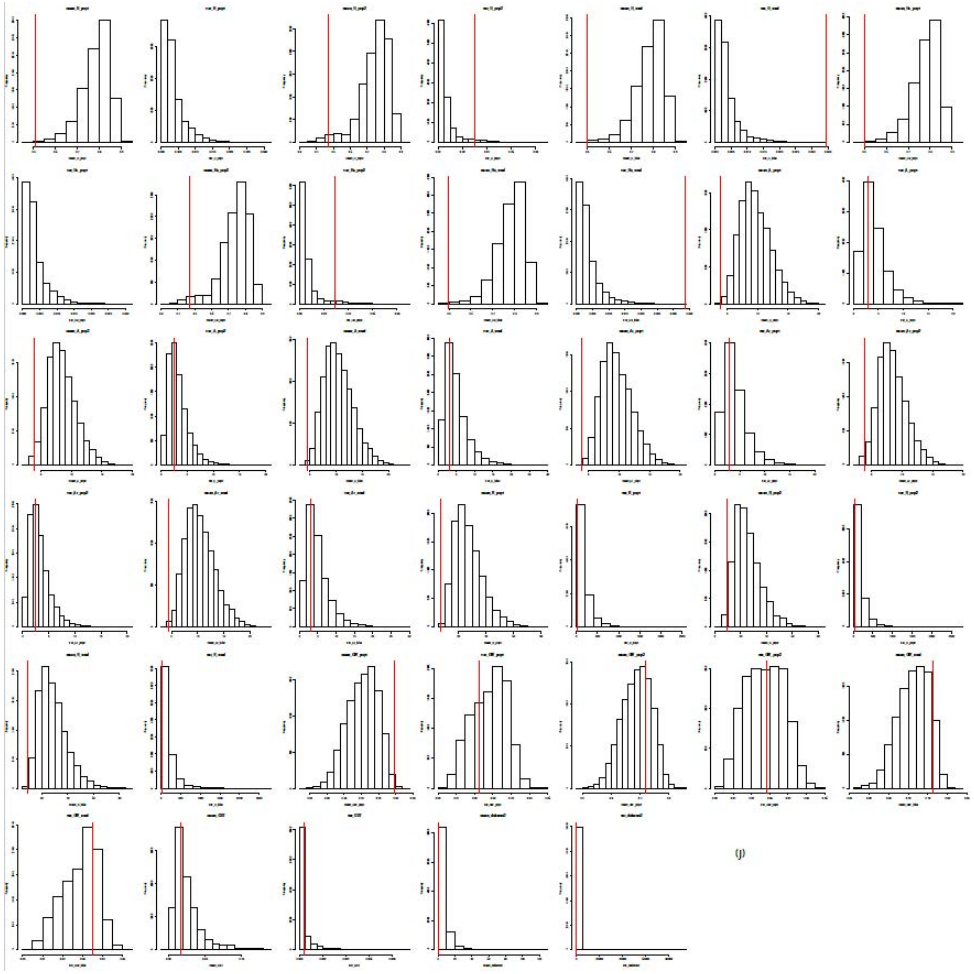
(G)



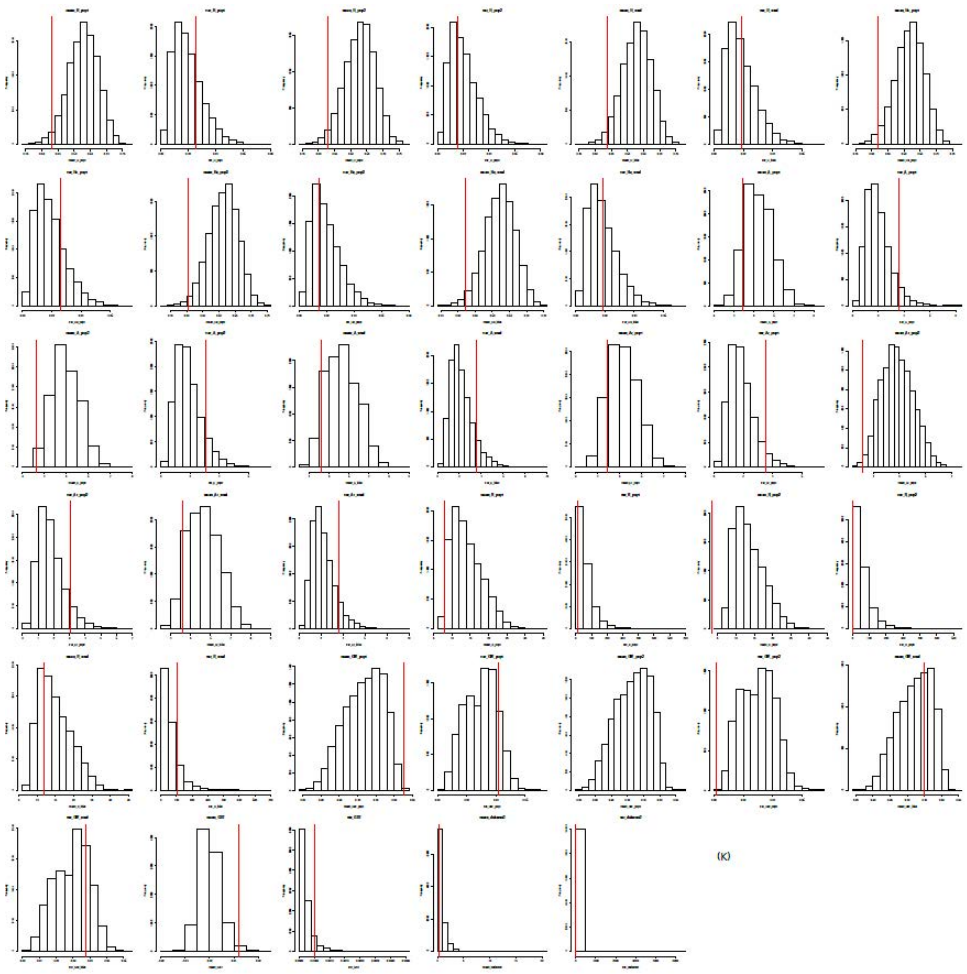
(H)



(i)



(j)



(k)

Chapter 4:

Understanding speciation: moving toward genomics

Article 3: Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and non-parasitic lamprey ecotypes

Quentin Rougemont, Pierre-Alexandre Gagnaire, Charles Perrier, Clémence Genthon, Anne-Laure Besnard, Sophie Launey, Guillaume Evanno

In prep for Molecular Ecology, abstract accepted for the special issue "molecular basis of adaptation and speciation"

Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and non-parasitic lamprey ecotypes

Quentin Rougemont[1,2], Pierre-Alexandre Gagnaire [3,4*], Charles Perrier[5*], Clémence Genthon[6], Anne-Laure Besnard[1,2], Sophie Launey[1,2], Guillaume Evanno[1,2]

Author for Correspondence: quentinrougemont@orange.fr

[1]INRA, UMR 985 Ecologie et Santé des Ecosystèmes, 35042 Rennes, France

[2]Agrocampus Ouest, UMR ESE, 65 rue de Saint-Brieuc, 35042 Rennes, France

[3]Institut des Sciences de l'Evolution (UMR 5554), CNRS-UM2-IRD, Place Eugène Bataillon, F-34095 Montpellier, France

[4]Station Méditerranéenne de l'Environnement Littoral, Université Montpellier 2, 2 Rue des Chantiers, F-34200 Sète, France

[5]CEFE-CNRS, Centre D'Ecologie Fonctionnelle et Evolutive, Route de Mende, 34090 Montpellier, France

[6]Plateforme génomique INRA GenoToul Chemin de Borderouge - Auzeville, 31320 Castanet-Tolosan France

* Both authors contributed equally to this work.

Abstract

Understanding the evolutionary mechanisms generating parallel genomic divergence patterns among replicate ecotype pairs remains an important challenge in speciation research. We investigated the genomic divergence between the anadromous parasitic river lamprey *(Lampetra fluviatilis)* and the freshwater-resident non-parasitic brook lamprey (*Lampetra planeri*) in nine population pairs displaying variable levels of geographic connectivity. We genotyped 338 individuals with RAD-sequencing and inferred the demographic divergence history of each population pair using a diffusion approximation method. Divergence patterns in geographically connected population pairs were better explained by introgression after secondary contact, whereas disconnected population pairs have retained a signal of ancient migration. In all ecotype pairs, models accounting for differential introgression among loci outperformed homogeneous migration models. Generating neutral predictions from the inferred divergence scenarios to detect highly differentiated markers identified greater proportions of outliers in disconnected population pairs than in connected pairs. However, increased similarity in the most divergent genomic regions was found among connected ecotypes pairs, indicating that gene flow was instrumental in generating parallelism at the molecular level. These results suggest that heterogeneous genomic differentiation and parallelism among replicate ecotype-pairs have partly emerged through local gene flow reduction in genomic islands.

Introduction

Understanding the mechanisms responsible for the build-up and maintenance of species divergence is of primary importance in evolutionary biology. In particular, it is now common to observe highly heterogeneous differentiation patterns across the genome. However it is not always straightforward to determine whether these patterns are due to semipermeable barriers to gene flow (Wu 2001; Harrison and Larson 2014, Harrison 1986) or to postspeciation selection at linked sites (Turner et Hahn 2010; Cruickshank and Hahn 2014). It is even less easy to understand which sequences of historical and selective events underlie these heterogeneous patterns of differentiation. For instance, many studies have used genome scans to investigate patterns of heterogeneous differentiation and identify so-called genomic islands of differentiation (Seehausen *et al.* 2014). However, accurately disentangling the relative influence of gene flow, historical processes and recombination rate variations (Barrett & Hoekstra 2011b; Nachman & Payseur 2012; Roesti *et al.* 2013; Ellegren 2014) on the genomic landscape of differentiation has remained highly challenging so far.

The study of replicated pairs of populations (e.g. Schluter & McPhail 1993; Nosil *et al.* 2002, 2009a; Berner *et al.* 2009; Kaeuffer *et al.* 2012; Gagnaire *et al.* 2013b) inhabiting different environments offer great opportunities to investigate all these processes for which variable outcomes are expected (Welch and Jiggins, 2014; Lindtke and Buerkle 2015). It is often hypothesized that these populations offer independent replicates of a repeated evolutionary response to similar ecological constraints, and thus provide ideal systems to study ecological speciation (e.g. Feder et al. 2012). Parallel phenotypic and genetic divergence among population pairs is generally interpreted as the outcome of convergent parallel evolution due to the repeated action of natural selection driving speciation (Schluter & Nagel 1995; Johannesson 2001).

However, different scenarios can lead to such patterns of parallel genomic divergence (Johannesson et al. 2010; Bierne et al. 2013). For instance, allopatric divergence followed by secondary contacts in multiple refugia can mimic or contribute to this pattern of parallel divergence but is difficult to distinguish from primary differentiation (Endler 1977, 1982; Barton and Hewitt 1985; Bierne *et al.* 2013). Therefore, understanding the origin of adaptive variation becomes a key issue in studies of parallel evolution, since there is a need to determine if divergence has been fuelled by new mutations, standing variation which arose by mutation or gene flow in the ancestral population, or recent secondary contact (Welch & Jiggins 2014). Reconstructing the demographic history of species divergence may help in identifying the evolutionary scenarios that can generate observed divergence patterns, and therefore has the potential to reveal how much parallel divergence patterns reflect parallel evolutionary histories. Moreover, understanding the demographic history of divergence may be a necessary requisite to the implementation of selection

detection tests, and may thus contribute to make better sense of genome scans (Nielsen *et al.* 2007, 2009; Li *et al.* 2012).

New methods, relying on full Bayesian likelihood approaches (Li & Durbin 2011; Mailund *et al.* 2012), approximating the likelihood through simulations (Tavare *et al.* 1997; Beaumont *et al.* 2002; Beaumont 2010) or computing composite likelihood from the site frequency spectrum (Williamson *et al.* 2005; Gutenkunst *et al.* 2009) have greatly helped testing increasingly complex hypotheses about demographic history. However several challenges remain. One particularly challenging task in reconstructing the history of species divergence is to integrate temporal variations in gene flow intensity as well as the possibility for heterogeneous amounts of gene flow across the genome. Only a few studies have taken this heterogeneity into account (e.g. Roux et al. 2013, 2014; Sousa et al. 2013; Tine et al. 2014) to address the effect of genetic barriers reducing the effective migration rate at linked neutral loci (Barton and Bengtsson 1986; Feder & Nosil 2010). Here, our goal was to address whether parallel phenotypic differentiation between pairs of lamprey ecotypes was accompanied by parallel genetic differentiation, and if these patterns have likely resulted from independent or shared divergence histories.

To address this issue, we used several population pairs of parasitic river lamprey (*Lampetra fluviatilis,* Linnaeus 1758) and non-parasitic brook lamprey (*Lampetra planeri,* Bloch 1784), which exhibit varying level of geographical connectivity (i.e. from complete sympatry to completely disjunct distributions). Lampreys are jawless vertebrates (agnathans) generally occurring as “paired” species (or ecotypes) with drastically divergent life-history strategies at the adult stage, with parasitic anadromous and nonparasitic freshwater resident forms. Relatively little is known about the speciation level among parasitic and freshwater resident ecotypes (Docker 2009). Most genetic studies have found little to moderate levels of genetic differentiation between ecotypes (Schreiber & Engelhorn 1998; Espanhol *et al.* 2007; Blank *et al.* 2008; Pereira *et al.* 2010; Bracken *et al.* 2015), but they usually did not distinguish sympatric and parapatric sites and they did not use enough markers to address the extent of divergence across the genome. Mateus et al. (2013) used RAD-Sequencing and found a strong genome-wide differentiation between ecotypes from a single population pair from the southern limit of the distributional range and provided a list of candidate genes putatively involved in adaptation to migratory *versus* resident life-styles, which could be implicated in reproductive isolation between ecotypes. However, this study provided limited insights into the historical, demographic and selective aspects underlying genetic divergence as it was on a single population pair. More recently, Rougemont et al (2015) studied ten population pairs located in the northern part of the distributional range and varying in their level of geographic connectivity. They showed that within-river opportunities for gene flow have a strong influence on the average level of differentiation. The parapatric pairs displayed stronger genetic differentiation than sympatric pairs, which were less divergent than the southern sympatric pair described in Mateus et al. (2013). They

thus hypothesize that incomplete reproductive isolation and gene flow have allowed stronger genetic introgression in northern compared to southern sympatric pairs, which would have important consequences for choosing the most relevant population pairs to disentangle the effects of selection from genetic drift during divergence. In the populations connected by gene flow, only genetic regions involved in divergence are expected to resist its homogenizing effect, and thus these populations would be the best candidates for the study of reproductive isolation (RI).

In this study, we used a model-based approach to reconstruct the divergence history of *L. fluviatilis* and *L. planeri* using RAD sequencing data. We took advantage of the original distribution of resident and migratory lamprey in both sympatry and parapatry in nine replicated pairs to first document levels of gene-flow between ecotypes and among population pairs. We then used these population pairs, in order to *i*) compare alternative models of demographic divergence history, *ii*) estimate the proportion of the genome experiencing reduced gene flow, *iii*) identify genomic markers showing particularly high differentiation between ecotypes, and *iv*) evaluate the extent of parallelism among replicate pairs of ecotypes.

Methods

Sampling

We focussed on 9 populations pairs of lampreys sampled in France including 4 sympatric and 5 parapatric pairs varying in their level of neutral genetic differentiation and geographic connectivity (Rougemont *et al.* 2015) (Fig 1). Populations were collected in sympatry in the Aa, Oir and Bethune and very close to each other in Bresle, where sampling sites were located only 8 km apart. Pairs from the Risle and Odon River were collected on the same river but were separated by dams, with *Lampetra planeri (Lp*) from the Odon being captured far upstream in sites unlikely to be colonized by *Lampetra fluviatilis (Lf*), whereas *Lp* on the Risle might still be connected to *Lf* through passive drift during flood events. Similarly, pairs from the Loire-Cens, Dordogne-Jalles, and Garonne-Saucats were strongly disconnected, with river lampreys collected in estuarine areas before spawning and brook lampreys collected in isolated upstream reaches not accessible to river lampreys. Details of sampling methods are provided in Rougemont *et al.* (2015). A total of 338 individuals were included in the present analysis (Table 1). Three sea lamprey (*Petromyzon marinus*) individuals were also included as outgroup to polarize SNPs.

Library preparation and Sequencing

Genomic DNA was extracted from each individual using individual extraction kits (NucleoSpin Tissue, Macherey Nagel) following manufacturer's recommended protocols. DNA quality was checked using a spectrophotometer (Nanodrop 2000, Thermo Scientific) and quantity was assessed

using a fluorimeter (Qubit 2.0) and standardized to 22 ng.µL^{-1}. DNA was then used to construct a total of 13 libraries, each composed of 48 randomly distributed individuals including other lamprey populations (not presented in this study) and following the protocol from Baird et al. (2008) using the restriction enzyme Sb*f*1. Samples were individually barcoded and paired-end sequenced on 8 lanes of a Illumina Hiseq 2500 (125-bp paired-end reads) and 5 lanes of a Illumina Hiseq 2000 (100-bp paired-end reads) at Montpellier Genomix and Genotoul sequencing platforms, respectively.

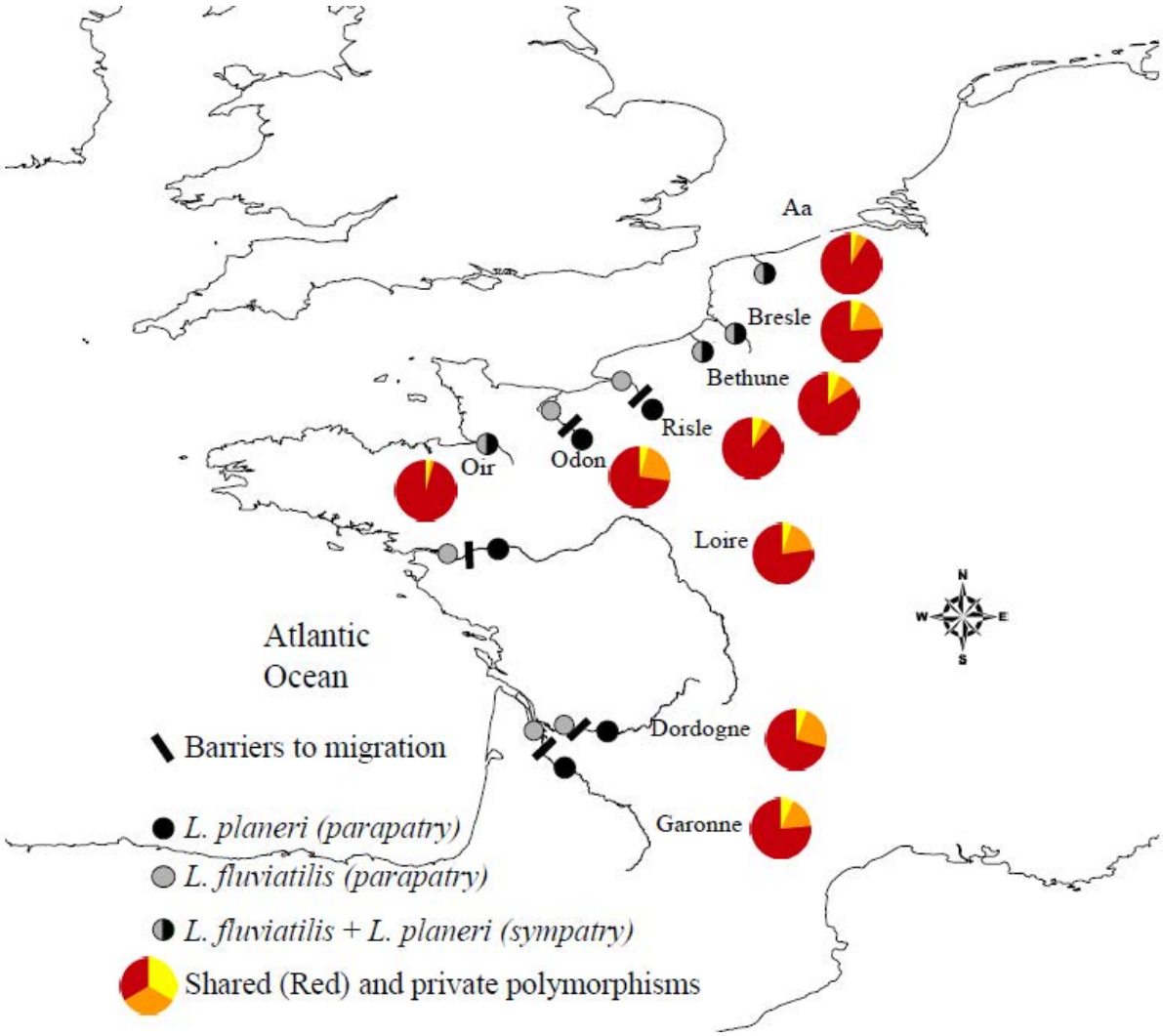


Figure 1: **Map of sampling site across the Atlantic and channel area**. River names match those given in Table 1. Number of shared (red) and private polymorphisms (orange = Lf, yellow = Lp)

Genotyping and Bioinformatics

Raw-reads were first demultiplexed using GBSX (Herten *et al.* 2015), filtered for overall quality, checked for the presence of a barcode using Cutadapt (Martin 2011) and trimmed to 85-bp. We then took advantage of our paired-end sequencing to remove PCR duplicates using Stacks' program clone_filter (Catchen *et al.* 2013). Our data contained an average of 38.5% of PCR duplicates that were removed. We subsequently used the Stacks pipeline to identify RAD loci from forwards reads using all individuals from all populations. We used ustacks with a minimum stack depth of 4 (m = 4) and a maximum of four mismatches (M = 4). These parameters were determined using replicated individuals included at random in each sequencing platform. This allowed us to minimize the presence of paralogs in the dataset while maximizing the genetic diversity obtained. Due to the large divergence with *P. marinus* (see below) and an incompletely assembled genome for this species (Smith *et al.* 2013) we performed a *de novo* alignment using cstacks. Sequences with less than 3 nucleotide differences were considered as homologs, resulting in a catalog containing 28,2610 RAD tags. Each individual was finally genotyped after matching its RAD data against the catalog, and the genotypes were exported in vcftools format (Danecek *et al.* 2011) for further filtering. Different datasets were created: for population genetics analyses we constructed nine pairwise datasets corresponding to each river and composed of the two ecotypes of brook and river lampreys; for demographic analyses we constructed another series of nine datasets with different filtering criteria (see below) and we ultimately created a global dataset to investigate patterns of global population structure. We applied a series of filtering steps as we were particularly concerned about recovering biologically meaningful SNPs. To do so the dataset was first split into nine datasets corresponding to each river. We further split each of them into two datasets corresponding to each ecotype. We kept markers genotyped in at least 80 % of the individuals in each population, with a minimum sequencing depth of 10 and a maximum of 100 (to avoid the inclusion of highly repetitive tags that could reflect paralogy). We then excluded loci deviating from Hardy Weinberg equilibrium assumptions by keeping only those loci with a p-value greater than 0.05. In order to limit the presence of weakly polymorphic loci being uninformative (Roesti *et al.* 2012b), we kept loci with a minor allele frequency (MAF) greater than 0.1 in at least one of the populations separately or with a MAF greater than 0.05 in each global dataset (containing the two ecotypes) for all analyses except for demographic inferences. This filtering step allowed us to recover polymorphism occurring at low frequency in some river lampreys but monomorphic in brook lampreys and *vice versa* in order to obtain an equivalent representation of low frequency polymorphisms present in each population. In a final filtering step, we excluded loci with an observed heterozygosity greater than 0.5 in both freshwater and migratory lamprey in each river (Hohenlohe *et al.* 2011). The nine 'pairwise' datasets obtained were then merged into a global dataset in which only markers genotyped in at least 70 % of the individuals across all populations

were kept to reduce the proportions of missing data. All filters were performed using VCFtools and R scripts embedded in custom bash scripts.

Population genetic differentiation, diversity and individual clustering

The level of genetic differentiation between ecotypes from a given river and over all populations was computed using F_{ST} (Weir & Cockerham 1984) for each locus in VCFtools. Negative values were set to zero to compute mean genome-wide F_{ST}. Significance of F_{ST} values and 95% confidence intervals were computed in R using bootstrap methods. We further investigated the proportions of shared polymorphisms between these pairs. To investigate the level of population clustering we first computed a matrix of Nei's genetic distance Da, (Nei *et al.* 1983) across all populations to create a UPGMA dendrogram describing broad patterns of genetic structure. Secondly we used Admixture (Alexander *et al.* 2009) to perform individual clustering and to estimate individual admixture proportions. Given the strong geographical component of genetic differentiation in *L. planeri* (see results), we were not able to distinguish the two morphs based on this full dataset. However, we were concerned about distinguishing putatively recent hybrids from pure individuals to avoid biases in our demographic inferences. We therefore used the 10% most highly differentiated SNPs that were shared between the four least differentiated pairs (i.e. global $F_{ST} < 0.10$), based on the distribution of between-pairs $F_{ST,}$ to improve our discrimination power between the two ecotypes and identify potential hybrids. This resulted in a set of 40 SNPs that were also shared between the remaining population pairs (see results). We then tested the ability of these markers to assign individuals to their respective ecotypes and to identify putative hybrid individuals (i.e. possible F1 and backcrosses) using Structure2.3.3 (Pritchard *et al.* 2000) and NewHybrids (Anderson & Thompson 2002). Details of Structure and NewHybrids settings are presented in supplementary materials. The robustness of inference from NewHybrids was evaluated using simulated data with HybridLab (Nielsen *et al.* 2006). Briefly, individuals with a q-value greater than 0.9 of being either river lamprey or brook lamprey were used to generate individual genotypes belonging to 4 different hybrid classes (F1; F2 and first generation backckrosses), which were then reclassified in NewHybrids (see supplementary Materials and Results). Since the subset of 40 SNPs offered a high power to identify hybrid genotypes (> 99%), we excluded all individuals identified as hybrids from the real dataset containing the nine populations pairs, and performed admixture analysis using the 40 markers and populations to better discriminate the ecotypes.

Demographic history of divergence

We used a modified version of the diffusion approximation method implemented in *δaδi* (Gutenkunst *et al.* 2009) to analyse the joint site frequency spectrum (JSFS) of river and brook lampreys in each population pair. Different models of demographic divergence were compared for

each pair and the model providing the best fit to the observed JSFS was determined using the Akaike Information Criterion (AIC). The main changes to the original *δaδi* program that were implemented in the modified version by Tine et al. (2014) include (*i*) the addition of two Simulated Annealing optimization phases prior to the BFGS to better explore the likelihood landscape and improve global convergence, (*ii*) the introduction of a parameter that accounts for the proportion of mis-oriented SNPs in the JSFS, and (*iii*) most importantly, the incorporation of varying migration rates across the genome to account for semi-permeability. This is done by defining two categories of loci occurring in proportions P and 1-P across the genome, the first one containing neutral loci that are exchanged between populations with a gross migration rate (m12 and m21), and the second category comprising selected and hitchhiker loci that experience a reduced effective migration rate (me12 and me21).

Seven alternatives models of speciation were fitted to the observed JSFS and compared (Fig 2), including the model of Strict Isolation (SI), three different models incorporating homogeneous migration along the genome (IM, AM and SC), and three models incorporating heterogeneous migration along the genome (IM2m, AM2m, SC2m). Gene flow was supposed to be either ongoing during the whole divergence history (Isolation-with-Migration models IM and IM2m), only present at the beginning of divergence (Ancient Migration models AM and AM2m), or starting after a period of complete isolation (Secondary Contact models SC and SC2m).

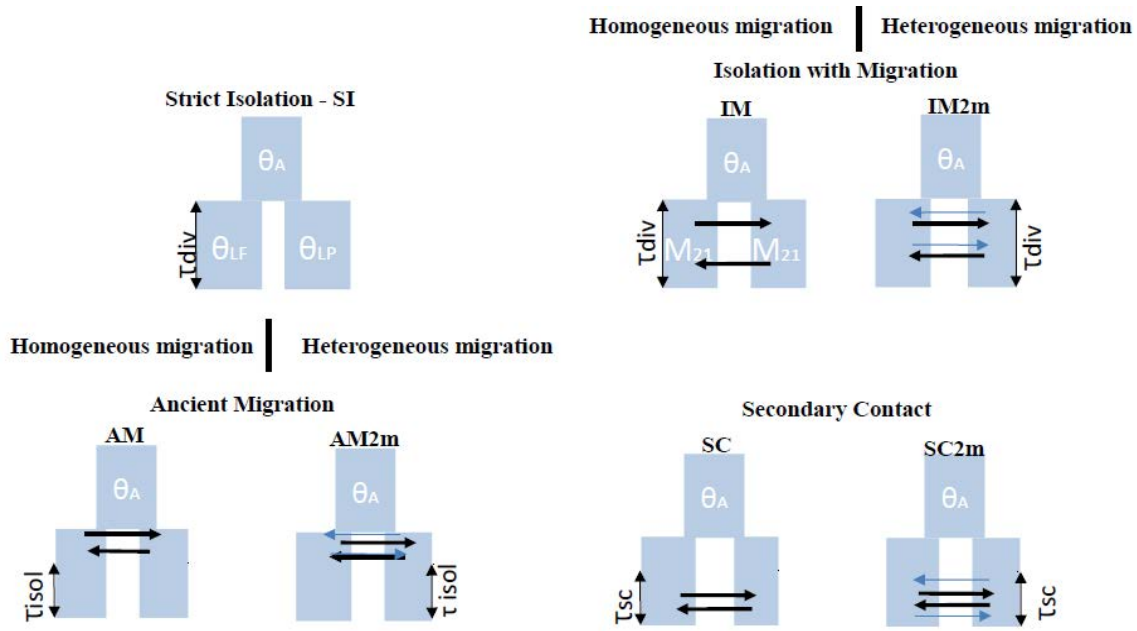


Figure 2: **Representation of the 7 demographic scenarios compared** (A) Seven models with different parameters are tested and compared. Strict Isolation (SI), isolation with constant migration (IM), ancient migration (AM) and secondary contact (SC). The following parameters are shared by all models: τ_{div}: number of generation since divergence time. θ_A, θ_{Lf} , θ_{Lp} : effective population size of the ancestral population, of *L. fluviatilis* and *L. planeri* respectively (in units of $4N_{ref}\mu$). τ_{isol} is the number of generations since the two ecotypes have stopped exchanging genes (in units of 2 N_{ref} generations). τ_{sc} is the number of generations since the two morphs have entered into a secondary contact after a

period of isolation. M_{12} and M_{21} represent the effective migration rates expressed in $2.N_{ref}m$ units per generation with *m* the proportion of population made of migrants from the other populations. In addition to these four models, three additional models incorporating heterogeneity in divergence along the genome were tested: isolation with heterogeneous migration (IM2m), ancient migration with heterogeneous effective migration along the genome (AM2m) and secondary contact with heterogeneous migration (SC2m)

The observed JSFS was built using a SNP dataset with no minor allele frequency threshold filter. We randomly sampled a single SNP per RAD tag to avoid as much as possible including linked SNPs in the spectrum, which produced JSFS composed of 5000 to 10000 SNPs depending on the analysed pair. We excluded the population from the Dordogne River as it contained only 7 individuals suitable for this analysis (one individual had a too low genotyping rate). A total of 25 independent runs were performed for each model to check for convergence. The run providing the lowest AIC for each model was kept for comparisons among models and parameter estimations. Demographic parameters were all scaled by theta in the ancestral population and were not converted into biological estimates as relevant estimates of mutation rate are not available in lampreys. In addition, estimating the total number of nucleotides that effectively contribute to the JSFS construction remains complicated when no reference genome is available. However, some parameter ratios which do not depend on theta or on the mutation rate can be used for interpretation, even though their absolute values must be interpreted very carefully. We used the folded JSFS for model choice because using the sea lampey as the outgroup resulted in a too low number of polymorphisms available. Nevertheless, we validated diffusion accuracy for estimates of migration rate and proportion of neutral loci using the unfolded JSFS. In these cases, the previously estimated time of separation, time of secondary contact or time of migration stop and effective population size were fixed based on estimates from the folded JSFS.

Detecting selection and measuring parallel differentiation

Our demographic investigations confirmed that some populations exchanged genes while others experienced almost no gene flow (see results). Consequently populations experiencing gene flow (identified in δaδi) were more appropriate to investigate the basis of reproductive isolation than populations with low gene flow as only regions involved in reproductive isolation should resist the homogenising effect of gene flow. In addition, this complex history challenges the use of classical model-based outlier detection tests which may result in high rates of false positives in such cases (Lotterhos & Whitlock 2014, 2015). To circumvent these problems, we took advantage of our previous analyses where model parameters were inferred from the best model identified, in order to simulate datasets using neutral gene flow estimates and subsequently compute the neutral envelop of F_{ST} under this model. Markers with F_{ST} values greater than this envelop were considered as putatively under direct or indirect selection. We used ms (Hudson 2002) and msstatsFst (Eckert & al.

2010) embedded within custom scripts to perform such analyses and compute Weir et Cockerham's F_{ST}. We generated 200 000 SNPs with the same number of individuals as those observed in our real dataset in each of our simulated datasets.

In order to test for parallelism in divergence we first counted the number of shared outliers (determined from our neutral model) between connected pairs (i.e. 10 comparisons) and performed randomizations tests to verify that this number was not due to chance alone. We subsequently constructed coplots of F_{ST}/F_{ST} between pairs of rivers to better visualize the extent of outlier sharing and parallelism. We then fitted linear models and compared their slope to investigate patterns and directionality of introgression. A slope of 1 indicates a similar level of differentiation while a slope significantly different from 1 indicates asymmetric introgression into the compared pairs. To validate our hypothesis that the most disconnected populations were less suited to investigate patterns of parallelism, we also performed coplots of F_{ST}/F_{ST} between pairs of disconnected rivers based on outliers identified from our neutral model. We then compared the distribution of correlation coefficients between connected pairs and disconnected populations pairs using simple boxplots. To further gain insight into allele frequency shifts we determined the frequency of the derived allele (using the *P. marinus* as the outgroup) when this information was available. We used a stringent approach in which the ancestral state of the allele was inferred only when fixed in *P. marinus.* We then tested for asymmetries in allelic frequency between the two morphs using t-tests. All analyses were performed using custom R scripts.

Results

Genome-wide diversity

1,054,852,013 reads were obtained after demultiplexing, representing an average of 2,747,010 reads per individual. A total of 12 individuals were excluded due to their low read numbers. 326 individuals were kept for subsequent analyses. After appropriate filtering, a total of 8,962 SNPs were kept for the analyses done over all the populations. The number of SNPs kept for pairwise population analyses ranged from 14,201 to 17,335 (table 1, table S1, Fig 4). Average expected heterozygosity was 0.296 in *Lf* and 0.264 in *Lp* and there were significant differences between ecotypes for all parapatric pairs (t-test, $p = 2.10^{-16}$) except the Risle (t-test, $p > 0.01$). There were significant differences in two sympatric pairs (the Aa and Bresle river (t-test, $p = 2.10^{-16}$). Overall *Lf* displayed 2.7 times more unique polymorphic sites than did *Lp*. As for heterozygosity, this pattern hides a more complex situation in which sympatric populations of migratory lampreys contained 1.8 times more private polymorphic loci than resident lamprey whereas parapatric migratory populations contained 3.2 times more polymorphic loci than isolated resident populations.

Based on their genetic characteristics, the Risle and Bresle River were two “outliers” with respect to the geographical context. The number of unique polymorphisms in migratory and resident lamprey was indeed similar on the Risle River (949 vs 901) whereas it was strongly asymmetric between the sympatric populations from the Bresle river (2758 vs 889 polymorphic only in migratory and resident populations respectively). In agreement with these results, histograms of minor allelic frequencies showed that most isolated brook lamprey population displayed an L-shaped distribution whereas allelic frequencies were more evenly distributed in connected *Lp* and in *Lf* populations (Fig S1).

1**Table 1: Pattern of genome-wide differentiation across the nine population pairs.** Situation indicates the location of the 2 populations in the watershed 2(S=sympatry, P=parapatry) see materials and methods for more information. Number of lampetra fluviatilis and Number of Lampetra planeri (N *Lf* / N *Lp*) 3are provided as well as the estimates of genome-wide differentiation from microsatellite markers. Expected Heterozygosity (Exp Het) in each populations 4are indicated as well as genome-wide F_{ST}, maximum F_{ST} value, number of F_{ST} reaching the maximum value of 1, the 90th quantile and the number of outliers 5found using either de 97% quantile method or the neutral model.

												Outliers
River	**Situation**	**N *Lf* / N *Lp***	**Fst microsat**	**N SNPs**	**N hybrids**	**Obs Het *Lf***	**Obs Het *Lp***	**Fst SNPs**	**Fst max**	**Fst = 1**	**90%Fst quantile**	**Neutral model**
OIR	S	28 / 25	0.030	16557	7	0.294	0.296	0.042	0.823	0	0.117	324
BET	S	15 /13	0.028	14199	0	0.297	0.293	0.053	1	21	0.168	643
RIS	P	22 / 16	0.040	16672	6	0.294	0.296	0.065	1	2	0.166	958
AA	S	24 /27	0.080	15405	9	0.292	0.280	0.076	0.821	0	0.169	883
BRE	S	21 /17	0.076	15511	0	0.300	0.257	0.143	1	24	0.315	2150
DOR	P	21 / 8	0.190	17330	0	0.293	0.237	0.150	1	23	0.312	NA*
GAR	P	21 / 15	0.087	16926	0	0.293	0.261	0.157	1	24	0.354	2483
LOI	P	21 /19	0.150	16407	1	0.300	0.254	0.153	1	5	0.327	2535
ODO	P	21 / 22	0.190	15908	0	0.305	0.247	0.207	1	29	0.404	3317

6* The neutral expectations for the DOR population pairs cannot be computed. For comparisons purposes, we used a quantile of Fst that provided similar 7number of outlier (i.e. quantile of Fst greater than 0.86, providing 2428 “outliers”

Population structure and introgression

Patterns of population structure inferred from an Admixture analysis, and neighbor-joining tree (Fig 3) revealed a strong geographical effect on the genome wide level of differentiation. Admixture analysis for K = 2 (Fig S2) did not allow differentiation of the two ecotypes and increasing values of K confirmed the clustering by geographical origin of the isolated brook lamprey populations rather than by ecotypes. This geographic trend was driven by the strong genetic structure of brook lamprey populations whereas migratory lampreys did not appear geographically structured. The optimal clustering K value determined using Admixture cross-validation procedure (K=7) allowed discrimination of each brook lamprey population except the sympatric populations from the Oir, Bethune and Risle River. Ultimately, for K = 10, all *Lp* populations but the Bethune River clustered separately and we found two clusters of river lampreys between which gene flow was reduced. The two clusters corresponded to the Channel area and the Atlantic area.

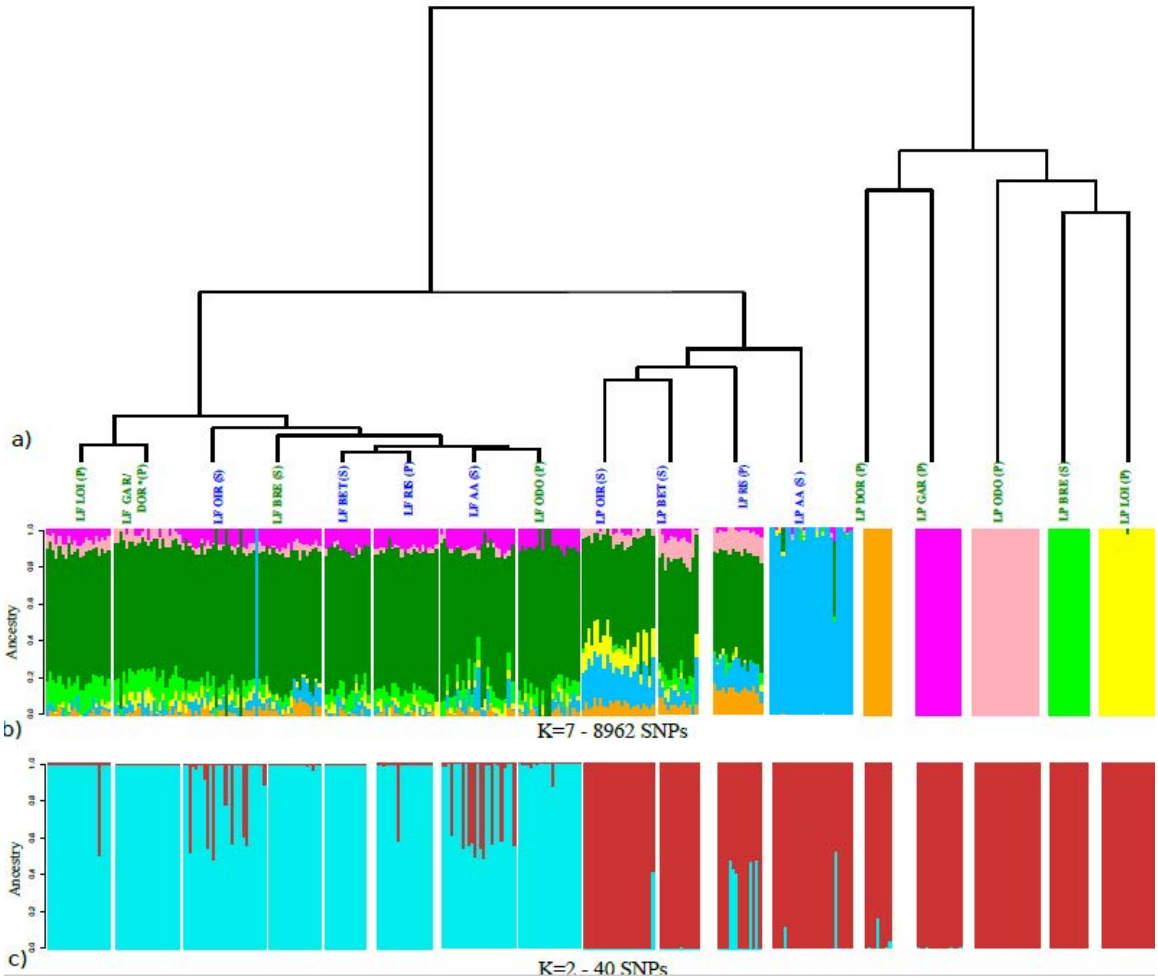


Figure 3: **Inference of population structure:** a) Hierarchical clustering and b) Admixture analysis performed on all individuals using the 8962 SNPs. C) Structure analysis on a subset of 40 highly discriminative SNPs. Blue = Connected population pairs, Green = Disconnected population pairs. P = Parapatric population, S = Sympatric population

The subset of 40 highly differentiated markers revealed a F_{ST} of 0.7 between *Lp* and *Lf*. Structure analysis indicated that this marker set was able to accurately discriminate the two morphs (mean q-value from Structure was 0.97 and 0.95, in river and brook lamprey respectively). Structure analysis also identified a total of 22 individuals (5 females, 9 males, 7 with undetermined sexes, Table S3) with mixed q-values resembling F1 and one individual of possible backcross origin. NewHybrids confirmed all Structure results and identified 20 F1 hybrids, 1 F2 and 1 backcross. The majority of individuals (64%) displayed a river lamprey phenotype and the 36% remaining a brook lamprey phenotype (Table S3). Simulated parents, F1 and backcrosses with *Lp* individuals using HybridLab were always correctly assigned into their respective categories at the q-value threshold greater than 0.9, yielding a detection power of 100% for these three hybrid classes (table S4). One individual backcrossed with *Lf* and 7 F2 were correctly classified as of hybrid origin but with a q-value threshold below 0.9, hence yielding a power of 99.6% and 97.6% respectively.

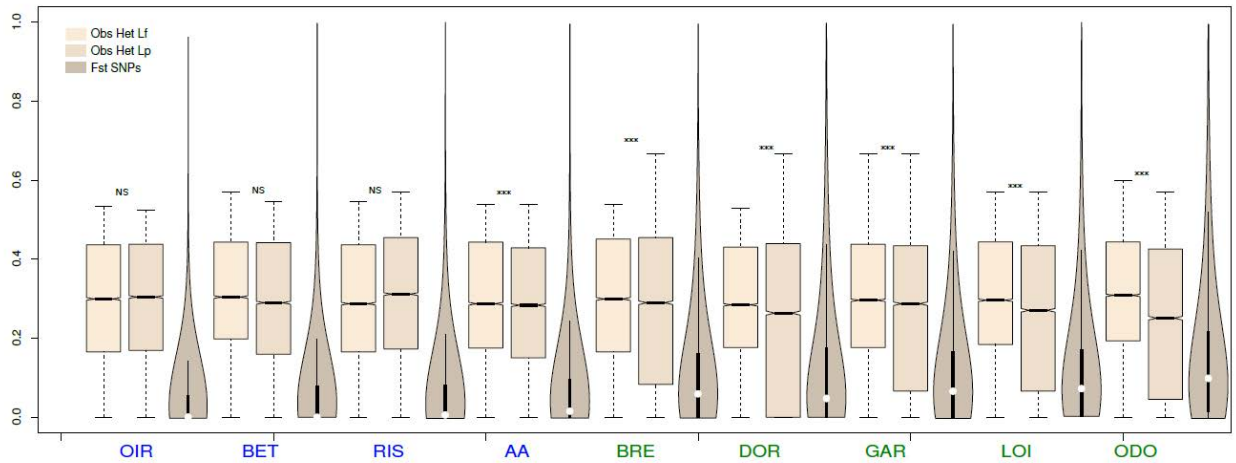


Figure 4: **Boxplot of observed Heterozygosity in each pair of river and brook lamprey and violin plot of F_{ST} between pairs.** Populations are sorted by increasing order of differentiation. Significance of difference in heterozygosity between river and brook lampreys (as measured by t-tests) are depicted by *** when significant with a *p-value* < 0.0001 or NS when non-significant. Blue = Connected population pairs, Green = Disconnected population pairs.

Pattern of genome-wide differentiation

After exclusion of hybrids, the global genome-wide differentiation between the two morphs was 0.128 (95%CI =0,124-0,131, P <0.00001) but hides very different levels of genetic differentiation across pairs as shown table 1. Most parapatric rivers were more genetically differentiated than sympatric pairs (Table 1). In sympatry, genome wide F_{ST} between the two morphs ranged from 0.045 in the Oir River to 0.135 in the Bresle River. F_{ST} in parapatry ranged from 0.06 in the Risle River to

0.20 in the Odon River. Global F_{ST} between river lamprey populations was 0.01 (95%CI=0.005-0.015, P <0.0001, ranging from F_{ST} = 0 to 0.024). In contrast, brook lamprey populations displayed a stronger genetic structure (F_{ST} = 0.167, 95%CI=0.162-0.173, P <0.0001) ranging from F_{ST} = 0.062 to 0.262. The genome-wide variance of F_{ST} increased with geographical isolation and displayed increasing values of the right tail of an L-shaped F_{ST} frequency distribution (Fig 4). Accordingly, the number of differentially fixed loci increased with genome-wide divergence between ecotypes from 0 in the Oir and Aa rivers to more than 20 on the Bresle, Dordogne and Odon. 21 fixed loci were found on the Bethune, despite low genome-wide divergence. Similarly the 90% F_{ST} quantile increased with increasing patterns of genome wide divergence and of increasing geographical isolation in all cases but the Bresle, which was moderately differentiated despite the close geographical proximity of river and brook lamprey samples.

Demographic history

Demographic inference results are provided detailed in Table 2, Fig 5 and Fig S3. The Strict Isolation model (SI) was weakly supported as it never captured asymmetries in the observed JSFSs and failed to recover the most differentiated SNPs. In comparison, the three models with homogeneous gene flow (AM, IM, SC) provided better fit to the data with a good prediction of asymmetric gene flow. However, in every case, incorporating heterogeneous gene-flow (AM2m, IM2m, SC2m) largely improved the fit (Table 2). In connected populations, under the AM (and AM2m) model, the timing of ancient migration stop (τ_{am}) reached a value of zero meaning that this model converged to the simple IM (and IM2m) model. Overall we repeatedly found that the secondary contact model incorporating heterogeneous migration rates (SC2m) provided the best fit to the observed JSFS in the four population for which F_{ST} was <0.10. In the sympatric population with a F_{ST} >0.10 (Bresle) the most likely scenario was IM2m but both the SC2m and AM2m produced very similar AIC. The 3 parapatric populations converged unambiguously to the AM2m. In each case the effective population size was larger in the migratory species than in the resident species as indicated by the ratio of (Ne *Lf*)/(Ne *Lp*) in which effective population size under the SC2m model was on average 8 times greater in *Lf* and 19 times greater in *Lf* under the AM2m model. Using unfolded JSFS allowed estimating gene flow parameters (*m, me,* P) with great accuracy. We observed asymmetries in gene flow with the ratio of *m21/m12* indicating a stronger migration from river lampreys to brook lampreys in all rivers except the Aa under the SC2m. When taking into account effective population size, ratios of Ne*m* indicated similar levels of migration in the Aa river and greater level of migration from river lampreys to brook lampreys in the remaining cases. Under the AM2m model, the inverse trend was observed with higher levels of migration from brook lampreys to river lampreys except in the Odon River where migration was not correctly estimated in brook lampreys. Ratio of τ_{sc}/τ_{split} indicated variable secondary contact times ranging from 7% of the total divergence time to 22% of the total divergence time. The proportion of loci freely exchanged ranged from 0.55 (Aa) to 0.68

(Risle) indicating that from 32 to 45% of loci displayed a reduced effective migration rate between the two morphs. In disconnected populations ratios of τ_{am}/τ_{split} suggest a very recent stop of gene-flow.

Table 2: Comparison of 7 different models obtained from *δaδi* between each pair using the folded JSFS and estimation of their demographic parameters.

The best models are in bold. An additional line (in italic) provides the results of parameter estimation using unfolded JSFS for each best model. The time of split τ_s, contact τ_{sc}, migration stop τ_{am} and effective population size (Nu1, Nu2) were fixed based on values inferred from the folded JSFS. Lines highlighted in grey represent models for which ΔAIC is < 10.
AIC = Akaike Information Criterion, MLE = Maximum Likelihood, Theta = 4 *Nrefμ,* effective mutation rate of the reference population, which here corresponds to the ancestral population. Nu *Lf* and Nu *Lp*: effective population size of *L. fluviatilis* and *L. planeri*. m12 and m21: migration from *L. planeri* to *L. fluviatilis* and from *L. fluviatilis* to *L. planeri*, respectively. me12 and me21: effective migration rate estimated in the most differentiated regions of the genome from *L. planeri* to *L. fluviatilis* and from *L. fluviatilis* to *L. planeri*, respectively. τ_s = Time of split of the ancestral population in the two daughter species. τ_{sc} and τ_{am}: Time of initiation of the secondary contact (SC and SC2m) and time of period with ancestral migration. P = proportion of the genome freely exchanged (1-P provides the proportion of the genome non-neutrally exchanged).

For each best model, *m12, m21, me12, me21* and *P* were estimated using the unfolded JSFS.

River	Model	AIC	log lik	theta	Nu Lf	Nu Lp	m12	m21	me12	me21	Ts	Tsc/am	P	O
OIR	SI	2217.3	-1105.7	2475.6	19.00	0.44					0.03			
OIR	AM	1697.5	-842.8	4532.6	18.42	0.39	19.34	4.84			3.80	3.80		
OIR	IM	1740.7	-865.3	1278.5	12.36	0.56	6.72	8.30			3.15			
OIR	SC	1733.6	-860.8	1189.3	12.80	0.53	6.74	8.66			1.22	1.17		
OIR	AM2m	1706.4	-844.2	381.7	7.82	2.06	0.09	37.77	1.32	2.17	9.95	9.95	0.27	
OIR	IM2m	1688.6	-836.3	697.0	7.42	1.60	10.95	3.36	0.86	2.72	8.23		0.70	
OIR	**SC2m**	**1663.5**	**-822.7**	**2125.5**	**1.97**	**0.55**	41.29	8.85	3.48	7.44	**0.49**	0.14	0.67	
OIR	**SC2m**	***789.7***	***-384.8***	***421.6***	***1.95***	***0.55***	***8.67***	***7.17***	***0.29***	***0.39***	***0.55***	***0.13***	***0.94***	***0.98***
BET	SI	1705.3	-849.7	1575.4	10.84	0.83					0.08			
BET	AM	1237.2	-612.6	863.3	19.16	0.98	3.63	2.25			2.31	2.31		
BET	IM	1326.9	-658.4	2027.9	8.69	0.43	8.86	3.22			3.80			
BET	SC	1199.5	-593.7	191.7	9.70	4.34	1.24	0.91			8.82	0.84		
BET	AM2m	1051.5	-516.7	668.2	19.38	0.58	0.63	39.03	4.99	2.71	9.56	9.54	0.14	
BET	IM2m	1067.2	-525.6	994.1	2.75	0.24	0.00	26.51	0.03	0.90	0.93		0.95	
BET	**SC2m**	**1028.4**	**-505.2**	**653.9**	**8.23**	**0.54**	16.33	0.18	0.00	9.74	**5.12**	**0.38**	0.59	
BET	**SC2m**	***563.3***	***-271.6***	***77.4***	***8.15***	***0.53***	***3.89***	***9.86***	***0.26***	***0.37***	***5.07***	***0.37***	***0.90***	*0.98*
RIS	SI	1478.4	-736.2	2098.4	19.96	0.50					0.05			
RIS	AM	1213.1	-600.5	1009.5	19.69	1.14	4.05	2.29			3.05	3.05		
RIS	IM	1162.2	-576.1	1965.8	9.98	0.36	8.43	7.72			0.91			
RIS	SC	1200.1	-594.1	1630.8	1.69	0.25	3.10	19.67			0.42	0.07		
RIS	AM2m	1038.3	-510.2	591.1	4.79	0.36	0.00	16.39	0.09	1.36	4.83	4.83	0.95	
RIS	IM2m	1033.3	-508.7	1068.8	2.70	0.19	0.00	29.22	0.11	2.59	1.98		0.95	
RIS	**SC2m**	**992.9**	**-487.4**	**2311.4**	**2.80**	**0.46**	59.88	0.00	1.75	9.42	**0.79**	**0.10**	0.68	
RIS	**SC2m**	***670.95***	***-325.47***	***351.81***	***2.77***	***0.45***	***5.71***	***7.01***	***0.76***	***0.29***	***0.78***	***0.10***	***0.94***	***0.98***
AA	SI	2215.7	-1104.9	1624.8	5.33	0.58					0.07			
AA	AM	1642.7	-815.4	2679.0	16.67	0.36	8.59	2.46			7.38	7.38		
AA	IM	1756.8	-873.4	595.3	4.14	0.62	0.93	4.68			2.59			
AA	SC	1677.0	-832.5	1016.0	2.20	0.86	5.22	2.40			0.78	0.20		
AA	AM2m	1501.5	-741.8	575.6	4.60	0.40	0.36	9.94	0.17	0.51	2.69	2.69	0.95	
AA	IM2m	1489.6	-736.8	1177.4	12.94	0.55	11.17	0.82	0.00	4.44	6.33		0.87	
AA	**SC2m**	**1475.1**	**-728.6**	**658.0**	**5.72**	**1.09**	0.59	3.16	9.72	0.00	**3.23**	**0.44**	0.45	
AA	**SC2m**	***528.2***	***-254.1***	***102.4***	***5.66***	***1.08***	***3.00***	***7.87***	***0.58***	***0.31***	***3.19***	***0.44***	***0.88***	***0.98***

(a) Connected Pairs

River	Model	AIC	log lik	theta	Nu lf	nuLp	m12	m21	me12	me21	Ts	Tsc/am	P	O
BRE	SI	1699.8	-846.9	1820.8	4.66	0.16					0.05			
BRE	AM	1465.6	-726.8	480.9	4.87	0.23	0.11	6.57			5.55	0.01		
BRE	IM	1475.5	-732.7	319.4	7.19	0.39	0.12	3.40			8.95			
BRE	SC	1475.3	-731.6	1364.1	1.66	0.08	0.60	16.96			0.66	0.18		
BRE	AM2m	1394.6	-688.3	1312.7	1.76	0.11	0.88	14.56	0.03	1.67	0.95	0.00	0.95	
BRE	**IM2m**	**1393.9**	**-688.9**	**418.4**	**5.30**	**0.39**	0.02	0.50	0.36	3.84	**6.53**		0.05	
BRE	**IM2m**	***635.3***	***-308.7***	***219.5***	***5.25***	***0.39***	***7.28***	***7.10***	***0.70***	***0.55***	***6.47***		***0.94***	*0.98*
BRE	SC2m	1395.6	-688.8	1359.8	1.65	0.12	1.22	12.29	0.00	1.46	0.00	0.86	0.95	
SAU	SI	2129.8	-1061.9	2586.6	1.73	0.16					0.04			
SAU	AM	1946.4	-967.2	656.3	4.73	0.37	0.25	5.34			6.10	0.03		
SAU	IM	2098.0	-1044.0	2560.0	1.52	0.17	0.80	3.65			0.07			
SAU	SC	2099.9	-1044.0	2560.6	1.55	0.16	0.72	3.63			0.00	0.06		
SAU	**AM2m**	**1863.6**	**-922.8**	**1621.2**	**1.91**	**0.15**	0.93	15.28	0.15	2.10	**1.52**	0.01	0.95	
SAU	**AM2m**	***1601.3***	***-790.7***	***474.0***	***1.89***	***0.15***	***3.61***	***5.18***	***0.46***	***0.71***	***1.51***	***0.01***	***0.83***	***0.98***
SAU	IM2m	1975.8	-979.9	1287.3	2.29	0.28	0.17	0.49	1.00	4.98	2.33		0.05	
SAU	SC2m	1973.5	-977.7	2041.3	1.50	0.16	1.38	8.78	0.17	0.89	0.00	0.69	0.95	
CEN	SI	2228.3	-1111.2	2675.7	2.05	0.10					0.03			
CEN	AM	2015.7	-1001.9	484.9	7.01	0.25	0.12	7.09			8.69	0.02		
CEN	IM	2176.0	-1083.0	531.0	6.18	0.36	0.33	3.09			7.90			
CEN	SC	2178.1	-1083.0	1562.5	2.11	0.12	0.96	9.20			0.06	1.64		
CEN	**AM2m**	**1889.3**	**-935.7**	**461.6**	**7.41**	**0.42**	0.01	0.53	0.62	4.99	**9.34**	**0.04**	0.05	
CEN	**AM2m**	***513.7***	***-246.9***	***320.0***	***7.33***	***0.41***	***6.10***	***4.55***	***0.62***	***0.69***	***9.24***	***0.04***	***0.85***	***0.98***
CEN	IM2m	2000.0	-992.0	2206.9	1.51	0.13	2.25	9.51	0.01	0.80	0.58		0.95	
CEN	SC2m	2003.8	-992.9	2100.0	1.68	0.11	2.17	12.01	0.00	1.55	0.07	0.64	0.95	
ODO	SI	3095.1	-1544.5	2855.6	1.27	0.06					0.02			
ODO	AM	2724.1	-1356.1	1109.8	3.28	0.08	0.12	14.89			3.23	0.01		
ODO	IM	2916.3	-1453.1	619.1	5.77	0.17	0.15	4.92			6.84			
ODO	SC	2893.6	-1440.8	527.6	5.75	0.15	0.19	5.50			8.66	0.59		
ODO	**AM2m**	**2648.1**	**-1315.1**	**538.3**	**7.75**	**0.27**	22.30	0.00	0.25	4.38	**9.41**	0.02	0.15	
ODO	**AM2m**	***1129.2***	***-554.6***	***232.7***	***7.67***	***0.26***	***2.93***	***5.90***	***0.58***	***0.64***	***9.31***	***0.02***	***0.89***	*0.98*
ODO	IM2m	2791.8	-1387.9	2105.0	1.75	0.05	0.84	17.43	0.08	1.53	0.79		0.95	
ODO	SC2m	2781.2	-1381.6	2263.8	1.56	0.05	1.34	19.70	0.00	1.87	0.43	0.11	0.95	

(b)Disconnected pairs

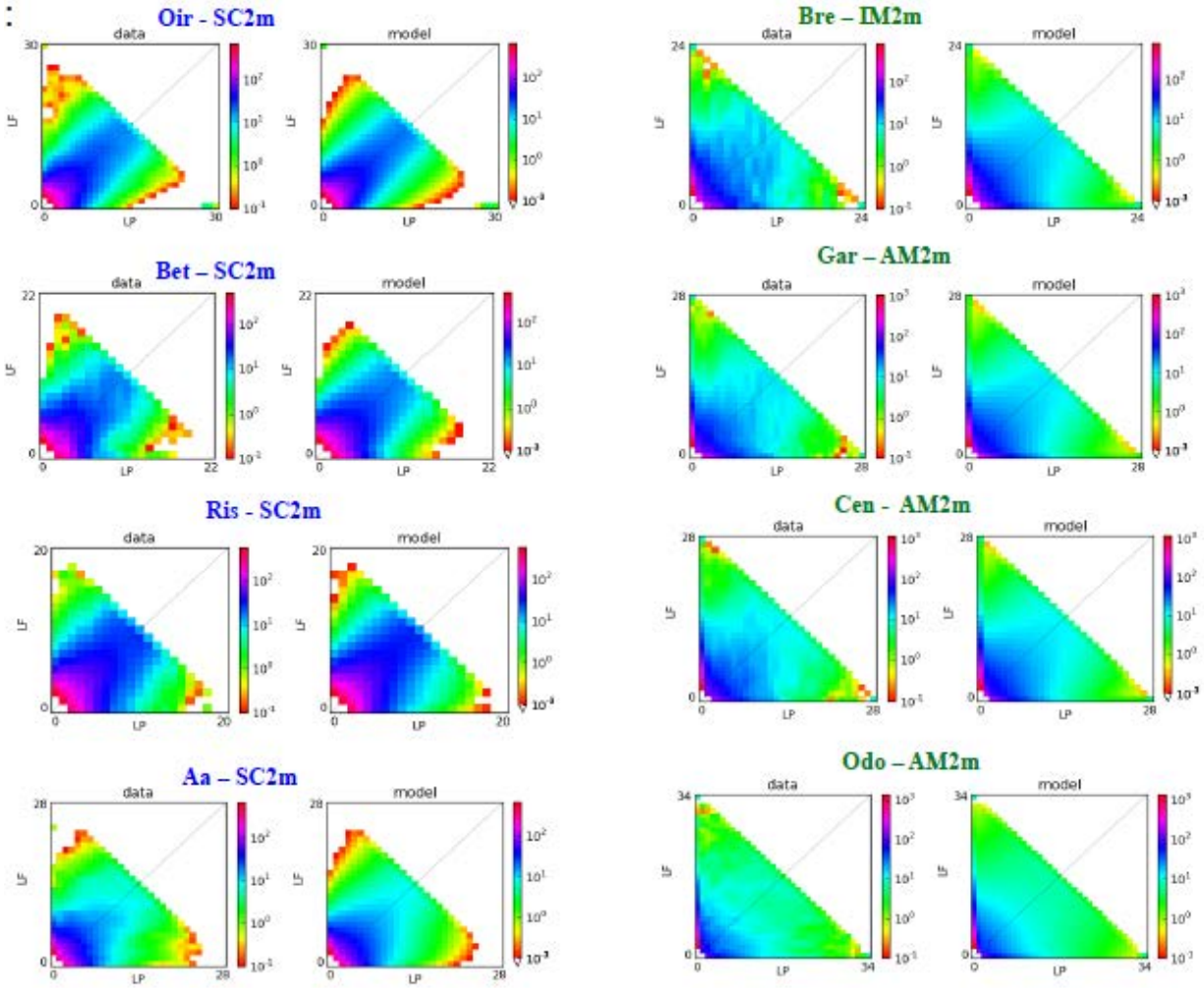


***Figure 5*: Results of the diffusion approximation models for eight pairs of populations**. The observed data and the best fitting models are displayed. Plots of all other models are provided in supplementary results together with their residuals. The Dordogne river was excluded at this stage due to the low number of individuals available (n=7). Blue = Connected population pairs, Green = Disconnected population pairs.

Inference of outliers and extent of parallel differentiation

Gene flow was always strong in connected populations, as opposed to completely disconnected populations, making the former good candidates to study the genetic architecture of reproductive isolation. In line with this result, we consistently observed a larger proportion of outlier SNPs in parapatry (Table 1). Five populations were used for these inferences (the Oir, Bet, Ris, Aa and Bre) (Table 1). Outliers SNPs represented from 1.96 to 5.75 % of the total dataset in each connected river resulting in an average of 4.55 % of SNPs shared between population pairs. The Bresle was again an "outlier" with a high number of SNPs departing from neutrality in this population (13.9%). Conversely, outlier SNPs represented from 14.7 to 20.9% of the dataset in disconnected pairs (Table 1).

We investigated the proportion of outlier loci that were shared across the five connected population pairs. In the ten F_{ST}/F_{ST} pairwise comparisons, the majority of putatively 'neutral' loci were shared across the pairs (mean=10,667). Simulations under the neutral model yielded a total of 28 outliers (24 independent) shared across the ten comparisons and an average of 100 SNPs (46 to 172) shared between pairs. In all cases these loci displayed high correlation in F_{ST} values (r range= 0.8-0.89, *p-value* <0.0001, see Fig 6). This amount of sharing was higher than expected by chance alone (1,000 permutations, all *p-values* < 0.0001). Inference from our neutral models in disconnected populations revealed a much higher number of putative outliers (2150 to 3317, Table 1, Fig 6). In addition, these populations displayed lower values of the correlation coefficients obtained from regression analysis than the connected populations (Fig 7) demonstrating that the level of parallelism was obscured by other processes than selection in parapatry. Using data from *P. marinus* allowed us to identify the derived allele (DAF) in some outliers for the connected pairs. The DAF was significantly greater in brook lampreys than in river lampreys (Fig 8, paired t-test, all *p-values* <0.0001).

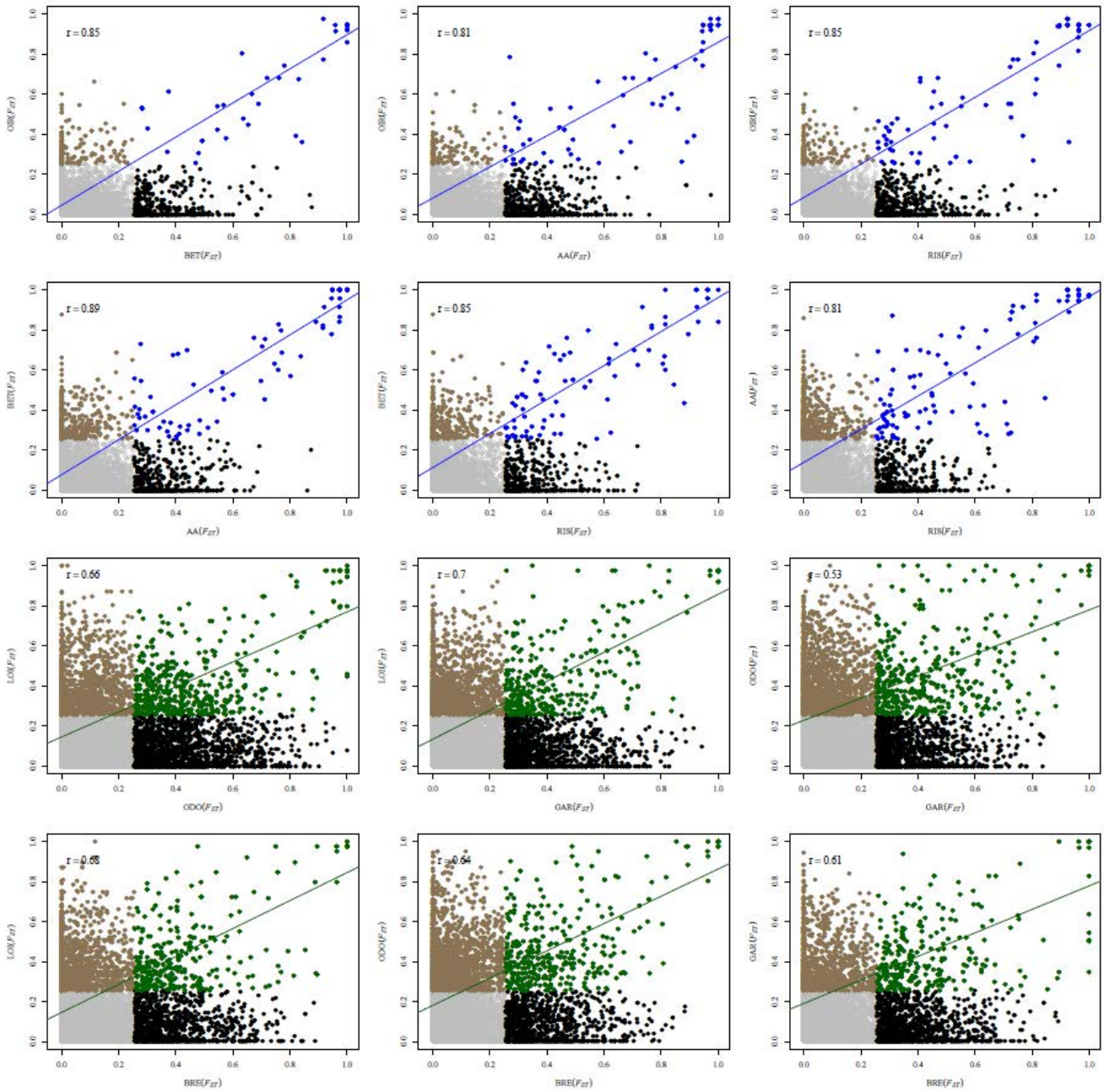


Figure 6 (above): **Coplots of parallelism in genome-wide divergence** based on the neutral model estimated from *δaδi* for connected and disconnected population pairs. F_{ST} values of putatively neutral markers are displayed with grey open circles. Black and maroon circles represent markers outside the neutral model not shared between population pairs and putatively under selection. Blue (connected population pairs) and green (disconnected population pairs) circles denote shared markers falling outside the neutral envelope (regression lines are provided for illustration).

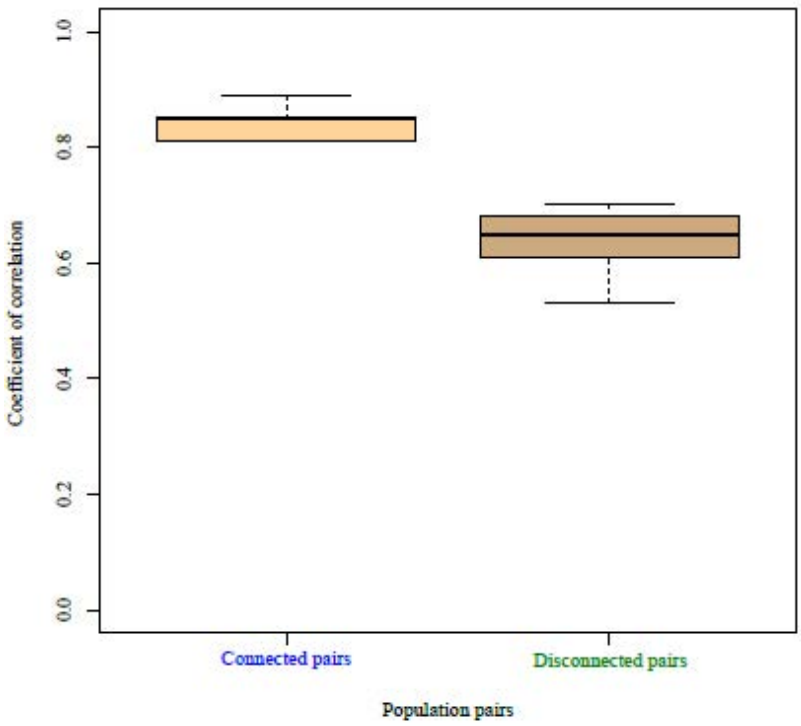
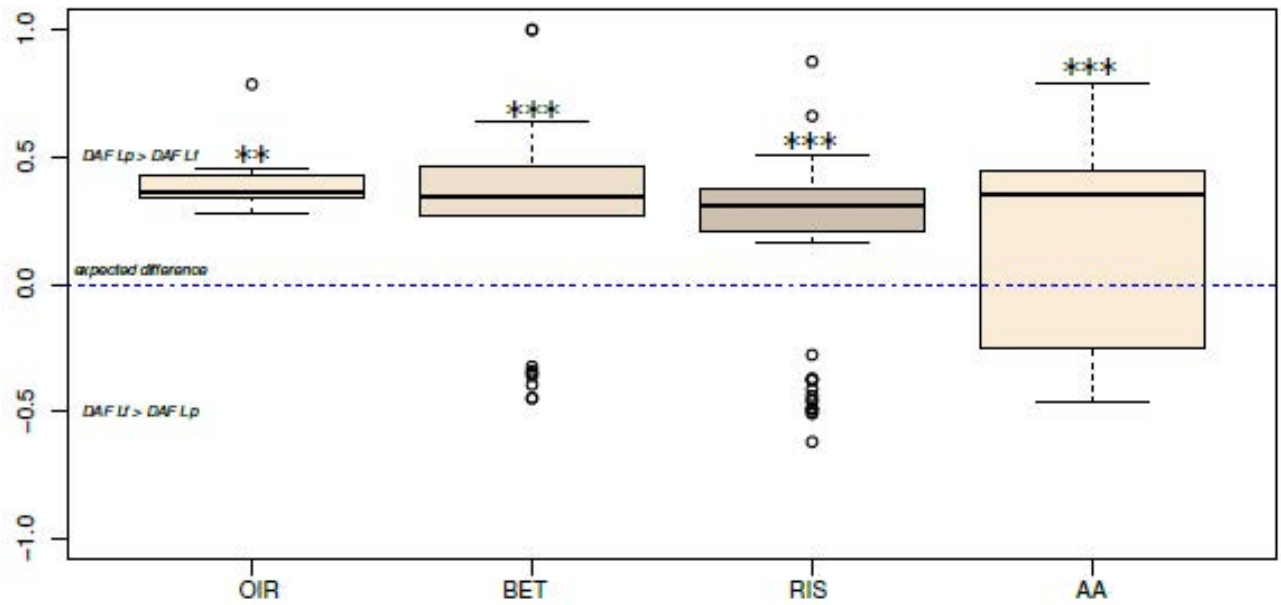


Figure 7: **Comparison of the distribution of the coefficient of correlation (r^2)** obtained from the linear regression of shared outlier F_{ST} in connected population (left) and disconnected populations (right).

Figure 8: **Derived Allele Frequency (DAF) distribution in *Lf* and *Lp* determined with the *P. marinus* outgroup**. Only markers previously detected as outliers base on our neutral model were kept for calculation.



Blast analysis

Blasts analyses (Altschul et al. 1997) of a set of 34 independent sequences (see supplementary methods) of parallel outliers resulted in the identification of 2 SNPs and their sequences (85pb) with significant hits (e-value < 10e-7, sequence identity greater than 93%) (Table S4). None matched to the sea lamprey contigs but we found significant matches to the Arctic lamprey *Lethentheron camtshaticum* draft genome (see Table S4). Among them, one gene known to play a key role in osmoregulation was found. This gene was also found to play a key role in the expression of GnRH, a key developmental hormone. In the second sequence, we also found key genes involved in the immune system, an axial patterning gene, a pineal gland specific opsin gene and a sodium channel gene that was also known to play a role in osmoregulation. Two studies (Mateus et al. 2013; Yamazaki and Nagai 2013) found similar results with their sequence matching to the same genes.

Discussion

Disentangling the relative influence of demographic and selective processes at the genomic level is a challenging issue in speciation studies. In particular, most studies that tested whether parallel phenotypic divergence is mirrored by genetic parallelism did not attempt to infer historical divergence scenarios to explicitly account for the confounding effects of demography in the detection of selection (Hohenlohe et al. 2010; Kaeuffer et al. 2012; Gagnaire et al. 2013; Roda et al. 2013; Butlin et al. 2014; Westram et al. 2014; Ravinet et al. 2015). Here, we compared contrasted demographic divergence models to determine the most likely evolutionary scenario underlying genome-wide patterns of divergence among nine pairs of *L. fluviatilis* and *L. planeri* displaying variable levels of geographic connectivity. Gene flow was strong in four connected pairs whereas the most geographically isolated populations were highly genetically differentiated and showed reduced contemporary gene flow. The most connected populations consistently revealed a signal of historical divergence followed by a recent secondary contact between *L. fluviatilis* and *L. planeri*. In addition, there was a higher degree of parallelism in the differentiation level of shared outliers in connected populations than in disconnected ones. Altogether, these results suggest that correlated genomic differentiation patterns among replicate ecotype pairs results from a common history of divergence and gene flow rather than independent gene reuse due to the repeated action of natural selection (Conte et al. 2012; Roesti et al. 2015).

Level of connectivity as a determinant of hybridization

Before analysing genomic differentiation patterns and demography, a prerequisite was to confirm that RAD sequencing analysis of population genetic diversity and structure corroborates recent findings based on neutral markers (Rougemont et al. 2015), and then to determine the most

appropriate populations for understanding the history of ecotypic divergence. First, we validated the high levels of gene flow among *L. fluviatilis* populations (Rougemont et al. 2015, Bracken et al. 2015) suggesting no homing behaviour, as observed in other migratory lamprey systems (Spice et al. 2012; Hess et al. 2013). Second, the admixture analysis (Fig. 3b) confirmed that *L. planeri* populations could be broadly classified into two categories: populations located in downstream areas and highly connected by gene flow to *L. fluviatilis*, and isolated upstream populations that currently do not exchange genes with *L. fluviatilis*. Ecotype pairs consisting of well-connected populations of *L. fluviatilis* and *L. planeri* consistently displayed low levels of genetic differentiation (genome-wide FST <0.10). In these pairs, populations of *L. planeri* also showed levels of heterozygosity similar to the migratory ecotype (table 1, Fig 2) and clustered together with *L. fluviatilis* in the population tree (Fig. 3a). Contrastingly, the most disconnected pairs displayed higher levels of genetic differentiation (genome-wide FST > 0.10). Populations of *L. planeri* had a lower heterozygosity, shared less polymorphism with river lamprey and formed a separate cluster on the population tree (Fig. 3a). Interestingly, two *L. planeri* populations (the Bresle and Risle) were outliers with respect to the geographic context (geographical connection or not), as they displayed lower polymorphism and high differentiation in sympatry (Bresle), or higher polymorphism and low differentiation in parapatry (Risle) compared to other populations in comparable geographical contexts. This suggests that the spatial distribution of these populations has changed through time, or, in the case of Risle, gene flow may occur despite the barriers to migration (e.g. through downstream migration of juvenile Lp or occasional upstream migration of adult Lf). The current geographical distribution of these populations might thus not reflect the demographic history that has shaped their genetic structure (Bierne et al. 2013).

Using a subset of 40 highly differentiated SNPs (FST = 0.7) shared among connected population pairs (identified as such prior to demographic analyses) allowed us to circumvent geographical effects and to distinguish the two ecotypes across all nine population pairs. The low genetic differentiation generally observed between *L. fluviatilis* and *L. planeri* has been interpreted either as a result of phenotypic plasticity (Beamish 1987; Yamazaki et al. 2006; April et al. 2011; Docker et al. 2012) or as evidence of a very recent divergence (Docker et al. 1999; Espanhol et al. 2007; Okada et al. 2010). However, these interpretations mostly relied on mitochondrial variation patterns (the most widely used genetic marker in lamprey studies so far), which can be obscured by introgression (Shaw 2002). Our results clearly refute the hypothesis of phenotypic plasticity within a single gene pool, and demonstrate that shared genetic differences exist among replicate pairs of brook-river lamprey ecotypes. The subset of highly discriminating markers also allowed us to document geographical hybridization patterns. In general, sympatric populations displayed higher proportions of hybrids compared to parapatric pairs. However, the Bresle River (i.e. the “outlier” sympatric pair) did not contain any hybrid whereas in the Risle River (i.e. the “outlier” parapatric

pair) a relatively high proportion of hybrids was observed. Genome-wide differentiation levels between *L. fluviatilis* and *L. planeri* thus appeared to be mostly determined by the degree of geographic connectivity, which directly influence the opportunity for gene exchange through hybridization. Among the 4 highly connected pairs, the Aa and Oir rivers displayed the largest proportions of F1 hybrids (Aa: 16%: Oir 13%). Evidence for frequent hybridization in these hotspots may be exacerbated by sampling effects, especially if individuals were collected in places where hybrids tend to occur at higher densities (Vines et al. 2015) as it can be the case during the simultaneous sampling of individual in the same nest. The comparatively small proportions of later generation hybrids that were detected (one backcross and one F2) may suggest some form of hybrid breakdown (i.e. selection against hybrids or reduced fertility), an hypothesis that would require a more extensive sampling in these hybrid zones to be validated. These results confirm the interest of these populations for the study of speciation in lampreys (Barton & Hewitt 1985; Abbott et al. 2013; Harrison & Larson 2014), and raise the necessity of investigating the effect of gene flow during the long term divergence history of lamprey ecotypes.

A globally shared history of divergence

Our demographic inferences provided new insights into the recent history of divergence, revealing a relatively longer period of divergence in allopatry compared to the subsequent recent episode of secondary contact. In accordance with previous studies (Roux et al. 2013, 2014; Tine et al. 2014), we found that integrating the heterogeneity of migration rate strongly improves models' prediction accuracy, thus supporting the importance of taking this source of variation into account for inferring the history of gene flow during speciation. Our analysis revealed two well supported scenarios (ΔAIC > 10) related to the degree of connectivity between *L. fluviatilis* and *L. planeri* populations. Connected populations pairs generally held a signal of secondary contact with heterogeneous introgression rates (SC2m), whereas disconnected populations pairs held a signal of recent divergence preceded by heterogeneous migration (AM2m). These contrasted divergence models may however not necessarily reflect radically different divergence histories. Indeed, the signal of a past secondary contact may have been lost or obscured by recent drift in parapatric populations. In any case, our results do not support the hypothesis of divergence with ongoing migration (IM/IM2m), except in the case of the Bresle River where it could not be excluded. However in this river, all models including heterogeneous migration rates displayed nearly similar supports (ΔAIC < 10 for AM2m and SC2m compared to IM2m). A possible explanation is that a large difference in effective population sizes during the initial phase of the secondary contact has facilitated gene swamping from the largest population into the small introgressed population.

In the connected populations where the secondary contact scenario was unambiguously detected, only regions involved in divergence are expected to resist the homogenizing effect of gene

flow. This makes these population pairs good models to study the evolution of reproductive isolation. These results, together with the finding of hybrids confirm that RI remains partial as suggested earlier (Rougemont et al. 2015). Thus, the genetic architecture of divergence between lamprey ecotypes seems mostly consistent with the existence of barrier loci that locally reduce the effective migration rate along the genome (Barton & Bengtson 1986; Feder et al. 2010). Parameters estimates in connected populations further suggest that the period of allopatry has been (on average) 8 times longer than the duration of secondary contact, which matches well the idea of a differential erosion of past genetic differentiation outside the direct vicinity of barrier loci. This seems also consistent with the overall low level of reproductive isolation measured experimentally by Rougemont et al. (2015). This interpretation is also compatible with the lack of divergence observed with mitochondrial markers (Espanhol et al. 2007; Blank et al. 2008) and the moderate level of divergence observed at neutral loci (Bracken et al. 2015; Rougemont et al. 2015). We also repeatedly found higher effective population size in *L. fluviatilis* compared to *L. planeri*, which may contribute to increased genome swamping in *L. planeri*. In these conditions, the signal of adaptation and reproductive isolation held by divergent loci can quickly be lost as may be the case in the least differentiated populations (eg. Oir) (Yeaman 2015).

Recent gene flow and the origin of genetic parallelism

Estimates from our genome scan model revealed a high proportion (between 2% to 14%) of loci that were departing from neutrality in each ecotype pair. The level of parallelism was significantly larger in connected pairs compared to disconnected pairs (Fig. 6 And Fig. 7), which is the exact pattern expected after differential introgression (Bierne et al. 2013). Thus, our results support that recent gene flow has played a key role in generating genetic parallelism, because effective migration is strongly reduced in some regions of the genome. These genomic regions experiencing restricted introgression, which are believed to harbour speciation genes, are best revealed in connected population pairs where the confounding effect of drift is less important than in disconnected populations. We found 28 loci shared across the 5 connected pairs. Such level of parallelism was greater than expected by chance, and in most cases stronger than that observed in several other systems (Gagnaire et al. 2013; Perrier et al. 2013; Westram et al. 2014; Ravinet et al. 2015). In addition, our demographic inferences suggested that about 6 to 12% of the sampled genome was not evolving neutrally in connected populations against 11 to 17% in disconnected ones, in which our modelling approach may underestimate the neutral variance in differentiation values. Different hypotheses can explain the partial genetic parallelism observed (black and maroon points in Fig. 6). First, such “private” outliers may belong to the peak-valley-peak signature of divergence between different freshwater derived populations left by a selective sweep or a recent spatial recolonization by the brook ecotype (Bierne 2010; Kim & Maruki 2011; Roesti et al. 2014). Second, coupling of endogenous barriers (e.g. Dobzhansky-Muller Incompatibilities) and exogenous barriers

may occur (Barton & de Cara 2009), resulting in complex architectures. Depending on the timing and duration of secondary contact, these architectures can be differentially eroded by gene flow, hence also contributing to partial parallelism among different pairs. Finally the very high number of outliers in disconnected populations can be simply explained by the major role of drift that impacts genome-wide differentiation among populations inhabiting fragmented networks such as rivers (Fourcade et al. 2013). In these conditions, genome scans are probably not suited to detect the genomic regions that were influenced by selection. On the other hand, these populations were extremely useful to highlight the instrumental role of gene flow in generating partial parallelism in the connected population pairs.

We also aimed at investigating the origin of genetic variation that contributes to parallel divergence patterns at the molecular level. The modelling approach indicated that parallelism was not generated by independent de novo mutations arising during divergence with gene flow. Other possible scenario may involve (i) secondary contact from multiple freshwater refugia, (ii) parallel gene reuse from standing variation present in the parasitic population, or (iii) secondary contact between parasitic and non-parasitic populations having diverged in different refugia, with a subsequent spread (spatial re-assortment) of alleles involved in non-parasitism in the neighbouring rivers (i.e. the transporter hypothesis, Schluter & Conte 2009; Bierne et al. 2013, Welch & Jiggins 2014). The scenario (ii) would imply divergence with gene flow, a model that was not supported in our demographic inferences. Without further support for the multiple refugia hypothesis, the hypothesis of a spatial re-assortment of ancestral variation by migration between rivers appears more parsimonious. This raises the question of whether *L. planeri* alleles are segregating at low frequency in *L. fluviatilis* populations at linkage equilibrium, or instead spread among rivers through hybrid genotypes. In any case, the repeated colonization of new rivers by a few individuals with brook-adapted alleles is expected to have driven rare mutations to high frequencies. This allele surfing effect (Travis et al. 2007; Excoffier & Ray 2008; Lehe et al. 2012) has been detected here since we found an excess of derived mutations reaching higher frequencies in *L. planeri* than *L. fluviatilis* populations for the most highly differentiated loci (Fig. 8). This last line of evidence supports a scenario involving a spread of founder genotypes from river to river during the post-glacial recolonization of rivers by the *L. planeri* ecotype.

The 28 strongly differentiated loci, repeatedly found across multiple locationsnine population pairs could be considered as good candidates implied in the divergence between *L. planeri* and *L. fluviatilis*. Accordingly, some of these outliers were located in genomic regions containing key genes of the GnRh2 family involved in maturation and growth, another gene involved in fast skeletal development and two genes involved in immunity. Interestingly, these genes were also identified as outliers between *L. planeri* and *L. fluviatilis* by Mateus et al. (2013) in a population pair from

Portugal, which further suggests a shared history of divergence and adaptation at a larger spatial scale.

Finally a last question is whether the heterogeneous patterns of differentiation found here between *L. planeri* and *L. fluviatilis* were exclusively compatible with the hypothesis of a semi-permeable barrier to gene flow (Wu 2001). An alternative hypothesis is post-speciation selection at linked sites that locally increased FST measures by reducing within population diversity (Cruickshank & Hahn 2014). Our demographic inferences were consolidated by the occurrence of hybrids which support a role for gene flow in eroding most of the differentiation across the genome. This interpretation is not incompatible with recombination rate variations across the genome favouring the accumulation of differentiation in regions of low recombination (Noor & Bennett 2009; Turner & Hahn 2010; Tine et al. 2014). However, given the high level of genetic differentiation observed in some regions, which contrast to the low levels of differentiation elsewhere, it appears unlikely that linked selection in low recombining regions would produce such heterogeneity alone. The two processes may act jointly to shape genomic divergence in lampreys, and would be better evaluated using a genetic map.

Conclusion and perspectives

Our results clearly support that parallel patterns of genetic divergence between *L. planeri* and *L. fluviatilis* can be caused by a common history of divergence initiated in allopatry, and then followed by secondary gene flow eroding past divergence at variable rates across the genome. The level of geographic connectivity between population pairs was a strong determinant of observed divergence patterns, with direct impacts on both demographic inference and genome scans. In particular, stronger drift in populations of small effective size could obscure signals of divergence and result in smaller level of parallelism. In these situations, gene swamping and further ecological divergence can also act to respectively erode divergent regions or create new regions of divergence that will not bear any signal of parallel divergence. Overall, our data support the idea that the speciation process is best studied in population pairs experiencing high levels of gene flow. *L. fluviatilis* and *L. planeri* were thus best described as partially reproductively isolated ecotypes. In addition, the use of replicated pairs enabled us to identify candidate regions under the direct or indirect effect of selection. Further investigations of these genomic regions will have to be performed in the future to determine their role in the evolution of life history divergence between *L. planeri* and *L. fluviatilis*. Finally, it would be particularly interesting to accurately quantify the relative contribution and interactions between recombination rate variations, selection at linked site (Charlesworth & Campos 2014) and differential introgression (Harrison 1986) on the heterogeneous patterns of differentiation. Full genome data combined with demographic modelling at the European scale and integrating local variations in effective population size could help address these issues.

Acknowledgments

We thank A. Oger who helped us collecting samples and extracting DNA. We thank F. Marchand, J. Tremblay, V. Dolo,Y. Salaville, R. Lemasquerier, V. Lauronce, C. Taverny, B. Rigault C. Rigaud, Y. Perraud, C. Perrier, G. Sanson, J.-L. Fagard and P. Domalain who helped us collect the samples. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing and storage resources. This study was funded by the European Regional Development Fund (Transnational program Interreg IV, Atlantic Aquatic Resource Conservation Project) and by the Office National de l'Eau et des Milieux Aquatiques (ONEMA).

Supporting Information

Supplementary Methods

Structure analysis

To discriminate the two ecotypes using the 40 informative SNPs and identify admixed individuals we first used the program STRUCTURE 2.3.3 (Pritchard *et al.* 2000). We performed five independent replicates at a fixed *k* value of 2 under the admixture model with correlated allele frequencies (Falush *et al.* 2003). Markov Chain Monte Carlo simulations (MCMC) used 100 000 burn-in followed by 100 000 iterations. We also computed 95% confidence interval to help determining admixture values. At this stage individuals displaying admixture q-value <0.90 and for with 95% did not reach the values of 1 were considered as of putative hybrid origin. We also excluded 8 individuals for which too many SNPs were not genotyped.

New Hybrid analysis

The software program New Hybrids (Anderson & Thompson 2002) was used to confirm Structure result and gain insights onto the putative hybrid category of admixed individuals. More specifically, we modelled the probability that individuals belong to one of the following categories: Pure river lamprey, pure brook lamprey, F1, F2, first generation backcrosses. We ran the software using a burn-in period of 500 000 MCMC followed by 1 million sweeps using Jeffreys-like priors for the distribution of allelic frequency and mixing proportions.

Hybrid lab

As noted by Gelman et al. 1995 and emphasized in Nielsen et al. 2005, the drawback of the two former Bayesian approach is that is that validity of the prior needs to be assessed, which requires a simulating approach. We consequently used Hybridlab (Nielsen *et al.* 2006) to generate simulated data and assess the power, accuracy and type I error of our New Hybrids to correctly assign individuals to their true categories. Power and accuracy were define following Vähä and Primmer (2006). (i.e. power=(number of correctly identified individuals to the category)/(true number of simulated individuals of that category) and accuracy=(number of correctly identified individuals in a category)/(total number of individuals assigned to this category) , the type error = (sum of falsely assigned hybrids)/ (total number of pure individuals simulated).

We simulated individuals belonging to the same category as those defined in New Hybrids (Pure Lf, Pure Lp, F1, F2, Backrosses with Lf and Backcross with Lp) using individuals with a q-value >0.90 of being pure *Lf* or *Lp*. For this purpose, we used only populations in which hybrids were found, that is the Aa, Oir and Risle. Given the very high power of our approach (see below) we did not perform simulations on the Loire River were only one hybrid was found. Given the modest sample size population we generated only one hundred individual of each category in each river.

Blast and annotation

We used the 29 sequences potentially under selection between the two ecotypes to perform BLAST (Altschul et al. 1990) analysis on the NCBI public database using an e-value threshold of 1×10^{-10}.

Among the 40 SNPs shared across the nine population pairs, 34 were located in independent reads and two of these reads yielded significant hits as provided in table S4.

Supplementary Results

Table S1: Summary of SNPs number at each filtering steps in each populations

	Number of SNPs								
Catalog	280 000								
Populations	**AA**	**BET**	**BRE**	**CEN**	**JAL**	**ODO**	**OIR**	**RIS**	**SAU**
	113469	96856	77718	60977	68515	77413	86412	90092	75884
Without Pos 59	62618	49714	51322	42418	42211	44425	53094	58644	45191
LP: 80% Genotyped & HWE > 0.05	31180	23165	34750	29312	26952	30243	31862	41005	30119
LF:	31180	23261	34472	28025	28244	28449	30934	38129	29009
MAF Globlal > 0.05 & MAF within pop > 0.1	15984	14542	15709	16881	17766	16231	16913	16902	17331
10 – 100 X	15951	14527	15633	16728	17672	16115	16857	16821	17235
Het <= 0.5	**15413**	**14201**	**15524**	**16410**	**17335**	**15921**	**16564**	**16681**	**16930**

Table S2: Hybrids profile

River	Phenotypic Status LF	LP	N	F1	F2	Backcross LP	Backcross LF
Aa	7	1	8	7	0	1	0
Oir	6	1	7	7	0	0	0
Risle	1	5	6	5	1	0	0
Loire	1		1	1	0	0	0

Table S3: NewHybrids performance on simulated data

Simulations are based on 100 individuals of each category except on the Risle river, where 50 individuals were generated, due to a small initial sample size. Simulations are performed on each river separately. Results are averaged over the three rivers.

	Inferred status									
True pedigree	Pure LF	Pure LP	F1	F2	Backcross Lf	Backcross Lp	Admixed hybrid	Type I error	Accuracy	Power
Pure LF	250						0	-	1	1
Pure LP		250					0	-	1	1
F1			249	0	0	0	1	-	1	0.998
F2			0	243	1	0	6	-	1	0.976
Backcross LF			0	0	249	0	1	-	0.992	0.996
BackcrossLP			0	0	0	250	0	-	0.996	1
								0.022	0.998	0.992

Table S4: Blasts and annotation results

Gene family	E-value	%Identity	Known Function	Ref	Accession Number	Sequence ID
Myosin Heavy Chain	3e-20	100	Contraction	1	AB720829.1	30676
GnRH2, Vasotocin, Protein tyrosin phosphatase (Ptpra) Precursor	3e-18	98	Maturation and growth	2	FJ195978.1	30676 23710
Hox Genes (Hox-delta 3, delta2, eta 4, epsilon 9, epsilon 10, gamma 9, Hox-zeta-1, Evx), zinc finger protein.	3e-5 to 1e-17	95 to 98	Development	3	KF318006 KF318008 KF318008 KF318011	30676
Mannose-binding lectin-associated serine protease	7e-14	96	Immunity	5	AB078894.1	30676
Pineal gland specific opsin gene	3e-12	93%	Photoreception	6	AB116381.1	30676
IVariable Lymphocyte Receptor (IVLR)	3e-6	97%	Immunity	7	AB275449.1	23710
Hedgehog	7e-7	95%	????	4	FP929026.1 FP929027.1	23710

References:

1 Ikeda, D., Y. Ono, S. Hirano, N. Kan-no, and S. Watabe. 2013. Lampreys have a single gene cluster for the fast skeletal myosin heavy chain gene family. PloS One 8:e85500.

2 Gwee, P.-C., B.-H. Tay, S. Brenner, and B. Venkatesh. 2009. Characterization of the neurohypophysial hormone gene loci in elephant shark and the Japanese lamprey: origin of the vertebrate neurohypophysial hormone genes. BMC Evol. Biol. 9:47.

3. Mehta, T. K., V. Ravi, S. Yamasaki, A. P. Lee, M. M. Lian, B.-H. Tay, S. Tohari, S. Yanai, A. Tay, S. Brenner, and B. Venkatesh. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proc. Natl. Acad. Sci. 110:16044–16049.

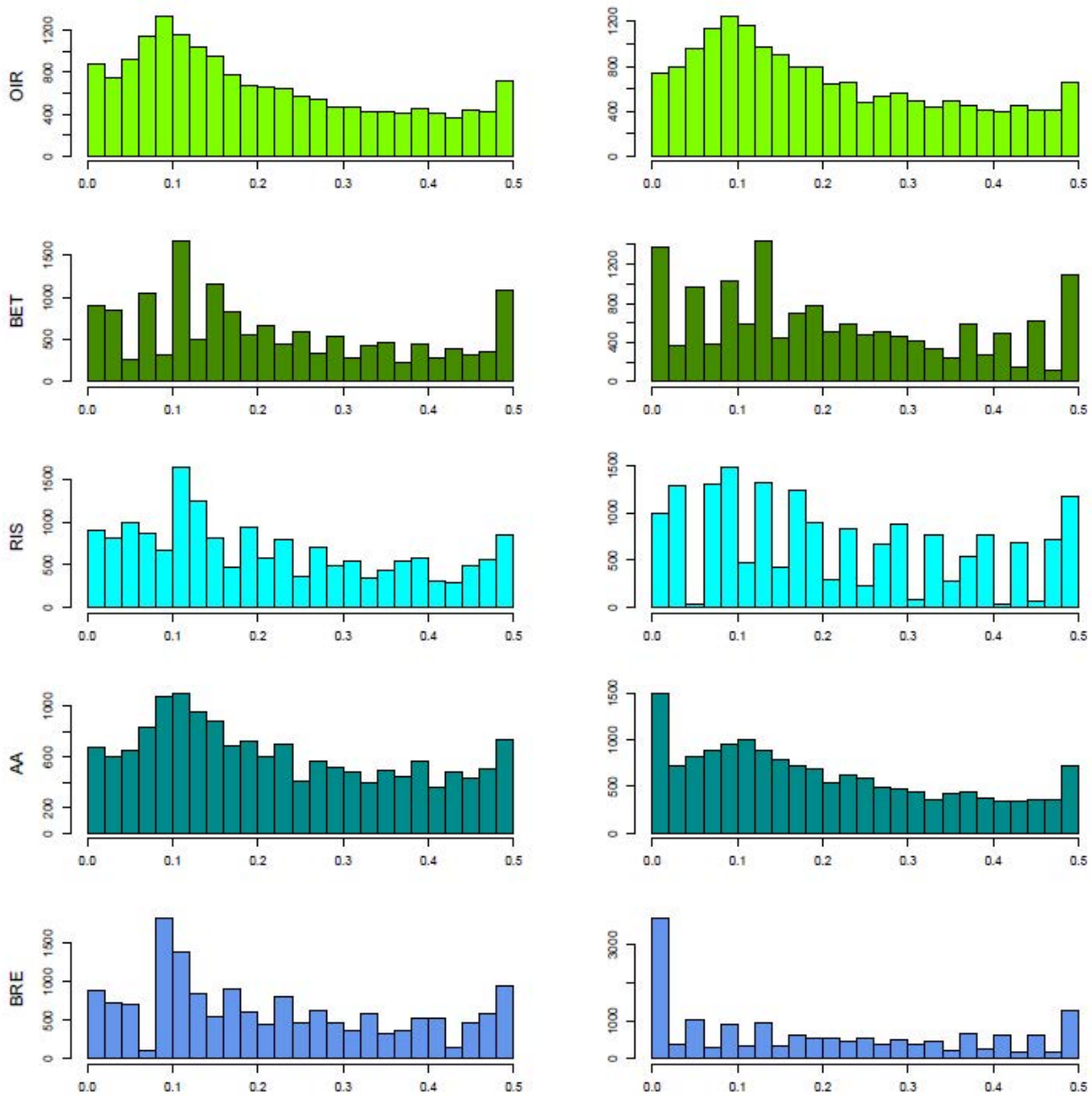
4. Kano, S., J.-H. Xiao, J. Osório, M. Ekker, Y. Hadzhiev, F. Müller, D. Casane, G. Magdelenat, and S. Rétaux. 2010. Two lamprey Hedgehog genes share non-coding regulatory sequences and expression patterns with gnathostome Hedgehogs. PloS One 5:e13332.

5. Endo, Y., M. Nonaka, H. Saiga, Y. Kakinuma, A. Matsushita, M. Takahashi, M. Matsushita, and T. Fujita. 2003. Origin of mannose-binding lectin-associated serine protease (MASP)-1 and MASP-3 involved in the lectin complement pathway traced back to the invertebrate, amphioxus. J. Immunol. Baltim. Md 1950 170:4701–4707.

6. Koyanagi, M., E. Kawano, Y. Kinugawa, T. Oishi, Y. Shichida, S. Tamotsu, and A. Terakita. 2004. Bistable UV pigment in the lamprey pineal. Proc. Natl. Acad. Sci. U. S. A. 101:6687–6691.

7. Nagawa, F., N. Kishishita, K. Shimizu, S. Hirose, M. Miyoshi, J. Nezu, T. Nishimura, H. Nishizumi, Y. Takahashi, S. Hashimoto, M. Takeuchi, A. Miyajima, T. Takemori, A. J. Otsuka, and H. Sakano. 2007. Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. Nat. Immunol. 8:206–213.

***Figure S1*: Distribtuions of allelic frequency**



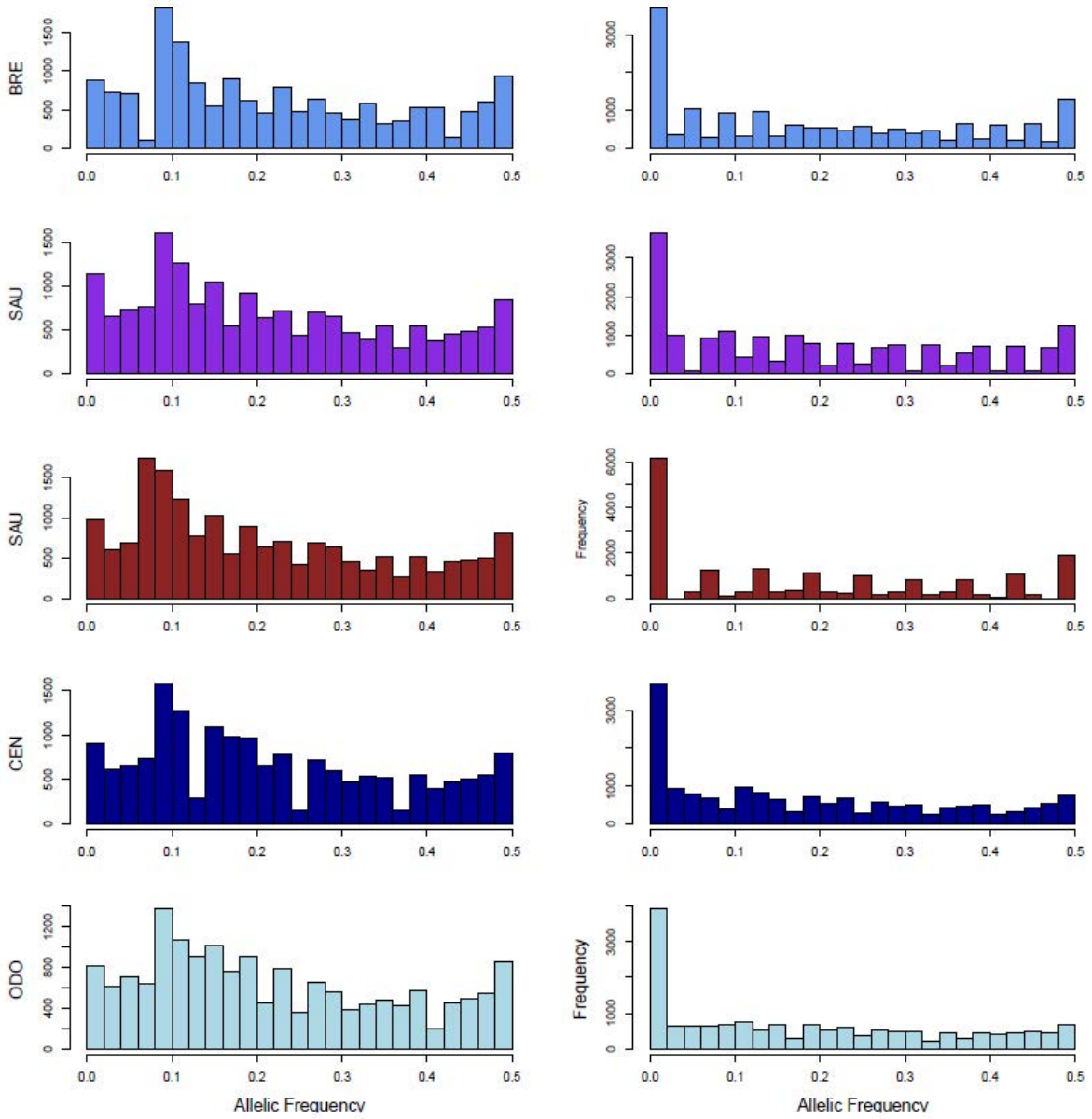

BRE
SAU
SAU
CEN
ODO
Frequency
Frequency
Allelic Frequency
Allelic Frequency

Figure S2: **Admixture Analysis for K = 2 Illustrating swamping at neutral sites and differentiation by geography**

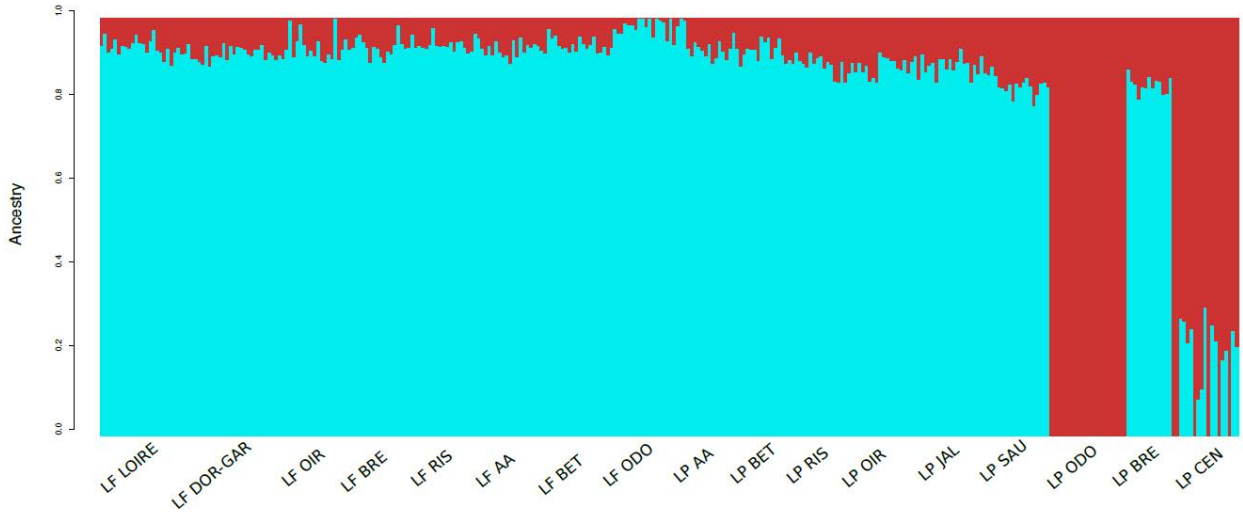
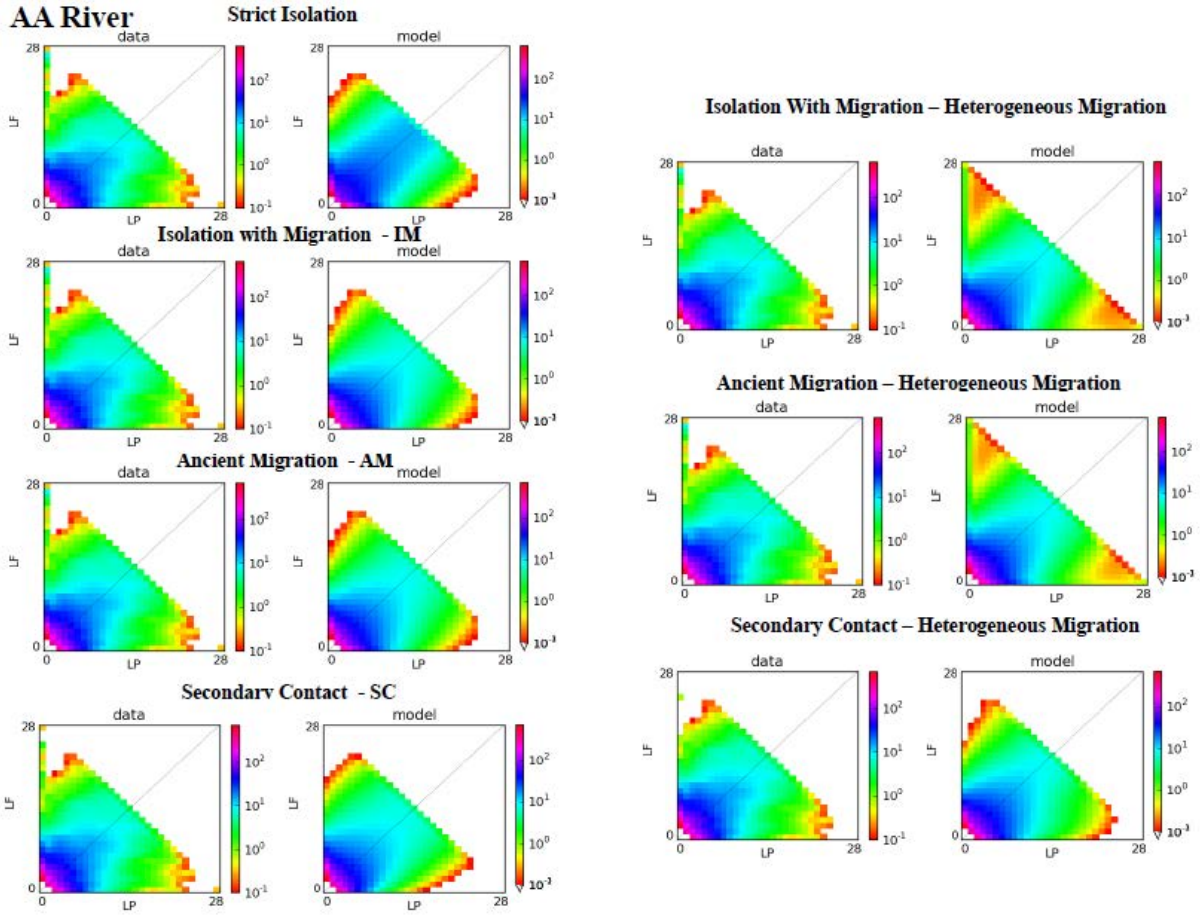
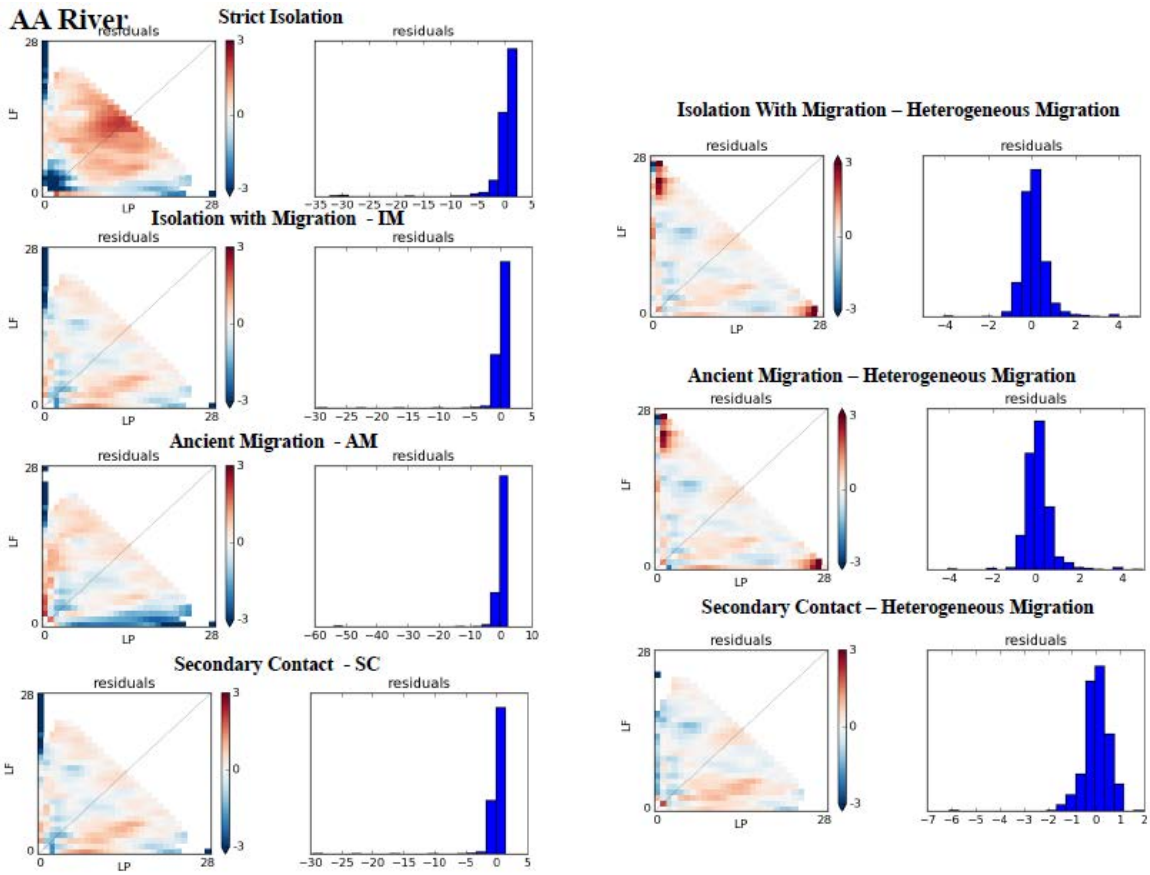
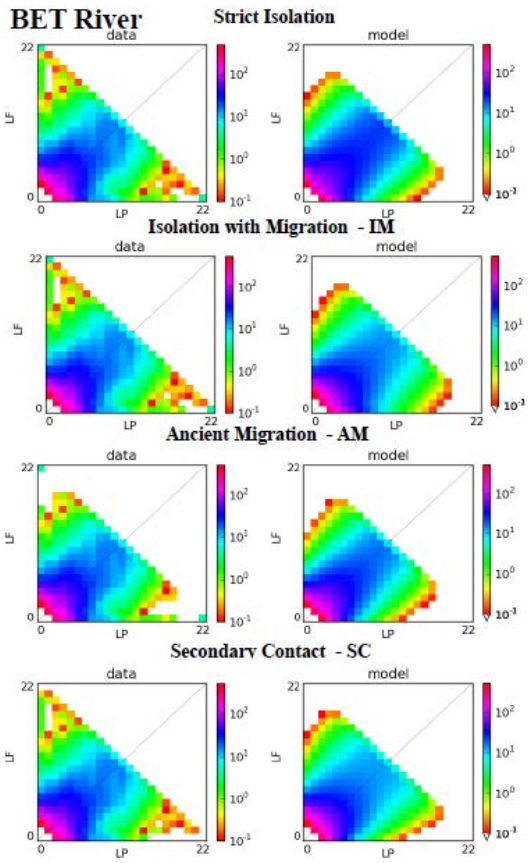
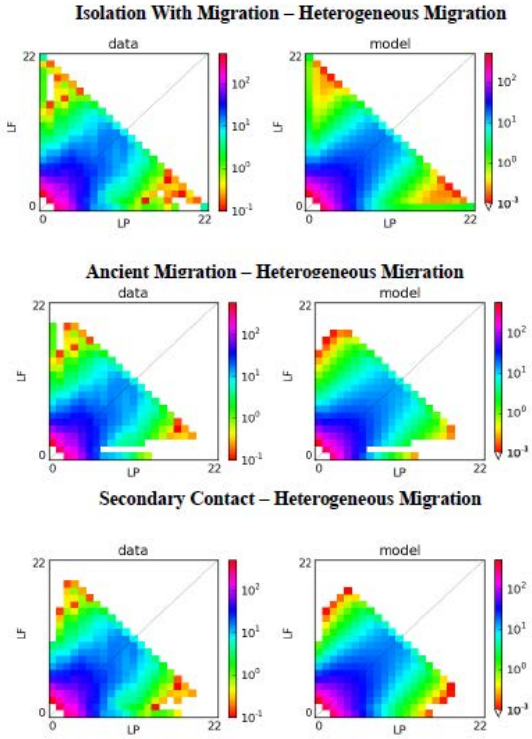


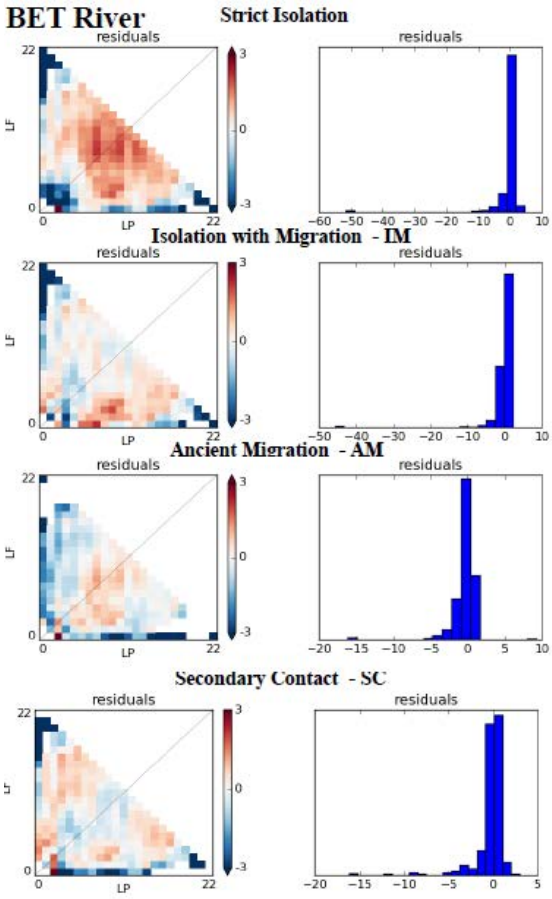
Figure S3: **Model fit and Residuals for each pairs of populations:**

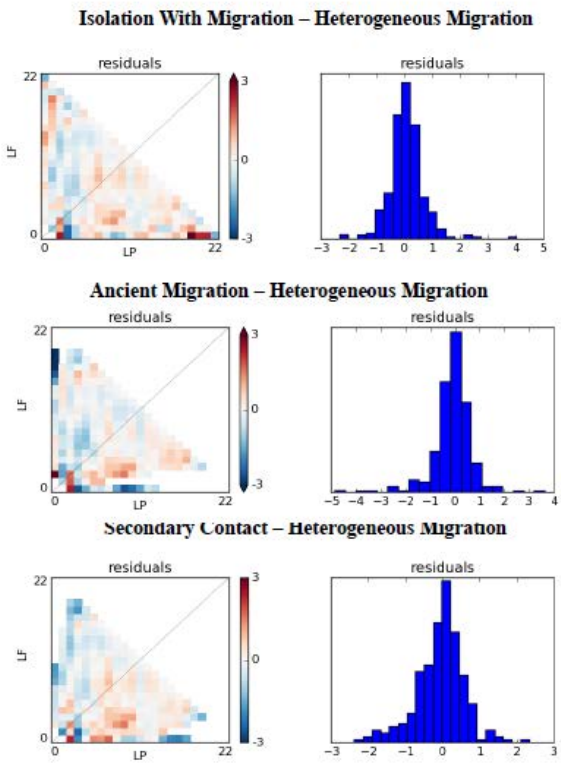

AA River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP

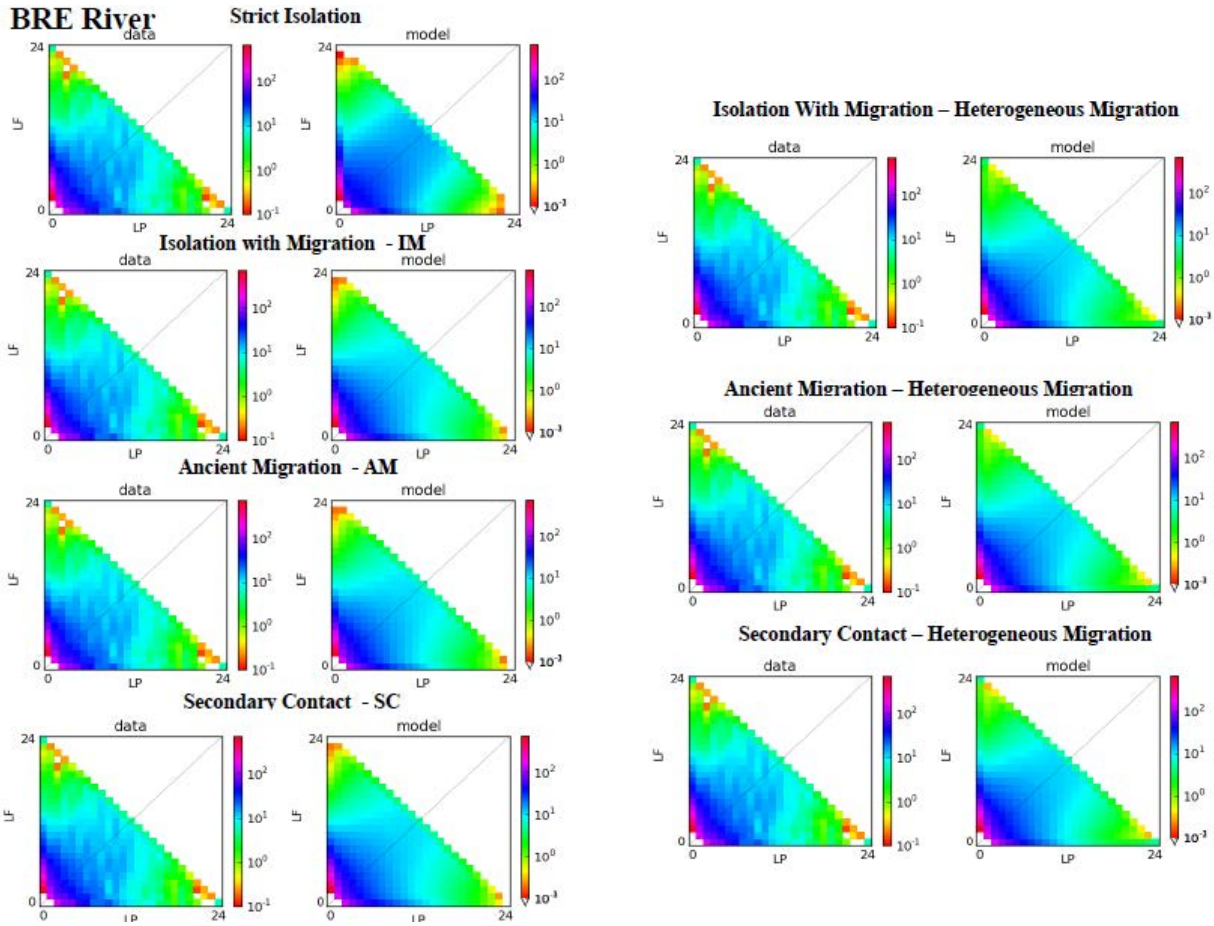

AA River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP

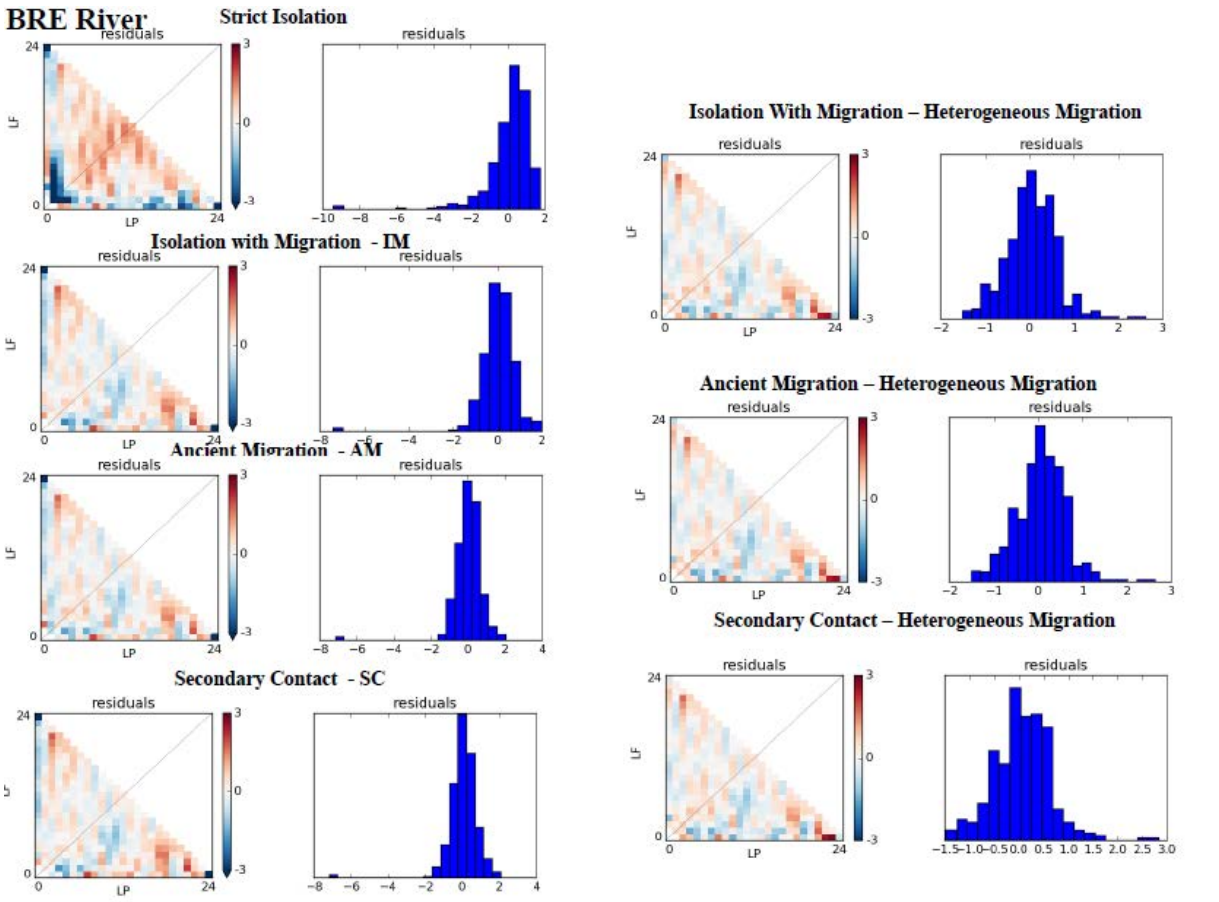

BET River
Strict Isolation
data
model
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
LF
LP

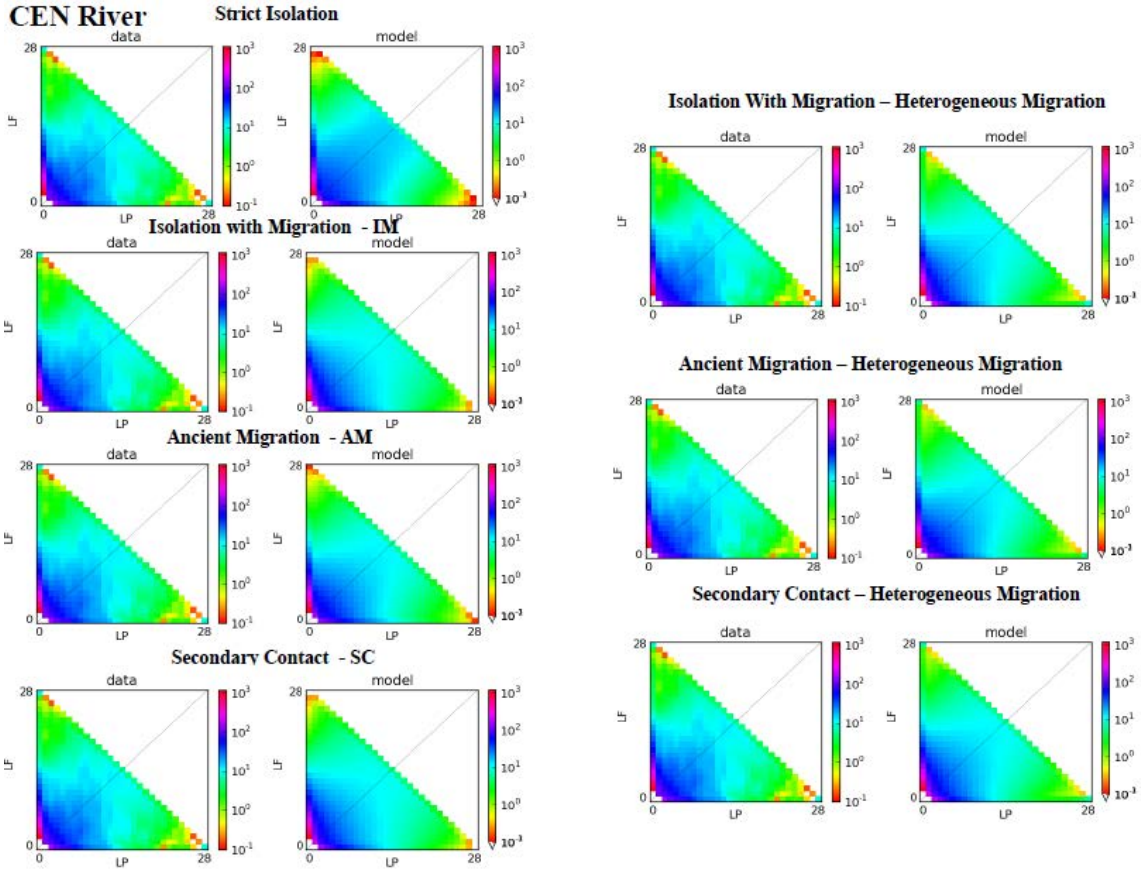

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP

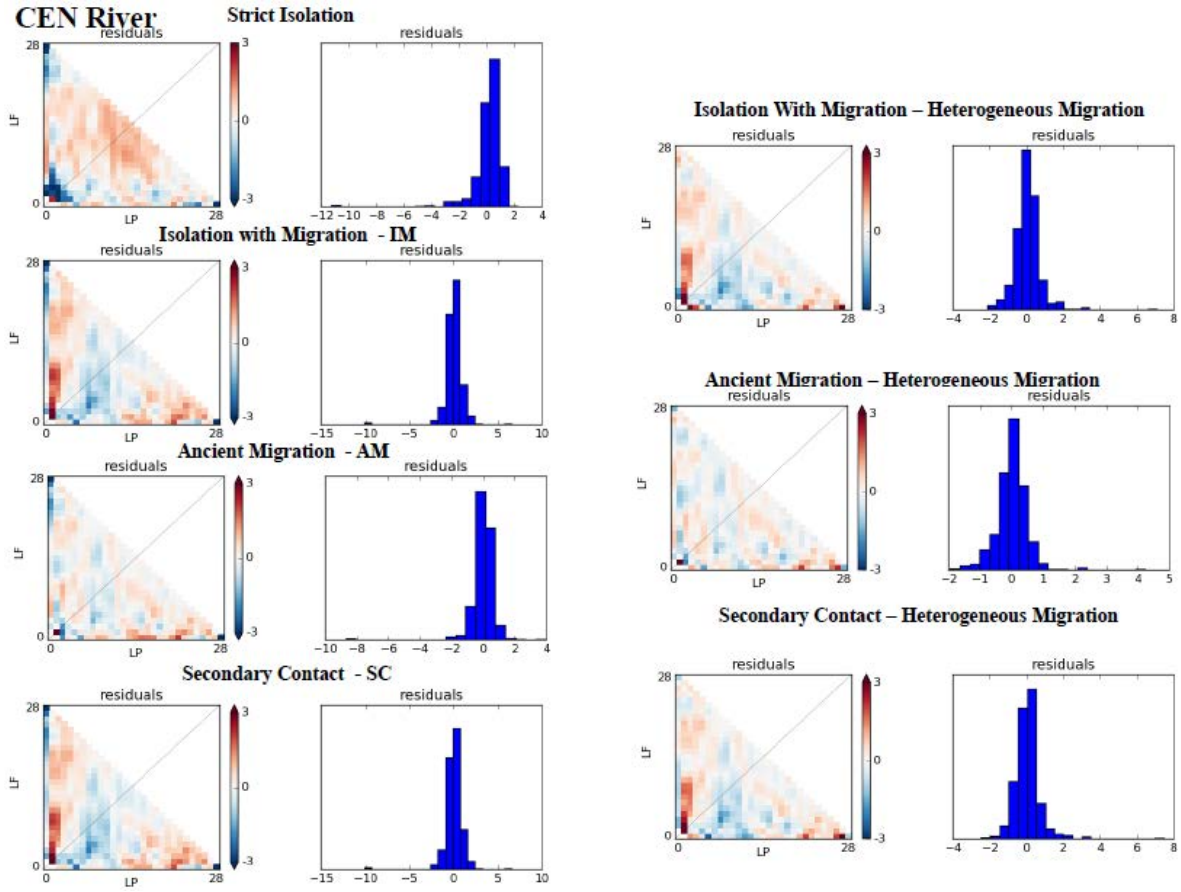

BET River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
residuals
LF
LP

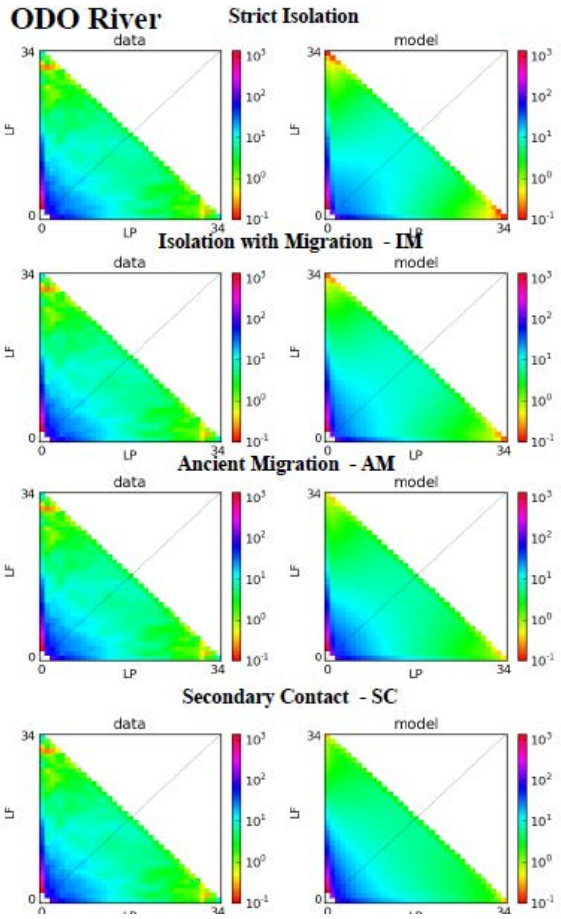

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP

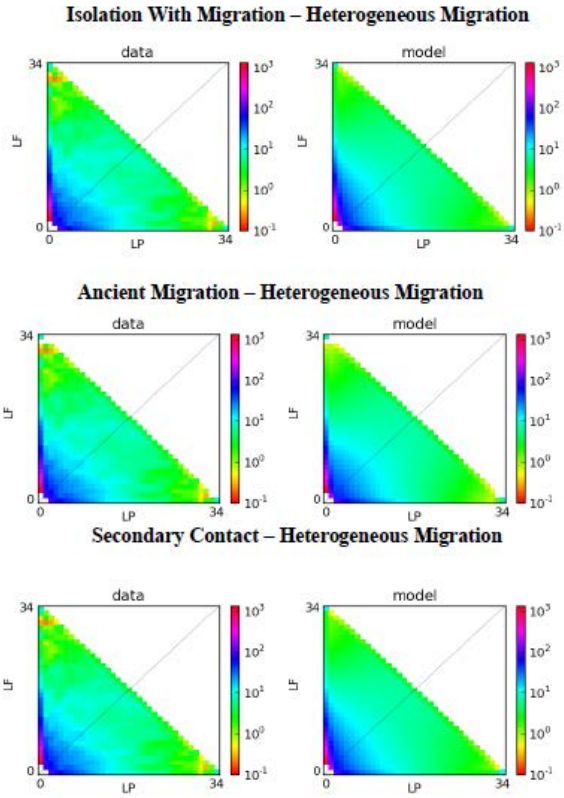

BRE River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP

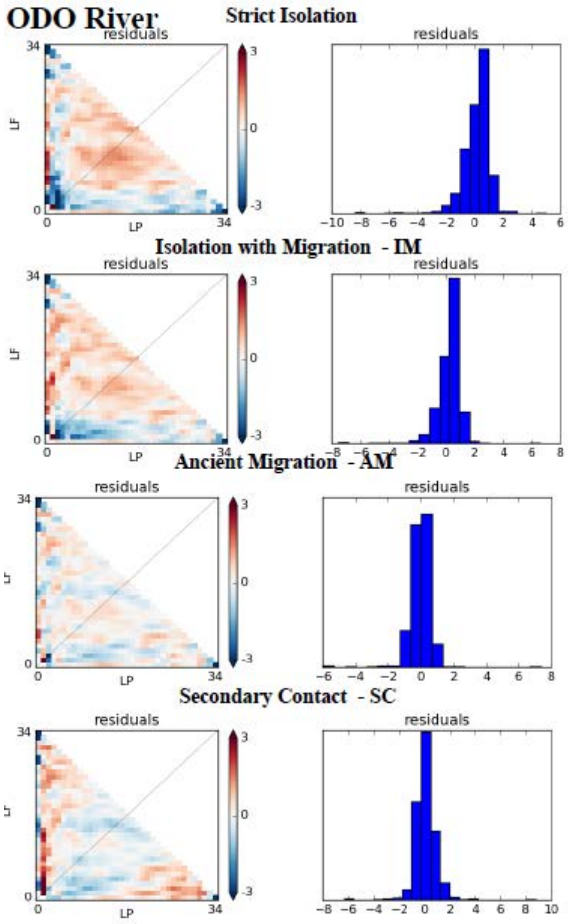

BRE River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP

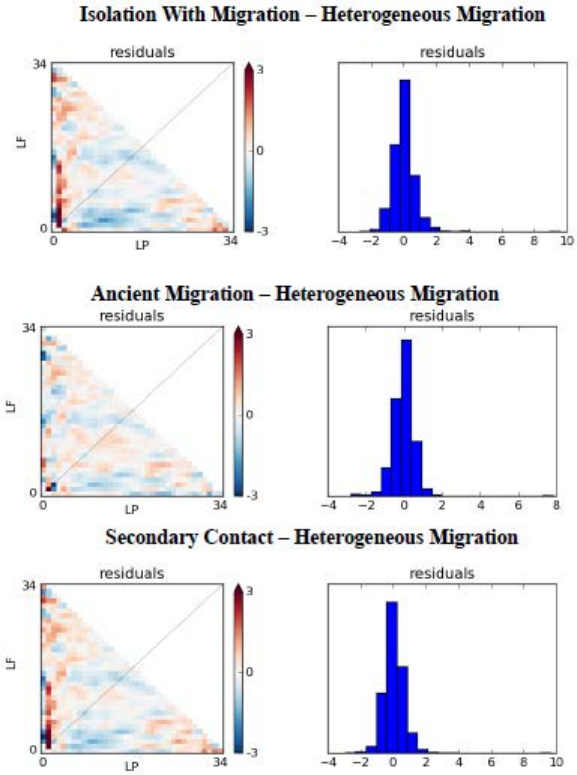

CEN River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LP
LF

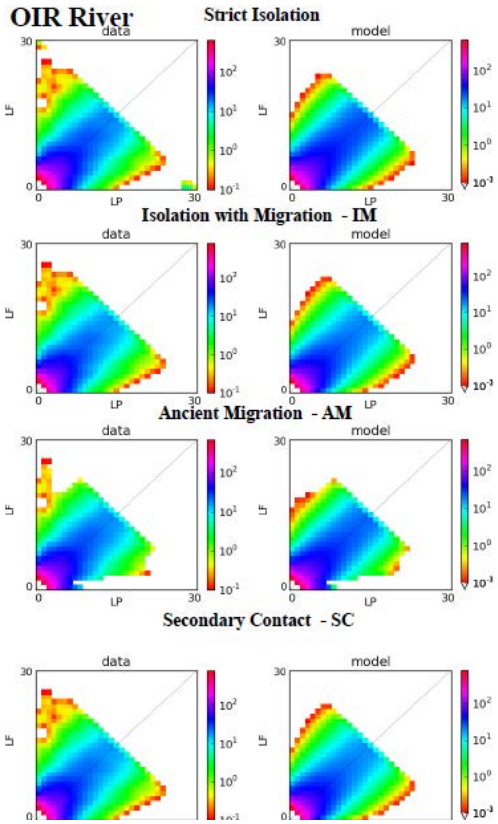

CEN River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LP
LF

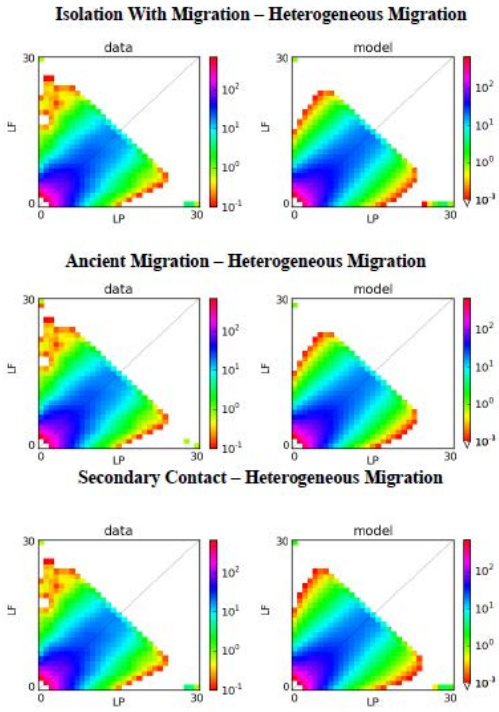

ODO River
Strict Isolation
data
model
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
LF
LP

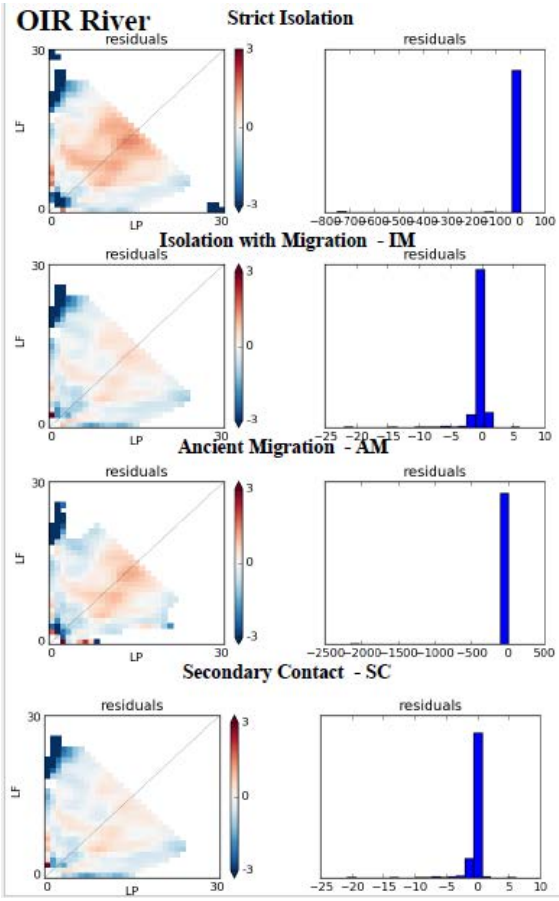

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP

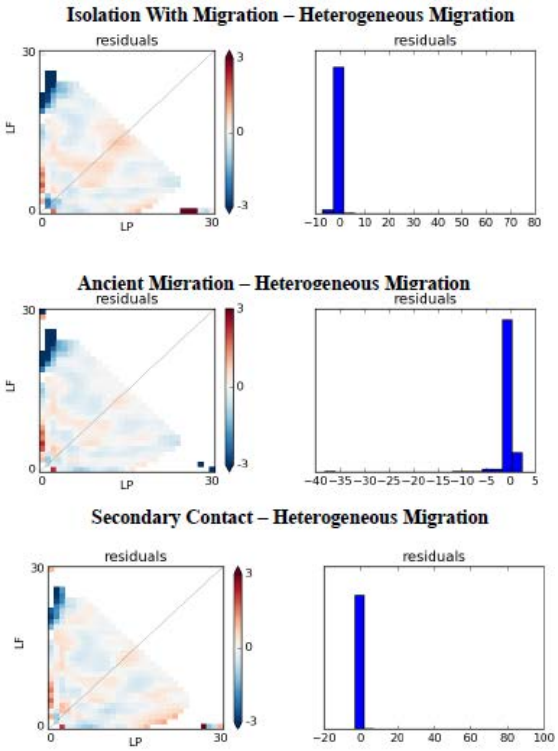

ODO River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
residuals
LF
LP

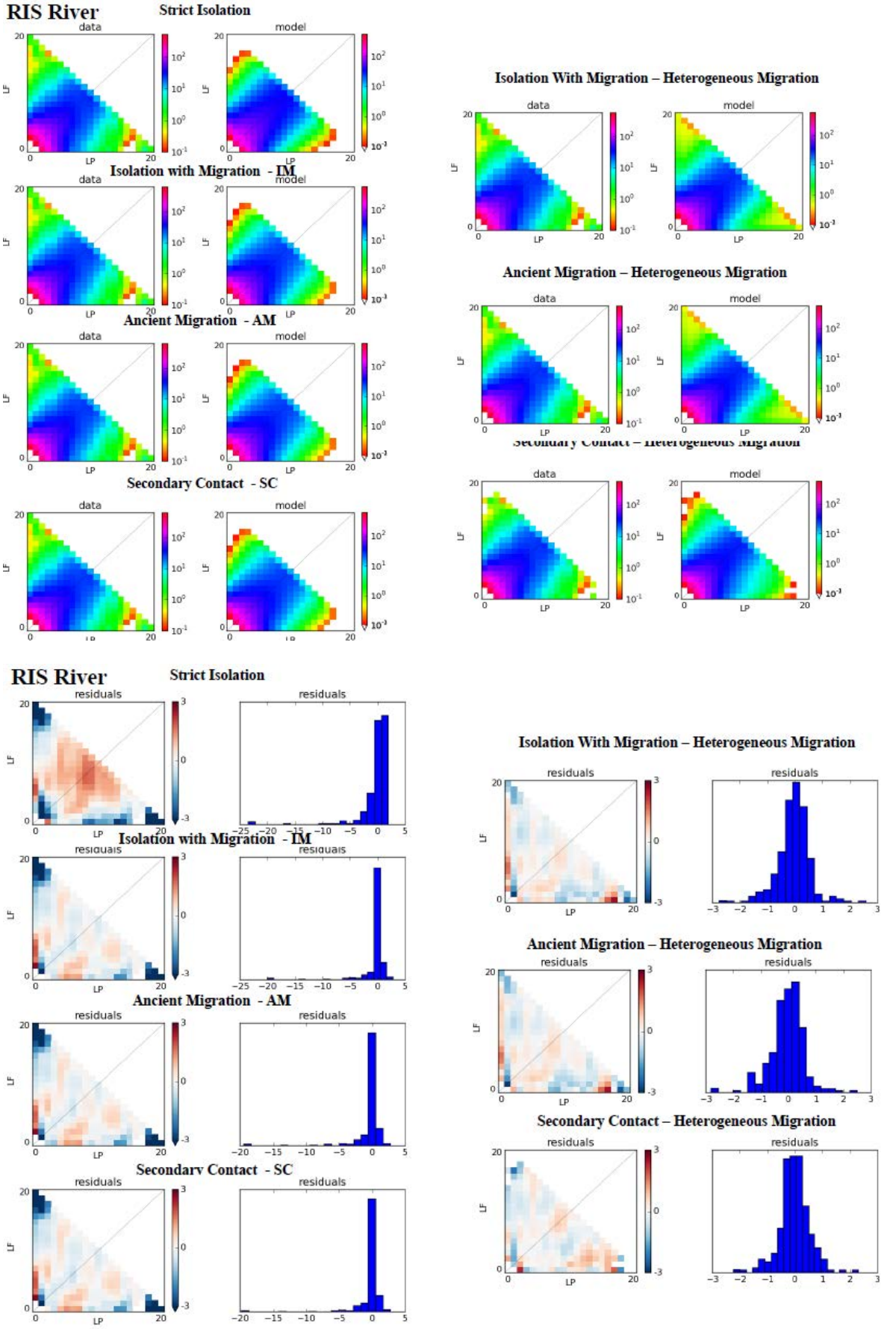

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP

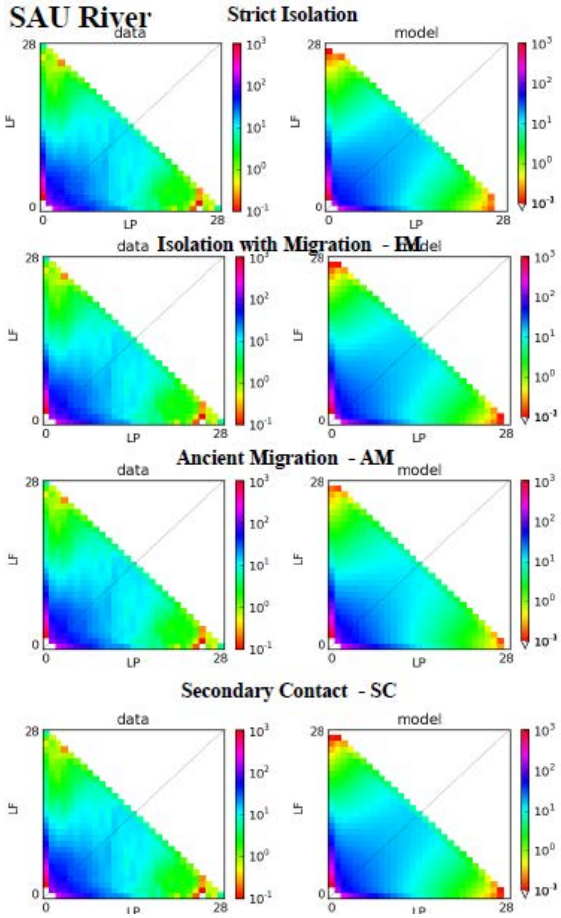

OIR River
Strict Isolation
data
model
LF
LP
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC

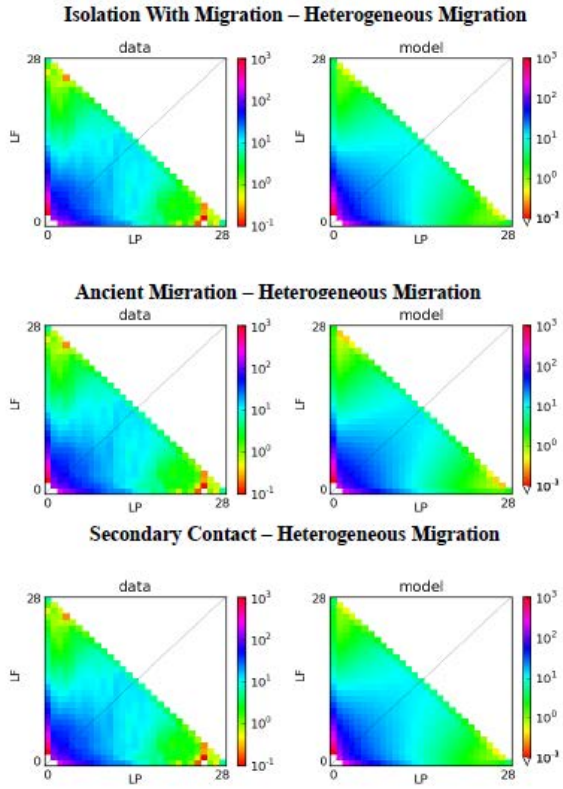

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP

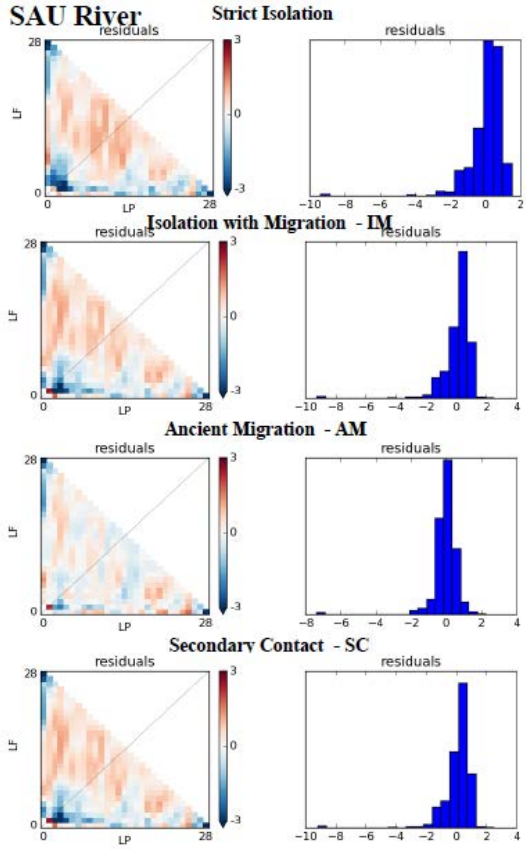

OIR River
Strict Isolation
residuals
LF
LP
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC

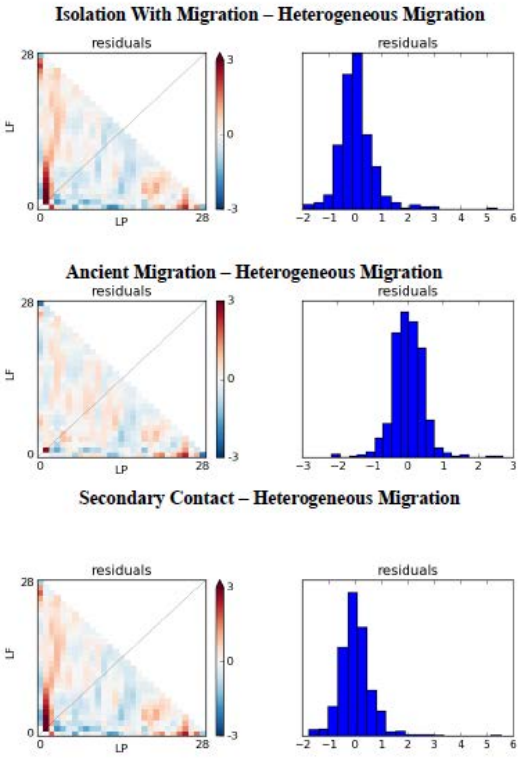

Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP


RIS River
Strict Isolation
data
model
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
LF
LP
residuals


SAU River
Strict Isolation
data
model
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
LF
LP


Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
data
model
LF
LP


SAU River
Strict Isolation
Isolation with Migration - IM
Ancient Migration - AM
Secondary Contact - SC
residuals
LF
LP


Isolation With Migration – Heterogeneous Migration
Ancient Migration – Heterogeneous Migration
Secondary Contact – Heterogeneous Migration
residuals
LF
LP

Chapter 5:

Effect of anthropogenic disturbance on population genetic diversity and structure of European brook lamprey

Article 4: Moderate effect of river fragmentation but strong influence of gene flow between ecotypes on the genetic diversity of brook lamprey populations

Quentin Rougemont, Victoria Dolo, Adrien Oger, Anne-Laure Besnard, Marie-Agnès Coutellec, Sophie Launey, Charles Perrier, Guillaume Evanno

In prep for Evolutionary Applications

Moderate effect of river fragmentation but strong influence of gene flow between ecotypes on the genetic diversity of brook lamprey populations

Quentin Rougemont [1,2], Victoria Dolo[1,2], Adrien Oger[1,2], Anne-Laure Besnard [1,2], Marie-Agnès Coutellec[1,2], Sophie Launey[1,2], Charles Perrier[3] ,Guillaume Evanno [1,2]

[1]INRA, UMR 985 Ecologie et Santé des Ecosystèmes, 35042 Rennes, France

[2]Agrocampus Ouest, UMR ESE, 65 rue de Saint-Brieuc, 35042 Rennes, France

Corresponding author: quentinrougemont@orange.fr

Abstract

Human induced river fragmentation can alter gene flow and have important effects on the spatial distribution of genetic diversity in aquatic species. In contrast, natural introgression between closely related taxa found in sympatry might increase the genetic diversity in some isolated populations. Here we investigated the effect of river fragmentation on the distribution of genetic diversity among brook lamprey populations (*Lampetra planeri*) and we tested for potential effects of introgression with the closely related species *L. fluviatilis*. We used 13 microsatellite markers to genotype 2472 individuals collected in 81 sites in 29 rivers from Western Europe and sampled upstream and downstream of barriers to migration. Results suggested a negative effect of the number and height of barriers only on allelic richness of *L. planeri* populations. A strong increase in downstream diversity was also observed suggesting a major impact of asymmetric gene flow in brook lampreys. *L. planeri* populations coexisting with *L. fluviatilis* consistently displayed higher levels of genetic diversity than allopatric populations, which may be due to introgression between the two ecotypes. These results have important implications for conservation strategies by showing that i) barriers have moderate negative effects on local genetic diversity of *L. planeri* populations and ii) both ecotypes should be jointly managed.

Keywords: Habitat fragmentation, migration-drift equilibrium, gene flow, *Lampetra* sp.

Introduction

Human activities strongly modify natural ecosystems (Vitousek *et al.* 1997b), and have strong impacts on evolutionary trajectories of wild species (Palumbi 2001). In particular, habitat fragmentation is a major threat on wild species (Vitousek *et al.* 1997b; Fahrig 2003). Habitat fragmentation can decrease dispersal rates, which reduces gene flow among subpopulations and ultimately decreases effective population size and genetic diversity (Frankham 1998, 2005; Couvet 2002; DiBattista 2008; Blanchet *et al.* 2010; Whiteley *et al.* 2013). Small populations are expected to experience stronger effects of genetic drift, potentially leading to higher levels of inbreeding (Frankham, 1995a,b, 1998) which can lead to local extinctions (Saccheri *et al.* 1998; Spielman *et al.* 2004). Habitat fragmentation has influenced the genetic composition of hundreds of species of birds (Alonso *et al.* 2009), fishes (Hänfling & Weetman 2006; Raeymaekers *et al.* 2008; Blanchet *et al.* 2010; Torterotot *et al.* 2014) and plants (Young *et al.* 1996). A key challenge is thus to accurately understand how human induced habitat fragmentation alters gene flow to better predict the future viability of populations and help design management strategies.

Freshwater ecosystems have been particularly affected by fragmentation worldwide (Dynesius & Nilsson 1994; Nilsson *et al.* 2005) due to the construction of dams, weirs, and to artificial modifications of river trajectories. Such fragmentation alters the possibility of gene flow between populations of aquatic organisms, so that upstream isolated populations are particularly exposed to genetic drift and its consequences on genetic diversity and ultimately inbreeding. This is particularly problematic in river systems that are naturally shaped as dendritic networks, which poses several challenges to traditional models used in population genetics such as Wright's (1951) infinite island model or the linear stepping stone model (Kimura & Weiss 1964). Indeed, migration is often expected to occur following water currents, generating pattern of asymmetric gene flow (Hänfling & Weetman 2006; Pollux *et al.* 2009) and structuring populations along linear networks. As a result, populations are structured following a source-sink model (Kawecki & Holt 2002; Hänfling & Weetman 2006) in which the genetic diversity of upstream source populations will be smaller than the downstream sink populations. Thus, upstream populations are naturally expected to deviate from the migration drift equilibrium. A recent study has investigated three possible processes responsible for this observed pattern of downstream increase in genetic diversity across taxa (Paz-Vinas *et al.* 2015): i) downstream biased dispersal generating downstream gene flow (Paz-Vinas *et al.* 2013), ii) increase in downstream habitat availability (e.g. Raeymaekers *et al.* 2008) and iii) upstream founding events with loss of genetic diversity (e.g. following postglacial colonization Cyr & Angers 2012). They showed that each of these processes can theoretically influence patterns of downstream increase in genetic diversity. In such conditions, human mediated alterations of habitat connectivity in rivers may obscure these patterns or exacerbate it. To date, most studies focused on delineating the effect of barriers on migration in species targeted by fisheries such as salmonids that display an important migratory ability (Morita & Yamamoto 2002; Yamamoto *et al.* 2004) and may thus not be the more relevant to study the

effect of fragmentation. Few empirical studies have focused on species with a low dispersal ability (e.g. Hänfling & Weetman 2006; Raeymaekers *et al.* 2008; Blanchet *et al.* 2010) whereas effects of fragmentation are expected to impact more strongly the genetic diversity of these species.

In addition, species can display various life history strategies or ecotypes that may differ in their dispersal capacity and thus be differentially affected by changes in habitat connectivity. For instance, in certain fish species some individuals are freshwater-resident while others are anadromous (i.e. reproduce in freshwater and juveniles migrate to sea for growth, Jonsson & Jonsson 1993; Dodson *et al.* 2013). Anadromous individuals can either display a homing behavior as they return back to their native river to spawn, or disperse into neighboring rivers, creating large opportunities for gene flow. Consequently, anadromous populations generally display lower levels of population genetic structure than resident populations (Hohenlohe *et al.* 2010; Spice *et al.* 2012; Hess *et al.* 2013; Rougemont *et al.* 2015, Quéméré et al. 2015). Resident populations are also expected to be more strongly impacted by barriers to migration, which can isolate populations located in upstream reaches (e.g. Perrier *et al.* 2013).

The European brook lamprey *Lampetra planeri* is a widespread freshwater resident species with a putatively low dispersal ability linked to its small size (15-22 cm) and its particular life cycle, as the adults stay only a few months in the river before they reproduce and subsequently die (Taverny & Élie 2010). It is closely related to the river lamprey *Lampetra fluviatilis* that is parasitic and anadromous at the adult stage. The two ecotypes share many similarities as they spend 3 to 6 years burrowed in the substrate of river beds. *L. fluviatilis* however is often located in lower parts of the river due to their low upstream migratory ability, particularly when confronted with barriers (Lucas *et al.* 2009; Russon *et al.* 2011; Kemp *et al.* 2011; Foulds & Lucas 2013). However, due to dispersal abilities through the marine environment and apparent moderate homing behavior, populations from nearby watersheds remain connected and display a low genetic differentiation (Bracken *et al.* 2015; Rougemont *et al.* 2015). In contrast, *L. planeri* has a highly reduced migratory behavior: it does not move outside its watershed and only migrates short distances within the river for breeding purposes (Malmqvist 1980). Thus, the most isolated brook lamprey populations located in the upper reaches of rivers can be strongly genetically differentiated from other populations either downstream or in other rivers (Pereira *et al.* 2010; Mateus *et al.* 2011; Bracken *et al.* 2015; Rougemont *et al.* 2015). These isolated populations often display a low genetic diversity (Rougemont et al. 2015a) as revealed for instance by low levels of expected heterozygosity, an indicator of effective population size (Nei & Takahata 1993). Brook lamprey populations living in sympatry with river lampreys have been found to display a higher level of genetic diversity than populations located in upstream reaches where river lampreys are absent (Rougemont et al. 2015). As a result, brook lamprey populations may benefit from the connection with *L. fluviatilis* that may act as a 'reservoir' of genetic diversity. To test this hypothesis one should compare the level of genetic diversity between rivers with only brook lampreys and those where both species coexist.

The main aims of this study were to understand: i) the role of river fragmentation on population genetic diversity and structure of *L. planeri*, ii) the spatial distribution of genetic diversity among *L. planeri* populations at a large scale by comparing populations from the UK, Ireland and France and iii) the possible role of *L. fluviatilis* in increasing genetic diversity in sympatric *L. planeri* populations *via* introgression. In order to assess the effect of downstream increase in genetic diversity and the potential role of migration barriers in this effect we performed extensive sampling of *L. planeri* upstream and downstream of barriers to migration in 29 rivers from three geographical regions: Brittany, Normandy and the Upper Rhône in France. Moreover, two watersheds were sampled more extensively in order to better test for a pattern of downstream increase in genetic diversity and fragmentation effects. To test the prediction that *L. planeri* populations found in sympatry with river lampreys will display greater levels of genetic diversity than populations where river lampreys are absent, we sampled populations of *L. planeri* in sympatry or parapatry with its sister species *L. fluviatilis* in Normandy and populations in Brittany where river lampreys are absent.

Materials and Methods

Sampling design

A total of 2472 lamprey individuals were captured at 81 sites spread over 29 rivers. All individuals were captured in 2013 and 2014. 228 *L. fluviatilis* were sampled in 8 rivers: 7 rivers from Normandy and one from the Loire, to better analyse the distribution of genetic diversity between *L. fluviatilis* and *L. planeri* at a global scale. We targeted *L. planeri* located upstream and downstream of a putative barrier to dispersal and, if possible, close to the barrier (less than 1 km upstream or downstream) to limit the effect of isolation by distance. In this study a site thus corresponds to a sampled point located either downstream or upstream of a barrier to migration. We considered all kind of barriers of moderate size (mean height = 1.07 m, min height = 0.25 m and max height = 5 m, table S1) that may limit the active dispersal of lampreys. In some cases, we were unable to capture lampreys immediately downstream or upstream of dams and some sampled points were separated by more than one obstacle. 2302 *L. planeri* lampreys were collected from 73 sites in four French major regions (Mediterranean area (Rhone), Normandy, Brittany and Upper Rhine), as well as 3 sites in the United Kingdom and Ireland to better understand the large scale distribution of genetic diversity. The 17 sites (n= 536 individuals) from Northern France and Normandy all occurred in sympatry or parapatry with *L. fluviatilis*. Conversely, the 32 sample sites from Brittany (n= 969 individuals) were completely allopatric. Indeed, despite being located along the coast of Atlantic, *L. fluviatilis* is currently not present in the rivers in this area. In addition, two sites (one upstream and one downstream of a migratory barrier) were sampled from a tributary of the Loire (the Cens River). In Brittany, for 2 rivers we failed to capture *L. planeri* both upstream and downstream of a barrier, so only 30 sites were suitable to study the effect of fragmentation in this area. Ultimately we sampled 18 sites (n= 575 individuals) from the Rhone

area and Upper Rhine to better capture the geographic distribution of genetic diversity. We focussed only on populations distributed in Northern France and along the Atlantic to study the effect of river fragmentation (see below and results). A total of 49 sites (30 from Britany, 2 from the Loire, 17 from Northern France, n=1440 individuals) were suitable for such analysis.

A fin was clipped on each specimen and preserved in 95% EtOH. Physical variables used as explanatory variables of genetic parameters included the number of obstacles, their cumulative height, the geographic distances between both sample points and the distance of the sampling point from the river source. Data about obstacle height were gathered from the French "Referentiel des Obstacles à l'Ecoulement". Geographic distances were computed using QGis 2.10.1. In addition, we performed two linear transects on two independent rivers (the Arz and Scorff) with a total of 8 and 7 sample sites respectively, to investigate the respective effects of obstacles and isolation by distance.

DNA extraction for microsatellite and genomic analysis

Microsatellite DNA extraction and genotyping

Genotyping was performed with 13 microsatellite markers specifically developed for *L. planeri* and *L. fluviatilis* after DNA extraction using a Chelex protocol modified from Estoup *et al.* (1996) and strictly following protocols of Gaigher *et al.* (2013) and Rougemont *et al.* (2015).

Statistical analysis of microsatellite data

Genetic diversity within populations

We tested deviations from Hardy-Weinberg equilibrium using GENEPOP 4.1.0 (Rousset 2008) exact tests with Bonferroni corrections (Rice, 1989, $\alpha = 0.05$) and computed the inbreeding coefficient (*Fis*) for each population using FSTAT 2.9.3(Goudet 2001). Genetic diversity indices were computed and included the number of alleles (An), Allelic richness (*Ar*), observed heterozygosity (*Hobs*) and expected heterozygosity (*Hnb*). We computed Loiselle relatedness coefficients (Loiselle *et al.* 1995) among individuals within each population using the software Genodive (Meirmans & Van Tienderen 2004).

Genetic structure among populations

We computed Weir & Cockerham's (1984) estimator of F_{ST} between all pairs of populations and used permutation tests with Bonferroni corrections to test for significance in FSTAT. We tested for global pairwise differences in F_{ST} between upstream and downstream sites and among the three major regions using permutations tested in FSTAT (10,000 TO 15,000 test in each cases) as well as pairwise t-test adjusted for multiple testing using FDR corrections in R. However, populations are expected to deviate from migration drift equilibrium and to show a downstream increase in genetic diversity resulting in biased F_{ST} that may reflect this gradient effect rather than true differences. As a result, we also used indices of genetic

differentiation that are independent from variations in genetic diversity among populations: the Jost D (Jost 2008) and Hedrick G'st (Hedrick 2005). We illustrated the distribution of pairwise F_{ST} and Jost D values in R using the heatmap.2 function implemented in the gplots package, which computes a heatmap together with a hierarchical clustering tree.

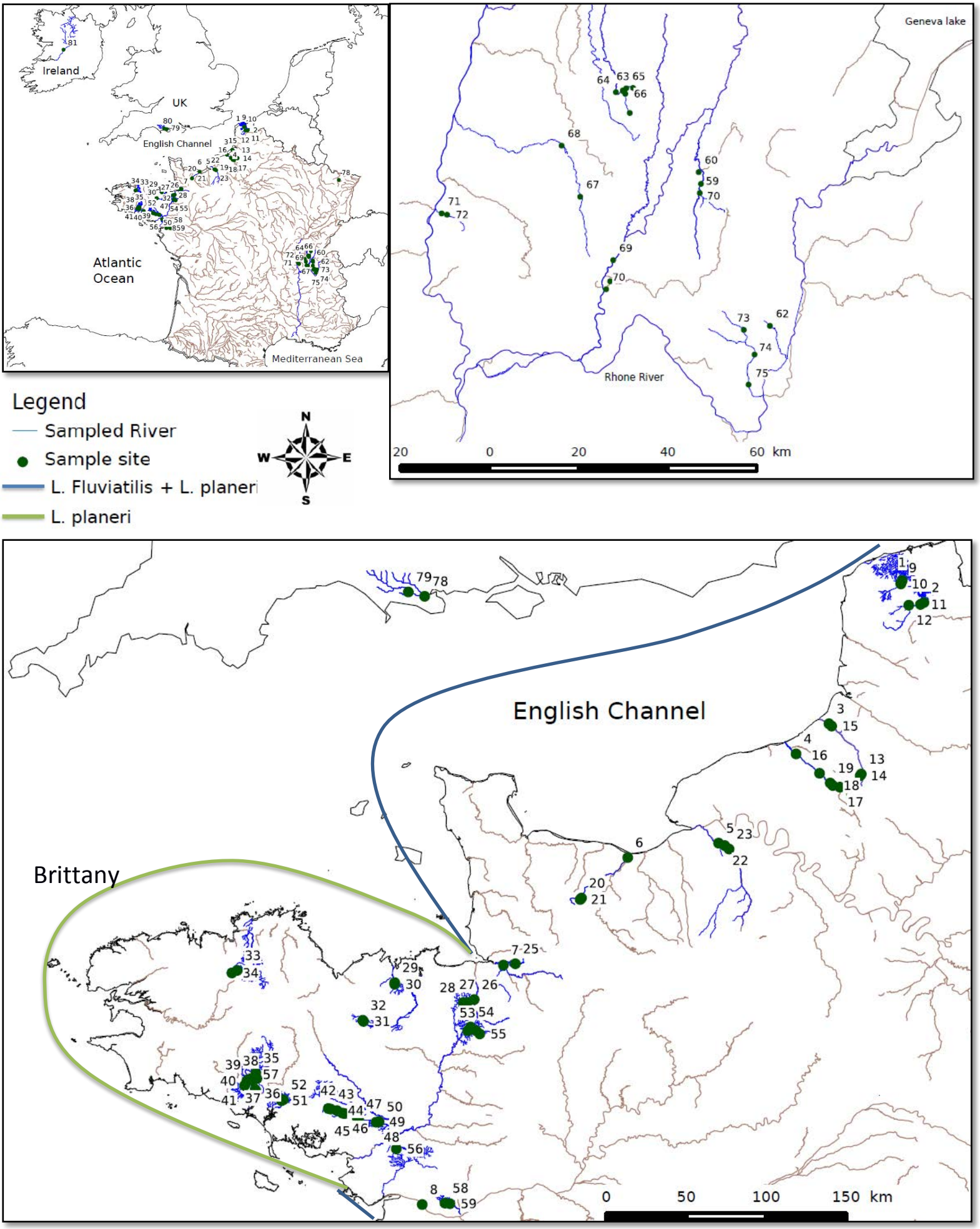


Figure 1: **Sampled areas** (Numbers correspond to sampling site, as provided in table S1)

The Bayesian clustering program STRUCTURE 2.3.3(Pritchard *et al.* 2000) was used to evaluate the number of clusters (*k*) from 1 to 49 in the complete dataset. 10 independent replicates per *k* value were performed. Markov Chain Monte Carlo simulations (MCMC) used 200 000 burn-in and 200 000 iterations under the admixture model with correlated allele frequencies (Falush *et al.* 2003). We used log likelihood Ln Pr(X|K) and the ΔK method (Evanno *et al.* 2005) to determine the number of clusters in STRUCTURE HARVESTER (Earl & vonHoldt 2012). Plots were drawn using DISTRUCT 1.1(Rosenberg 2004). Due to the high differentiation of the populations of the upper Rhone we tested for population substructure by splitting the dataset in three sub-dataset: i) upper Rhône separately (575 *L. planeri* individuals), ii) Brittany and iii) Normandy (*L. fluviatilis* and *L. planeri)* Ireland and the UK together (1727 *L. planeri* individuals).

Landscape genetic analyses

Different approaches were used to test the effect of fragmentation in each site. In a first approach, based on linear mixed models (LMM hereafter) we used only sites that were truly independent (i.e. a sampling site was never used in more than one comparison), leading to a total of 25 independent upstream/downstream pairs of sites. The *Hnb, Ar* and relatedness differential between upstream and downstream sites were used as estimator of difference in genetic diversity within each river. We used the linearized genetic distance F_{ST} /(1- F_{ST}) (Rousset 1997) between both sites in each river. The number of obstacles, cumulative height, and geographic distance between sites were treated as fixed effects while the river was treated as a random effect. We also integrated the distance to the source as a co-variable in our models. Models were tested using likelihood ratio tests and AIC as implemented in the MASS (Venables et Ripley 2002), lme4 and nlme packages in the R software (R Core Team, 2015). Due to the very low genetic diversity of the Upper Rhone (see results) and an expected lack of power (e.g. the downstream populations of Upper Rhone displayed less genetic diversity than the upstream populations of Atlantic and northern France), it was not included in the analysis.

In addition to this analysis, we tested the prediction of an increase of neutral genetic diversity downstream (Raeymaekers *et al.* 2008; Blanchet *et al.* 2010; Paz-Vinas *et al.* 2013). We used point estimates of *Ar, Hnb* and Relatedness in linear mixed models with a Gaussian error family. In these cases we included all upstream and downstream sample sites from all rivers (ranging from 1 site per river to up to 8 sites). The distance to the source was used as predictor variables (fixed effect) while the river was considered as a random effect. In this case again we did not include the Upper Rhone.

In addition, we tested for a pattern of IBD, and tested to which extent it was affected by the presence of obstacles on the Crano and Arz River (Brittany region) where more sites had been sampled (7 and 8 sites respectively). We used Mantel and partial Mantel tests in R. We constructed matrices of

linearized F_{ST} (G'_{ST} and D), *Ar* and Hnb differentials and matrices of pairwise geographical distance, number of obstacles and cumulated height for each river.

Results

Genetic diversity within populations

*F*is was only statistically significant (Table S1) in four populations: the downstream sites on the Leguer River (Brittany), the downstream sites on the Oignin, Calonne and Neyrieux Rivers (Upper Rhône).

Levels of allelic richness (*Ar)* based on a minimal sample size of 11 varied from 1.180 (Reyssouze River, Upper Rhone) to 3.791 (Bethune River, Normandy) and from 1.20 to 3.85 for the mean number of alleles per locus (Table 1, Table S1, Table S2). Levels of *Hnb* averaged over all loci per population also varied substantially, ranging from 0.011 (Reyssouze River) to 0.563 (Aa River, Normandy). Population of the Upper Rhine displayed similar level of diversity to those of Britany (Table 1). On average populations of *L. fluviatilis* were significantly more diversified than populations of *L. planeri* in terms of allelic richness ($p < 0.0042$, 15 000 permutations) and expected heterozygosity ($p < 0.0057$, 15 000 permutations) (See also Figure 4). However there was important regional difference in patterns of genetic diversity. In Normandy, *L. fluviatilis* populations were not different from downstream *L. planeri* populations in terms of expected heterozygosity and relatedness (fdr corrected t.test > 0.05) except for allelic richness ($p = 0.049$) (see table S3 and Figure 4). However, the genetic diversity of *L. fluviatilis* populations was systematically higher than the one of upstream *L. planeri* populations of Normandy (i.e. parapatric) and from the neighboring *L. planeri* populations of Brittany (all $p<0.05$, see tableS3 and Figure 4). Levels of genetic diversity of the Frome (UK) and Shannon (Ireland) populations (Table 1) were similar to those observed in Normandy.

Comparisons among geographical areas revealed a lower genetic diversity of *L. planeri* populations from the Upper Rhône compared to Brittany and Normandy (Figure 4, Table S3).

Genetic differentiation among populations

Global F_{ST} over the full dataset was 0.377 (95%IC = 0.334-0.418). When excluding *L. fluviatilis*, global F_{ST} was 0.394 (95%IC=0.353-0.437) (Table 1). Figure 2 illustrates 2 main groups of populations: the Upper Rhône population *vs* all other populations. Populations of *L. fluviatilis* were weakly differentiated ($F_{ST} = 0.003$). Populations of *L. planeri* were significantly differentiated from populations of *L. fluviatilis* ($p < 0.0003$, 6000 permutations). *L. planeri* populations were generally highly differentiated from one another. The highest F_{ST} was observed between the Reyssouze River and the Moulin du Rocher River (Brittany) with a value of 0.90. The lowest F_{ST} was 0 between several sites (Figure 2 and table S4). Average pairwise F_{ST} between upstream and downstream sites within a river was 0.025, with a minimal value of zero and a

maximal value of 0.095 on the Crano between two sites located near the river source and without obstacles (Table S4). Populations from the Upper Rhine, Frome and Shannon Rivers were moderately differentiated from *L. fluviatilis* (Table S4). The Frome Downstream in particular displayed modest differentiation from *L. fluviatilis.* Pairwise F_{ST} between upstream and downstream sites was significant in 8 out of 43 pairwise comparisons. Results from both Hedrick G_{ST} and Jost D (data not shown) were largely similar to those from Weir and Cockerham F_{ST}.

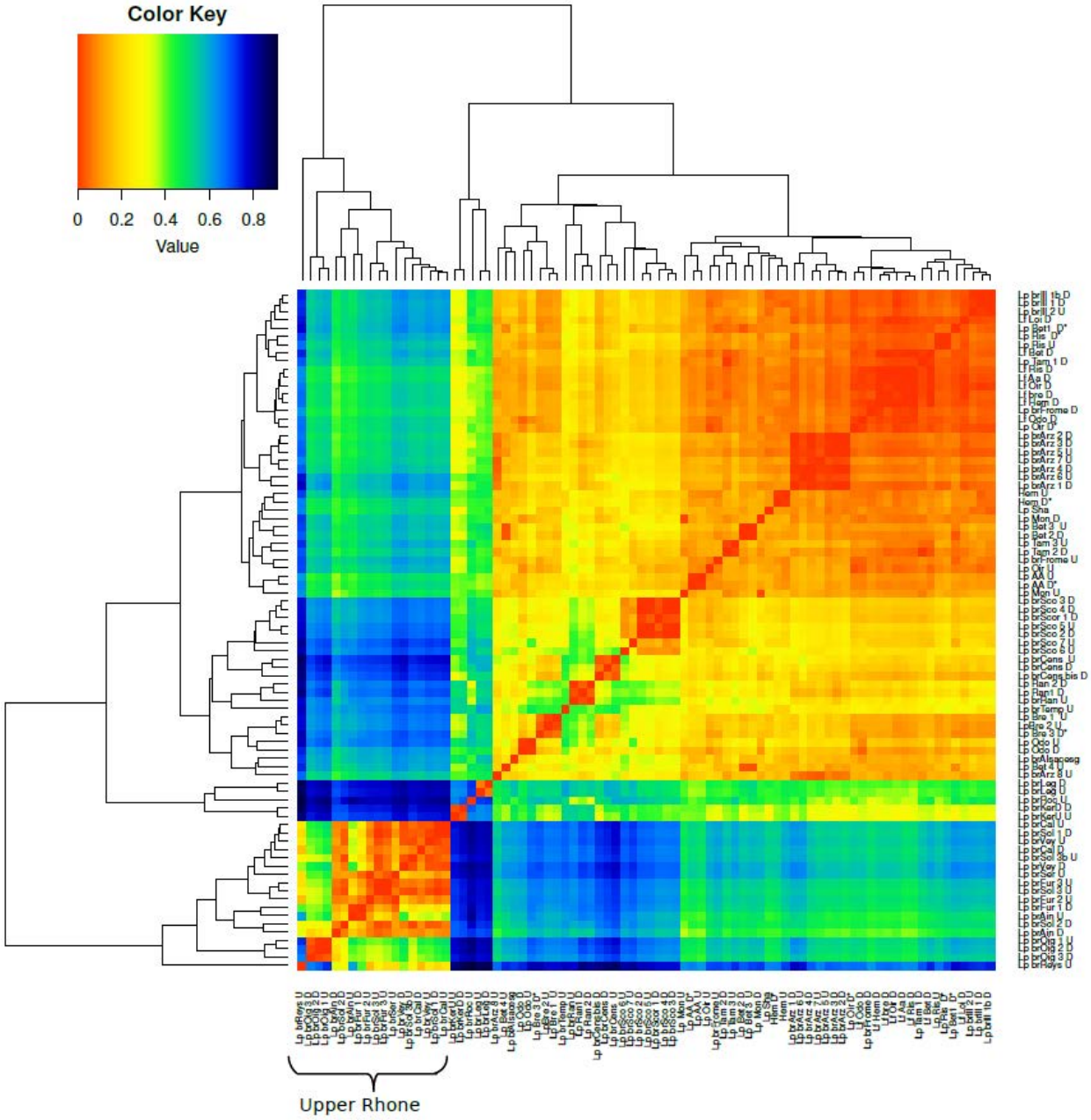


Figure 2: **Heatmap and dendrogram of F_{ST} among the 81 populations.**

Structure analysis indicated an optimal clustering solution for K = 22 (Table S5) based on the mean Ln (K) while the ΔK produced multiple peaks and reach a maximum at K =36. Investigating different levels of population structure in between these values indicated that clustering values in between 20 and 30 were more realistic, while greater values were less biologically interpretable (Figure 2). In this case, Structure analysis indicated two clustering solution, one for K = 4 and one for K = 6. The ΔK produced multiple peaks of small values with the highest peak occurring at K = 6. The clustering solution for K = 4 and K =6 allow to separate the Ange-Oignin river (from the same watershed) from the 5 remaining cluster. All remaining rivers displayed mixed membership contribution to these 5 groups (Figure S1). Plots with increasing values of K (up to six) resulted in increasing admixture values between these populations. For Brittany Lp populations, likelihood values reached a plateau for K values in between 10 and 16 whereas the ΔK produced multiple small peaks, with the highest peak obtained for K =21. In general, results based on mean likelihood values appeared more biologically realistic with K = 10 corresponding to a grouping of all rivers independently, while increasing values of K resulted in increased admixture values within rivers. For Normandy *L. fluviatilis* and *L. planeri* populations, two optimal clustering solutions were found for K = 6 and 8 (table S5). Figure 2 clearly demonstrated that *L. fluviatilis* were admixed whereas *L. planeri* form distinct clusters on the Aa, Hem, Bresle, Bethune and Odon River but were more admixed on the Oir and Risle River. In addition, the downstream population of the Bethune River also displayed mixed membership probabilities.

Table 1: Summary statistics of genetic diversity of *L. planeri* and *L. fluviatilis* populations for each geographic area. N = Number of individuals, NbA = Number of Alleles (averaged of all loci), Ar = Allelic Richness, He = unbiased Expected Heterozygosity, Ho = Observed Heterozygosity

	N	NbA	Ar	He	Ho	F_{ST} [95%IC]
Global	2268	2.73	2.43	0.354	0.344	0.377 [0.334 – 0.418]
L. fluviatilis (Normandy)	209	3.84	3.39	0.505	0.491	0.003 [0 – 0.007]
L. planeri (all regions)	2059	2.65	2.37	0.344	0.334	0.396 [0.353 – 0.437]
Normandy. *L. planeri*	574	3.15	2.88	0.435	0.440	0.139 [0.113 – 0.170]
Britanny & Normandy *L. planeri*	1474	2.94	2.62	0.416	0.411	0.241 [0.207 – 0.284]
Britanny *L. planeri*	910	2.88	2.58	0.406	0.396	0.279 [0.235 – 0.334]
Upper Rhone *L. planeri*	575	1.76	1.52	0.111	0.089	0.249 [0.047 – 0.317]
UK *L. planeri*	83	3.5	2.96	0.476	0.463	NA
Ireland *L. planeri*	48	3.31	2.79	0.458	0.453	NA
Upper Rhine *L. planeri*	36	3.00	2.58	0.355	0.323	NA

Finally, finer analyses of populations structure in the Crano-St Sauveur (n=7 sites) and Arz River (n=8 sites) revealed different patterns of admixture. On the Crano-St Sauveur, two distinct upstream tributaries formed distinct clusters, whereas downstream populations displayed increased admixture values (Figure S2). On the Arz, the source population formed a distinct cluster while downstream populations were admixed (Figure S2).

Landscape genetics

Table 2: Effect of landscape fragmentation on genetic diversity and differentiation between pairs of sites. Mixed Linear Model result based on populations of Northern and Western France (25 pairs of sites)

		Ar Differential					He differential				
model	**effect tested**	**AIC**	**df**	**estimate**	**Chi 2**	**P**	**AIC**	**df**	**estimate**	**Chi 2**	**P**
1		-15.75	11	0.159			-96.06	11	0.044		
2	obstacle height	-14.31	5	-0.006	11.44	***0.0433***	-96.61	6	-0.008	9.452	0.092
3	number of obstacles	-12.30	6	-0.072	13.45	***0.0195***	-96.31	6	-0.034	9.752	0.083
4	distance	-22.85	6	0.004	2.90	0.7152	-97.55	6	-0.001	8.513	0.130
5	distance to the source	-14.01	6	-0.006	11.73	***0.0386***	-96.42	6	-0.0004	9.646	0.086

		Relatedness Differential					Fst/(1-Fst)				
model	**effect tested**	**AIC**	**df**	**estimate**	**Chi 2**	**P**	**AIC**	**df**	**estimate**	**Chi 2**	**P**
1		-66.228	11	-0.044			-88.35	11	0.044		
2	obstacle height	-71.347	6	0.009	4.8807	0.4306	-95.21	6	-0.014	3.144	0.6778
3	number of obstacles	-71.362	6	0.0228	4.8655	0.4325	-94.61	6	-0.007	3.738	0.5877
4	distance	-72.598	6	0.0004	3.6298	0.6038	-96.63	6	-0.00002	1.718	0.8867
5	distance to the source	-72.046	6	0.0015	4.1812	0.5236	-72.05	6	-0.0008	26.3	***7.79E-005***

Tests of isolation by distance indicated a significant relationship between distance and linearized genetic differentiation in the upper Rhone area ($r_{spearman}$= 0.469, p=2e^{-4}). The pattern of isolation by distance was less pronounced in Brittany, but still significant ($r_{spearman}$ = 0.188, p = 0.016). In contrast, isolation by distance was not significant in Normandy (r= 0.145, p = 0.143). This absence of correlation was largely driven by the lack of genetic differentiation between the upstream/downstream populations of the Oir River as compared to the remaining *L. planeri* Normandy populations. When this population was removed, the pattern of IBD was the strongest in Normandy populations ($r_{spearman}$= 0.55, p=1e^{-4}). To gain further insights into the evolutionary relationships among watersheds from coastal areas either connected with river lampreys (i.e. Normandy) or disconnected (i.e. Brittany), we tested the pattern of IBD by keeping one site by river. In this case IBD remained significant in Normandy ($r_{spearman}$= 0.43, p=0.042), while there was no more significant relationship in Brittany ($r_{spearman}$= -0.0208, p=0.53).

Mixed linear models performed on 25 pairs of sites from Normandy and Brittany revealed a significant influence of the number of obstacles and cumulative height on allelic richness differentials (Table 2). The geographic distance between sites did not influence patterns of genetic diversity or differentiation but the distance to the source of the upstream site had a strong influence on allelic richness and genetic differentiation (F_{ST}) (Table 2).

Investigating patterns of downstream increase in genetic diversity revealed a strong influence of the distance to the source on allelic richness (DF= 23, F-value = 46.5, p<0.0001), expected heterozygosity (DF= 23, F-value = 26.2, p<0.0001) and relatedness (DF= 23, F-value = 22.723, p<0.0001).

Mantel tests and partial Mantel tests on the two rivers where a sufficient number of sites were available (Arz and Crano) indicated different influences of distance and obstacle-related variables. On the Arz River, all variables significantly influenced allelic richness (Table 3) whereas it was influenced solely by distance on the Crano. The extent of pairwise differentiation was also influenced by distance in the Crano River whereas this pattern was only revealed in the Arz when the influence of the number of obstacles was controlled for (table 3).

We also investigated the intensity of gene flow between upstream and downstream sites using BayesAss (Wilson & Rannala 2003) but failed to obtain reliable results as confidence intervals were too large (data not shown)

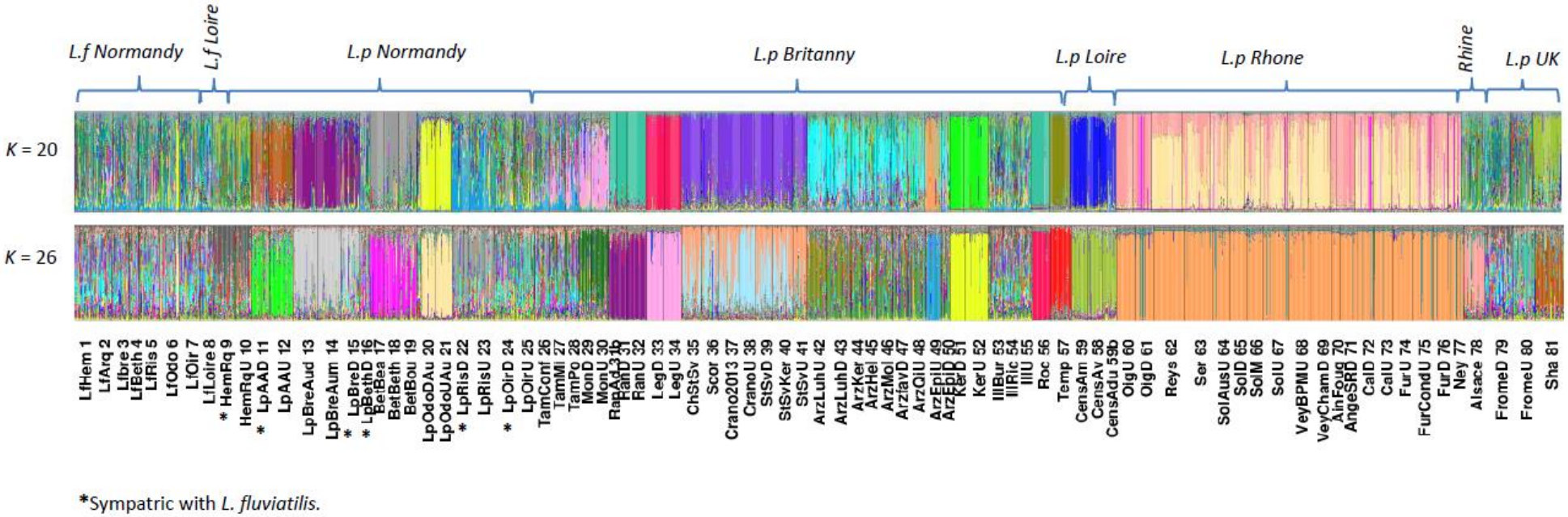


Figure 3: **Patterns of genetic structure inferred from the full dataset**

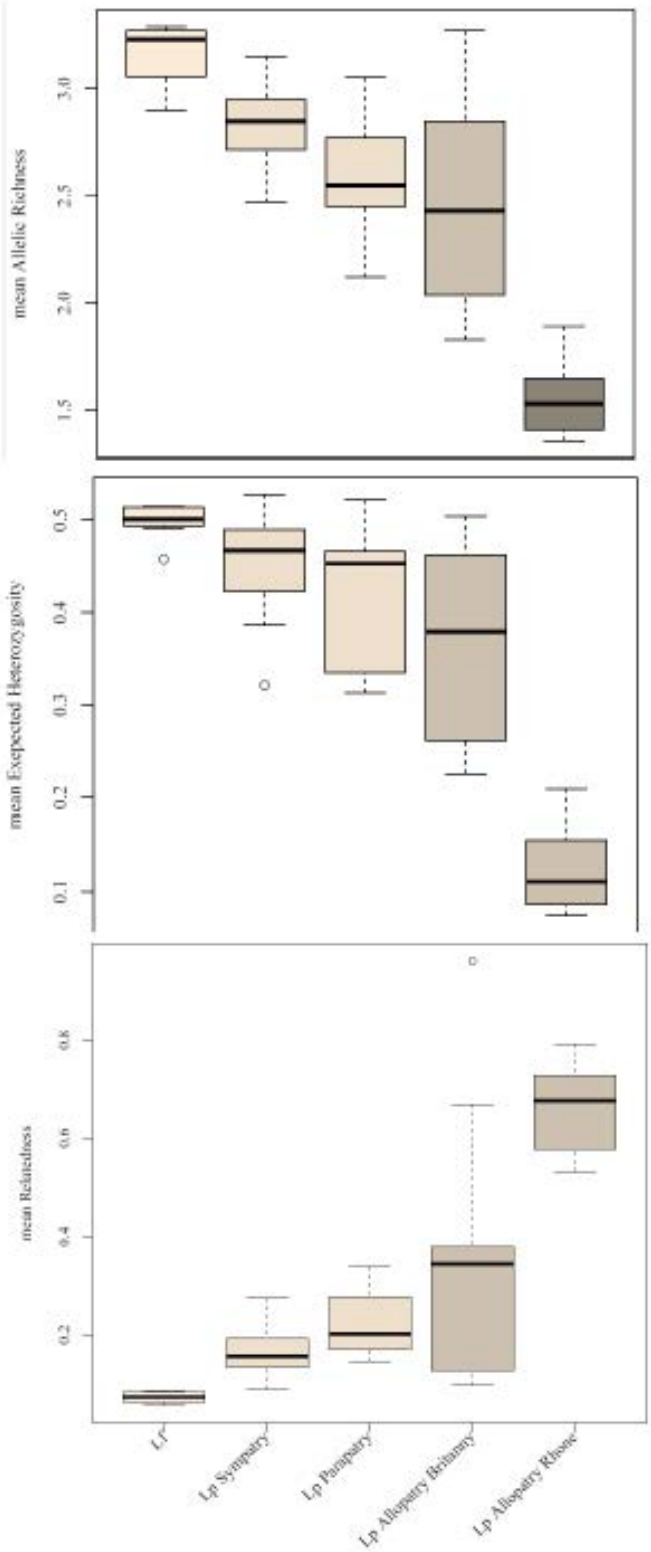


Figure 4: **Comparisons of the mean allelic richness (upper boxplots), expected heterozygosity (middle) and levels of relatedness. Left parts depict *L. fluviatilis.*** The three other pair of boxplot depicts difference in each region and compares upstream and downstream site.

Discussion

The goal of this study was threefold: testing the effect of river fragmentation on patterns of population genetic diversity and structure, investigating the spatial distribution of genetic diversity in *L. planeri* populations at a large scale, and exploring the potential influence of the presence of *L. fluviatilis* on the level of genetic diversity of *L. planeri.* We used *L. planeri* as a model to test the effect of fragmentation as this species displays a reduced migratory behaviour and, in particular, a limited ability to move upstream (Malmqvist, 1983). Our large dataset composed of sympatric, parapatric and allopatric populations of *L. planeri*, located in downstream and isolated areas of different watersheds revealed a key role of *L. fluviatilis* in maintaining genetic diversity of *L. planeri* populations in the lower part of rivers where they co-occur. Strikingly different levels of genetic diversity and population genetic structure were observed in different regions. We found limited evidence for the effect of fragmentation on genetic diversity and differentiation of populations and models indicated that the distance to the source was a more pertinent variable to explain patterns of genetic diversity and relatedness of individuals within populations.

Impacts of river fragmentation on distribution of genetic diversity and differentiation

While several studies have reported strong impact of barriers to migration on either genetic diversity and/or structure (Hänfling & Weetman 2006; Leclerc *et al.* 2008; Raeymaekers *et al.* 2008; Blanchet *et al.* 2010; Faulks *et al.* 2010; Torterotot *et al.* 2014; Gouskov *et al.* 2015), here evidence for "negative" impacts was less obvious.

Investigating the effect of fragmentation on levels of population genetic diversity and structure revealed that population genetic diversity and relatedness was mostly affected by its distance to the source, as upstream populations showed lower levels of allelic richness and heterozygosity and higher levels of relatedness (Figure 4 and Table 5). This so called downstream increase in genetic diversity is expected in riverine habitat (Morrissey & de Kerckhove 2009; Paz-Vinas & Blanchet 2015) and is frequently observed in empirical studies (e.g. Hänfling & Weetman 2006; Torterotot *et al.* 2014; Gouskov *et al.* 2015). Detailed investigations on the Arz River provided strong evidence for an increased downstream allelic richness and this pattern was also significantly influenced by all other physical variables. On the Crano, increase in genetic diversity was not influenced by geographic variables other than distance. A recent simulation study investigated the underlying processes that can generate this pattern (Paz-Vinas & Blanchet 2015). Among the three proposed processes, it appears likely that downstream dispersal plays a key role in *L. planeri*. Indeed the long larval stage of lampreys buried in the soft substrate of river beds (up to five years, Hardisty & Potter, 1971) may lead to their passive downstream drift (Dawson et al. 2015) during flood events,

which may explain the increased diversity downstream. The low genetic diversity and high genetic differentiation of the populations collected close to the source on the Arz, Crano, St Sauveur and Bethune Rivers support this hypothesis and suggest that these upstream populations form source populations. Bayesian clustering analysis (fig S2) of the St-Sauveur-Crano river system (the Crano is a small stream flowing in the St Sauveur) revealed another important pattern explaining the increase in downstream genetic diversity *via* admixture among individuals originating from different upstream sites. The two upstream populations of the St Sauveur and Crano Rivers form two genetically distinct clusters (F_{ST} = 0.265*)* and individuals located downstream the Crano appear admixed, probably having a shared ancestry stemming from these two source populations (and possibly from other unsampled populations). Last, finer analysis of the effect of distance and number of obstacles revealed exponential decrease in relatedness (and exponential increase with allelic richness) with distance from the source (Figure S4). A simple log linear model captured more variance (R^2=41%) than did the linear model (R^2 = 30%) which is another evidence suggesting that individuals within the most upstream populations are more related. These populations may display small effective population sizes (estimations of *Ne* produced too large confidence intervals, results not shown) and eventually suffer from inbreeding. The second process that may have generated low upstream genetic diversity is the occurrence of bottlenecks following usptream river colonization after glacial retreats (Hewitt 1996; Taberlet *et al.* 1998). It remains unclear so far whether freshwater populations have recovered from ancestral bottlenecks and disentangling the two hypotheses will require further data.

As previously suggested, independent drift of resident populations can explain our inability to detect any global signal. A pattern of independent drift of freshwater populations was expected as demonstrated in a simulation study of river colonization of freshwater and marine *Gasterosteus aculeatus* populations (Bierne et al. 2013). Note also that the strongest differentiation between upstream and downstream sites occurred on the Crano river (F_{ST} = 0.095) between two sites located in the headwater of the river, but not separated by any physical barriers. Finer investigations on the Crano and the Arz revealed a significant effect of distance on differentiation in the Crano River suggesting migration-drift equilibrium in this river. On the Arz River on the contrary, the effect of distance was only revealed when controlled for the effect of obstacle number. Finally, our results do not allow drawing any general conclusion on the impact of obstacle height and number on the extent of genetic differentiation. Such impacts may be best revealed by focusing on a single catchment and with bigger obstacles to migration (eg Raeymaekers *et al.* 2008; Blanchet *et al.* 2010; Gouskov *et al.* 2015). In addition, we investigated the impact of obstacles of small to moderate size and it is possible that these obstacles do not influence the downstream passive drift of lamprey larvae, which may be sufficient to homogenize populations and obscure patterns of differentiation (Faubet *et al.* 2007).

River lamprey as a source of genetic diversity for resident lampreys

Understanding the evolutionary relationship between parasitic and nonparasitic lamprey is a long standing debate (Docker, 2009; 2015). Recent evidence (Bracken *et al.* 2015; Rougemont *et al.* 2015a) have shown that gene flow is ongoing between *L. fluviatilis* and *L. planeri,* lowering their level of genetic differentiation at a genome-wide scale. Notably, Rougemont et al (2015b) suggested the occurrence of asymmetric introgression from anadromous to freshwater populations following secondary contact. Such introgression from a large marine population toward freshwater populations is also known to occur in *Gasterosteus aculeatus* (Hohenlohe *et al.* 2010, 2012). Here, genetic analyses of populations of *L. planeri* and *L. fluviatilis* in sympatry (on the same nest), in parapatry (where the two species co-occur in the same watershed but are geographically separated by impassable dams) and in allopatry (in coastal rivers where *L. fluviatilis* is absent) revealed that allopatric *L. planeri* lamprey displayed a lower genetic diversity than sympatric and parapatric populations (Figure 4). Additionally, Bayesian clustering analysis revealed higher levels of admixture in sympatry than in allopatry and parapatry (Figure 3), confirming findings of contemporary introgression (Espanhol *et al.* 2007; Bracken *et al.* 2015; Rougemont *et al.* 2015). In contrast, the Bayesian clustering analysis also confirmed that each *L. planeri* population formed an independent genetic cluster. In addition, we found a significant pattern of IBD in the connected pairs of *L. planeri* (i.e. populations of downstream areas in Normandy) whereas populations from Brittany were not at migration-drift equilibrium. This result further suggests that the current genetic makeup of *L. planeri* populations in Normandy is influenced by ongoing gene flow with *L. fluviatilis*. In the absence of inter-basin gene flow mediated by *L. fluviatilis,* populations of Brittany evolved independently from each other and are not globally at migration drift equilibrium (which does not imply that populations within rivers deviate from this equilibrium). Populations from the Upper Rhone area, together with the few populations from the Rhine, Ireland and United Kingdom were also very informative with regards to the historical divergence of the two forms. First, populations from Ireland, the United Kingdom and the Rhine displayed moderate levels of genetic differentiation, suggesting that these populations have undergone gene exchange with *L. fluviatilis* and share a more recent history with populations of *L. planeri* from Normandy, than did populations of *L. planeri* from Brittany. In contrast, all populations from the Upper Rhone area displayed a highly reduced genetic diversity and were strongly differentiated (e.g. Fig 2) from all other populations. Different complementary hypotheses can be made to explain such a result. First, there is evidence for at least three major evolutionary lineages existing in *L. planeri* (Espanhol *et al.* 2007). It is thus possible that colonization of the Mediterranean area (Upper Rhone region) following postglacial colonization (the usual pattern in European fish species, Bernatchez & Wilson 1998) was due to a different lineage than the one having colonized the Atlantic and Channel areas. In these conditions it is possible that our

microsatellite marker set (originally developed using *L. planeri* and *L. fluviatilis* samples from the Atlantic and channel area) is not the most appropriate to perform accurate population genetic inference of Rhône samples. The low performance of Structure analysis confirms this lack of power. Second, *L. fluviatilis* no longer colonises this area and was already reported to be declining during the last century (Bernard 1909; Gensoul 1907). Consequently, it is possible that the history of divergence between Mediterranean and Atlantic populations was initiated a long time ago and that gene flow between neighbouring rivers of the Mediterranean area was further reduced during the last century.

Conservation implications

Fragmentation of rivers may impact lamprey populations, especially the most upstream populations that do not receive migrants from downstream sites. Whether the most isolated populations from headwaters suffer a mutation load and greater extinctions risk would require further investigations. It is now known from theory and empirical data that small isolated populations suffer greater inbreeding and extinction risks (Lynch 1991; Higgins & Lynch 2001; Spielman *et al.* 2004; Frankham 2005, 2015). On the other hand it is not clear if maintaining a possibility for upstream migration by removing obstacles may help prevent the loss of genetic diversity in the source populations of *L. planeri* through the beneficial effects of gene flow (Frankham 2015). The source populations on the Crano and St Sauveur, as well as the observed genetic differentiation on the Tamoute River despite the absence of migratory barrier illustrate this problem well.

Importantly, our study revealed positive impacts of the presence of *L. fluviatilis* in maintaining genetic diversity in sympatric populations of *L. planeri*. However, in Europe, the *L. fluviatilis* abundance has strongly declined in some areas (Maitland et al. 2015) and it is now considered as Vulnerable in France in the IUCN red list (UICN France et al. 2010). In addition the low upstream migratory ability of the anadromous ecotype often restricts its distribution to downstream areas where *L. planeri* are often less abundant. In terms of conservation priority it appears fundamental to first ensure that *L. fluviatilis* will have access to upstream reaches of rivers. This will benefit both the river and *L. planeri* in sympatric and parapatric areas. In these areas a joint regional management of the two ecotypes could be envisioned, whereas in allopatric areas, a management at the river scale may be more parsimonious.

Conclusion

We have shown here that impacts of barriers to migration were modest on the extent of genetic differentiation, but we provided evidence that headwater populations of *L. planeri* displayed reduced genetic diversity, high levels of relatedness and were the most genetically differentiated. Given the strong asymmetric downstream gene flow (probably due to passive drift of larvae) it is not clear whether restoring the possibility for upstream migration could have beneficial impact through facilitating gene flow from downstream populations. In addition, we have shown that populations of *L. planeri* in sympatric areas displayed higher levels of genetic diversity probably due to introgression from *L. fluviatilis*. Potential strong gene flow or even genome swamping from anadromous population to resident populations is fundamental in maintaining genetic diversity of *L. planeri*. In addition it may play a key role through adaptive introgression and also through the movement of freshwater adaptive alleles between adjacent rivers (i.e. the transporter hypothesis (Schluter & Conte 2009), a topic that will definitely require further investigations. Further investigations about the effective population size of populations along river networks based on genome wide data and tests of mutation load in isolated populations may provide additional cues for conservation management purposes.

Table S1: Characteristics of each obstacle

Code_ROE	River	Area	X_L2E	Y_L2E	Obstacle Name	height	remarks
ROE25520	Montafilan	Britanny	265308	2397805	Camboeuf	1	
NA	Rance	Britanny	245314	2374669	Chaos Quemelin	0.8	
ROE39631	Saint Emilion	Britanny	166120	240980	Etang Beffou	2.8	
ROE32742	St Sauveur	Britanny	171360	233878	Moulin de Tronchateau	3.5	
ROE32752	St Sauveur	Britanny	172165	2334535	Moulin de Mélien	1.3	
ROE32765	St Sauveur	Britanny	173067	2336925	Moulin du Moustoir	0.6	
ROE327936	St Sauveur	Britanny	173783	2337323	Moulin de Restraudan	0	
ROE32891	St Sauveur	Britanny	174115	2337490	Moulin de Kerviden	0.25	
ROE232898	St Sauveur	Britanny	174260	2337575	Moulin de Kerviden	1.2	
ROE32907	St Sauveur/Crano	Britanny	176120	2337455	Prise d'eau pont en daul	0	removed
ROE12583	St Sauveur/Crano	Britanny	176620	2337155	Prise d'eau kerhault	0	removed
ROE32910	St Sauveur	Britanny	176935	2339390	Moulin de Becherel	3	
ROE32911	St Sauveur	Britanny	177450	2340353	Moulin de Malachappe	0.2	
ROE32924	Kerousseau	Britanny	169665	2328173	Moulin de Kerousseau	1.85	
ROE40867	l'Arz	Britanny	223686	2319252	Moulin de Luhan	0.85	
NA	l'Arz	Britanny	229200	2318057	Moulin de Kerfily	0.6	removed
ROE15566	l'Arz	Britanny	233819	2316207	Moulin du Helfau	0.4	
ROE15866	l'Arz	Britanny	236914	2314780	Moulin du Pont de Molac	0	removed
ROE11672	l'Arz	Britanny	239735	2313380	Moulin de Larré	0.6	
ROE11666	l'Arz	Britanny	241203	2312630	Moulin de l'échange	0	removed
ROE67738	l'Arz	Britanny	242436	2312782	Seuil de Jeaugeage du Pont du Favre	0.25	
ROE11661	l'Arz	Britanny	244499	2312319	Moulin du Bois Bréhan	0.6	
ROE11652	l'Arz	Britanny	245574	2312462	Moulin de Bragou	1.5	
ROE11641	l'Arz	Britanny	248223	2311557	Moulin de Quenelet	0.8	
ROE11637	l'Arz	Britanny	248765	2311882	Clapet de la ville Boury	0	removed
ROE11632	l'Arz	Britanny	249782	2311738	Moulin d'Arz	0.65	
ROE62238	l'Arz	Britanny	250405	2311378	Ancien moulin de quiquéma	0	removed
ROE11626	l'Arz	Britanny	231416	2311280	Moulin de l'éthier	0.3	
ROE11619	l'Arz	Britanny	253659	2310760	Moulin du Quiban	1.1	
ROE11479	l'Arz	Britanny	255312	2311194	Gué de l'épine	0.85	
NA	Moulin du Rocher	Britanny	266169	2293175	NA	3	
ROE41477	Kergroix	Britanny	195130	2324501	Moulin Castellin	0.4	
ROE13091	Oir	Normandy	338931	2410294	Mongothier	1.3	
ROE8503	Oir	Normandy	333847	2409329	Cerisel	2	
ROE12912	Oir	Normandy	333298	2409315	Moulins des geins	1.5	
ROE69925	Cens	Atlantic	296962	2259848		1.25	
ROE18864	Illet	Britanny	3111647	2369451	Moulin de Piguel	0	removed
ROE22554	Illet	Britanny	311023	2369195	Moulin de la Hurlais	0	removed
ROE 22520	Odon	Normandy	382038	2451484	NA	1.97	
ROE27745	Risle	Normandy	472498	2483963	Ouvrage le foll	2.36	
ROE27745	Risle	Normandy	472498	2483963	Seuil des échauds	0.7	
ROE233	Risle	Normandy	471746	2484213	Ouvrage les 3 Moulins	0.75	
ROE44020	Bresle	Normandy	557144	2528609	Microcentrale de la chaussée	2.1	
ROE15296	Bethune	Normandy	543078	2520496	Moulin à huile	2.69	
ROE15264	Bethune	Normandy	591416	6956015	Moulin de Saint Saire	1.5	
ROE15278	Hem	Normandy	582197	2648778	Minoterie de Recques sur Hem	3.5	
ROE15259	Hem	Normandy	582867	2650269	Moulin Bleu	0.7	
ROE35628	Aa	Normandy	596555	2636417	Moulin Snick	0.8	
ROE34472	Aa	Normandy	596214	2636130	Moulin Marin	0.4	
ROE27362	Aa	Normandy	595480	2635993	Moulin Fer Blanc	0.25	
ROE27360	Aa	Normandy	595098	2635917	Moulin de Westhore	0.2	
ROE27357	Aa	Normandy	594336	2635310	Moulin de Wins	2	
ROE35613	Aa	Normandy	593728	2634920	Ancien Moulin	0.2	
ROE27354	Aa	Normandy	593316	2635084	Seuil de Gondardene	NA	
ROE27353	Aa	Normandy	592675	2635011	Le choquet	0.1	

ROE27349	Aa	Normandy	591909	2634711	Paepterie	2.5	
ROE27345	Aa	Normandy	591909	2634711	Seuil du Cours leulieux 2	0.6	
ROE27344	Aa	Normandy	591077	2634659	Moulin Pidoux	1	
ROE27343	Aa	Normandy	590682	2634481	A26	0.8	
ROE27339	Aa	Normandy	590375	2634608	Moulin de Confosse 1	2.5	
ROE27341	Aa	Normandy	590297	2634567	Moulin de Confosse 2	1.6	
ROE27364	Aa	Normandy	588619	2634767	Vannage de Ferssinghen	1.8	
ROE27365	Aa	Normandy	588603	2634932	Moulin Colbert	0.2	
ROE27367	Aa	Normandy	587951	2635053	Seuil Pourdrerie	0.25	
ROE27376	Aa	Normandy	587295	2635039	Ancien Moulin Roland	NA	
ROE48991	Ange-Oignin	Upper rhone	895980	6562533	Station Jeaugeage Diren	0.8	
ROE48995	Ange-Oignin	Upper rhone	896223	6.6E+07	Bassins Autoroute	2	
ROE48997	Ange-Oignin	Upper rhone	896877	6566296	Radier de Pont du Mollard	0.3	
ROE41251	Veyle	Upper rhone	820719	2129572	Bel Air 03	0.32	
ROE46812	Veyle	Upper rhone	820724	2129689	Bel Air 02	0.28	
ROE41498	Veyle	Upper rhone	820672	2129784	Bel Air 02	0.19	
ROE41207	Veyle	Upper rhone	820741	2129884	Moulin Longchamp	0.28	
ROE46810	Veyle	Upper rhone	820393	2130170	Moulin Maillet01	0.73	
ROE41490	Veyle	Upper rhone	820440	2130644	Moulin Maillet02	0.73	
ROE41496	Veyle	Upper rhone	820475	2132075	Gourion Voie SNCF	0.35	
ROE46808	Veyle	Upper rhone	820201	2132181	Colon	0.56	
ROE46807	Veyle	Upper rhone	819910	2132932	La fretaz	1.57	
ROE46805	Veyle	Upper rhone	819707	2133858	La Vermée	0.44	
ROE41505	Veyle	Upper rhone	NA	NA	La Viergé	0	removed
ROE41508	Veyle	Upper rhone	NA	NA	Granges Blanches 03	0	removed
ROE41510	Veyle	Upper rhone	819305	2135380	Granges Blanches 01	0.1	
ROE41512	Veyle	Upper rhone	NA	NA	Granges Blanches 02	0	removed
ROE46804	Veyle	Upper rhone	819149	2136908	Le Chatelard Moulin Neuf	0.4	
ROE46803	Veyle	Upper rhone	819019	2137098	Moulin Neuf	1.6	
ROE46802	Veyle	Upper rhone	NA	NA	Les combes	0	removed
ROE41530	Veyle	Upper rhone	NA	NA	Confluence gravier	0	removed
ROE46801	Veyle	Upper rhone	NA	NA	Barrage Chamambard	0	removed
ROE46800	Veyle	Upper rhone	817437	2139890	Barrage Chamambard	1	
ROE46799	Veyle	Upper rhone	816964	2140464	Barrage Moumin de Loyasse	0.95	
ROE54149	Solnan	Upper rhone	830594	2153393	Prairie de Presle	0.5	
ROE54150	Solnan	Upper rhone	830789	2153540	Les cailloux	0.5	
ROE564305	Solnan	Upper rhone	829248	2152253	Prise d'eau du Moulin des Ponts	1	
ROE564306	Solnan	Upper rhone	829334	2152223	Moulin des Ponts	1	
ROE564304	Solnan	Upper rhone	NA	NA	Barrage du Moulin des Ponts	0	removed
ROE27801	Calonne	Upper rhone	790063	2125213	Moulin Crozet	1.3	
ROE27807	Calonne	Upper rhone	790151	2125265	Moulin Crozet	1.9	
ROE42374	Furans	Upper rhone	NA	NA	Passage à Gué d'Arbigneu	0	removed
ROE42376	Furans	Upper rhone	858555	2087896	Ancien Moulin du Pont	0.6	
ROE42379	Furans	Upper rhone	NA	NA	Barrage de la pie	0	removed
ROE42381	Furans	Upper rhone	859609	2095141	Passage à Gué voie ferré	0.4	
ROE42387	Furans	Upper rhone	859106	2095496	barrage Roissey	1.1	
ROE42390	Furans	Upper rhone	858250	2095721	Cheminet	1	
ROE48987	Furans	Upper rhone	857858	2096360	Pris d'eau pisciculture	2.1	

Table S2: synthesis of population genetic diversity indices

Pop	Area	river	sympatry	id	N	NbA	Ar	He	Ho	Fis	Kinship
Lf Hem	Normandy	Hem	yes	1	30	3.77	3.29	0.504	0.497	0.013	0.072
Lf Arq	Normandy	Aa	yes	2	34	3.85	3.26	0.514	0.504	0.019	0.059
Lf bre	Normandy	bresle	yes	3	30	3.69	3.24	0.497	0.476	0.044	0.057
Lf Beth	Normandy	bethune	yes	4	17	3.54	3.28	0.514	0.476	0.078	0.086
Lf Ris	Normandy	risle	yes	5	35	3.77	3.22	0.515	0.503	0.024	0.066
Lf Odo	Normandy	odon	no	6	30	3.77	3.11	0.491	0.470	0.042	0.082
Lf Oir	Normandy	oir	yes	7	30	3.46	3.00	0.498	0.514	-0.034	0.073
Lf Loire	Normandy	loire	no	8	19	3.08	2.90	0.458	0.465	-0.015	0.084
HemRq	Normandy	Hem	yes	9	39	3.46	2.86	0.477	0.469	0.017	0.164
HemRqU	Normandy	Hem	no	10	26	2.92	2.71	0.471	0.504	-0.071	0.172
LpAAD	Normandy	Aa	yes	11	30	3.23	2.94	0.528	**0.563**	-0.068	0.148
LpAAU	Normandy	Aa	no	12	39	3.46	3.05	0.522	0.492	0.059	0.168
LpBreAud	Normandy	bresle	no	13	40	2.92	2.48	0.335	0.335	0.000	0.295
LpBreAum	Normandy	bresle	no	14	40	2.54	2.36	0.335	0.321	0.044	0.276
LpBreD	Normandy	bresle	yes	15	31	2.77	2.47	0.321	0.351	-0.095	0.273
LpBethD	Normandy	bethune	yes	16	17	3.38	3.15	0.472	0.482	-0.023	0.116
BetBea	Normandy	bethune	no	17	23	2.62	2.45	0.440	0.428	0.026	0.266
BetBeth	Normandy	bethune	no	18	25	2.77	2.55	0.453	0.452	0.001	0.201
BetBou	Normandy	bethune	no	19	30	3.15	2.84	0.462	0.446	0.034	0.158
LpOdoDAu	Normandy	odon	no	20	30	3.00	2.59	0.386	0.403	-0.046	0.221
LpOdoUAu	Normandy	odon	no	21	30	2.23	2.12	0.313	0.352	-0.124	0.338
LpRisD	Normandy	risle	no	22	28	3.08	2.84	0.460	0.435	0.055	0.152
LpRisU	Normandy	risle	no	23	43	3.77	3.05	0.466	0.435	0.066	0.145
LpOirD	Normandy	Oir	yes	24	34	3.31	2.96	0.503	0.538	-0.071	0.089
LpOirU	Normandy	Oir	no	25	31	3.00	2.77	0.458	0.466	-0.018	0.159
TamConf	Britanny	Tamoute	no	26	25	3.46	3.14	0.502	0.465	0.076	0.082
TamMi	Britanny	Tamoute	no	27	26	3.31	3.05	0.486	0.445	0.087	0.145
TamPo	Britanny	Tamoute	no	28	24	3.31	3.07	0.444	0.417	0.064	0.185
MonD	Britanny	Montafillan	no	29	20	2.77	2.69	0.453	0.492	-0.089	0.128
MonU	Britanny	Montafillan	no	30	29	3.00	2.74	0.479	0.459	0.043	0.134
RanAd	Britanny	Rance	no	31b	13	2.23	2.21	0.339	0.362	-0.074	0.348
RanD	Britanny	Rance	no	31	17	2.08	2.04	0.326	0.327	-0.003	0.378
RanU	Britanny	Rance	no	32	32	2.15	1.95	0.300	0.324	-0.082	0.398
LegD	Britanny	Leguer	no	33	30	2.46	2.07	0.239	0.178	***0.259***	0.960
LegU	Britanny	Leguer	no	34	29	1.92	1.77	0.245	0.224	0.087	0.916
ChStSv	Britanny	Scorff	no	38	32	2.62	2.43	0.438	0.438	0.002	0.387
Scor	Britanny	Scorff	no	41	30	3.15	2.65	0.419	0.354	0.157	0.340
Crano2013	Britanny	Scorff	no	37	35	2.62	2.34	0.421	0.444	-0.055	0.455
CranoU	Britanny	Scorff	no	36	25	2.23	2.13	0.367	0.408	-0.112	0.470
StSvD	Britanny	Scorff	no	40	30	2.62	2.37	0.418	0.406	0.029	0.434
StSvKer	Britanny	Scorff	no	39	33	2.54	2.38	0.438	0.469	-0.071	0.420
StSvU	Britanny	Scorff	no	35	24	2.31	2.20	0.369	0.359	0.028	0.450
ArzLuhU	Britanny	Arz	no	42	32	2.85	2.57	0.441	0.459	-0.041	0.199

ArzLuhD	Britanny	Arz	no	43	40	3.38	2.91	0.479	0.471	0.016	0.151
ArzKer	Britanny	Arz	no	44	17	3.08	2.90	0.494	0.549	-0.115	0.172
ArzHel	Britanny	Arz	no	45	26	3.15	2.96	0.518	0.503	-0.03	0.077
ArzMol	Britanny	Arz	no	46	26	3.15	2.92	0.493	0.445	0.118	0.142
ArzfavD	Britanny	Arz	no	47	26	3.46	3.08	0.516	0.477	0.078	0.109
ArzQiU	Britanny	Arz	no	48	30	3.69	3.33	0.524	0.448	0.148	0.106
ArzEpiU	Britanny	Arz	no	49	27	3.69	3.30	0.454	0.384	0.156	0.097
ArzEpiD	Britanny	Arz	no	50	13	3.31	3.27	0.504	0.466	0.078	0.669
KerD	Britanny	Kergroix	no	51	25	2.00	1.83	0.225	0.228	-0.016	0.589
KerU	Britanny	Kergroix	no	52	39	2.46	1.95	0.243	0.237	0.023	0.113
IllBur	Britanny	Illet	no	53	25	3.08	2.85	0.462	0.450	0.026	0.126
IllRic	Britanny	Illet	no	54	22	3.15	2.97	0.474	0.490	-0.034	0.128
IllU	Britanny	Illet	no	55	26	3.00	2.78	0.463	0.454	0.021	0.630
Roc	Britanny	Rocher	no	56	31	1.54	1.40	0.105	0.131	-0.252	0.470
Temp	Britanny	Temple	no	57	34	2.62	2.27	0.330	0.320	0.030	0.387
CensAm	Atlantic	Cens-Loire	no	58	27	2.31	2.01	0.239	0.224	0.060	0.363
CensAv	Atlantic	Cens-Loire	no	59	24	2.23	1.99	0.261	0.224	0.144	0.305
CensAdu	Atlantic	Cens-Loire	no	58b	25	2.54	2.25	0.297	0.323	-0.087	0.687
OigU	Upper Rhone	Oignin	no	60	30	1.23	1.18	0.055	0.050	0.088	0.614
OigD	Upper Rhone	Oignin	no	61	30	1.69	1.54	0.110	0.045	***0.598***	0.918
Reys	Upper Rhone	Reysouze	no	62	51	1.38	1.17	0.019	0.011	0.447	0.735
Ser	Upper Rhone	Séran	no	63	47	1.77	1.41	0.110	0.101	0.084	0.610
SolAusU	Upper Rhone	Solnan	no	64	28	1.77	1.59	0.127	0.110	0.140	0.696
SolD	Upper Rhone	Solnan	no	65	29	1.69	1.44	0.091	0.076	0.167	0.533
SolM	Upper Rhone	Solnan	no	66	26	1.85	1.70	0.171	0.139	0.188	0.722
SolU	Upper Rhone	Solnan	no	67	44	1.85	1.66	0.137	0.121	0.122	0.728
VeyBPMU	Upper Rhone	Veyle	no	68	36	1.69	1.50	0.101	0.090	0.112	0.791
VeyChamD	Upper Rhone	Veyle	no	69	35	1.54	1.36	0.081	0.073	0.105	0.724
AinFoug	Upper Rhone	Ain	no	70	13	1.38	1.36	0.102	0.110	-0.074	0.661
AngeSRD	Upper Rhone	Oignin	no	71	30	1.62	1.38	0.074	0.068	0.081	0.761
CalD	Upper Rhone	Calonne	no	72	30	1.69	1.52	0.111	0.057	***0.495***	0.765
CalU	Upper Rhone	Calonne	no	73	30	1.62	1.39	0.085	0.062	0.274	0.707
FurU	Upper Rhone	Furans	no	74	32	1.54	1.47	0.134	0.111	0.170	0.714
FurCondU	Upper Rhone	Furans	no	75	33	2.08	1.74	0.142	0.114	0.198	0.692
FurD	Upper Rhone	Furans	no	76	30	1.85	1.60	0.139	0.123	0.113	0.539
Ney	Upper Rhone	Ain	no	77	21	2.08	1.89	0.209	0.137	***0.349***	
Alsace	Upper Rhine	Fischbaechel	no	78	36	3.00	2.58	0.355	0.323	0.090	

FromeD	UK	Frome	no	79	48	3.69	3.02	0.490	0.505	-0.031
FromeU	UK	Frome	no	80	35	3.31	2.92	0.463	0.422	0.091
Sha	Ireland	Shannon	no	81	48	3.31	2.79	0.458	0.453	0.011

Table S3: Population genetic diversity for each locus

See electronic files

Table S4: Results of pairwise t.test for difference in Allelic Richness, Expected Heterozygosity and relatedness (FDR adjusted) a= *L. fluviatilis, b= L. planeri* downstream (Normandy), *c = L. planeri upstream (Normandy), d= L. planeri* downstream (Britanny), *e = L. planeri upstream (Britanny), f = L. planeri* downstream (Ain), *g = L. planeri upstream (Ain),*

	Allelic Richness					
pop	a	b	c	d	e	f
b	0.51433	-	-	-	-	-
c	0.10327	0.3435	-	-	-	-
d	**0.00138**	**0.01077**	0.10327	-	-	-
e	**0.00013**	**0.00127**	**0.0179**	0.45148	-	-
f	**2.70E-013**	**1.70E-012**	**2.10E-011**	**4.40E-009**	**7.00E-008**	-
g	**6.30E-014**	**2.90E-013**	**3.20E-012**	**8.00E-010**	**1.50E-008**	0.9626

	Expected Heterozygosity					
pop	a	b	c	d	e	f
b	0.04888	-	-	-	-	-
c	**0.00123**	0.18619	-	-	-	-
d	**0.000036**	**0.02123**	0.29686	-	-	-
e	**2.70E-008**	**0.000058**	**0.00331**	**0.04355**	-	-
f	**3.20E-013**	**3.30E-010**	**1.70E-008**	**3.70E-007**	**3.60E-004**	-
g	**7.50E-015**	**6.70E-012**	**3.30E-010**	**8.30E-009**	**1.60E-005**	0.48228

	Relatedness Coefficient					
pop	a	b	c	d	e	f
b	0.28943	-	-	-	-	-
c	0.07053	0.44487	-	-	-	-
d	**0.0009**	**0.02244**	0.1057	-	-	-
e	**0.000038**	**0.00152**	**0.00999**	0.30633	-	-
f	**5.10E-009**	**2.60E-007**	**1.90E-006**	**2.90E-004**	**0.00544**	-
g	**9.90E-011**	**5.10E-009**	**2.30E-008**	**5.40E-006**	**2.00E-004**	0.37356

Table S5: Matrix of Pairwise fst:

See electronic Supplementary files

Table S6: K and Delta K Values in the different areas

Area	K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
Whole	2	10	-51236.55	33.6364	NA	NA	NA
Whole	3	10	-49833.71	116.8824	1402.84	21.43	0.183347
Whole	4	10	-48452.3	137.2631	1381.41	351.52	2.560922
Whole	5	10	-47422.41	360.665	1029.89	265.28	0.73553
Whole	6	10	-46657.8	588.9301	764.61	95.9	0.162838
Whole	7	10	-45797.29	266.753	860.51	23.33	0.087459
Whole	8	10	-44960.11	331.2437	837.18	395.06	1.192657
Whole	9	10	-44517.99	81.8777	442.12	90.94	1.110681
Whole	10	10	-44166.81	115.0596	351.18	71.17	0.618549
Whole	11	10	-43886.8	142.9302	280.01	115.45	0.807737
Whole	12	10	-43491.34	202.6882	395.46	193.69	0.955606
Whole	13	10	-43289.57	150.8811	201.77	24.26	0.160789
Whole	14	10	-43063.54	238.337	226.03	113.46	0.476049
Whole	15	10	-42724.05	167.4499	339.49	175.55	1.048373
Whole	16	10	-42560.11	204.1735	163.94	126.06	0.617416
Whole	17	10	-42270.11	89.252	290	282.26	3.162506
Whole	18	10	-42262.37	152.8393	7.74	190.38	1.245622
Whole	19	10	-42445.01	416.2834	-182.64	402.08	0.965881
Whole	**20**	**10**	**-42225.57**	**313.3949**	**219.44**	**318.88**	**1.017502**
Whole	21	10	-42325.01	304.1504	-99.44	207.21	0.681275
Whole	**22**	**10**	**-42217.24**	**297.217**	**107.77**	**159.32**	**0.536039**
Whole	23	10	-42268.79	230.2911	-51.55	23.28	0.101089
Whole	24	10	-42343.62	306.2431	-74.83	132.16	0.431553
Whole	25	10	-42286.29	257.4501	57.33	45.48	0.176656
Whole	**26**	**10**	**-42274.44**	**263.107**	**11.85**	**767.64**	**2.917597**
Whole	27	10	-43030.23	1355.9896	-755.79	1067.46	0.787218
Whole	28	10	-42718.56	308.079	311.67	595.44	1.932751
Whole	29	10	-43002.33	671.5214	-283.77	482.31	0.718235
Whole	30	10	-42803.79	206.7787	198.54	445.44	2.154187
Whole	31	10	-43050.69	448.5969	-246.9	286.83	0.639394
Whole	32	10	-43010.76	334.3169	39.93	283.91	0.849224
Whole	33	10	-43254.74	283.8797	-243.98	142.32	0.501339

Whole	**34**	**10**	**-43356.4**	**375.714**	**-101.66**	**2201.63**	**5.859855**
Whole	35	10	-45659.69	6538.803	-2303.29	4299.5	0.657536
Whole	**36**	**10**	**-43663.48**	**264.3743**	**1996.21**	**2102.46**	**7.952588**
Whole	37	10	-43769.73	230.296	-106.25	179.31	0.778607
Whole	38	10	-44055.29	508.8362	-285.56	2554	5.019297
Whole	39	10	-46894.85	8688.7813	-2839.56	5063.22	0.582731
Whole	40	10	-44671.19	283.4375	2223.66	2229.82	7.867062
Whole	41	10	-44677.35	472.3086	-6.16	127.98	0.270967
Whole	42	10	-44811.49	575.4831	-134.14	1046.68	1.818785
Whole	43	10	-45992.31	3100.018	-1180.82	1893.41	0.610774
Whole	44	10	-45279.72	638.35	712.59	1181.78	1.851304
Whole	45	10	-45748.91	867.523	-469.19	621.07	0.715912
Whole	46	10	-45597.03	465.016	151.88	309.45	0.665461
Whole	47	10	-45754.6	618.2294	-157.57	444.11	0.718358
Whole	48	10	-46356.28	973.2859	-601.68	485.19	0.498507
Whole	49	10	-46472.77	685.6923	-116.49	NA	NA

Area	K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
AIN	2	10	-3342.17	2.6289	NA	NA	NA
AIN	3	10	-3266.06	95.4	76.11	135.48	1.420125
AIN	**4**	**10**	**-3054.47**	**168.2079**	**211.59**	**788.95**	**4.690328**
AIN	5	10	-3631.83	1534.5069	-577.36	1240.15	0.808175
AIN	**6**	**10**	**-2969.04**	**33.3695**	**662.79**	**617.8**	**18.513901**
AIN	7	10	-2924.05	29.1361	44.99	227.36	7.803391
AIN	**8**	**10**	**-3106.42**	**16.6814**	**-182.37**	**222.5**	**13.338172**
AIN	9	10	-3066.29	24.6844	40.13	63.37	2.567205
AIN	10	10	-3089.53	33.0915	-23.24	24.65	0.744904
AIN	11	10	-3088.12	41.8663	1.41	11.31	0.270146
AIN	12	10	-3075.4	36.4818	12.72	32.88	0.901271
AIN	13	10	-3095.56	28.0957	-20.16	56.33	2.004932
AIN	14	10	-3172.05	45.0292	-76.49	41.44	0.920292
AIN	15	10	-3289.98	76.1441	-117.93	18.49	0.242829
AIN	16	10	-3389.42	53.7612	-99.44	73.9	1.374597
AIN	17	10	-3414.96	47.1226	-25.54	89.39	1.896965
AIN	18	10	-3529.89	99.2124	-114.93	94.12	0.948672
AIN	19	10	-3550.7	92.5576	-20.81	37.55	0.405693
AIN	20	10	-3533.96	125.9373	16.74	NA	NA

Area	K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
Britanny	2	9	-20967.8444	125.6852	NA	NA	NA
Britanny	3	10	-19620.05	270.4727	1347.794444	394.7	1.459297
Britanny	4	9	-18666.9556	85.435	953.094444	10.05	0.117633
Britanny	5	9	-17703.8111	61.387	963.144444	581.755556	9.476856
Britanny	6	9	-17322.4222	151.4653	381.388889	20.4	0.134684
Britanny	7	9	-16961.4333	129.331	360.988889	138.333333	1.069607
Britanny	8	9	-16738.7778	265.8443	222.655556	11.777778	0.044303
Britanny	9	9	-16527.9	127.6357	210.877778	134.566667	1.054303
Britanny	10	9	-16451.5889	60.1105	76.311111	58.552222	0.974076
Britanny	**11**	**10**	**-16433.83**	**13.8073**	**17.758889**	**58.241111**	**4.21813**
Britanny	12	10	-16357.83	65.4586	76	112.7	1.7217
Britanny	**13**	**10**	**-16394.53**	**28.2714**	**-36.7**	**92.64**	**3.276807**
Britanny	14	10	-16338.59	100.7041	55.94	71.47	0.709703
Britanny	**15**	**10**	**-16211.18**	**35.2316**	**127.41**	**247.15**	**7.015008**
Britanny	16	10	-16330.92	156.8005	-119.74	106.41	0.678633
Britanny	17	10	-16344.25	241.2342	-13.33	26.33	0.109147
Britanny	18	10	-16331.25	172.9993	13	9.19	0.053122
Britanny	19	10	-16327.44	42.5705	3.81	135	3.17121
Britanny	20	10	-16458.63	183.778	-131.19	152.53	0.829969
Britanny	21	10	-16437.29	109.7405	21.34	2316.17	21.10588
Britanny	22	10	-18732.12	6906.2805	-2294.83	4422.72	0.640391
Britanny	**23**	**10**	**-16604.23**	**139.2101**	**2127.89**	**2226.98**	**15.997263**
Britanny	24	10	-16703.32	133.2807	-99.09	98.12	0.736191
Britanny	25	10	-16704.29	117.8404	-0.97	245.97	2.087314
Britanny	26	10	-16951.23	219.6445	-246.94	160.14	0.729087
Britanny	27	10	-17038.03	206.4291	-86.8	30.42	0.147363
Britanny	28	10	-17094.41	148.107	-56.38	145.74	0.984019
Britanny	29	10	-17296.53	143.4651	-202.12	NA	NA

Area	K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
Normandy	2	10	-17334.46	2.4985	NA	NA	NA
Normandy	3	10	-17008.7	12.6167	325.76	18.16	1.439367
Normandy	4	10	-16701.1	22.3663	307.6	35.12	1.57022
Normandy	5	10	-16358.38	3.6651	342.72	101.23	27.620054
Normandy	**6**	**10**	**-16116.89**	**6.4309**	**241.49**	**186.57**	**29.011463**
Normandy	7	10	-16061.97	11.0253	54.92	53.31	4.83523
Normandy	**8**	**10**	**-15953.74**	**7.6072**	**108.23**	**221.11**	**29.065925**
Normandy	9	10	-16066.62	40.3092	-112.88	83.54	2.072479
Normandy	10	10	-16263.04	111.419	-196.42	87.83	0.788286
Normandy	11	10	-16371.63	268.2642	-108.59	76.98	0.286956
Normandy	12	10	-16403.24	311.0902	-31.61	87.37	0.280851
Normandy	13	10	-16522.22	318.5147	-118.98	72.86	0.228749
Normandy	14	10	-16568.34	121.3671	-46.12	188.31	1.551574
Normandy	15	10	-16802.77	132.2892	-234.43	140.52	1.062218
Normandy	16	10	-16896.68	147.077	-93.91	179.51	1.220517
Normandy	17	10	-17170.1	261.6498	-273.42	29.85	0.114084
Normandy	18	10	-17413.67	238.9002	-243.57	24.9	0.104228
Normandy	19	10	-17632.34	221.9244	-218.67	25.1	0.113102
Normandy	20	10	-17825.91	306.2531	-193.57	NA	NA

Figure S1: Structure Results for the Rhone area

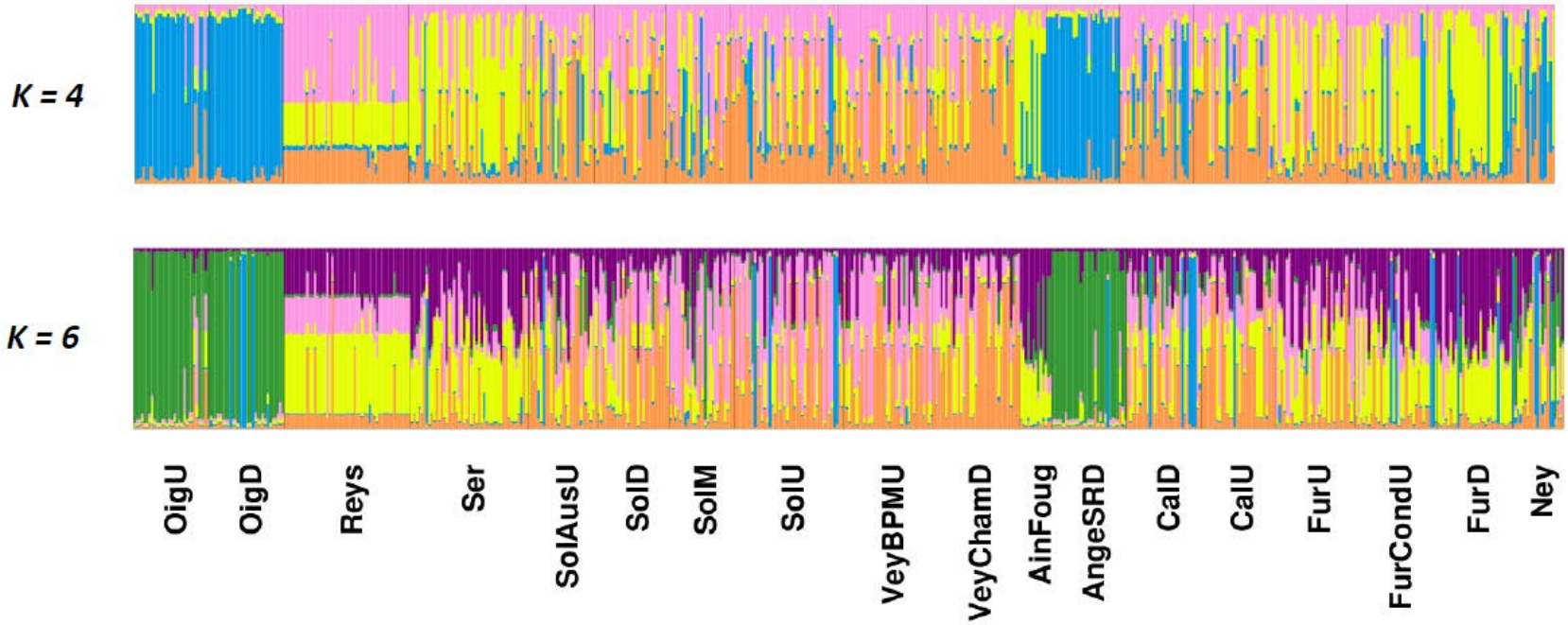
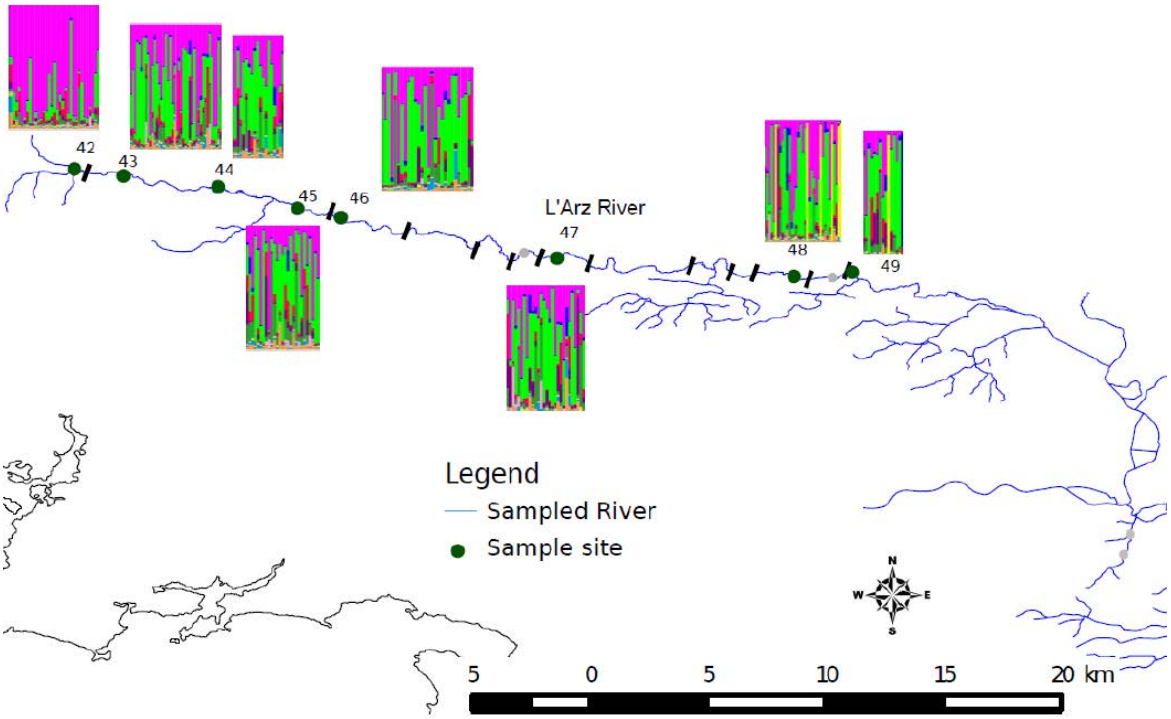
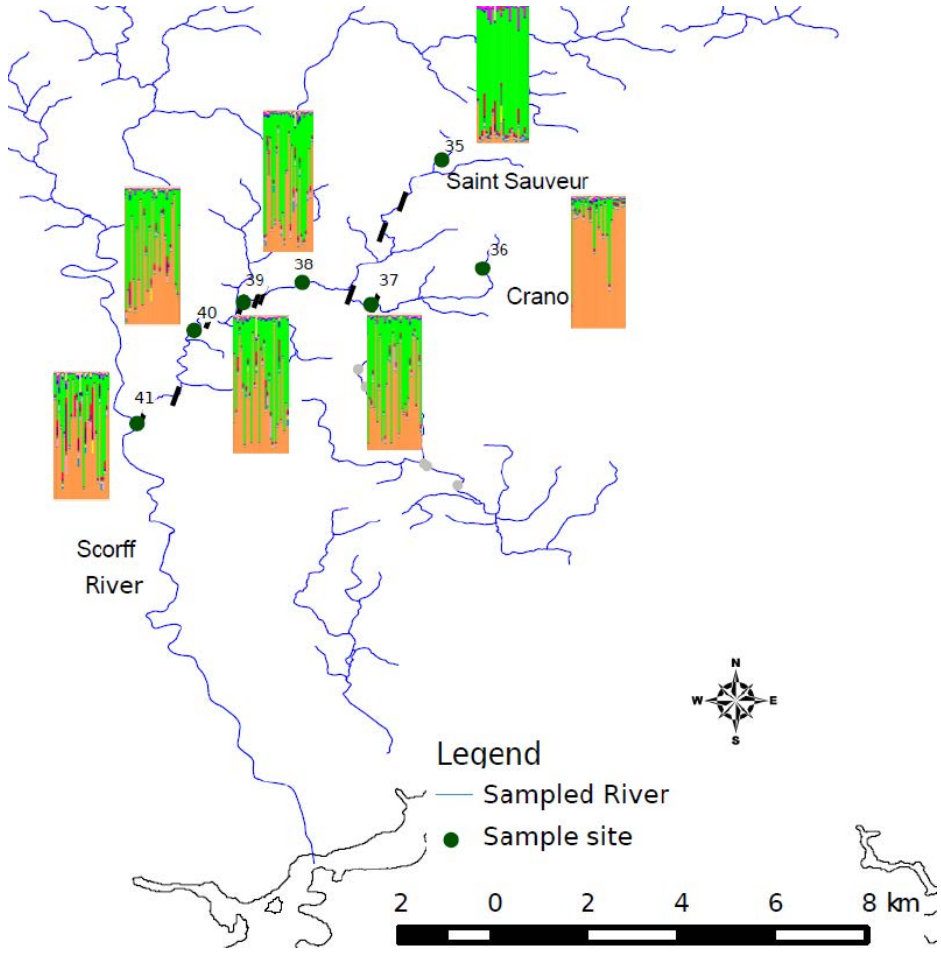


Figure S2: Sampling along the Arz (8 sites) and Crano river (7 sites) and levels of admixture

(a) Arz River



(b) Crano



Chapter 6

Discussion & perspectives

In this thesis I used lampreys of the genus *Lampetra* as a new model to study the genetic underpinnings of species divergence and ultimately speciation. Specifically I addressed the following questions:

- How (did) divergence proceed between the European *L. planeri* and *L. fluviatilis?*

This question was addressed *via* a multidisciplinary approach including experimental tests of reproductive isolation, investigations of gene flow in natural populations based on microsatellite and genome wide data and simulations to investigate the demographic history of divergence.

Results have demonstrated that *L. fluviatilis* and *L. planeri* display very a low level of RI when crossed experimentally and were able to produce viable F1 hybrids in semi-natural conditions. This low level of RI was mirrored by strong levels of gene flow measured either by microsatellite data or with a genome-wide RAD sequencing approach. Approximate Bayesian computation and diffusion approximations suggested that contemporary levels of introgression between *L. fluviatilis* and *L. planeri* were more likely to have emerged following a secondary contact, as divergence was likely to have been initiated in allopatry rather than in the face of continuous gene flow. The level of geographic connectivity has a strong influence on the level of gene flow: populations highly connected in areas of sympatry (or hybrid zones) generally displayed high levels of gene flow whereas populations disconnected, especially those occurring in different branches of the river network, were highly differentiated suggesting that no gene flow was currently ongoing;

- What are the effects of natural or human-induced river fragmentation on the genetic integrity of *L. planeri* populations*?*

I used a landscape genetics approach to test the effect of distance, barriers to migration as well as admixture between ecotypes on the spatial distribution of genetic diversity among *L. planeri* populations.

Extensive sampling of *L. planeri* in fragmented areas suggested limited evidence for a role of human induced fragmentation (at least when obstacles are small) but demonstrated a higher influence of asymmetric gene flow on the extent of genetic differentiation and diversity. It seems that the most isolated population displayed high level of relatedness and low levels of genetic diversity and this decreased exponentially when moving away from the headwaters.

Below I review some of these findings, discuss some limitations of the methods used and propose new avenues of research. Specifically, to complement and synthesize the discussions that were developed in each chapter, I will discuss four issues*: i*) The challenges in measuring hybrid fitness and investigating the strength of pre- and post-zygotic barriers ii) the importance of the current and historical geographic settings of populations divergence (i.e. importance of biogeography), iii) the

complexity of histories of divergence and iv) the interest of better understanding how isolated *L. planeri* populations evolved.

1. The challenges in measuring hybrid fitness and investigating the strength of pre and postzygotic barrier

A first major step when investigating levels of RI is to identify barriers to gene flow, and if possible to quantify their strength and order of appearance (Coyne & Orr, 2004). Relatively little was known on the level of RI in *L. planeri* and *L. fluviatilis*. Hume *et al.* (2013) recently demonstrated that viable F1s could be artificially produced, but they did not control for fertilization rates in their estimates of hatching rates, potentially resulting in bias in measures of post-zygotic reproductive isolation. Here, I separated the two effects and demonstrated that both fertilization success and hatching rates were close to 100% (Rougemont *et al.* 2015, Chapter 2). These results suggested no intrinsic hybrid inviability at an *early developmental stage only.* Breakup of co-adapted gene complexes generally occurs in F2s (Edmands 1999) and heterosis is more generally expected and frequently observed in F1s. In addition extrinsic postzygotic isolation (ecological or hybrid inferiority, see figure 1 in chapter 1) may occur and could not be measured. Getting further insights into the level of reproductive isolation would require rearing the F1 hybrids up to maturity, and I participated in the ongoing development of new rearing methods. A major improvement was performed in early life-stages feeding techniques, which led to survival rates close to 100% after about 3 months. Unfortunately, developing a rearing method up to maturity requires many trials and thus a lot of technical and human resources, which may not be easily available.

To complement this approach, I tested whether *L. planeri* and *L. fluviatilis* were *effectively* able to mate in semi-natural conditions. This was a necessary step since in the artificial crosses ova and sperm were mated artificially, and other premating barriers may exist in the wild (for instance, behavioral barriers). This experiment demonstrated that *L. planeri* males were indeed able to mate with *L. fluviatilis* and produce viable offspring at an early developmental stage (Rougemont et al. 2015, Chapter 2). Of course, this does not exclude that in the wild other isolating barriers can occur, such as temporal isolation or local habitat isolation. In addition, the least genetically differentiated population pair from the Oir River (pairwise F_{ST} range: 0.008 to 0.032) was used in this experiment. In this river, *L. fluviatilis* are significantly smaller (224 mm, $p<0.001$) than the average size of *L. fluviatilis* across the studied range (mean = 288 mm). It would thus be interesting to investigate whether *L. planeri* males are able to mate with females of bigger size. In addition, using males and females of the two species together may result in more matings of males with their conspecific females rather than with females of the other species, resulting in more realistic estimates of intra- and interspecific reproductive success.

Despite the current difficulty in measuring long term levels of reproductive isolation, lampreys are a great model to test other reproductive barriers. For instance, as most species with external fertilization, lampreys allow for experimental measures of levels of conspecific sperm precedence (CSP), a form of postmating prezygotic barrier (Coyne & Orr, 2004, see also Chapter 1). During my PhD I began to explore levels of sperm competition between *L. fluviatilis* and *L. planeri*. The preliminary question was whether it was possible to accurately measure some parameters of sperm quality, such as motility, sperm concentration and the level of viability. First results (not shown here) revealed that sperm of good quality could be obtained even from adults that had been kept for several days in a hatchery tank, and had already been used for experimental crosses. Differences (that could not be statistically tested due to small sample size) were observed between sperm of *L. fluviatilis* and *L. planeri*. In particular, sperm of *L. fluviatilis* from the Oir river displayed lower quality (lower motility, viability and concentration) than sperm of *L. planeri* from the same river. These results will open the way to address exciting questions about CSP. For instance we could test whether sperm from *L. planeri* sneaker males are more competitive than sperm from allopatric populations of *L. planeri* and *L. fluviatilis* males.

2. The importance of the geographical context in maintaining gene flow between lamprey ecotypes

It was necessary to validate whether the experimental measurement of low RI was reflected by contemporary gene flow in wild populations. A set of ten pairs of *L. fluviatilis* and *L. planeri* populations were thus sampled in different rivers and genotyped with microsatellite markers. Five pairs were putatively connected by gene flow as individuals were sampled simultaneously on the same nests in the Aa, Oir and Bethune Rivers, and at close proximity on the Hem and Bresle. Five other populations were parapatric: the Odon and Risle on the same river and the Garonne-Saucats, Dordogne-Jalles and Loire-Cens watersheds. The results indicated that gene flow could be strong between both ecotypes in sympatry (Oir and Bethune F_{ST} varied from 0.0076 to 0.0323) but the extent of gene flow was lower in the Aa River (F_{ST} = 0.08). However the inter-population differentiation was weaker in sympatric population pairs than in parapatric populations pairs (except the Risle where we hypothesized asymmetric downstream gene flow) for which reductions in gene flow was stronger (F_{ST} ranged between 0.10 and 0.20). Based on these results, the genetic differentiation was sometimes too strong in sympatry to validate the hypothesis of a single population displaying phenotypic plasticity for the life history strategy as suggested earlier (e.g. Beamish 1987, see also review in Docker 2009). Instead, I hypothesized that some genetic barriers may act to reduce effective gene flow in parts of the genome, but these barriers may act over a small proportion of the genome so that they could not be identified based on a set of 13 microsatellite markers. A first question that arose was whether this ongoing gene flow in spite of significant differentiation was a sign of primary or secondary intergradation (Harrison, 2011). Interestingly, the

least differentiated population, the Oir river (F_{ST} 2010 = 0.048, F_{ST} 2011 = 0.0323, F_{ST} 2014 = 0.008), is located at the transition zone between Normandy and Brittany, which corresponds to the western limit of distribution of *L. fluviatilis* in Northern France, whereas *L. planeri* are widespread in Brittany (see Figure 1 in chapter 4). Overall, our populations displayed similar levels of genetic and genome wide divergence as other fish models of speciation (Table 1).

Table 1 : Comparison of patterns of genetic differentiation observed in our study system and in other fish species.

Species/Ecotype Pair	Sampling	F_{ST}	*References*
Parasitic-nonparasitic *L. planeri* and *L. fluviatilis*	9 -10 pairs	µsat: 0.0076 – 0.192 rad: 0.042 – 0.207	
Limnethic-Benthic Whitefish	4-5 pairs	µsat :0.058 – 0.256 AFLP:0.042 – 0.22 RAD: 0.008 – 0.216	Campbell & Bernatchez (2004) Gagnaire et al. 2013
Lake-Stream Stickleback	4-6 pairs	µsat : 0 – 0.21 RAD: 0 – 0.149	Berner et al. 2008, 2009 Roesti et al. 2012
Anadromous-Resident Stickleback		0.0462 – 0.1391	Hohenlohe et al. 2012
Parasitic-nonparasitic Silver and Northern brook lamprey	3 pairs	0 – 0.143	Docker et al. 2012

The low levels of expected heterozygosity and allelic richness observed in upstream isolated populations (see also section 5 below) were particularly suggestive of a departure from demographic equilibrium in these populations. Such a pattern can be explained by two non-exclusive processes: random genetic drift in finite populations and non-recovery of historical bottlenecks following colonization of the river with few founder individuals. These results pointed to the necessity of testing histories of divergence.

In parallel to investigations of historical scenarios with microsatellite data (see below) a RAD-sequencing genotyping on a subset of nine pairs of population was undertaken. The goals were (i) to confirm the patterns of genetic structure given by microsatellite markers and especially (ii) to obtain a better understanding of the process of speciation at play between lamprey ecotypes, and (iii) to find genomic regions of high differentiation.

The RAD sequencing approach allowed validation of the high levels of gene flow but also to discrimination between the two species and putative hybrids with nearly 100% accuracy (Chapter 4). However, most hybrids were identified as F1s, and very few as F2s or backcrosses, which questioned the survival of hybrids beyond the F1 stage. The current weakness of such hybrid detection is that it remains purely statistical. In addition, only a small number of individuals were initially sampled in each pair (approximately 20 per river and ecotype). As a consequence no conclusion can be drawn on whether the absence of F2 and backcross hybrids reflects strong counter-selection due to some genetic incompatibility or is simply a consequence of our small sample size. A simulation approach (e.g. using Nemo (Guillaume & Rougemont 2006)) could be developed but would require larger sample sizes of parental populations in hybrid zones. For instance, simulating diverging populations and incorporating various proportions of DMI across the genome could help estimate the expected proportion of hybrids as a function of the number of endogenous barriers.

Our results of genetic structure within and between population pairs of lampreys based on RAD-seq data were in line with the microsatellite based estimates, but very different from those of Mateus *et al.* (2013), as they did not reveal strong divergence between both ecotypes. Indeed, Mateus et al. (2013) found a strong genome wide divergence in a southern sympatric pair using a RAD-seq approach. Leaving aside the problems pointed out in chapter 1, their results raised the possibility that southern populations of *L. fluviatilis* and *L. planeri* may have an older divergence time than northern populations. This hypothesis is supported by the fact that 1) several lineages of *L. planeri* coexist in the Iberian peninsula (Mateus et al. 2011, 2012; Perreira et al. 2010; Espanhol et al. 2007) and 2) the Iberian peninsula could have been an ancient refugium during the last glacial period, as for other species of fishes and vertebrates (Taberlet et al 1998; Hewitt 1999).

While recent improvements in genome wide analyses have greatly improved the knowledge of the process of speciation, it seems to me that the old but important debate surrounding the geographical context of speciation has been put aside in favor of other lines of research. In particular, the argument that the geographical context of speciation was unimportant (Butlin *et al.* 2008), the simplifying hypothesis of the speciation with gene flow literature (Smadja & Butlin 2011) or the tendency to focus mainly on rapid adaptation in the face of gene flow may lead to misleading conclusions. Recent investigations on the architecture of genomic divergence in sunflower have demonstrated that genomic islands of speciation were not affected by geography but that the functional architecture of the genome was more important, as expected from population genetic theory (Charlesworth *et al.* 1997a; Charlesworth 1998, 2012; Noor & Bennett 2009). However, based on results of this thesis, I suggest that the geography of speciation and spatial arrangement of populations do matter, especially in species with small *Ne* and in species arranged along linear networks (Fourcade *et al.* 2013). Indeed, in these conditions demographic disequilibria can generate false patterns of high genome wide differentiation and eventually generate genomic islands of

differentiation that have little to do with the process of speciation. On the other hand, studying species undergoing the homogenizing effect of gene flow allows highlighting key processes of speciation and revealing genomic regions truly involved in speciation. The combined use of parapatric and sympatric population pairs of lampreys was instrumental to better understand the process of speciation throughout this thesis (Chapter 2 and 4). As a consequence, accurately taking into account variation in recombination rates, selection at linked sites due to positive selection and background selection (BGS) will probably be the cornerstone of future studies aiming at better understanding how speciation proceeds throughout the genome. Full genome sequencing in hybrid zones (e.g. Aa and Oir river) combined with new modelling approaches (eg. Harris & Nielsen 2013b; Sedghifar *et al.* 2015) should help further understand speciation in lampreys.

3. The complexity of histories of divergence

a. Secondary contact, ancient migration or ongoing ecological differentiation?

Investigating the history of divergence in *Lp* and *Lf s*uggested that the two ecotypes probably display some genetic barriers maintaining their genetic integrity in the face of strong gene flow. The scenarios tested in an ABC framework and using diffusion approximation now strongly suggest that the species have initially diverged in allopatry. First, using a small set of loci I demonstrated, in line with old theory (Endler 1977, 1982; Harrison 2012) and recent investigations (Bierne *et al.* 2013), that distinguishing between primary *versus* secondary differentiation with a small number of putatively neutral loci was difficult (Chapter 3). In light of these findings it would be interesting to review studies based on the IM/IMa2 approaches (Hey & Nielsen 2004; Hey 2010) which often have found evidence for divergence in the face of ongoing gene flow and to test whether primary versus secondary differentiation can be distinguished.

To circumvent this problem, moving toward a genomic approach and incorporating heterogeneity of migration rates along the genome clearly improved the results. One key finding highlighting the importance of studying sympatric populations is that of a secondary contact scenario in hybrid zones (after having removed putative hybrid individuals). In this situation, it is more likely that only genomic regions truly involved in reproductive isolation were revealed by the eroding effect of gene flow. Being able to use *Petromyzon marinus* genome data to polarize SNPs also yielded interesting results with regards to estimation of demographic parameters. In particular, the proportion of the genome under neutrality was estimated to be around 90%. This result, combined with other evidence accumulated so far suggest that barrier loci may not form strong genomic islands acting over large regions but may only act on small parts of the genome. Unfortunately in the absence of a full genome sequence we cannot locate these regions. Another important result was that of asymmetric introgression from the anadromous (large) populations to the small resident

populations. This pattern is expected from theory (Barton 1986 in Bierne *et al*. 2013) and may lead to genome swamping of resident populations, as might be the case in the Oir River.

On the other hand, the inference of a scenario of ancient migration with a relatively high fit to the data in parapatric population pairs was unexpected but nevertheless interesting. I hypothesized that this inference is in fact wrong because of the inadequacy of the JSFS in populations with strong deviations from demographic equilibrium. This disequilibrium may exist in isolated *L. planeri* populations, and could generate spurious patterns that are not accurately captured by the δaδi method used here. Indeed a major hypothesis in δaδi is a large *N*e, which may not apply to isolated *L. planeri* populations. In addition, it is still not clear to me how to integrate demographic perturbations (e.g. post divergence bottlenecks) in models of speciation and it is not clear how recent bottlenecks may affect the JSFS. A recent study has suggested some bias in joint inferences of selection and demography from the JSFS (Mathew & Jensen 2015). Further development and theoretical modeling are thus certainly required.

Incorporating variations of *Ne* across the genome is also important as reflects genomic variation in selective processes, the so called selection at linked site process (Cutter & Payseur 2013), due either to positively selected mutations (selective sweep) (Cutter & Payseur 2013) or deleterious mutations and their subsequent elimination a process called background selection (BGS) (Charlesworth *et al.* 1993; Charlesworth 2012). Such selective effects result in reduced variability, reduced efficacy of selection and materialize as a reduction of effective population size at the site (a process called Hill-Robertson effect)(Hill & Robertson 1966). This process occurs particularly in genomic regions of low recombination (Cutter & Payseur 2013). In these regions the effects are larger, drift is increased, and ultimately genetic differentiation as measured by F_{ST} is increased (Cruickshank & Hahn 2014). This is very important to take into account in the speciation research and investigation of genomic islands. Indeed, the first observation of genomic islands (Turner *et al.* 2005), was interpreted as “speciation islands” protected from gene flow by selection. It was then suggested that during speciation with gene flow (under primary differentiation) genomic islands could be generated by selective sweeps and hitchhiking of neutral variants (Via 2012) while the remainder of the genome will be freely exchanged. Upon secondary contact, differentiation islands are revealed by the eroding effect of gene flow in regions unrelated to speciation. This was one of the concluding results from Chapter 4. This can also be related to the concept of barrier semi-permeability described in Chapter 1. Now in the light of the theory presented above it appears necessary to test whether the two processes can jointly act to model patterns of genome wide differentiation.

To do so, and bypass some of the assumptions of δaδi (large *Ne* and independence of the SNPs) that may not be fulfilled in the populations studied, I propose to analyze the RAD-sequencing data in an ABC framework incorporating both heterogeneity of migration rates along the genome but also

heterogeneity of effective population size *Ne* along the genome to better take into account selection at linked sites (see also Chapter 4). This work is ongoing but could not be completed in the thesis' timeframe; I will only present here the methodology and preliminary results (Annex 1). To do so I have used the same ABC approach as the one developed by Roux *et al*. (2013, 2014) incorporating heterogeneous *m* through a beta distribution subsequently rescaled. In the new development of C. Roux (unpublished), heterogeneity of Ne is drawn from a distribution 1 – α where the α distribution is the proportion of "neutral" loci sharing the same Ne. The 1 – α loci are then rescaled by a beta distribution, to incorporate heterogeneity. I skip the methodological implementation which is the same as in Roux *et al*. (2013) but with modified scripts and rescaled priors to fit biological parameters of lamprey populations. Ongoing ABC analyses (not shown here) performed on sympatric populations suggest that taking this parameter into account is very important in model inferences. In addition, simple model comparisons of the traditional scenarios have allowed me to validate previous δaδi inferences when comparing the 3 traditional models of AM, IM and SC with heterogeneous migration. For instance on the Aa, posterior probability of AM is 0, P(IM) is 0.160 and P(SC) is 0.84. Similar results were observed in the Bethune population. Interestingly, in the Oir, the method does not seem to be able to distinguish between IM and SC, the posterior probability being 50/50. It is possible that we are reaching the limit of the ABC and coalescent methods in this case of very low population differentiation. This is expected under a mode with a short period of divergence followed by a longer period of secondary contact where barriers to gene flow may couple or scatter depending on several conditions (see Bierne *et al.* 2013). Future work based on haplotypes to estimate introgression from tract length should help validating our hypothesis of secondary introgression and may help unravel the complexity of the history of lamprey divergence (Harris & Nielsen 2013b).

These estimations of *Ne* for each locus are important because although barrier semi-permeability and gene flow are certainly instrumental in revealing islands of divergence (or at least barrier loci formed by DMI), the role of selection at linked sites in regions of low recombination cannot be neglected as a force shaping patterns of differentiation. Finally, given the suggestion that individuals carry massive amounts of deleterious mutations (Charlesworth 2012), the effect of background selection can be expected to impact a wide range of organisms and should no longer be neglected (Ewing & Jensen 2015). In lampreys, disentangling the relative contributions of selection at linked sites and barrier semi-permeability due to speciation will require further genomic resources such as full genome sequences and further estimations of parameters that measure absolute divergence (e.g. D_{xy}) rather than F_{ST} alone (but see Chapter 1). Given the current lack of such resources, we are currently developing a linkage map that could help understand the origin of barrier loci.

Finally, based on the joint analysis of sympatric and parapatric populations, I conclude that the demographic history of *L. fluviatilis* and *L. planeri* may well be more complicated than a single event

of allopatry and introgression. Given the known alternation of contractions and expansions of species ranges (Hewitt 2000) it is possible that populations have undergone several periods of contact and isolation. The intriguing pattern of species distribution that we observed in Brittany (Figure 1 of Chapter 5) may be related to historical processes as this region was proposed as a refugial area for Atlantic salmon (Finnegan *et al.* 2013). The Channel was a massive fluvial river flowing in Northern France until recently (20 000 years ago). It was a major drainage system in Europe flowing during glacial retreats according to a cycle occurring every 100 000 yr during the last million years. (Lericolais *et al.* 2003; Toucane 2008). During these periods it is possible that populations have had large opportunities to alternatively exchange genes and be separated by glacial sheets, ultimately largely influencing their current genetic makeup.

Ultimately, I hope to develop a more integrative ABC approach that would allow putting all pairs of population pairs from all rivers together. Such an approach might be the best solution to infer a global scenario that would make sense at the scale of the species distribution range. This may help better estimate demographic parameters that are shared by all populations. Such an approach will require some developments that were beyond the scope of the thesis.

b. Historical divergence and reproductive barriers: endogenous or exogenous?

Under the secondary contact model, endogenous barriers to gene flow such as DMIs presented in Chapter 4 are expected to accumulate across the genome, as opposed to the ecological speciation scenario proposed in lampreys (Salewski 2003) in which exogenous barriers should accumulate as a result of divergent ecological selection (Rundle et Nosil 2005). As discussed in Chapter 1, the distinction between the two scenarios is fundamental for our understanding of the speciation process. In addition, genome scan results are often interpreted as evidence for the ongoing action of exogenous selection. Unfortunately, there was insufficient time during my thesis to further validate the hypothesis that the observed barriers are endogenous. To address this issue, I started the development of a RAD derived linkage map. Its ongoing construction (see Annex 2) will allow (hopefully) location of the most differentiated markers and to test whether outlier loci systematically deviate from expected Mendelian segregation ratios. If this is the case, then it is most likely that segregation distortion patterns are caused by endogenous selection against hybrids. Hundreds of distorted markers have already been identified. If some of them might be technical noise, it is however possible that others will be linked to RI. To push investigations a step further a modelling approach should be performed to simulate linkage maps containing variable number of DMIs (T. Leroy Personnal Communication). The simulation approach developed by T. Leroy demonstrated that the linkage map integrity can be strongly affected by the history of diverging populations. In particular, the distortion of map length should be proportional to time since divergence and accumulation of genetic incompatibilities.

4. The genetic consequences of spatial isolation in *L. planeri* populations

a. A low effect of human induced fragmentation

While many studies have reported negative effects of human induced fragmentation on species diversity and differentiation (e. g. Hanfling et Weetman 2006; Rayemakers *et al*. 2008; Alp *et al*. 2012; Torterotot *et al* 2014), investigations in lampreys did not reveal such strong negative effects. Instead, I found a downstream increase in genetic diversity (DIGD, Paz-Vinaz *et al*. 2015) which can be caused by passive drift of ammocoetes. This passive downstream drift may well generate sufficient gene flow to circumvent the establishment of strong genetic differences in populations located downstream of barriers. However, the barriers to migration that were investigated here are small (0.20-2 meters high, representative of the most widespread type of barrier across rivers in France). The effects may be stronger if taller barriers were investigated. The results are suggestive of source-sink dynamics with the most upstream populations acting as "source" populations. The low diversity of these upstream populations (that never receive migrants), their high level of differentiation (particularly noticeable in the Arz, Crano and Bethune Rivers where multiple sites were available) and this even when no obstacle was present, represent cumulative evidence suggesting that such a source-sink dynamic is the most parsimonious hypothesis. The results suggest also a more important role of isolation by distance rather than physical barriers in shaping patterns of genetic diversity and structure. However this study also suffered from several biases that may contribute to the lack of significant differences. First, populations living in independent coastal drainages evolved more or less independently, as found especially in Brittany. In these conditions, each population may have drifted independently and this signal may be much stronger in shaping population differences than the effect of obstacles tested on populations sampled in each river. Thus I argue that more meaningful results could be obtained by focusing on a single (or two) watershed and by performing extensive sampling of a single river and its tributaries. Most studies that found significant influences of barriers to migration focused on multiple sample sites within a single catchment (eg. Raeyemaekers *et al*. 2008; Torterotot *et al*. 2014; Gouskov *et al*. 2015). In this case, the main difficulty is to find statistically valid methods that correct for non-independency among sites, as the number of obstacles and several variables will be correlated with geographic distance, and accurately quantifying their respective effects in shaping population structure and diversity may be challenging. Finally, a solution could be to rely upon simulation tools (e.g. ABC) to accurately quantify gene flow and size of populations located upstream and downstream of barriers to migration. An approach focusing on a single catchment could be performed in the near future in the well characterized Sélune River. Indeed several obstacles exist on this watershed and the major dams (35 meters high) will be removed soon. Monitoring the evolution of population genetic structure in this catchment before and after dam removal may provide very useful information to improve future actions aiming at restoring river connectivity.

b. River lamprey as a "reservoir" of genetic diversity

Investigating spatial patterns of genetic diversity and differentiation at a larger scale revealed interesting results. First, populations of *L. planeri* located in streams where *L. fluviatilis* were present displayed higher genetic diversity levels than completely allopatric population such as those located in Brittany where the *L. fluviatilis* is rarely reported. This raises two questions: why *L. fluviatilis* does not currently migrate to spawn in Brittanny as this species is present in nearby rivers? Is this best explained by historical processes or ecological factors? I currently feel that it is very difficult to give a firm reply because of the complex patterns of genetic diversity observed in Brittany. The second important result was linked to the very high differentiation observed in the Upper Rhône. This may reflect either some kind of ascertainment bias linked to the development of our microsatellite markers or the fact that populations from this area represent different lineages. Moving toward a genomic approach in these populations, and more generally, performing a detailed genomic study at the European scale should help unravel the evolutionary history of divergence between *L. fluviatilis* and *L. planeri*.

c. Mutation accumulations in isolated populations: do lampreys suffer from a high drift load?

The loss of genetic diversity due to drift in finite populations can lead to increased genetic risks (Lynch *et al.* 1995; Frankham 1998, 2005). Indeed, effects of drift increase inversely as a function of ½ *N*e. In particular, drift decreases the efficiency of positive selection and increases the accumulation of deleterious mutations when $s < ½ N$e. As a consequence, accumulation of weakly deleterious mutations is expected in small populations, such as isolated populations of *L. planeri*. Additionally, drift leads to increased variance in allele frequencies and increased homozygosity, and since most deleterious alleles are recessive or partially recessive (Charlesworth *et al.* 1993), this will lead to a reduction in average fitness. Weakly deleterious mutations and increased homozygosity together result in a 'drift load'. This load can have profound evolutionary consequences on the adaptive potential of small populations. These small populations should show a low inbreeding depression (reduced fitness of inbred individuals as compared to a randomly mating population) according to population genetics theory: in small populations, partial purging efficiently removes recessive deleterious (especially lethal) mutations exposed in a homozygous state (Glémin 2003). Mildly deleterious mutations can also be fixed by drift and those fixed mutations do not contribute to fitness differences between inbred populations and randomly mating ones (Lynch *et al.* 1995; Glémin 2003). As a consequence, crosses between genetically differentiated isolated populations can lead to heterosis (hybrid vigor) of offspring relative to progeny from random mating within each parental population. The outcrossing of individuals will result in increased heterozygosity and subsequent masking of recessive or partially recessive deleterious alleles. As a consequence, offspring will be

heterozygotes at each of these fixed sites and are expected to display higher fitness than parental populations. In theory, the smaller the populations the higher the heterosis should be (Glémin *et al.* 2003). On the other hand, breakup of co-adapted gene complexes and epistatic interactions between distant populations may lead to outbreeding depression and smaller fitness (Lynch 1991; Tallmon *et al.* 2004; see also Chapter 1). In light of this theory, and given the observations of reduced genetic diversity in small isolated lamprey populations, I tested the existence of a drift load by performing crosses among distantly related populations of *L. planeri* and between *L. fluviatilis* for which no load was expected. The preliminary results of this experiment are no evidence for heterosis and are presented in Annex 3.

d. The importance of asymmetric gene flow in isolated populations: moving toward a genomic approach

In Chapter 5 I began to address issues related to the conservation of *L. fluviatilis* and *L. planeri* in fragmented rivers and across a variety of geographical contexts.

Using ABC on the set of 13 microsatellite markers and on RAD data I attempted to estimate the asymmetry of migration between populations located upstream and downstream of barriers. A set of 4 population pairs was randomly chosen and genetic characteristics of these populations are presented in Table 2.

Table 2: Comparison of RAD sequencing and microsatellite summary statistics in four population pairs

Riv	*n* U/D	Fst µsat	*n* U/D RAD	nbSNPs	Fst SNPs	Het µsat D / U	Exp Het µsat D/U	Het SNPs D / U	Het Exp D/U
ARZ	32/40	0.0350	17/13	14254	0.0332	0.471 / 0.459	0.479 / 0.441	0.296 / 0.287	0.336 / 0.327
RAN	32/18	0.0071	22/18	10215	0.0325	0.327 / 0.323	0.326 / 0.299	0.278 / 0.262	0.337 / 0.333
SCO	30/30	0.0128	15/14	17050	0.0367	0.354 /0. 406	0.418 / 0.418	0.266 / 0.285	0.339 / 0.338
TAM	24/25	0.0149	19/18	12427	0.0322	0.449 / 0.417	0.4864 / 0.444	0.247 / 0.267	0.294 / 0.307

Riv = river (ARZ= Arz, RAN= Rance, SCO = Scorff, TAM = Tamoute) U = Upstream population, D = Downstream population, *n* = number of individual sampled, *90*th F_{ST} = upper quantile (90%) of the F_{ST} distribution. Het = Observed Heterozygosity, Het Exp = Expected Heterozygosity.,

Results of parameter estimation (untransformed) are provided in Table 3 below.

Table 3: Estimates of migration rate and effective population size from ABC approaches

River	Markers	**N1 [95%CI]**	**N2[95%CI]**	**Nanc[95%CI]**	**T[95%CI]**	**M1[95%CI]**	**M2[95%CI]**
ARZ	SNPs	1.24 [1.08 - 1.59]	0.03 [0.03 - 0.04]	9.80 [9.74 - 9.89]	0.05 [0.045 - 0.054]	**30 [30 - 30]**	**21.23 [18.92 - 23.65]**
ARZ	microsat[1]	0.95 [0.40 - 1.90]	0.54 [0.38 - 0.79]	1.86 [1.70 - 2.05]	7.07 [5.45 - 9.82]	16.40 [13.67 - 18.10]	8.40 [4.50 - 12.60]
ARZ	microsat[2]	1.11 [0.22 - 2.69]	0.51 [0.31 - 0.89]	0.85 [0.11 - 2.06]	14.18 [4.14 - 22.20]	14.77 [6.90 - 19.46]	9.92 [1.64 - 17.95]
RAN	SNPs	4.11 [2.69 - 6.38]	4.86 [3.63 - 6.68]	8.51 [7.62 - 9.33]	0.094 [0.072 - 0.124]	**29.12 [26.54 - 29.99]**	**26 [18.27 - 29.58]**
RAN	microsat	1.15 [0.51 - 2.42]	0.47 [0.14 - 1.47]	2.15 [1.55 - 2.62]	5.46 [3.65 - 7.62]	9.85 [10.88 - 20.36]	5.46 [3.65 - 7.61]
RAN	microsat	0.89 [0.16 - 2.58]	0.52 [0.16 - 1.55]	1.76 [0.37 - 2.83]	15.70 [5.20 - 22.75]	12.30 [4.16 - 18.40]	8.62 [1.36 - 17.10]
TAM	SNPs	4.65 [4.12 - 5.23]	5.06 [3.39 - 6.78]	7.20 [6.00 - 8.16]	18 [17.26 - 18.70]	**9.88 [7.11 - 12.69]**	**0.89 [0.45 - 1.65]**
TAM	microsat	0.88 [0.53 - 1.31]	1.28 [1.2 - 1.37]	2.07 [1.77 - 2.33]	11.82 [11.15 - 12.47]	7.01 [4.96 - 8.69]	10.86 [8.84 - 12.88]
TAM	microsat	0095 [0.40 - 1.76]	1.18 [0.43 - 2.30]	1.47 [0.47 - 2.43]	12.96 [5.16 - 19.97]	10.33 [3.87 - 16.29]	12.68 [5.22 - 18.72]
SCO	SNPs	0.23 [0.101 - 0.548]	0.119 [0.08 - 0.183]	9.95 [9.88 - 9.99]	0.318 [0.229 - 0.4507]	**27.88 [24 - 29.75]**	**22 [16.17 - 26.57]**
SCO	microsat	1.07 [0.68 - 1.66]	0.713 [0.337 - 1.47]	1.96 [1.56 - 2.26]	13.15 [9.72 - 16.32]	18.10 [14.77 - 19.54]	10.53 [4.58 - 16.10]
SCO	microsat	1.18 [0.32 - 2.55]	0.60 [0.19 - 1.81]	1.78 [0.55 - 2.72]	12.55 [2.13-22.79]	17.15 [7.88 - 19.77]	8.79 [0.69 - 18.77]

Parameter estimates for RAD sequencing made using de best demographic scenario (Heterogeneous *N* and *M).*
N1 (N2) = downstream (upstream) ratios of effective population size, Nanc = ratios of effective population size T = ratio of Tsplit/4Nref, M1 = migration from pop2 to 1 M2 = migration from pop 1 to 2.
[1]: Estimations performed with 200 posterior samples
[2]: Estimations performed with 500 posterior samples

From a management point of view, one may be interested in estimating migration *M.* The striking result here is that when retaining the same number of posterior samples for parameter estimation, confidence intervals are very wide in microsatellite data while they were often narrow using RAD data. The second result is that migration downstream was really high and often reaches the maximum values used in prior parameters (ie. *M* = 30). While this urged me to used wider priors, this also demonstrates that migration is asymmetric and strong from upstream sites to downstream sites. Note also that upstream migration was strong but slightly reduced as CIs never reached the maximum value of 30. The Tamoute, where the two populations are *not* separated by any obstacle was a notable exception to our results. Migration was reduced as compared to the three other cases and strongly asymmetric. However, this can be explained by isolation by distance as the two sites were separated by 3.5 km whereas sites were much closer in the three other rivers.

General conclusion

This thesis investigates in details the process of speciation in European lampreys and provides an overview of the level of population genetic structure both between and within species. It brings some key results with regards to the level of speciation, by clearly suggesting that *Lampetra fluviatilis* and *L. planeri* are ecotypes of a single species, as revealed by the very low level of reproductive isolation, but these ecotypes display partial reproductive isolation, as suggested by genetic analyses. Both genetic and genomic analyses revealed the importance of the geographical context in understanding the speciation process. Besides, demographic inferences revealed the importance of historical processes during species formation. The identification of local hybrid zones was fundamental to better understand the process of speciation. Some key findings emerged with regards to current questions from the speciation literature.

First, it appears that European lamprey paired species, which were initially proposed as a model of "ecological speciation" (Salewski, 2003), have in fact diverged following a period of allopatry. A key prediction of allopatric isolation is the accumulation of Dobzhansky-Muller incompatibilities (DMI) that will form reproductive barriers to gene flow. However, inferences revealed high contemporary gene flow. One future question will thus be to understand how DMI interact upon secondary contact. The variable levels of genetic differentiation observed in some sympatric rivers, suggest an erosion of DMI in certain populations (Oir River) but not in others (Aa and Bresle River). A hypothesis for this contrasted situation is that coupling of some incompatibilities occurs in some rivers, but not in others where barriers may break down (Barton & de Cara, 2009). This coupling can be affected by various factors such as migration rate or drift of resident populations (Bierne *et al.* 2013). This may contribute to explain the variable levels of genetic parallelism observed across the studied populations. Importantly, inferences of parallelism upon secondary contact cast some doubts about the claimed role of natural selection driving parallel evolution of phenotypes in the face of gene flow (ex: Butlin *et al.* 2014).

A second major result is the importance of heterogeneity of gene flow across the genome. Taking this factor into account has not only allowed us to refine demographic inferences, but also provided key insights into the debate about the role of genomic islands of divergence. Indeed, genomic islands were proposed to emerge as byproducts of ecological selection during speciation with gene flow. Here, inferring a secondary contact suggests that heterogeneous divergence was initiated by the accumulation of DMI that subsequently formed local barriers to gene flow and resulted in increased genetic differentiation in parts of the genome, while the remainder is *currently* homogenized by gene flow. Obviously, this does not mean that ecological selection or past local adaptation have not played a role or do not currently operate, but these factors are unlikely to have been the initiators of

divergence. More recently, Cruickshank & Hahn (2014), in line with population genetics theory, proposed that genomics islands originate as a result of background selection after divergence in regions of low recombination (i.e. postspeciation selection at linked sites). In *Lampetra*, demographic inferences suggest a role for genomic islands during divergence and not after speciation. As mentioned above however, If DMI accumulate in allopatry then it is possible that following episodes of strong gene flow some of them are removed, hence generating genomic heterogeneity in differentiation. Besides, exploratory analyses of background selection in isolated populations, suggest that this process does play a key role in shaping genetic architecture within populations of lampreys. As a consequence, heterogeneous differentiation may emerge due to selection at linked sites within populations. I hope that we will be able to perform further investigations on this topic, both within and between lamprey ecotypes. Combining demographic inferences and analyses of full genome data could help unravel the role of this process in the evolution of populations. Another question that will be hopefully addressed is the localization and extent of the genomic islands of divergence. Using linkage mapping and demographic inference should allow us to address this question.

A third key finding related to the strong level of gene flow was that the genome of resident *L. planeri* may be swamped by neutral alleles of *L. fluviatilis*. This swamping seems to contribute to the maintenance of genetic diversity of *L. planeri* that otherwise evolve by drift and display small effective population sizes. *L. fluviatilis* not only form a “reservoir “of genetic diversity for brook lampreys but may also play a key role in promoting the transport of adaptive alleles between neighboring freshwater populations. From a conservation point of view, maintaining connectivity and access to spawning habitat of *L. fluviatilis* may be fundamental for the maintenance of the adaptive potential of the two ecotypes.

Appendices

Appendix 1: Exploring the effect of selection at linked genomic sites

Here I present some additional investigations related to selection at linked sites. They are not related to the current debates about speciation islands and deserts of recombination (Cruickshank et Hahn, 2014). Instead, I had the idea to study the effect of background selection within isolated *L. planeri* populations displaying interesting conditions of small size and potential accumulations of several mildly deleterious mutations

The initial idea was to evaluate the utility of combining ABC and RAD data to measure the asymmetry of gene flow. To quantify gene flow and ratios of effective population size, the same ABC approach as previously developed in Chapter 3 was used for microsatellite data. I attempted to estimate asymmetric gene flow using the traditional island models. For RAD data, I took benefit of scripts of C. roux to compare alternative island models. More specifically I compared the four following scenarios:

(1) homogenous effective population size *Ne* and homogeneous migration *m*
(2) Heterogeneous *m* and homogeneous *Ne;*
(3) Heterogeneous Ne and homogenous m;
(4) Heterogeneous *m* and heterogeneous Ne.

My reasoning was that isolated populations of brook lamprey may undergo the effect of BGS, which may translate into low *N*e in regions of low recombination whereas few traces of positive selection were expected. In this condition, migration across the genome may be mostly homogeneous while *Ne* can (eventually) be heterogenous. Here fascinating results emerged with regards to model comparison from these RAD data. The most likely model was largely that of heterogeneous migration and heterogeneous Ne (Table 1).

Table 1: Model Comparisons for RAD sequencing

	scenario			
River	**1**	**2**	**3**	**4**
ARZ	0.00	0.01	0.08	**0.91**
RAN	0.00	0.03	0.21	**0.76**
SCO	0.00	0.06	0.03	**0.90**
TAM	0.00	0.05	0.09	**0.85**

River abbreviation: Arz = Arz, Ran = Rance, Sco = Scorff, Tam = Tamoute

Scenario 1: homogenous effective population size *Ne* and homogeneous migration *m, scenario 2: H*eterogeneous *m* and homogeneous *Ne, Scenario 3:* Heterogeneous Ne and homogenous m; Scenario 4: Heterogeneous *m* and heterogeneous Ne

However, when I attempted to discriminate the role of variation in Ne and variation in Me, the best supported model was consistently the one incorporating variation in Ne (p (*Ne variable*) >= 0.80 against p (*M variable*) =0.15-20). While I struggle to understand these results, I suggest that, in light of the theory of selection at linked sites, this may be explained by BGS reducing *Ne* in regions of low recombination, as hypothesized by Charlesworth (2012).

In this scenario, lamprey populations of small size (small *Ne*) could, under the effects of genetic drift, accumulate deleterious mutations. In principle, background selection needs to be strong to efficiently eliminate these mutations (i.e. populations will have a significant load if s < ½ Ne) (Whitlock *et al*. 2000, Whitlock et Burger 2004). BGS will have a key role in reducing genetic diversity at linked loci (Charlesworth *et al*. 1993, Charlesworth, 2012). This reduction in nucleotide diversity (π/π_0) at a neutral locus due to BGS is a function of the number of deleterious mutations (μd), the selection (s) against heterozygotes for deleterious alleles, and recombination rates (c). These effects can be summed across linked sites that mutate into deleterious alleles according to Nordborg *et al*. 1996:

$\pi/\pi 0$ ~exp - $\exp - \sum_i (\mu \frac{di}{Si})/\{1 + \frac{ci(1-si)}{si}\}^2$. This implies reduction of *Ne* along the genome at linked sites, especially in regions of low recombination. Although this is just a hypothesis; I feel that further work on this topic and its evolutionary consequences is urgently needed. An additional question that arises in lampreys is the role of the very high number of chromosomes (n = 168) in this process. Is recombination increased? How do BGS and recombination jointly shape the genome? What is the role of positive selection? How do they act within isolated ecotypes and between connected ecotypes? How do they confound signals of positive selection? All these topics will require new research.

Appendix 2: Development of a hybrid linkage map: mapping endogenous and exogenous barrier

We crossed (in-vitro fertilization) a female river lamprey from the Loire River with a male brook lamprey from the Oir River to produce F1 offspring. They were reared until they reached a size sufficient for extraction of DNA suitable in quantity and quality for subsequent RAD sequencing. A total of 90 offspring was sequenced, and parents were sequenced twice. RAD sequencing and genotyping were performed as described in Chapter 4. A total of 10 700 markers were available and 78 larvae were suitable for construction of the linkage map. We developed a python script to determine which markers were suitable for linkage mapping, to calculate distortion segregation errors and to provide suitable data for subsequent mapping. Markers suitable were of 4 types and coded as follows: homozygous in the female and heterozygous in the male (segregating 1:1); homozygous in the male and heterozygous in the female (segregating 1:1); homozygous in both parents (segregating 1:1:1:1), heterozygous in both parents (segregating 1:2:1). CarthaGene (Givry *et al.* 2005) was used for linkage mapping using outbred data of a type F2 intercross with unknown phase. Different tests were performed to optimize the map. In a final run CarthaGene was used with a LOD score of 4 and a distance threshold of 0.2. We constructed a female map, a male map and a consensus map. We performed Chi^2 test to detect significant deviations from Mendelian inheritance at the 5% level followed by standard Bonferonni correction for multiple testing.

Preliminary results: From the 5185 markers that could be mapped unambiguously, 95 linkage groups were obtained when removing groups with less than six markers. The average number of SNPs per group was 66 (median = 60 min = 6 max=208).

Removing groups with less than 4 markers resulted in 110 linkage groups and removing groups containing less than three markers results in 129 groups. This linkage map will allow positioning of outliers and study of the patterns of LD to determine the origin of barriers to reproductive isolation.

Appendix 3: No evidence for heterosis in small and isolated brook lamprey populations

Species extinction can be influenced by stochastic demographic processes generating fluctuations in population size. Such fluctuations are currently largely enhanced by human perturbations of natural ecosystems (Vitousek *et al.* 1997; Palumbi 2001). Habitat fragmentation is one of the most important threats, reducing gene flow, decreasing genetic diversity and hence effective population size (Lynch *et al.* 1995; Frankham 1998, 2005). In these conditions, small populations will undergo stronger effects of genetic drift than populations of large effective size as the strength of drift increases inversely with the size of the populations as a function of ½ *N*e. Ultimately, drift can lead to higher levels of inbreeding (Frankham, 1995a,b; 1998) and results in local extinctions (Saccheri *et al.* 1998; Spielman *et al.* 2004).

Genetic drift also decreases the efficiency of positive selection and leads to the accumulation of deleterious mutations when $s < ½ N$e. As a consequence, the accumulation of weakly deleterious mutations is expected to be stronger in small populations. Secondly, drift increases the variance in allele frequencies and increases homozygosity, and since most deleterious alleles are recessive or partially recessive, this will lead to a reduction in average population fitness due to a 'drift load'. This load can have profound evolutionary consequences on the adaptive potential of small populations. In these populations, partial purging can efficiently remove recessive deleterious mutations (especially lethal ones) exposed in a homozygous state (e.g. Glémin 2003). This should result in a low inbreeding depression that is the reduced fitness of inbred individuals as compared to a randomly mating population (Charlesworth & Charlesworth 1999). However, mildly deleterious mutations could be fixed by drift as they do not contribute to fitness differences between inbred and outbred individuals (Lynch *et al.* 1995; Glémin 2003). Such mutations can contribute to fitness differences between individuals originating from within and among population crosses.

In particular, crosses between genetically differentiated isolated populations can lead to heterosis (hybrid vigor) of offspring relative to progeny from random mating within each parental population. The outcrossing of individuals will result in increased heterozygosity and subsequent masking of recessive or partially recessive deleterious alleles. Offspring are then expected to display higher fitness than parental populations. In theory, the smaller the population, the higher the heterosis should be (Glémin *et al.* 2003). On the other hand, breakup of co-adapted gene complexes and epistatic interactions between distant populations may lead to outbreeding depression and smaller fitness in inter-population crosses (Lynch 1991; Tallmon *et al.* 2004). Importantly, outbreeding depression is expected to occur mainly in F2s and subsequent generations whereas heterosis is already observed in F1 hybrids.

The brook lamprey *(Lampetra planeri*) is a jawless vertebrate that may allow testing hypotheses about the evolution of a drift load in small and isolated populations. It is a non parasitic,

strictly freshwater resident species that displays a low migratory and dispersal ability (Malmqvist, 1980)and its populations can be isolated in upstream headwaters of rivers resulting in strong genetic differentiation (Rougemont et al. 2015). *L. planeri* is closely related to the anadromous and parasitic river lamprey *L. fluviatilis* whose populations are weakly differentiated and display higher levels of genetic diversity (Docker 2009). We have previously demonstrated that isolated *L. planeri* populations show a very reduced genetic diversity as measured by expected heterozygosity, a proxy of effective size (*Ne*) (Nei & Takahata 1993) and allelic richness (Rougemont et al. 2015). In contrast, populations of the same species occurring in sympatry with *L. fluviatilis* displayed higher levels of diversity and higher Ne. *L. planeri* populations living in rivers where *L. fluviatilis* is absent are most prone to demographic instability (i.e. bottlenecks) and may undergo strong effects of genetic drift even in non-fragmented habitats (Rougemont et al. in prep). Altogether this makes *L. planeri* an interesting model to investigate the effect of random drift load in natural populations. In contrast, genetically more diverse *L. fluviatilis* populations should not display a high drift load.

The aim of the present study was to compare the strength of heterosis between isolated populations of *L. planeri* and *L. fluviatilis* populations. We also sampled a *L. planeri* population living in sympatry with *L. fluviatilis* as it may display a lower heterosis effect than isolated *L. planeri* populations.

We performed artificial crosses within and between-populations and within and between-species to estimate heterosis and outbreeding depression, based on relative performances of offspring resulting from these crosses. Crosses among the most isolated and genetically differentiated populations allowed us to evaluate the importance of the load in these populations. Theoretically, offspring from between-population crosses should display higher performance than individuals from within-population crosses.

Methods

Sampling and crosses:

5 L. planeri populations were sampled in isolated rivers (Montafilan, Rance, Rocher, Cens and Odon) and two populations of *L. fluviatilis* were sampled in the Loire and Oir river. We also collected individuals in the sympatric and weakly differentiated Oir *L. planeri* population. 3 to 10 individuals from each sex were sampled with electrofishing in each population. They were then kept in the laboratory and crossed with *in vitro* fertilization following the partially factorial design presented below (Figure 1). All individuals did not mature at the same time so we aimed at producing at least three intra-population and three inter-population crosses with progenitors from each population leading to a total of 102 crosses. The fertilized eggs were reared in 24 well plates at constant 12°C (±1°C) (Rougemont et al. 2015) and we measured the survival and length of larvae at hatching.

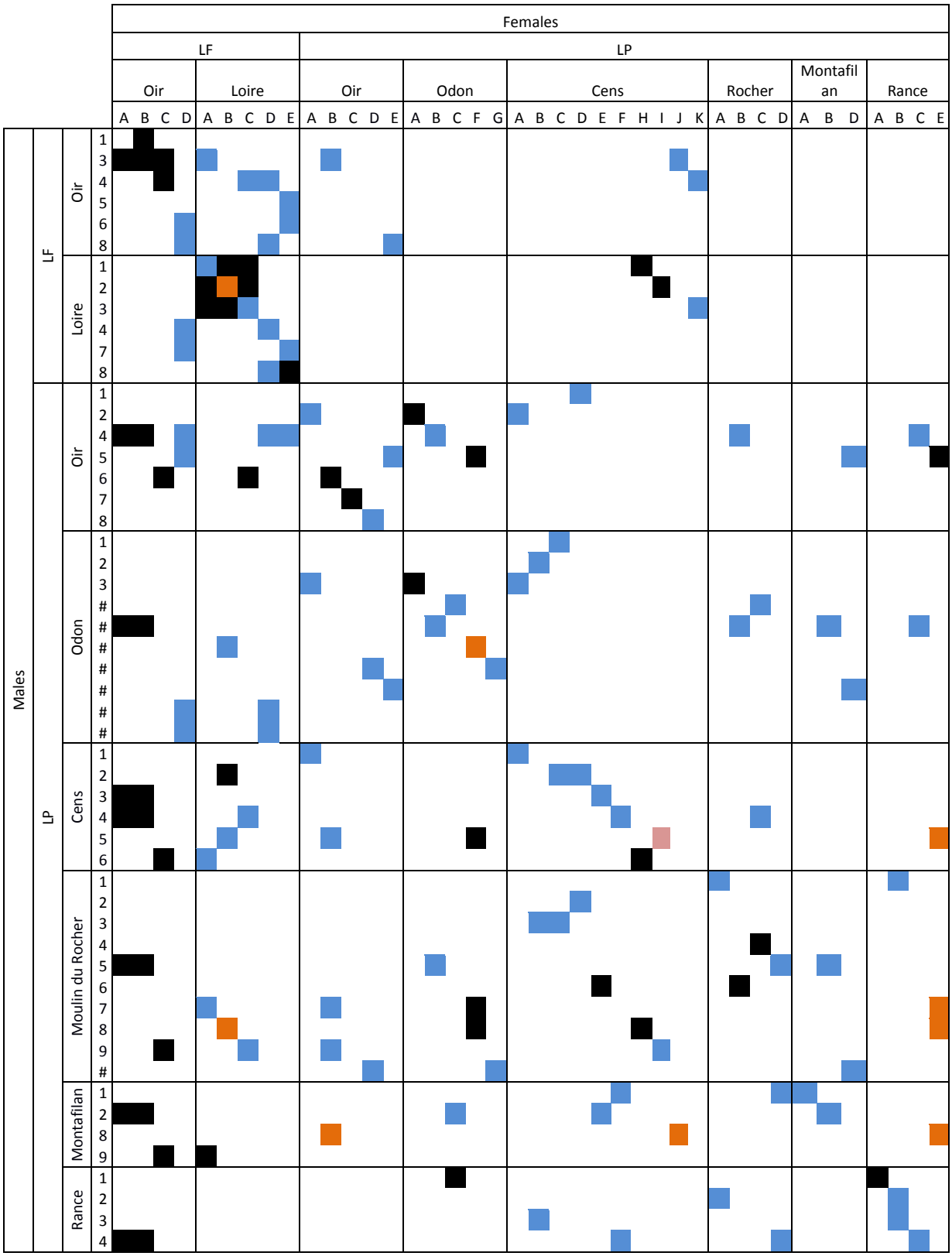


Fig. 1: Matrix of crosses performed between six *L. planeri* populations and two *L. fluviatilis* populations (females are identified by letters and males by numbers). Within and among pouplation crosses were dependent on the timing of maturation of individuals, which varied both within and

among populations. In black: All crosses. In blue: crosses used for hatching and length analysis. In orange: crosses used only for length measurement. In pink: crosses used only for hatching rates.

Data analysis

The data were analyzed using Generalized Linear mixed effect Models (GLMM) and Linear Mixed effect Models (LMM). We tested the influence of the cross type on larval hatching rate (binomial error family) and length (Gaussian error family. The cross type was a factor with six modalities: within *L. planeri* populations, among *L. planeri* populations, within *L. fluviatilis* populations, among *L. fluviatilis* population, hybrid from a *L. fluviatilis* female and hybrid from a *L. planeri* female. Sire and dam effects were treated as random effects nested within their respective population of origin. The date of cross and the shelf where the eggs were located during incubation were also treated as random effects. GLMMs were performed using either all data (Table 1) or only crosses among *L. planeri* (Table 2)

Results and discussion

The average hatching success over all crosses was 89.5% (±16%). The average hatching success in crosses between *L. planeri* was 93.9% (±12%) while it was 83.1% (±12%) in *L. fluviatilis* crosses.

The mean size of offspring over all crosses was 6.81 mm (±0.51 mm). Larvae from crosses between *L. planeri* had a mean size of 6.78 mm (±0.51 mm) and those from *L. fluviatilis* crosses had a mean size of 6.91 mm (±0.50 mm). Results of GLMMs and LMMs are given in tables 1 and 2 and reveal significant effects of cross type as well as the paternal and maternal identity of offspring.

Table 1: Results of GLMMs and LMM testing the effect of cross type on survival and length of the larvae. Tests are performed including *L. planeri* and *L. fluviatilis.* AIC value, degree of freedom (df), Chi2 and P-values are provided. Significance: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$)

		Survival				length			
model	**effect tested**	**AIC**	**df**	**Chi 2**	**P**	**AIC**	**df**	**Chi 2**	**P**
1		433.263	9			22572.19	13		
2	cross type	439.455	4	16.191	***0.0063 *****	22597	8	12.294	***0.0309 ****
3	dam id	566.22	8	134.96	***2.10e-16 ******	22761.37	12	191.97	***2.10e-16 ******
4	sire id	441.835	8	10.572	***0.0011 *****	22616.85	12	45.722	***2.10e-16 ******
5	pop dam	--	--	--	--	22572.14	12	1.2961	0.255
6	pop sire	--	--	--	--	22570.69	12	0.2891	0.591
7	date	434.349	8	3.0857	0.079	22570.41	12	0.1822	0.669
8	bloc					22697.9	12	127.47	***2.10e-16 ******

Multiple comparison tests showed that the survival of offspring from crosses between populations of *L. planeri* was significantly higher than the survival of offspring from crosses between females of *L. fluviatilis* and males of *L. planeri* (p=0.048; figure 2). All other pairwise comparisons were not significant (Figure 2). Clearly, offspring from *L. fluviatilis* females displayed a higher variance in survival (Figure 2).

We also found that cross type had a significant effect on the mean size of offspring (table 1). Multiple comparison tests indicated first that the mean size of offspring from *L. planeri* females x *L. fluviatilis* males were significantly larger than offspring from other crosses among and within populations of *L. planeri* and from *L. fluviatilis* females x *L. planeri* males. However these crosses were not different from within and between-population crosses of *L. fluviatilis* (figure 3). Second, offspring from crosses using *L. planeri* males and females were significantly smaller than crosses between populations of *L. fluviatilis*. We also found a significant correlation between female size and offspring size when considering only crosses from *L. planeri* (r=0.46; p=0.016). However there was no evident relationship between other cross types and parental size.

Table 2: Results of GLMMs and LMM testing the effect of cross type on survival and length of the larvae using *L. planeri* crosses only.

		Survival				length			
model	**effect tested**	**AIC**	**df**	**Chi 2**	**P**	**AIC**	**df**	**Chi 2**	**P**
1		260.01	5			14615	13		
2	cross type	267.58	4	9.5757	***0.001972 *****	14619	8	6.2115	***0.01269***
3	dam id	281.96	4		***9.9e-7 ******	14744	12	131.61	***2.2e-16 ******
4	sire id	262.81	4	4.8	***0.02846 ****	14642	12	28.82	***7.943e-08 ******
5	pop dam	--	--	--	--	14616	12	2.973	0.08467
6	pop sire	--	--	--	--	14613	12	0.764	0.382
7	date	258.01	4	0	0.9999	14613	12	0.0377	0.8461
8	bloc					14684	12	70.963	***2.2e-16 ******

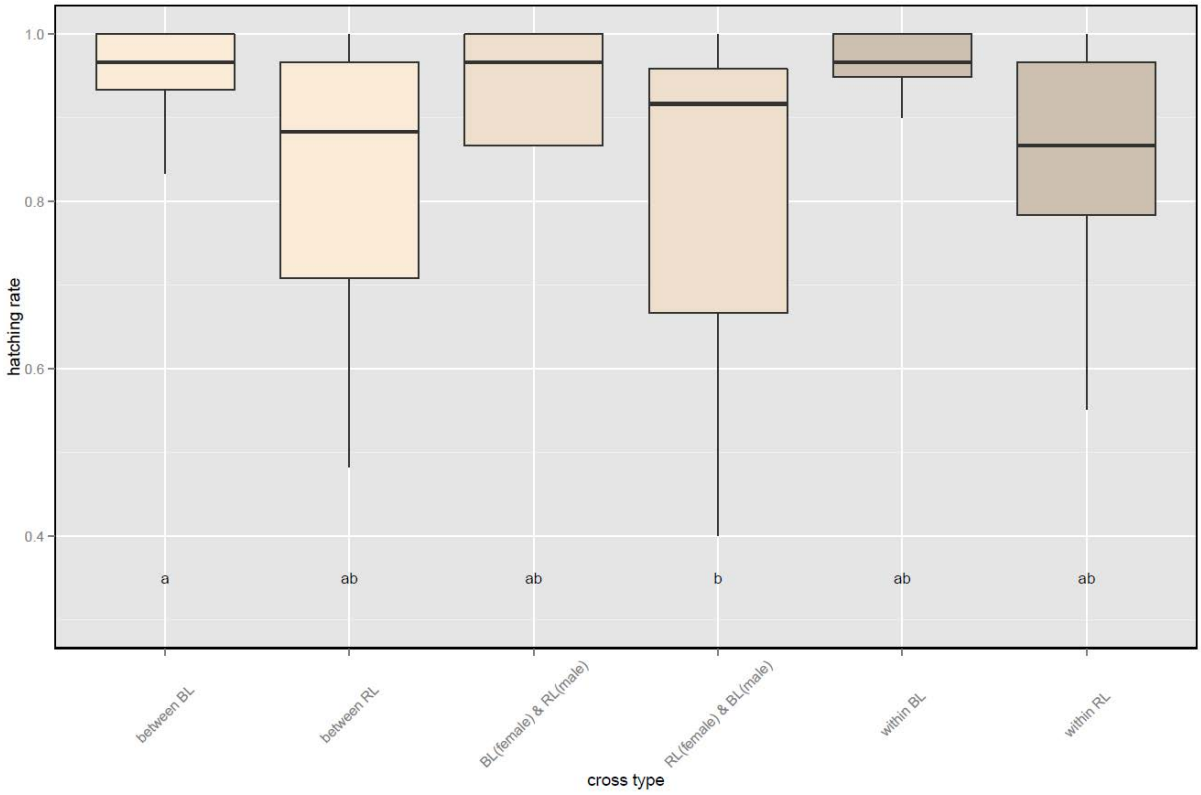


Figure 2: Box plot of hatching rates by cross type (BL: brook lampret, RL: river lamprey), letters (a, b and c) indicate significant difference between crosses.

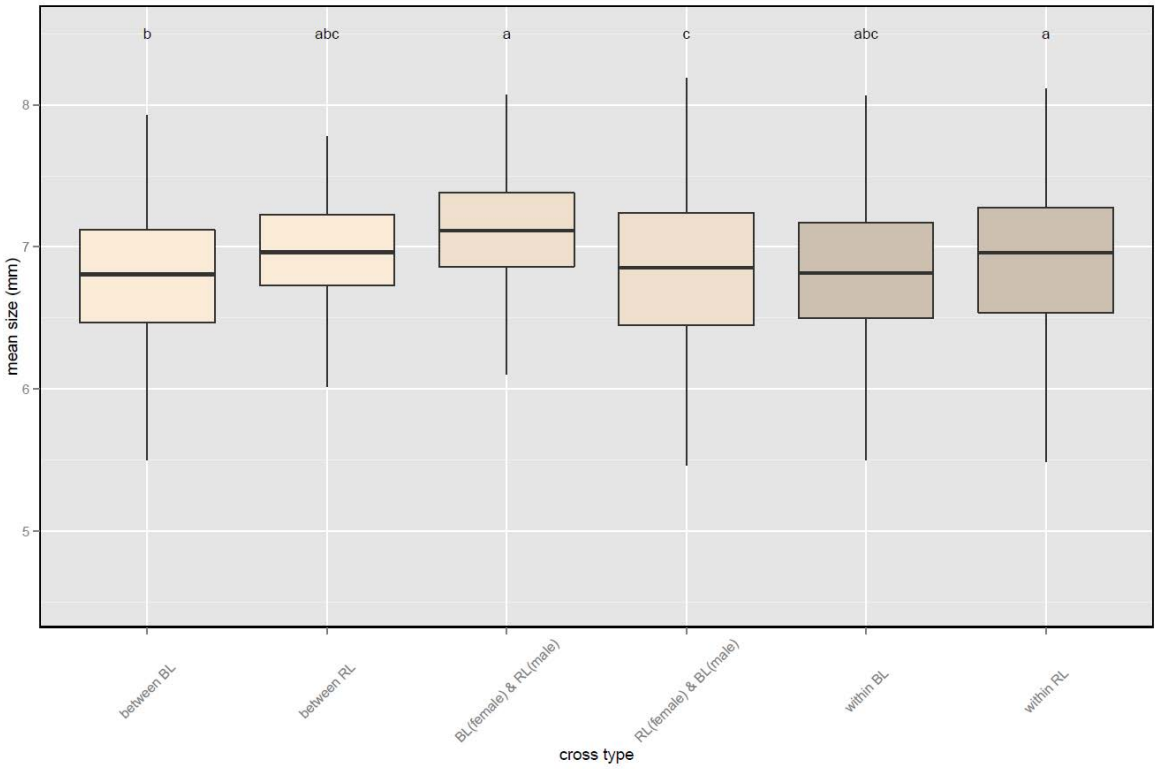


Figure 3: Box plot of larval size by cross type, letters (a, b and c) indicate significant differences between crosses.

Overall, analyses did not show any significant differences in survival rate or larval size between populations of *L. planeri* and between populations of *L. fluviatilis* indicating no clear effect of heterosis or outbreeding depression among populations of the same species. Conversely, our analyses indicated differences in interspecific hybrid performances and at least one of the species used in crosses. Offspring survival rates from *L. fluviatilis* female x *L. planeri* male crosses were significantly lower than those from crosses using males and females of *L. planeri* from different populations but they were not different from crosses using *L. fluviatilis* (either between or within-population crosses).

Larval size of hybrid offspring from *L. planeri* female x *L. fluviatilis* male crosses was significantly higher than size of intra-specific crosses revealing a tendency for heterosis for this trait. Finally, we found significant genetic effects within populations as reflected by significant dam and sire effects. However such effects may be primarily driven by maternal effects as revealed by the positive relationship between female size and offspring larval length in *L. planeri*. Using half-sib crossing designs may allow disentangling these (environmental) maternal from purely additive genetic effects.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Aldenhoven JT, Miller MA, Corneli PS, Shapiro MD (2010) Phylogeography of ninespine sticklebacks (Pungitius pungitius) in North America: glacial refugia and the origins of adaptive traits. *Molecular Ecology*, **19**, 4061–4076.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*.

Allen Orr H (2001) The genetics of species differences. *Trends in Ecology & Evolution*, **16**, 343–350.

Alonso H, Matias R, Granadeiro JP, Catry P (2009) Moult strategies of Cory's Shearwaters Calonectris diomedea borealis: the influence of colony location, sex and individual breeding status. *Journal of Ornithology*, **150**, 329–337.

Anderson EC, Thompson EA (2002) A Model-Based Method for Identifying Species Hybrids Using Multilocus Genetic Data. *Genetics*, **160**, 1217–1229.

April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America's freshwater fishes. *Proceedings of the National Academy of Sciences*, **108**, 10602–10607.

Arnold ML, Martin NH (2009) Adaptation by introgression. *Journal of Biology*, **8**, 82.

Arnqvist G (1998) Comparative evidence for the evolution of genitalia by sexual selection. *Nature*, **393**, 784–786.

Arnqvist G, Edvardsson M, Friberg U, Nilsson T (2000) Sexual conflict promotes speciation in insects. *Proceedings of the National Academy of Sciences*, **97**, 10460–10464.

Arnqvist G, Rowe L, Krupa JJ, Sih A (1996) Assortative mating by size: A meta-analysis of mating patterns in water striders. *Evolutionary Ecology*, **10**, 265–284.

Aspinwall N (1974) Genetic Analysis of North American Populations of the Pink Salmon, Oncorhynchus gorbuscha, Possible Evidence for the Neutral Mutation-Random Drift Hypothesis. *Evolution*, **28**, 295–305.

Atwell S, Huang YS, Vilhjálmsson BJ *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Bank C, Bürger R, Hermisson J (2012) The Limits to Parapatric Speciation: Dobzhansky–Muller Incompatibilities in a Continent–Island Model. *Genetics*, **191**, 845–863.

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population-genetic selection inference. *bioRxiv*, 009654.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, **439**, 719–723.

Barnosky AD, Matzke N, Tomiya S *et al.* (2011) Has the Earth/'s sixth mass extinction already arrived? *Nature*, **471**, 51–57.

Barrett RDH, Hoekstra HE (2011a) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews. Genetics*, **12**, 767–780.

Barrett RDH, Hoekstra HE (2011b) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barson N, Aykanat T, Hindar K *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in Atlantic salmon. *bioRxiv*, 024695.

Bartels H, Docker MF, Krappe M *et al.* (2015) Variations in the presence of chloride cells in the gills of lampreys (Petromyzontiformes) and their evolutionary implications. *Journal of Fish Biology*, **86**, 1421–1428.

Bartels H, Fazekas U, Youson JH, Potter IC (2011) Changes in the cellular composition of the gill epithelium during the life cycle of a nonparasitic lamprey: functional and evolutionary implications. *Canadian Journal of Zoology*, **89**, 538–545.

Barton N (1979) Heredity - Abstract of article: The dynamics of hybrid zones. *Heredity*, **43**, 341–359.

Barton NH (1983) Multilocus Clines. *Evolution*, **37**, 454.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57 (Pt 3)**, 357–376.

Barton NH, de Cara MAR (2009) The evolution of strong reproductive isolation. *Evolution; International Journal of Organic Evolution*, **63**, 1171–1190.

Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [stat]*.

Beamish RJ (1987) Evidence that Parasitic and Nonparasitic Life History Types are Produced by One Population of Lamprey. *Canadian Journal of Fisheries and Aquatic Sciences*, **44**, 1779–1782.

Beamish RJ, Neville C-EM (1992) The Importance of Size as an Isolating Mechanism in Lampreys. *Copeia*, **1992**, 191.

Beaumont MA (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

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

Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **263**, 1619–1626.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population. *Genetics*, **162**, 2025–2035.

Berg JJ, Coop G (2014) A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet*, **10**, e1004412.

Bernard, H., 1909 Bulletin de la Société des Sciences Naturelle et d'Archeologie de l'Ain. 57 , Imprimerie du Journal, Bourg.

Bernatchez L, Renaut S, Whiteley AR *et al.* (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 1783–1800.

Bernatchez L, Vuorinen JA, Bodaly RA, Dodson JJ (1996) Genetic Evidence for Reproductive Isolation and Multiple Origins of Sympatric Trophic Ecotypes of Whitefish (Coregonus). *Evolution*, **50**, 624–635.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology*, **7**, 431–452.

Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake-stream transitions. *Evolution; International Journal of Organic Evolution*, **63**, 1740–1753.

Beysard M, Heckel G (2014) Structure and dynamics of hybrid zones at different stages of speciation in the common vole (Microtus arvalis). *Molecular Ecology*, **23**, 673–687.

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, Bonhomme F, Boudry P, Szulkin M, David P (2006) Fitness landscapes support the dominance theory of post-zygotic isolation in the mussels Mytilus edulis and M. galloprovincialis. *Proceedings of the Royal Society B: Biological Sciences*, **273**, 1253–1260.

Bierne N, David P, Boudry P, Bonhomme F (2002) Assortative fertilization and selection at larval stage in the mussels Mytilus edulis and M. galloprovincialis. *Evolution; International Journal of Organic Evolution*, **56**, 292–298.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bierne N, Gagnaire PA, & David P (2013) The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology,* **59,** 72–86.

Blanchet S, Rey O, Etienne R, Lek S, Loot G (2010) Species-specific responses to landscape fragmentation: implications for management strategies. *Evolutionary Applications*, **3**, 291–304.

Blank M, Jürss K, Bastrop R (2008) A mitochondrial multigene approach contributing to the systematics of the brook and river lampreys and the phylogenetic position of Eudontomyzon mariae. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 2780–2790.

Blum MGB, Francois O (2010) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, **20**, 63–73.

Bolnick DI (2011) Sympatric Speciation in Threespine Stickleback: Why Not? *International Journal of Ecology*, **2011**, e942847.

Bolnick DI, Fitzpatrick BM (2007) Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 459–487.

Bomblies K, Weigel D (2010) Arabidopsis and relatives as models for the study of genetic and genomic incompatibilities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1815–1823.

Bonhomme M, Chevalet C, Servin B *et al.* (2010) Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics*, **186**, 241–262.

Bracken FSA, Hoelzel AR, Hume JB, Lucas MC (2015) Contrasting population genetic structure among freshwater-resident and anadromous lampreys: the role of demographic history, differential dispersal and anthropogenic barriers to movement. *Molecular Ecology*, **24**, 1188–1204.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **513**, 375–381.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Brook BW, Tonkyn DW, O'Grady JJ, Frankham R (2002) Contribution of inbreeding to extinction risk in threatened species. Conservation Ecology, 6, 16.

Brykov A, Polyakova N, Skurikhina LA, Kukhlevsky AD (1996) Geographical and temporal mitochondrial DNA variability in populations of pink salmon. *Journal of Fish Biology*, **48**, 899–909.

Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, gr.196485.115.

Butlin RK, Galindo J, Grahame JW (2008) Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2997–3007.

Butlin RK, Saura M, Charrier G *et al.* (2014a) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution; International Journal of Organic Evolution*, **68**, 935–949.

Butlin RK, Saura M, Charrier G *et al.* (2014b) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution; International Journal of Organic Evolution*, **68**, 935–949.

Carneiro M, Albert FW, Afonso S *et al.* (2014) The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit. *PLoS Genet*, **10**, e1003519.

Carneiro M, Blanco-Aguiar JA, Villafuerte R, Ferrand N, Nachman MW (2010) Speciation in the European Rabbit (oryctolagus Cuniculus): Islands of Differentiation on the X Chromosome and Autosomes. *Evolution*, **64**, 3443–3460.

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

Chang AS, Noor M (2004) Conspecific sperm precedence in sister species of drosophila with overlapping ranges. *Evolution*, **58**, 781–789.

Charles K, Guyomard R, Hoyheim B, Ombredane D, Baglinière J-L (2005) Lack of genetic differentiation between anadromous and resident sympatric brown trout (Salmo trutta) in a Normandy population. *Aquatic Living Resources*, **18**, 65–69.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Charlesworth B (2012) The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics*, **190**, 5–22.

Charlesworth B, Campos JL (2014) The Relations Between Recombination Rate and Patterns of Molecular Variation and Evolution in Drosophila. *Annual Review of Genetics*, **48**, 383–403.

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

Charlesworth B, Nordborg M, Charlesworth D (1997a) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research*, **70**, 155–174.

Charlesworth B, Nordborg M, Charlesworth D (1997b) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155–174.

Ciereszko A, Glogowski J, Dabrowski K (2000) Fertilization in landlocked sea lamprey: storage of gametes, optimal sperm: egg ratio, and methods of assessing fertilization success. *Journal of Fish Biology*, **56**, 495–505.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. *Science*, **307**, 1928–1933.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Cornuet J-M, Ravigné V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**, 401.

Couvet D (2002) Deleterious Effects of Restricted Gene Flow in Fragmented Populations. *Conservation Biology*, **16**, 369–376.

Coyne, J. A. & Orr, H. A. (1989). Patterns of speciation in drosophila. *Evolution*, **2,** 362–381.

Coyne JA, Orr HA (1998) The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **353**, 287–305.

Coyne JA, Orr HA (2004) *Speciation*. W.H. Freeman.Sunderland MA

Cross TF, Mills CPR, de Courcy Williams M (1992) An intensive study of allozyme variation in freshwater resident and anadromous trout, Salmo trutta L., in western Ireland*. *Journal of Fish Biology*, **40**, 25–32.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, **25**, 410–418.

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

Cutler DR, Edwards TC, Beard KH *et al.* (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews. Genetics*, **14**, 262–274.

Cyr F, Angers B (2012) Historical process lead to false genetic signal of current connectivity among populations. *Genetica*, **139**, 1417–1428.

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

Daniel J. Howard PGG (1998) Conspecific Sperm Precedence is an Effective Barrier to Hybridization Between Closely Related Species. *Evolution*, **52**, 511.

Davis AW, Wu CI (1996) The broom of the sorcerer's apprentice: the fine structure of a chromosomal region causing reproductive isolation between two sibling species of Drosophila. *Genetics*, **143**, 1287–1298.

Dawson HA, Quintella BR, Almeida PR, Treble AJ, Jolley JC (2015) The ecology of larval and metamorphosing lampreys. In Lampreys: Biology Conservation and Control. M.F. Docker (ed). Fish & Fisheries Serie 37.

Deiner K, Garza JC, Coey R, Girman DJ (2007) Population structure and genetic diversity of trout (Oncorhynchus mykiss) above and below natural and man-made barriers in the Russian River, California. *Conservation Genetics*, **8**, 437–454.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Dey A, Jeon Y, Wang G-X, Cutter AD (2012) Global Population Genetic Structure of Caenorhabditis remanei Reveals Incipient Speciation. *Genetics*, **191**, 1257–1269.

DiBattista JD (2008) Patterns of genetic variation in anthropogenically impacted populations. *Conservation Genetics*, **9**, 141–156.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature*, **400**, 354–357.

Docker MF (2009) A review of the evolution of nonparasitism in lampreys and an update of the paired species concept. , 71–114.

Docker MF, Heath DD (2003) Genetic comparison between sympatric anadromous steelhead and freshwater resident rainbow trout in British Columbia, Canada. *Conservation Genetics*, **4**, 227–231.

Docker MF, Mandrak NE, Heath DD (2012) Contemporary gene flow between "paired" silver (Ichthyomyzon unicuspis) and northern brook (I. fossor) lampreys: implications for conservation. *Conservation Genetics*, **13**, 823–835.

Docker MF, Youson JH, Beamish RJ, Devlin RH (1999) Phylogeny of the lamprey genus Lampetra inferred from mitochondrial cytochrome b and ND3 gene sequences. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 2340–2349.

Dodson JJ, Aubin-Horth N, Thériault V, Páez DJ (2013) The evolutionary ecology of alternative migratory tactics in salmonid fishes. *Biological reviews of the Cambridge Philosophical Society*, **88**, 602–625.

Doebeli M (1996) A quantitative genetic competition model for sympatric speciation. *Journal of Evolutionary Biology*, **9**, 893–909.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *arXiv:1504.04543 [q-bio]*.

Duvaux L, Belkhir K, Boulesteix M, Boursot P (2011) Isolation and gene flow: inferring the speciation history of European house mice. *Molecular Ecology*, **20**, 5248–5264.

Dynesius M, Nilsson C (1994) Fragmentation and Flow Regulation of River Systems in the Northern Third of the World. *Science*, **266**, 753–762.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Eberhard WG (2010) Evolution of genitalia: theories, evidence, and new directions. *Genetica*, **138**, 5–18.

Edmands S (1999) Heterosis and Outbreeding Depression in Interpopulation Crosses Spanning a Wide Range of Divergence. *Evolution*, **53**, 1757.

Edwards SV (2009) Is a New and General Theory of Molecular Systematics Emerging? *Evolution*, **63**, 1–19.

Elias M, Faria R, Gompert Z, Hendry A (2012) Factors Influencing Progress toward Ecological Speciation. *International Journal of Ecology*, **2012** doi:10.1155/2012/235010.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

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

Elmer KR, Fan S, Kusche H *et al.* (2014) Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nature Communications*, **5**, 5168.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press.NJ

Endler JA (1982) Problems in Distinguishing Historical from Ecological Factors in Biogeography. *American Zoologist*, **22**, 441–452.

Espanhol R, Almeida PR, Alves MJ (2007) Evolutionary history of lamprey paired species Lampetra fluviatilis (L.) and Lampetra planeri (Bloch) as inferred from mitochondrial DNA variation. *Molecular ecology*, **16**, 1909–1924.

Estoup A et al. (1996) Rapid one-tube DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Molecular Marine Biology and Biotechnology*, **5**, 295–298.

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

Ewing GB, Jensen JD (2015) The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, n/a–n/a.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Fagundes NJR, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, **104**, 17614–17619.

Fahrig L (2003) Effects of Habitat Fragmentation on Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 487–515.

Falush D, Stephens M, Pritchard JK (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, **164**, 1567–1587.

Faubet P, Waples RS, Gaggiotti OE (2007) Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular ecology*, **16**, 1149–1166.

Faulks LK, Gilligan DM, Beheregaray LB (2010) Islands of water in a sea of dry land: hydrological regime predicts genetic diversity and dispersal in a widespread fish from Australia's arid zone, the golden perch (Macquaria ambigua). *Molecular Ecology*, **19**, 4723–4737.

Feder JL, Berlocher SH, Roethele JB *et al.* (2003) Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 10314–10319.

Feder JL, Bush GL (1989) Heredity - Abstract of article: Gene frequency clines for host races of Rhagoletis pomonella in the Midwestern United States. *Heredity*, **63**, 245–266.

Feder JL, Egan SP, Nosil P (2012a) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Feder JL, Egan SP, Nosil P (2012b) The genomics of speciation-with-gene-flow. *Trends in genetics: TIG*, **28**, 342–350.

Feder JL, Gejji R, Yeaman S, Nosil P (2012c) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 461–474.

Feder JL, Nosil P (2010) The Efficacy of Divergence Hitchhiking in Generating Genomic Islands During Ecological Speciation. *Evolution*, **64**, 1729–1747.

Feder JL, Xie X, Rull J *et al.* (2005) Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in Rhagoletis. *Proceedings of the National Academy of Sciences*, **102**, 6573–6580.

Felsenstein J (1981) Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? *Evolution*, **35**, 124–138.

Felsenstein J (1982) How can we infer geography and history from gene frequencies? *Journal of Theoretical Biology*, **96**, 9–20.

Ferchaud A-L, Hansen MM (2015) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: threespine sticklebacks in divergent environments. *Molecular Ecology*, n/a–n/a.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, **31**, 1275–1291.

Ffrench-Constant RH, Anthony N, Aronstein K, Rocheleau T, Stilwell G (2000) Cyclodiene insecticide resistance: from molecular to population genetics. *Annual Review of Entomology*, **45**, 449–466.

Finnegan AK, Griffiths AM, King RA *et al.* (2013) Use of multiple markers demonstrates a cryptic western refugium and postglacial colonisation routes of Atlantic salmon (Salmo salar L.) in northwest Europe. *Heredity*, **111**, 34–43.

Fitzpatrick BM, Fordyce JA, Gavrilets S (2008) What, if anything, is sympatric speciation? *Journal of evolutionary biology*, **21**, 1452–1459.

Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074–4088.

Fleming IA (1996) Reproductive strategies of Atlantic salmon: ecology and evolution. *Reviews in Fish Biology and Fisheries*, **6**, 379–416.

Fogarty ND, Lowenberg M, Ojima MN, Knowlton N, Levitan DR (2012) Asymmetric conspecific sperm precedence in relation to spawning times in the Montastraea annularis species complex (Cnidaria: Scleractinia). *Journal of Evolutionary Biology*, **25**, 2481–2488.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Foulds WL, Lucas MC (2013) Extreme inefficiency of two conventional, technical fishways used by European river lamprey (Lampetra fluviatilis). *Ecological Engineering*, **58**, 423–433.

Fourcade Y, Chaput-Bardy A, Secondi J, Fleurant C, Lemaire C (2013) Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. *Molecular Ecology*, **22**, 2065–2073.

Frankham R (1998) Inbreeding and Extinction: Island Populations. *Conservation Biology*, **12**, 665–675.

Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131–140.

Frankham R (2015) Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, **24**, 2610–2618.

Frankham R, Ballou JD, Dudash MR *et al.* (2012) Implications of different species concepts for conserving biodiversity. *Biological Conservation*, **153**, 25–31.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution*, **30**, 1687–1699.

Gage MJG, Stockley P, Parker GA (1995) Effects of Alternative Male Mating Strategies on Characteristics of Sperm Production in the Atlantic Salmon (Salmo salar): Theoretical and Empirical Investigations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **350**, 391–399.

Gagnaire P-A, Normandeau E, Pavey SA, Bernatchez L (2013a) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (Coregonus clupeaformis). *Molecular Ecology*, **22**, 3036–3048.

Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L (2013b) The Genetic Architecture of Reproductive Isolation During Speciation-with-Gene-Flow in Lake Whitefish Species Pairs Assessed by Rad Sequencing. *Evolution*, **67**, 2483–2497.

Gaigher A, Launey S, Lasne E, Besnard A-L, Evanno G (2013) Characterization of thirteen microsatellite markers in river and brook lampreys (Lampetra fluviatilis and L. planeri). *Conservation Genetics Resources*, **5**, 141–143.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Gautier M, Vitalis R (2012) rehh : An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, bts115.

Gautier M, Vitalis R (2013) Inferring Population Histories Using Genome-Wide Allele Frequency Data. *Molecular Biology and Evolution*, **30**, 654–668.

Gavrilets S (2003) Perspective: models of speciation: what have we learned in 40 years? *Evolution; international journal of organic evolution*, **57**, 2197–2215.

Gavrilets S (2014) Models of Speciation: Where Are We Now? *Journal of Heredity*, **105**, 743–755.

Gensoul, J., 1907 Monographie des poisons d'eau douce de Saone et Loire. Extrait du Bulletin de la

société d'histoire naturelle d'Autun. 20è Bulletin.

Geraldes A, Basset P, Smith KL, Nachman MW (2011) Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination. *Molecular Ecology*, **20**, 4722–4736.

Geyer LB, Palumbi SR (2005) Conspecific sperm precedence in two species of tropical sea urchins. *Evolution; International Journal of Organic Evolution*, **59**, 97–105.

Gillespie JH (2000) Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. *Genetics*, **155**, 909–919.

Gillespie JH (2001) Is the population size of a species relevant to its evolution? *Evolution; International Journal of Organic Evolution*, **55**, 2161–2169.

Givry S de, Bouchez M, Chabrier P, Milan D, Schiex T (2005) Carh ta Gene: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics*, **21**, 1703–1704.

Glémin S (2003) How Are Deleterious Mutations Purged? Drift Versus Nonrandom Mating. *Evolution*, **57**, 2678–2687.

Glémin S, Ronfort J, Bataillon T (2003) Patterns of inbreeding depression and architecture of the load in subdivided populations. *Genetics*, **165**, 2193–2212.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6723–6727.

Gomez-Uchida D, Knight TW, Ruzzante DE (2009) Interaction of landscape and life history attributes on genetic diversity, neutral divergence and gene flow in a pristine community of salmonids. *Molecular Ecology*, **18**, 4854–4869.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html. Updated from Goudet (1995).

Gouskov A, Reyes M, Bitterlin L, Vorburger C (2015) Fish population genetic structure shaped by hydroelectric power plants in the upper Rhine catchment. *Evolutionary Applications*, n/a–n/a.

Green SA, Bronner ME (2014) The lamprey: a jawless vertebrate model system for examining origin of the neural crest and other vertebrate traits. *Differentiation; Research in Biological Diversity*, **87**, 44–51.

Gross, M.R. 1984. Sunfish, salmon, and the evolution of alternatives reproductives strategies and tactics in fishes, p55-75, in Potts G. And Wootton R.J. Editors. Fish reproduction: strategies and tactics. Academic Press, London, England.

Gross JB, Borowsky R, Tabin CJ (2009) A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax mexicanus. *PLoS Genet*, **5**, e1000326.

Guillaume F, Rougemont J (2006) Nemo: an evolutionary and population genetics programming framework. *Bioinformatics (Oxford, England)*, **22**, 2556–2557.

Günther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, genetics.113.152462.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, **5**, e1000695.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010) Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 2459–2468.

Hänfling B, Weetman D (2006) Concordant Genetic Estimators of Migration Reveal Anthropogenically Enhanced Source-Sink Population Structure in the River Sculpin, Cottus gobio. *Genetics*, **173**, 1487–1501.

Han CS, Jablonski PG, Kim B, Park FC (2010) Size-assortative mating and sexual size dimorphism are predictable from simple mechanics of mate-grasping behavior. *BMC Evolutionary Biology*, **10**, 359.

Hardisty MW (1944) The Life History and Growth of the Brook Lamprey (Lampetra planeri). *Journal of Animal Ecology*, **13**, 110–122.

Hardisty, M.W. I.C. Potter Paired species M.W. Hardisty, I.C. Potter (Eds.), The Biology of Lampreys Vol. I, Academic Press, London (1971) New York

Harris K, Nielsen R (2013a) Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genet*, **9**, e1003521.

Harris K, Nielsen R (2013b) Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genet*, **9**, e1003521.

Harrison RG (2012) The Language of Speciation. *Evolution*, **66**, 3643–3657.

Harrison RG, Larson EL (2014) Hybridization, Introgression, and the Nature of Species Boundaries. *Journal of Heredity*, **105**, 795–809.

Hatfield T, Schluter D (1999) Ecological Speciation in Sticklebacks: Environment-Dependent Hybrid Fitness. *Evolution*, **53**, 866–873.

Hausdorf B (2011) Progress toward a general species concept. *Evolution; International Journal of Organic Evolution*, **65**, 923–931.

Hawks J, Cochran G (2006) Dynamics of Adaptive Introgression from Archaic to Modern Humans. *PaleoAnthropology*, **2006**, 101–115.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, **32**, 835–845.

Hedrick PW (2005) A Standardized Genetic Differentiation Measure. *Evolution*, **59**, 1633–1638.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, **22**, 4606–4618.

Heimberg AM, Cowper-Sal·lari R, Sémon M, Donoghue PCJ, Peterson KJ (2010) microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proceedings of the National Academy of Sciences*, **107**, 19379–19383.

Hendry AP (2009) Ecological speciation! Or the lack thereof?This Perspective is based on the author's J.C. Stevenson Memorial Lecture delivered at the Canadian Conference for Fisheries Research in Halifax, Nova Scotia, January 2008. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 1383–1398.

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.

Herten K, Hestand MS, Vermeesch JR, Houdt JKV (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, **16**, 73.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **359**, 183–195; discussion 195.

Hewitt GM (2011) Quaternary phylogeography: the roots of hybrid zones. *Genetica*, **139**, 617–638.

Hey J (2010) Isolation with Migration Models for More Than Two Populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hey J, Nielsen R (2004) Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, **167**, 747–760.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, **104**, 2785–2790.

Higgins K, Lynch M (2001) Metapopulation extinction caused by mutation accumulation. *Proceedings of the National Academy of Sciences*, **98**, 2928–2933.

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetics Research*, **8**, 269–294.

Hindar, K., B. Jonsson, N. Ryman, and G. Stahl. 1991 Genetic relationships among landlocked, resident, and anadromous Brown Trout, Salmo trutta L. Heredity. 66: 83-91.

Hindorff LA, Sethupathy P, Junkins HA *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362–9367.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. *Science*, **313**, 101–104.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genet*, **6**, e1000862.

Hohenlohe PA, Catchen J, Cresko WA (2012) Population Genomic Analysis of Model and Nonmodel Organisms Using Sequenced RAD Tags. In: *Data Production and Analysis in Population*

Genomics Methods in Molecular Biology. (eds Pompanon F, Bonin A), pp. 235–260. Humana Press.

Howard DJ (1999) Conspecific Sperm and Pollen Precedence and Speciation. *Annual Review of Ecology and Systematics*, **30**, 109–132.

Howard DJ, Marshall JL, Hampton DD *et al.* (2002) The genetics of reproductive isolation: a retrospective and prospective look with comments on ground crickets. *The American Naturalist*, **159 Suppl 3**, S8–S21.

Hubbs, I.C. Potter Distribution, phylogeny and taxonomy M.W. Hardisty, I.C. Potter (Eds.), The Biology of Lampreys Vol. I, Academic Press, London (1971)

Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huggins RJ, Thompson A (1970) Communal spawning of brook and river lampreys, Lampetra planeri Bloch and Lampetra fluviatilis L. *Journal of Fish Biology*, **2**, 53–54.

Hughes AL (2012) Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity*, **108**, 347–353.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *The Annals of Applied Statistics*, **2**, 841–860.

Ishwaran H. and Kogalur U.B. (2015). Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.6.1.

Ishwaran H. and Kogalur U.B. (2007). Random survival forests for R. *R News* **7,**2, 25--31.

Ishwaran H., Kogalur U.B., Blackstone E.H. and Lauer M.S. (2008). Random survival forests. *Ann. Appl. Statist*. **2**,3, 841--860.

Jensen JD (2014) On the unfounded enthusiasm for soft selective sweeps. *Nature Communications*, **5**.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature*, **411**, 302–305.

Johannesson K (2001) Parallel speciation: a key to sympatric divergence. *Trends in Ecology & Evolution*, **16**, 148–153.

Johannesson K (2010) Are we analyzing speciation without prejudice? *Annals of the New York Academy of Sciences*, **1206**, 143–149.

Johannesson K, Panova M, Kemppainen P *et al.* (2010) Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1735–1747.

Johnson NS, Yun S-S, Thompson HT, Brant CO, Li W (2009) A synthesized pheromone induces upstream movement in female sea lamprey and summons them into traps. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 1021–1026.

Jones FC, Chan YF, Schmutz J *et al.* (2012) A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current biology: CB*, **22**, 83–90.

Jones AG, Moore GI, Kvarnemo C, Walker D, Avise JC (2003) Sympatric speciation as a consequence of male pregnancy in seahorses. *Proceedings of the National Academy of Sciences*, **100**, 6598–6603.

Jonsson B, Jonsson N (1993) Partial migration: niche shift versus sexual maturation in fishes. *Reviews in Fish Biology and Fisheries*, **3**, 348–365.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kaeuffer R, Peichel CL, Bolnick DI, Hendry AP (2012) Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. *Evolution; International Journal of Organic Evolution*, **66**, 402–418.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099–1106.

Kautt AF, Elmer KR, Meyer A (2012) Genomic signatures of divergent selection and speciation patterns in a "natural experiment", the young parallel radiations of Nicaraguan crater lake cichlid fishes. *Molecular Ecology*, **21**, 4770–4786.

Kawecki TJ, Holt RD (2002) Evolutionary consequences of asymmetric dispersal rates. *The American Naturalist*, **160**, 333–347.

Kemp PS, Russon IJ, Vowles AS, Lucas MC (2011) The influence of discharge and temperature on the ability of upstream migrant adult river lamprey (Lampetra fluviatilis) to pass experimental overshot and undershot weirs. *River Research and Applications*, **27**, 488–498.

Kim Y, Stephan W (2002) Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. *Genetics*, **160**, 765–777.

Kimura M, Weiss GH (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*, **49**, 561–576.

Kirkpatrick M, Ravigné V (2002) Speciation by Natural and Sexual Selection: Models and Experiments. *The American Naturalist*, **159**, S22–S35.

Klopfstein S, Currat M, Excoffier L (2006) The Fate of Mutations Surfing on the Wave of a Range Expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Knebelsberger T, Dunz AR, Neumann D, Geiger MF (2014) Molecular Diversity of Germany's Freshwater Fishes and Lampreys assessed by DNA barcoding. *Molecular Ecology Resources*, n/a–n/a.

Knebelsberger T, Dunz AR, Neumann D, Geiger MF (2015) Molecular diversity of Germany's freshwater fishes and lampreys assessed by DNA barcoding. *Molecular Ecology Resources*, **15**, 562–572.

Kremer A, Le Corre V (2012) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, **108**, 375–385.

Kruuk LE, Baird SJ, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–1971.

Kuraku S, Kuratani S (2006) Time Scale for Cyclostome Evolution Inferred with a Phylogenetic Diagnosis of Hagfish and Lamprey cDNA Sequences. *Zoological Science*, **23**, 1053–1064.

Kuratani S, Kuraku S, Murakami Y (2002) Lamprey as an evo-devo model: lessons from comparative embryology and molecular phylogenetics. *Genesis (New York, N.Y.: 2000)*, **34**, 175–183.

Lachaise D, Cariou M-L, David JR *et al.* (1988) Historical Biogeography of the Drosophila melanogaster Species Subgroup. In: *Evolutionary Biology* Evolutionary Biology. (eds Hecht MK, Wallace B, Prance GT), pp. 159–225. Springer US.

Lagadec R, Laguerre L, Menuet A *et al.* (2015) The ancestral role of nodal signalling in breaking L/R symmetry in the vertebrate forebrain. *Nature Communications*, **6**.

Langerhans RB, Gifford ME, Joseph EO (2007) Ecological Speciation in Gambusia Fishes. *Evolution*, **61**, 2056–2074.

Lang NJ, Roe KJ, Renaud CB *et al.* (2009) Novel relationships among lampreys (Petromyzontiformes) revealed by a taxonomically comprehensive molecular data set. *American Fisheries Society Symposium*, **72**, 41–55.

Laporte M, Rogers SM, Dion-Côté A-M *et al.* (2015) RAD-QTL Mapping Reveals Both Genome-Level Parallelism and Different Genetic Architecture Underlying the Evolution of Body Shape in Lake Whitefish (Coregonus clupeaformis) Species Pairs. *G3 (Bethesda, Md.)*, **5**, 1481–1491.

Larson EL, Hume GL, Andrés JA, Harrison RG (2012) Post-mating prezygotic barriers to gene exchange between hybridizing field crickets. *Journal of Evolutionary Biology*, **25**, 174–186.

Lasne E, Sabatié M-R, Evanno G (2010) Communal spawning of brook and river lampreys (Lampetra planeri and *L. fluviatilis*) is common in the Oir River (France). *Ecology of Freshwater Fish*, **19**, 323–325.

Leclerc E, Mailhot Y, Mingelbier M, Bernatchez L (2008) The landscape genetics of yellow perch (Perca flavescens) in a large fluvial ecosystem. *Molecular Ecology*, **17**, 1702–1717.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait land trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Lericolais G, Auffret J-P, Bourillet J-F (2003) The Quaternary Channel River: seismic stratigraphy of its palaeo-valleys and deeps. *Journal of Quaternary Science*, **18**, 245–260.

LeVasseur-Viens H, l&#xe8, ne *et al.* (2014) Individual Genetic Contributions to Genital Shape Variation between Drosophila simulans and D. mauritiana, Individual Genetic Contributions to Genital Shape Variation between Drosophila simulans and D. mauritiana. *International Journal of Evolutionary Biology, International Journal of Evolutionary Biology*, **2014, 2014**, e808247.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Liang M, Nielsen R (2014a) The Lengths of Admixture Tracts. *Genetics*, **197**, 953–967.

Liang M, Nielsen R (2014b) Understanding Admixture Fractions. *bioRxiv*, 008078.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li J, Li H, Jakobsson M *et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, **21**, 28–44.

Lindtke D, Buerkle CA (2015) The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution*, n/a–n/a.

Li WH, Nei M (1977) Persistence of common alleles in two related populations or species. *Genetics*, **86**, 901–914.

Liu J, Mercer JM, Stam LF *et al.* (1996) Genetic Analysis of a Morphological Shape Difference in the Male Genitalia of Drosophila Simulans and D. Mauritiana. *Genetics*, **142**, 1129–1145.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial Genetic Structure of a Tropical Understory Shrub, Psychotria officinalis (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 3009–3021.

Lucas MC, Bubb DH, Jang M-H, Ha K, Masters JEG (2009) Availability of and access to critical habitats in regulated rivers: effects of low-head barriers on threatened lampreys. *Freshwater Biology*, **54**, 621–634.

Ludlow AM, Magurran AE (2006) Gametic isolation in guppies (Poecilia reticulata). *Proceedings. Biological Sciences / The Royal Society*, **273**, 2477–2482.

Lu J, Li W-H, Wu C-I (2003) Comment on "Chromosomal Speciation and Molecular Divergence-Accelerated Evolution in Rearranged Chromosomes." *Science*, **302**, 988–988.

Lynch M (1991) The Genetic Interpretation of Inbreeding Depression and Outbreeding Depression. *Evolution*, **45**, 622.

Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, **104**, 8597–8604.

Lynch M, Conery J, Burger R (1995) Mutational Meltdowns in Sexual Populations. *Evolution*, **49**, 1067–1080.

Maheshwari S, Barbash DA (2011) The Genetics of Hybrid Incompatibilities. *Annual Review of Genetics*, **45**, 331–355.

Mailund T, Halager AE, Westergaard M *et al.* (2012) A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS Genet*, **8**, e1003125.

Maitland PS (1980) Review of the Ecology of Lampreys in Northern Europe. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 1944–1952.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**, 229–237.

Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2971–2986.

Malmqvist B (1980) The spawning migration of the brook lamprey, Lampetra planeri Bloch, in a South Swedish stream. *Journal of Fish Biology*, **16**, 105–114.

Manier MK, Lüpold S, Belote JM *et al.* (2013) Postcopulatory sexual selection generates speciation phenotypes in Drosophila. *Current biology: CB*, **23**, 1853–1862.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in genetics: TIG*, **24**, 133–141.

Marie Curie SPECIATION Network, Butlin R, Debelle A *et al.* (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27–39.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**.

Martin CH (2013) Strong assortative mating by diet, color, size, and morphology but limited progress toward sympatric speciation in a classic example: Cameroon crater lake cichlids. *Evolution; International Journal of Organic Evolution*, **67**, 2114–2123.

Martin CH, Cutler JS, Friel JP *et al.* (2015) Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, n/a–n/a.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, gr.159426.113.

Masly JP, Dalton JE, Srivastava S, Chen L, Arbeitman MN (2011) The genetic basis of rapidly evolving male genital morphology in Drosophila. *Genetics*, **189**, 357–374.

Masly JP, Masly JP (2011) 170 Years of "Lock-and-Key": Genital Morphology and Reproductive Isolation, 170 Years of "Lock-and-Key": Genital Morphology and Reproductive Isolation. *International Journal of Evolutionary Biology, International Journal of Evolutionary Biology*, **2012, 2012**, e247352.

Masly JP, Presgraves DC (2007) High-resolution genome-wide dissection of the two rules of speciation in Drosophila. *PLoS biology*, **5**, e243.

Mateus CS, Almeida PR, Quintella BR, Alves MJ (2011) MtDNA markers reveal the existence of allopatric evolutionary lineages in the threatened lampreys Lampetra fluviatilis (L.) and Lampetra planeri (Bloch) in the Iberian glacial refugium. *Conservation Genetics*, **12**, 1061–1074.

Mateus CS, RodrguezMuoz R, Quintella BR, Alves MJ, Almeida PR (2012) REVIEW Lampreys of the Iberian Peninsula: distribution, population status and conservation. *Endangered Species Research*, **16**, 183–198.

Mateus CS, Stange M, Berner D *et al.* (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology*, **23**, R649–R650.

Mathew LA, Jensen JD (2015) Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography. *Frontiers in Genetics*, **6**.

Matute DR, Butler IA, Turissini DA, Coyne JA (2010) A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science (New York, N.Y.)*, **329**, 1518–1521.

Mayden R.L. (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Species: the Units of Diversity (eds. Claridge MA, Dawah HA & Wilson MR), pp. 381–424. Chapman & Hall, London

Mayr E. (1942) *Systematics and the Origin of Species*. Columbia University Press, New York.

McCairns RJS, Bernatchez L (2008) Landscape genetic analyses reveal cryptic population structure and putative selection gradients in a large-scale estuarine environment. *Molecular Ecology*, **17**, 3901–3916.

McKinnon JS, Mori S, Blackman BK *et al.* (2004) Evidence for ecology's role in speciation. *Nature*, **429**, 294–298.

McNeil CL, Bain CL, Macdonald SJ (2011) Multiple Quantitative Trait Loci Influence the Shape of a Male-Specific Genital Structure in Drosophila melanogaster. *G3 (Bethesda, Md.)*, **1**, 343–351.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.

Meirmans PG, Van Tienderen PH (2004) genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Mendelson TC, Imhoff VE, Venditti JJ (2007) The Accumulation of Reproductive Barriers During Speciation: Postmating Barriers in Two Behaviorally Isolated Species of Darters (percidae: Etheostoma). *Evolution*, **61**, 2596–2606.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, **28**, 659–669.

Michel AP, Sim S, Powell THQ *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*, **107**, 9724–9729.

Morita K, Yamamoto S (2002) Effects of Habitat Fragmentation by Damming on the Persistence of Stream-Dwelling Charr Populations. *Conservation Biology*, **16**, 1318–1323.

Morrissey MB, de Kerckhove DT (2009) The maintenance of genetic variation due to asymmetric gene flow in dendritic metapopulations. *The American Naturalist*, **174**, 875–889.

Moyle LC, Nakazato T (2009) Complex Epistasis for Dobzhansky–Muller Hybrid Incompatibility in Solanum. *Genetics*, **181**, 347–351.

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 409–421.

Nadachowska-Brzyska K, Burri R, Olason PI *et al.* (2013) Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS genetics*, **9**, e1003942.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343–353.

Nagel L, Schluter D (1998) Body Size, Natural Selection, and Speciation in Sticklebacks. *Evolution*, **52**, 209–218.

Nance HA, Klimley P, Galván-Magaña F, Martínez-Ortíz J, Marko PB (2011) Demographic Processes Underlying Subtle Patterns of Population Structure in the Scalloped Hammerhead Shark, Sphyrna lewini. *PLoS ONE*, **6**, e21459.

Nater A, Greminger MP, Arora N *et al.* (2015) Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation. *Molecular Ecology*, **24**, 310–327.

Navarro A, Barton NH (2003a) Accumulating Postzygotic Isolation Genes in Parapatry: A New Twist on Chromosomal Speciation. *Evolution*, **57**, 447–459.

Navarro A, Barton NH (2003b) Chromosomal Speciation and Molecular Divergence--Accelerated Evolution in Rearranged Chromosomes. *Science*, **300**, 321–324.

Navarro A, Marquès-Bonet T, Barton NH (2003) Response to Comment on “Chromosomal Speciation and Molecular Divergence-Accelerated Evolution in Rearranged Chromosomes.” *Science*, **302**, 988–988.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Nei M, Maruyama T (1975) Letters to the editors: Lewontin-Krakauer test for neutral genes. *Genetics*, **80**, 395.

Nei M, Nozawa M (2011) Roles of Mutation and Selection in Speciation: From Hugo de Vries to the Modern Genomic Era. *Genome Biology and Evolution*, **3**, 812–829.

Nei M, Suzuki Y, Nozawa M (2010) The Neutral Theory of Molecular Evolution in the Genomic Era. *Annual Review of Genomics and Human Genetics*, **11**, 265–289.

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *Journal of Molecular Evolution*, **19**, 153–170.

Nei M, Takahata N (1993) Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, **37**, 240–244.

Nielsen EE, Bach LA, Kotlicki P (2006) hybridlab (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes*, **6**, 971–973.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews. Genetics*, **8**, 857–868.

Nielsen R, Hubisz MJ, Hellmann I *et al.* (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, **19**, 838–849.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Niemiller ML, Fitzpatrick BM, Miller BT (2008) Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. *Molecular Ecology*, **17**, 2258–2275.

Nilsson C, Reidy CA, Dynesius M, Revenga C (2005) Fragmentation and Flow Regulation of the World's Large River Systems. *Science*, **308**, 405–408.

Noor M a. F, 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 (2012) *Ecological Speciation*. Oxford Series in Ecology and Evolution.Oxford.

Nosil P, Crespi BJ, Sandoval CP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440–443.

Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 332–342.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009a) Divergent selection and heterogeneous genomic divergence. *Molecular ecology*, **18**, 375–402.

Nosil P, Harmon LJ, Seehausen O (2009b) Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, **24**, 145–156.

Okada K, Yamazaki Y, Yokobori S, Wada H (2010) Repetitive sequences in the lamprey mitochondrial DNA control region and speciation of Lethenteron. *Gene*, **465**, 45–52.

Orr HA (1995) The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, **139**, 1805–1813.

Orr HA (2005) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, **6**, 119–127.

Orr HA, Irving S (2001) Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization. *Genetics*, **158**, 1089–1100.

Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating dobzhansky-muller incompatibilities. *Evolution*, **55**, 1085–1094.

Palmer ME, Feldman MW (2009) Dynamics of Hybrid Incompatibility in Gene Networks in a Constant Environment. *Evolution*, **63**, 418–431.

Palumbi SR (2001) Humans as the world's greatest evolutionary force. *Science (New York, N.Y.)*, **293**, 1786–1790.

Paz-Vinas I, Blanchet S (2015) Dendritic connectivity shapes spatial patterns of genetic diversity: a simulation-based study. *Journal of Evolutionary Biology*, **28**, 986–994.

Paz-Vinas I, Loot G, Stevens VM, Blanchet S (2015) Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems. *Molecular Ecology*, **24**, 4586–4604.

Paz-Vinas I, Quéméré E, Chikhi L, Loot G, Blanchet S (2013) The demographic history of populations experiencing asymmetric gene flow: combining simulated and empirical data. *Molecular Ecology*, **22**, 3279–3291.

Peakall R, Smouse P (2012) GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics*, bts460.

Pearce RJ, Pota H, Evehe M-SB *et al.* (2009) Multiple Origins and Regional Dispersal of Resistant dhps in African Plasmodium falciparum Malaria. *PLoS Med*, **6**, e1000055.

Pearse DE, Hayes SA, Bond MH *et al.* (2009) Over the falls? Rapid evolution of ecotypic differentiation in steelhead/rainbow trout (Oncorhynchus mykiss). *The Journal of Heredity*, **100**, 515–525.

Pereira AM, Robalo JI, Freyhof J *et al.* (2010) Phylogeographical analysis reveals multiple conservation units in brook lampreys Lampetra planeri of Portuguese streams. *Journal of Fish Biology*, **77**, 361–371.

Perrier C, Bourret V, Kent MP, Bernatchez L (2013) Parallel and nonparallel genome-wide divergence among replicate population pairs of freshwater and anadromous Atlantic salmon. *Molecular Ecology*, **22**, 5577–5593.

Pettengill JB, Moeller DA (2012) Phylogeography of speciation: allopatric divergence and secondary contact between outcrossing and selfing Clarkia. *Molecular Ecology*, **21**, 4578–4592.

Pettersson JCE, Hansen MM, Bohlin T (2001) Does dispersal from landlocked trout explain the coexistence of resident and migratory trout females in a small stream? *Journal of Fish Biology*, **58**, 487–495.

Piavis GW, Howell JH, Smith AJ (1970) *Experimental hybridization among five species of lampreys from the Great Lakes*. United States Geological Survey.

Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*, **8**, e1002967.

Pinho C, Harris DJ, Ferrand N (2008) Non-equilibrium estimates of gene flow inferred from nuclear genealogies suggest that Iberian and North African wall lizards (Podarcis spp.) are an assemblage of incipient species. *BMC Evolutionary Biology*, **8**, 63.

Pinho C, Hey J (2010) Divergence with Gene Flow: Models and Data. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 215–230.

Pollux BJA, Luteijn A, Van Groenendael JM, Ouborg NJ (2009) Gene flow and genetic structure of the aquatic macrophyte Sparganium emersum in a linear unidirectional river. *Freshwater Biology*, **54**, 64–76.

Potter IC (1980) Ecology of Larval and Metamorphosing Lampreys. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 1641–1657.

Potter MWH& IC, Potter IC (1971) *The Biology of Lampreys. Volume 1*. Academic Press, London, New York.

Powell THQ, Hood GR, Murphy MO *et al.* (2013) Genetic Divergence Along the Speciation Continuum: The Transition from Host Race to Species in Rhagoletis (diptera: Tephritidae). *Evolution*, **67**, 2561–2576.

Præbel K, Knudsen R, Siwertsson A *et al.* (2013) Ecological speciation in postglacial European whitefish: rapid adaptive radiations into the littoral, pelagic, and profundal lake habitats. *Ecology and Evolution*, **3**, 4970–4986.

Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, **11**, 175–180.

Price CS, Kim CH, Gronlund CJ, Coyne JA (2001) Cryptic reproductive isolation in the Drosophila simulans species complex. *Evolution; International Journal of Organic Evolution*, **55**, 81–92.

Pritchard JK, Di Rienzo A (2010) Adaptation – not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology: CB*, **20**, R208–215.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pudlo P, Marin J-M, Estoup A *et al.* (2014) ABC model choice via random forests. *arXiv:1406.6288 [q-bio, stat]*.

R Core Team (2013). — European Environment Agency (EEA). *Available at http://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006* (accessed August 3, 2015).

Rabosky DL (2013) Diversity-Dependence, Ecological Speciation, and the Role of Competition in Macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 481–502.

Rabosky DL, Matute DR (2013) Macroevolutionary speciation rates are decoupled from the evolution of intrinsic reproductive isolation in Drosophila and birds. *Proceedings of the National Academy of Sciences*, **110**, 15354–15359.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, **16**, 359–371.

Raeymaekers JAM, Konijnendijk N, Larmuseau MHD *et al.* (2014) A gene with major phenotypic effects as a target for selection vs. homogenizing gene flow. *Molecular Ecology*, **23**, 162–181.

Raeymaekers JAM, Maes GE, Geldof S *et al.* (2008) Modeling genetic connectivity in sticklebacks as a guideline for river restoration. *Evolutionary Applications*, **1**, 475–488.

Raeymaekers JAM, Van Houdt JKJ, Larmuseau MHD, Geldof S, Volckaert FAM (2007) Divergent selection as revealed by P(ST) and QTL-based F(ST) in three-spined stickleback (Gasterosteus aculeatus) populations along a coastal-inland gradient. *Molecular Ecology*, **16**, 891–905.

Ralph P, Coop G (2010) Parallel Adaptation: One or Many Waves of Advance of an Advantageous Allele? *Genetics*, **186**, 647–668.

Ravinet M, Westram A, Johannesson K *et al.* (2015) Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. *Molecular Ecology*, n/a–n/a.

Reis-Santos P, McCormick SD, Wilson JM (2008) Ionoregulatory changes during metamorphosis and salinity exposure of juvenile sea lamprey (Petromyzon marinus L.). *The Journal of Experimental Biology*, **211**, 978–988.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). *Molecular Ecology*, **20**, 545–559.

Rice WR (1984) Disruptive Selection on Habitat Preference and the Evolution of Reproductive Isolation: A Simulation Study. *Evolution*, **38**, 1251–1260.

Rice WR (1989) Analyzing Tables of Statistical Tests. *Evolution*, **43**, 223.

Rice AM, Leichty AR, Pfennig DW (2009) Parallel evolution and ecological selection: replicated character displacement in spadefoot toads. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 4189–4196.

Rieseberg LH (2009) Evolution: Replacing Genes and Traits through Hybridization. *Current Biology*, **19**, R119–R122.

Robertson A (1975) Letters to the editors: Remarks on the Lewontin-Krakauer test. *Genetics*, **80**, 396.

Rockman MV (2012) The Qtn Program and the Alleles That Matter for Evolution: All That's Gold Does Not Glitter. *Evolution*, **66**, 1–17.

Roda F, Ambrose L, Walter GM *et al.* (2013) Genomic evidence for the parallel evolution of coastal forms in the Senecio lautus complex. *Molecular Ecology*, **22**, 2941–2952.

Rodríguez-Muñoz R, Tregenza T (2009) Genetic compatibility and hatching success in the sea lamprey (Petromyzon marinus). *Biology Letters*, **5**, 286–288.

Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D (2014) The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, **23**, 3944–3956.

Roesti M, Hendry AP, Salzburger W, Berner D (2012a) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome—patterns and consequences. *Molecular Ecology*, **22**, 3014–3027.

Roesti M, Salzburger W, Berner D (2012b) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (Coregonus sp. Salmonidae) species pairs. *Molecular biology and evolution*, **24**, 1423–1438.

Rosenberg NA (2004) distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical Divergence and Gene Flow in the Genus Zea. *Genetics*, **181**, 1399–1413.

Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of Polymorphism and Demographic History in Natural Populations of Arabidopsis lyrata. *PLoS ONE*, **3**, e2411.

Rougemont Q, Gaigher A, Lasne E *et al.* (2015) Low reproductive isolation and highly variable levels of gene flow reveal limited progress toward speciation between European river and brook lampreys. *Journal of Evolutionary Biology*, n/a–n/a.

Rousset F (1997) Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics*, **145**, 1219–1228.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Roux C, Fraïsse C, Castric V *et al.* (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. *Journal of Evolutionary Biology*, **27**, 1662–1675.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. *Molecular Biology and Evolution*, mst066. **7,** 1574-1587

Rundle HD (2002) A test of ecologically dependent postmating isolation between sympatric sticklebacks. *Evolution; international journal of organic evolution*, **56**, 322–329.

Rundle HD, Nagel L, Boughman JW, Schluter D (2000) Natural Selection and Parallel Speciation in Sympatric Sticklebacks. *Science*, **287**, 306–308.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Russon IJ, Kemp PS, Lucas MC (2011) Gauging weirs impede the upstream migration of adult river lamprey Lampetra fluviatilis. *Fisheries Management and Ecology*, **18**, 201–210.

Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Saccheri I, Kuussaari M, Kankare M *et al.* (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature*, **392**, 491–494.

Salewski V (2003) Satellite species in lampreys: a worldwide trend for ecological speciation in sympatry? *Journal of Fish Biology*, **63**, 267–279.

Sasabe M, Takami Y, Sota T (2007) The genetic basis of interspecific differences in genital morphology of closely related carabid beetles. *Heredity*, **98**, 385–391.

Sasabe M, Takami Y, Sota T (2010) QTL for the species-specific male and female genital morphologies in Ohomopterus ground beetles. *Molecular Ecology*, **19**, 5231–5239.

Schierup, M. H., & Christiansen. F.B. 1996. Inbreeding depression and outbreeding depression in plants. *Heredity* **77,**461–468.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science (New York, N.Y.)*, **323**, 737–741.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106 Suppl 1**, 9955–9962.

Schluter D, McPhail JD (1992) Ecological Character Displacement and Speciation in Sticklebacks. *The American Naturalist*, **140**, 85–108.

Schluter D, McPhail JD (1993) Character displacement and replicate adaptive radiation. *Trends in Ecology & Evolution*, **8**, 197–200.

Schluter D, Nagel LM (1995) Parallel Speciation by Natural Selection. *The American Naturalist*, **146**, 292–301.

Schreiber A, Engelhorn R (1998) Population genetics of a cyclostome species pair, river lamprey (Lampetra fluviatilis L.) and brook lamprey (Lampetra planeri Bloch). *Journal of Zoological Systematics and Evolutionary Research*, **36**, 85–99.

Schrider DR, Mendes FK, Hahn MW, Kern AD (2015) Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. *Genetics*, genetics.115.174912.

Sedghifar A, Brandvain Y, Ralph P, Coop G (2015) The Spatial Mixing of Genomes in Secondary Contact Zones. *Genetics*, genetics.115.179838.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Servedio MR, Noor MAF (2003) THE ROLE OF REINFORCEMENT IN SPECIATION: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 339–364.

Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: “magic” but not rare? *Trends in Ecology & Evolution*, **26**, 389–397.

Shaw KL (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16122–16127.

Shimeld SM, Donoghue PCJ (2012) Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). *Development*, **139**, 2091–2099.

Shimoda N, Knapik EW, Ziniti J *et al.* (1999) Zebrafish Genetic Map with 2000 Microsatellite Markers. *Genomics*, **58**, 219–232.

Singh ND, Jensen JD, Clark AG, Aquadro CF (2013) Inferences of demography and selection in an African population of Drosophila melanogaster. *Genetics*, **193**, 215–228.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Smith JJ, Kuraku S, Holt C *et al.* (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nature Genetics*, **45**, 415–421.

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The Biology of Speciation. *Evolution*, **64**, 295–315.

Sobel JM, Streisfeld MA (2015) Strong premating reproductive isolation drives incipient speciation in Mimulus aurantiacus. *Evolution*, **69**, 447–461.

Sorensen PW, Fine JM, Dvornikovs V *et al.* (2005) Mixture of new sulfated steroids functions as a migratory pheromone in the sea lamprey. *Nature Chemical Biology*, **1**, 324–328.

Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick Insect Genomes Reveal Natural Selection’s Role in Parallel Speciation. *Science*, **344**, 738–742.

Sorrells TR, Johnson AD (2015) Making sense of transcription networks. *Cell*, **161**, 714–723.

Sousa VC, Carneiro M, Ferrand N, Hey J (2013) Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. *Genetics*, **194**, 211–233.

Spice EK, Goodman DH, Reid SB, Docker MF (2012a) Neither philopatric nor panmictic: microsatellite and mtDNA evidence suggests lack of natal homing but limits to dispersal in Pacific lamprey. *Molecular Ecology*, **21**, 2916–2930.

Spice EK, Goodman DH, Reid SB, Docker MF (2012b) Neither philopatric nor panmictic: microsatellite and mtDNA evidence suggests lack of natal homing but limits to dispersal in Pacific lamprey. *Molecular Ecology*, **21**, 2916–2930.

Spielman D, Brook BW, Frankham R (2004) Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15261–15264.

Stefansson, S.O., Björnsson, B.Th., Ebbesson, L.O.E. and McCormick, S.D. 2008. Smoltification. In: Fish Larval Physiology (Finn, R.N. and Kapoor, B.G., Eds). Science Publishers, Inc. Enfield (NH) & IBH Publishing Co. Pvt. Ltd., New Delhi ISBN 978-1-57808-388-6.

Steinberg EK, Lindner KR, Gallea J *et al.* (2002) Rates and Patterns of Microsatellite Mutations in Pink Salmon. *Molecular Biology and Evolution*, **19**, 1198–1202.

Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive Variation in Beach Mice Produced by Two Interacting Pigmentation Genes. *PLoS Biol*, **5**, e219.

Strasburg JL, Rieseberg LH (2008) Molecular demographic history of the annual sunflowers Helianthus annuus and H. petiolaris--large effective population sizes and rates of long-term gene flow. *Evolution; International Journal of Organic Evolution*, **62**, 1936–1950.

Strasburg JL, Rieseberg LH (2010) How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. *Molecular Biology and Evolution*, **27**, 297–310.

Strasburg JL, Rieseberg LH (2011) Interpreting the estimated timing of migration events between hybridizing species. *Molecular Ecology*, **20**, 2353–2366.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tallmon DA, Luikart G, Waples RS (2004) The alluring simplicity and complex reality of genetic rescue. *Trends in Ecology & Evolution*, **19**, 489–496.

Tao Y, Chen S, Hartl DL, Laurie CC (2003a) Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. *Genetics*, **164**, 1383–1397.

Tao Y, Zeng Z-B, Li J, Hartl DL, Laurie CC (2003b) Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana. II. Mapping hybrid male sterility loci on the third chromosome. *Genetics*, **164**, 1399–1418.

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring Coalescence Times from DNA Sequence Data. *Genetics*, **145**, 505–518.

Taverny C, Élie P (2010) *Les lamproies en Europe de l'Ouest: écophases, espèces et habitats*. Editions Quae.

Templeton AR (1980) The theory of speciation via the founder principle. *Genetics*, **94**, 1011–1038.

Templeton AR (2008) The reality and importance of founder speciation in evolution. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **30**, 470–479.

Templeton AR, Robertson RJ, Brisson J, Strasburg J (2001) Disrupting evolutionary processes: The effect of habitat fragmentation on collared lizards in the Missouri Ozarks. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5426–5432.

Thrower F, Iii CG, Nielsen J, Joyce J (2004) A Comparison of Genetic Variation Between an Anadromous Steelhead, Oncorhynchus mykiss, Population and Seven Derived Populations Sequestered in Freshwater for 70 Years. *Environmental Biology of Fishes*, **69**, 111–125.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**. **5**. doi:10.1038/ncomms6770

Ting C-T, Takahashi A, Wu C-I (2001) Incipient speciation by sexual isolation in Drosophila: Concurrent evolution at multiple loci. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6709–6713.

Torterotot J-B, Perrier C, Bergeron NE, Bernatchez L (2014) Influence of Forest Road Culverts and Waterfalls on the Fine-Scale Distribution of Brook Trout Genetic Diversity in a Boreal Watershed. *Transactions of the American Fisheries Society*, **143**, 1577–1591.

Toucane S (2008) Reconstruction des transferts sédimentaires en provenance du système glaciaire de mer d'irlande et du paleo-fleuve manche au cours des derniers cycles climatiques, These de Doctorat. Univ. Bordeaux 370pp

Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology & Evolution*, **16**, 330–343.

Turelli M, Orr HA (1995) The dominance theory of Haldane's rule. *Genetics*, **140**, 389–402.

Turelli M, Orr HA (2000) Dominance, epistasis and the genetics of postzygotic isolation. *Genetics*, **154**, 1663–1679.

Turner TL, Hahn MW (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, **19**, 848–850.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic Islands of Speciation in Anopheles gambiae. *PLoS Biol*, **3**, e285.

Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends in ecology & evolution*, **16**, 381–390.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**, 451–460.

Via S, Conte G, Mason-Foley C, Mills K (2012) Localizing FST outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. *Molecular Ecology*, **21**, 5546–5560.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of Variation Across Marker Loci as Evidence of Selection. *Genetics*, **158**, 1811–1823.

Vitalis R, Gautier M, Dawson KJ, Beaumont MA (2014) Detecting and measuring selection from gene frequency data. *Genetics*, **196**, 799–817.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997a) Human Domination of Earth's Ecosystems. *Science*, **277**, 494–499.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997b) Human Domination of Earth's Ecosystems. *Science*, **277**, 494–499.

Vladykov VD, Kott E (1979) Satellite species among the holarctic lampreys (Petromyzonidae). *Canadian Journal of Zoology*, **57**, 860–867.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biol*, **4**, e72.

Wakeley J (2008) *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colo.

Waldman J, Grunwald C, Wirgin I (2008) Sea lamprey Petromyzon marinus: an exception to the rule of homing in anadromous fishes. *Biology Letters*, **4**, 659–662.

Wang RJ, White MA, Payseur BA (2015) The Pace of Hybrid Incompatibility Evolution in House Mice. *Genetics*, **201**, 229–242.

Waser NM, Price MV (1985) The effect of nectar guides on pollinator preference: experimental studies with a montane herb. *Oecologia*, **67**, 121–126.

Waser NM, Price MV (1994) Crossing-Distance Effects in Delphinium nelsonii: Outbreeding and Inbreeding Depression in Progeny Fitness. *Evolution*, **48**, 842.

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. *Molecular Ecology*, **23**, 3935–3937.

Weissenberg, R. 1925. Fluss- und Bachneuunauge (*Lampetra fluviatilis* L. und *Lampetra planeri* Bloch), ein morphologisch-biologischer *Vergleich. Zool. Anz* . **63** :293–306.

West-Eberhard MJ (1983) Sexual Selection, Social Competition, and Speciation. *The Quarterly Review of Biology*, **58**, 155–183.

Westram AM, Galindo J, Alm Rosenblad M *et al.* (2014) Do the same genes underlie parallel phenotypic divergence in different Littorina saxatilis populations? *Molecular Ecology*, **23**, 4603–4616.

Whiteley AR, Coombs JA, Hudy M *et al.* (2013) Fragmentation and patch size shape genetic structure of brook trout populations. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 678–688.

Wiley C, Qvarnström A, Andersson G, Borge T, Saetre G-P (2009) Postzygotic isolation over multiple generations of hybrid descendents in a natural hybrid zone: how well do single-generation estimates reflect reproductive isolation? *Evolution; International Journal of Organic Evolution*, **63**, 1731–1739.

Williams TH, Mendelson TC (2014) Quantifying Reproductive Barriers in a Sympatric Pair of Darter Species. *Evolutionary Biology*, **41**, 212–220.

Williamson SH, Hernandez R, Fledel-Alon A *et al.* (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, **102**, 7882–7887.

Wilson GA, Rannala B (2003) Bayesian Inference of Recent Migration Rates Using Multilocus Genotypes. *Genetics*, **163**, 1177–1191.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu CI, Hollocher H, Begun DJ *et al.* (1995) Sexual isolation in Drosophila melanogaster: a possible case of incipient speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 2519–2523.

Yamamoto S, Morita K, Koizumi I, Maekawa K (2004) Genetic Differentiation of White-Spotted Charr (Salvelinus leucomaenis) Populations After Habitat Fragmentation: Spatial–Temporal Changes in Gene Frequencies. *Conservation Genetics*, **5**, 529–538.

Yamazaki Y, Yokoyama R, Nishida M, Goto A (2006) Taxonomy and molecular phylogeny of Lethenteron lampreys in eastern Eurasia. *Journal of Fish Biology*, **68**, 251–269.

Yang J, Benyamin B, McEvoy BP *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

Yeaman S, Otto SP (2011) Establishment and Maintenance of Adaptive Genetic Divergence Under Migration, Selection, and Drift. *Evolution*, **65**, 2123–2129.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution; international journal of organic evolution*, **65**, 1897–1911.

Yeates SE, Diamond SE, Einum S *et al.* (2013) Cryptic Choice of Conspecific Sperm Controlled by the Impact of Ovarian Fluid on Sperm Swimming Behavior. *Evolution*, **67**, 3523–3536.

Young A, Boyle T, Brown T (1996) The population genetic consequences of habitat fragmentation for plants. *Trends in Ecology & Evolution*, **11**, 413–418.

Yue GH, David L, Orban L (2006) Mutation rate and pattern of microsatellites in common carp (Cyprinus carpio L.). *Genetica*, **129**, 329–331.

Zanandrea GSJ (1959) Speciation among Lampreys. *Nature*, **184**, 380–380.

Zeng Z-B, Liu J, Stam LF *et al.* (2000) Genetic Architecture of a Morphological Shape Difference Between Two Drosophila Species. *Genetics*, **154**, 299–310.

Zhang J, Wang X, Podlaha O (2004) Testing the Chromosomal Speciation Hypothesis for Humans and Chimpanzees. *Genome Research*, **14**, 845–851.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer New York, New York, NY.

Scientific Activities

***Additional Papers* (Not in the thesis)**

J. Martin, **Q. Rougemont,** H. Drouineau, S. Launey, P. Jattteau, G.Bareille, S. Berail, C. Pécheyran, E. Feunteun, S. Roques, D. Clavé, D.J. Nachon, C. Antunes, M. Mota, E.Réveillac and F. Daverat. (2015). Dispersal capacities of anadromous Allis shad population inferred from a coupled genetic and otolith approach. **Canadian Journal of Fisheries and Aquatic Science**. 72(7): 991-1003.

R. Lagadec, L. Laguerre, A. Menuet, A. Amara, C. Rocancourt, P. Péricard, B.G. Godard, M. C. Rodicio, I. Rodriguez-Moldes, H. Mayeur, **Q. Rougemont**, S. Mazan and A. Boutet. (2015). The ancestral role of Nodal signaling in breaking L/R symmetry in the vertebrate forebrain, **Nature Communication**, 6:6686.

Q. Rougemont, A.L. Besnard, J.L. Baglinière, and S. Launey. (2014). Characterization of thirteen new microsatellite markers for allis shad (Alosa alosa) and twaite shad (Alosa fallax). **Conservation Genetic Ressources**. 7(1): 259-261

Oral Communications

Q. Rougemont, Launey, S. Gaigher, A. Besnard, A-L. Lasne, E. Evanno, G. Genome Wide Divergence Among population pairs of parasitic and non-parasitic lampreys (*Lampetra fluviatilis* and *L. planeri*). 144th Annual Meeting of the **American Fisheries Society**, Québec City, Canada, 17th-21st August 2014.

Q. Rougemont, Launey, S. Gaigher, A. Besnard, A-L. Lasne, E. Evanno, G. Evidence for low reproductive isolation and strong gene flow between the river lamprey (Lampetra fluviatilis) and brook lamprey (Lampetra planeri). **International Conference of the Institute of Fisheries Management**, York, UK 6th-8th 2014. May.

Q. Rougemont, Launey, S. Gaigher, A. Besnard, A-L. Lasne, E. Evanno, G. Dispersal, gene flow and ecological speciation in lamprey. 14th Congress of the **European Society for Evolutionary Biology**, Lisbonn, Portugal, 19-24 August 2013.

Posters

Q. Rougemont, C. Perrier, A. Gaigher, S. Launey, G. Evanno. Genome-wide divergence and demographic history of the parasitic and non-parasitic European lampreys (Lampetra planeri and Lampetra fluviatilis). **European Society for Evolutionary Biology**. Lausanne, Switzerland, 10-14th August 2015.

Q. Rougemont, C. Perrier, A. Gaigher, S. Launey, G. Evanno. Genome-wide divergence parasitic and non-parasitic European lampreys (*Lampetra planeri* and *Lampetra fluviatilis*). **SMBE satellite meeting on Biological Adaptation.** Montpellier 25-28th May 2015.

Q. Rougemont, A. Gaigher, S. Launey, G. Evanno. Genome-wide divergence parasitic and non-parasitic European lampreys (*Lampetra planeri* and *Lampetra fluviatilis*). **Workshop: Genomics of the Speciation Continuum.** Fribourg 4-5th September 2014.

Q. Rougemont, S. Launey, G. Evanno. Evolution of anadromy in lampreys. PhD Student Internal Seminar INRA. Dourdan, France, 15-17th January 2013.

Teaching Experience

Courses of Statistics, Geographic Information System and Ecology To Bachelor students, University of Rennes I – 64h.

Intern supervision

Co-supervision of a Master 2 thesis: Population genetics of brook lampreys, impact of dams on gene flow and test for outbreeding depression

International mobility at the University of Lausanne, deportment of Ecology and Evolution supervised by J. Goudet (Switzerland) (December 2014 – March 2015).

Member of Agreenium International Research School (EIR-A) since 2012 (http://www.agreenium.org/)

Training undertaken

Writing scientific papers (Montpellier – February 2013)

Transcriptomics Analysis in R (Agrocampus – Ouest, September 2013)

Linux Programming (Montpellier - February2014)

Initiation to Python (Le Rheu – February 2014)

Gallaxy and Linux Programming (Roscoff – October 2014)

Population Genetics of Marine Organisms (Sete – February 2015)

Fellowships

2014-2015: Travel Grant (2000€): Rennes Metropoles

2014-2015: Travel Grant (2000€): INRA – EIR-A

ANNEXE 2 (Modèle dernière page de thèse)

VU :

Le Directeur de Thèse
Guillaume Evanno

VU :

Le Responsable de l'École Doctorale

VU pour autorisation de soutenance

Rennes, le

Le Président de l'Université de Rennes 1

Guy CATHELINEAU

VU après soutenance pour autorisation de publication :

Le Président de Jury,
(Nom et Prénom)

Résumé étendu en Français

Evolution de la divergence entre la lamproie fluviatile (*Lampetra fluviatilis*) et la lamproie de Planer (*L. planeri*) inférée par des approches expérimentales et de génomique des populations

Q. Rougemont

Comprendre les processus à l'origine de la formation des espèces est un enjeu fondamental de la biologie évolutive. Selon le concept biologique de l'espèce, une espèce correspond à un groupe d'individus se reproduisant entre eux et isolé reproductivement d'autres groupes. Selon cette définition l'isolement reproducteur est une propriété du génome et les hybrides ne devraient pas exister dans la nature. Hors, l'hybridation adaptative et le flux de gènes interspécifique sont communément observés. Le concept génique de l'espèce en revanche permet de prendre en compte la nature semi-perméable de la barrière aux flux de gènes si bien qu'une grande partie du génome peut être librement échangée entre espèces alors que d'autres régions restent imperméables au flux de gènes. Cette thèse étudie le processus de spéciation entre la lamproie fluviatile *(Lampetra fluviatilis)* et la lamproie de planer (*L. planeri*) en mettant en lumière la notion de barrière semi-perméable aux flux de gènes. Ces deux espèces présentent des stratégies d'histoire de vie extrêmement différentes : *L. fluviatilis* est parasite et anadrome alors que *L. planeri* n'est pas parasite et reste strictement dulcicole. Toutefois, leur degré d'isolement reproducteur et leur histoire de divergence demeurent méconnus. Le but majeur est de comprendre le processus de divergence historique ou en cours entre *L. fluviatilis* et *L. planeri.* Cette question a été étudiée par une approche multidisciplinaire combinant des mesures expérimentales de l'isolement reproducteur et des inférences du flux de gènes en populations naturelles basées sur des marqueurs 'neutres' microsatellite et des marqueurs RAD. Des approches par simulations ont aussi été implémentées afin de reconstruire le processus historique de divergence sous-jacent. Un second objectif, de nature plus appliqué en lien avec les thématiques de biologie de la conservation visait à déterminer l'impact des barrières naturelles et de la fragmentation anthropique sur l'intégrité génétique des populations de *L. planeri.* Une approche de génétique du paysage testant les effets de la distance, des barrières à la migration ainsi que de l'admixture avec l'écotype de *L. fluviatilis* sur la distribution spatiale de la diversité génétique des *L. planeri* a été développée.

Le degré d'isolement reproducteur entre *L. planeri* et *L. fluviatilis* a été mesuré expérimentalement ainsi que dans les populations naturelles (Chapitre 2). Nous avons tout d'abord mesuré l'isolement reproducteur post-zygotique par fécondation *in vitro* entre mâles et femelles des deux espèces et avons mis en évidence un isolement reproducteur très faible avec des taux de

fécondation et de survie proche de 100 % des croisements interspécifiques. Afin de vérifier la capacité des mâles de *L. planeri à* s'accoupler avec des femelles de *L. fluviatilis* et à produire des descendants viables, des expériences de croisements en condition semi-naturelles ont été effectuées. Les analyses de parenté montrent que les mâles sont capable de s'accoupler avec les femelles *L. fluviatilis* même s'il est possible qu'un isolement par la taille existe. Les analyses effectuées en populations naturelles sont basées sur un total de 10 paires de populations dont 5 sympatriques et 5 parapatriques. Les résultats montrent des degrés de flux de gènes variables en sympatrie avec une paires très faiblement différenciée (F_{ST} = 0.008) et d'autres paires plus fortement différenciées (F_{ST} = 0.08). A partir de ces résultats, nous proposons l'existence d'un gradient de divergence avec certaines populations correspondant à deux écotypes d'une seule espèce et d'autres populations correspondant à des écotypes à isolement reproducteur partiel. Ces résultats suggèrent que les barrières génétiques endogènes peuvent exister mais ne réduisent la migration efficace que sur des faibles proportions du génome. L'analyse des populations parapatriques montre une forte différentiation des populations de *L. planeri* due à l'effet combiné de l'isolement géographique et des barrières anthropiques à la migration. Alors que chaque population de *L. planeri* de chaque cours d'eau forme un cluster indépendant, les populations de *L. fluviatilis* forment deux clusters proches de la panmixie, de manière similaires aux observations chez les écotypes d'épinoche marine et résidente *Gasterosteus aculeatus.* A partir de ces résultats, nous formulons deux hypothèses d'histoires démographiques qui ont pu générer ces patterns : soit la sélection écologique en cours génère une différenciation génétique en présence de flux de gènes, soit les populations ont initialement divergé en allopatrie, accumulé un certains nombre d'incompatibilités génétiques et échangent à nouveau des gènes suite à un ou plusieurs contact(s) secondaire(s). Nos résultats suggèrent par ailleurs que le contexte géographique joue un rôle majeur dans la divergence des populations, surtout pour les populations de petite taille efficace pouvant facilement être affecté par la dérive génétique.

Dans le chapitre 3, nous utilisons les populations connectées par flux de gènes précédemment identifiées pour tester différents scénario d'histoire démographiques. Nous testons 5 scénarios à partir d'une approche d'Approximate Bayesian computation (ABC). Le premier scénario correspond à une séparation d'une population ancestrale en deux, suivie d'un isolement strict sans flux de gènes par la suite (SI). Le second scénario est celui d'une spéciation sympatrique avec décroissance du flux de gènes progressive puis arrêt total d'échanges de gènes (AM, ancient migration). Le troisième scénario correspond à une divergence de la population ancestrale en deux populations qui échangent continuellement des gènes (IM, isolement avec migration). Le quatrième scénario est celui d'une divergence allopatrique pour une durée variable suivie par des flux de gènes

lors d'un contact secondaire entres les 2 populations ayant plus ou moins divergées (SC). Le cinquième scénario finalement correspond au maintien d'une seule population à l'équilibre (scénario de panmixie). Les résultats suggèrent que dans la plupart des cas il est difficile de distinguer de manière robuste le scénario d'isolement avec migration de celui impliquant des contacts secondaires que ce soit avec une approche classique d'ABC ou avec une approche basée sur l'utilisation des forêts aléatoires. Dans le cas de l'Oir, où la population est la moins différenciée génétiquement, le scénario de panmixie est le plus probable. Il s'agit aussi de la population ou *L. fluvatilis* est significativement plus petite que les autres populations de *L. fluviatilis* étudiées, ce qui pourrait faciliter le flux de gènes contemporain entre les deux formes. Notre capacité à distinguer entre IM et SC peut s'expliquer simplement par le fait que dans un signal de divergence allopatrique suivi d'un contact secondaire les marqueurs de types 'neutres' telles que les microsatellites convergent vers un signal d'équilibre migration dérive après un certain temps de contact. Si la durée de divergence allopatrique a été trop courte relativement à la période de flux de gènes, le signal de divergence historique est simplement perdu.

Dans le chapitre 4 nous utilisons une approche génomique afin d'apporter davantage de résolution dans le but de mieux comprendre l'histoire de la divergence chez *Lampetra.* Par ailleurs nous exploitons nos différents répliquas de paires d'espèces pour tester l'étendue du parallélisme génétique dans le contexte des barrières semi-perméables au flux de gènes. Des analyses de séquençage RAD ont été effectuées sur 9 paires de populations sympatriques et pararatriques puis des scénarios de divergences ont été testés à l'aide d'approximations de diffusion du spectre joint des fréquences alléliques. Les analyses génomiques confirment les patterns globaux de divergence observés à l'aide des marqueurs microsatellite, tout en apportant une résolution beaucoup plus fine et de nombreux éléments supplémentaires dans la compréhension des processus de divergence. En premier lieu, les résultats permettent pour la premières fois de clairement discriminer les écotypes migrateurs et résident et d'identifier de manière statistique des hybrides F1, F2 et backcrosses avec une grande robustesse. D'autre part, les analyses d'approximation de diffusion suggèrent deux scénarios de divergences différents : les populations connectées par le flux de gènes ont généralement divergé en allopatrie suivie de contacts secondaires résultant en un parallélisme génomique partiel entre répliquas de paires de populations. Une hétérogénéité de la divergence génomique a démontré que les îlots génomiques de différenciation ne résultaient pas de l’action récente de la divergence écologique mais avaient plus probablement été générés suite à l'accumulation d'incompatibilités génétiques en allopatrie et étaient bien révélées par l'action du flux de gènes récent dans les populations où les deux écotypes étaient bien connectés. Au contraire, un degré de parallélisme moins important a été observé dans les populations parapatriques et un

scénario de divergence sympatrique suivi d'un arrêt récent de la migration a été révélé dans ces populations. Il est possible que le signal de divergence obtenu dans ce cas soit aussi biaisé par l'histoire récente de dérive dans les populations de petite taille efficace. Par ailleurs, l'utilisation de nos modèle démographiques pour générer des prédictions neutres ont permis d'identifier un plus grand nombre de marqueurs hautement différenciés génétiquement dans les populations parapatriques que dans les populations sympatriques et représentant des 'outliers' potentiels. Toutefois, une plus grande similarité dans les régions divergentes est détectée chez les populations sympatriques, mettant encore une fois en évidence le rôle fondamental du flux de gènes pour révéler l'histoire commune et le parallélisme partiel à l'échelle moléculaire. Bien que le rôle potentiel de la sélection sur des sites liés, en particulier dans les régions de faible recombinaison ne puisse être rejeté, ces résultats suggèrent fortement que l'hétérogénéité de la différenciation et du parallélisme génétique ont été générés par des réductions locales de flux de gènes dans les îlots génomiques.

Dans le chapitre 5 nous avons testé l'influence de la fragmentation anthropique des cours d'eau sur la diversité génétique des populations de *L. planeri* à travers un échantillonnage à large échelle de 81 populations situées en amont et aval d'obstacles et dans des contextes géographiques bien différents. Plus précisément, 2472 individus ont été collectés dans 29 cours d'eau en Normandie et Nord de la France en sympatrie et parapatrie avec *L. fluviatilis,* en Bretagne et dans le Haut Rhône en allopatrie avec *L. fluviatilis.* Les individus ont été génotypés à l'aide de 13 marqueurs microsatellites. Les résultats ont démontré un effet négatif du nombre d'obstacles et de leur hauteur uniquement sur la richesse allélique des populations de *L. planeri.* Par ailleurs, une augmentation significative et forte de diversité génétique vers l'aval est observée suggérant un rôle majeur de l'asymétrie du flux de gène, corroboré par des analyses génomiques sur 4 paires de populations. Les populations de *L. planeri* possèdent une diversité génétique plus grande lorsque le flux de gènes avec *L. fluviatilis* dans les parties aval des cours d'eau est possible, ce qui n'est pas le cas des populations allopatriques. L'hypothèse explicative la plus probable est une introgression des populations de *L. planeri* par les *L. fluviatilis* de taille efficace plus grande. Ces résultats ont une implication importante en ce qui concerne la biologie de la conservation en démontrant *i)* un impact modéré des barrières à la migration sur la diversité génétique locale des populations de *L. planeri* et *ii*) que les deux ecotypes devraient idéalement être gérés conjointement lorsqu'ils coexistent en sympatrie.

Pour conclure, nous avons notamment mis en évidence l'importance de la nature semi -perméable des barrières au flux génique, engendrant une forte hétérogénéité de la différenciation à l'échelle génomique. Nos analyses montrent que les îlots génomiques partagés entre toutes les

paires ont été révélé grâce à l'effet érosif du flux de gènes suite à des contacts secondaires et n'ont pas émergé suite à des événements indépendants de spéciation sympatrique dûs à la sélection écologique. Nos résultats n'excluent pas un rôle majeur de la sélection d'arrière-plan dans les régions de faibles recombinaison et des analyses en cours suggèrent même que ce processus est important entre populations d'un même écotype. Nous avons par ailleurs démontré la difficulté à discriminer avec précision l'histoire évolutive de la divergence entre les deux écotypes de *Lampetra* que ce soit avec un nombre faible de marqueurs ou à une échelle génomique. Il est probable que la véritable histoire démographique soit beaucoup plus compliquée que ce que les dernières méthodes statistiques permettent de tester actuellement. Par ailleurs, une cartographie génétique en cours de développement permettra de mieux connaître la distribution des zones du génome les plus différenciées.

Globalement cette thèse a démontré que les paires d’écotypes parasites et non-parasites de lamproies représentent un excellent modèle d’étude de la spéciation et notamment de l’architecture génomique de la divergence.

Evolution de la divergence entre la lamproie fluviatile (*Lampetra fluviatilis*) et la lamproie de planer (*L. planeri*) inférée par des approches expérimentales et de génomique des populations

Quentin Rougemont – UMR 985 ESE INRA – AGROCAMPUS OUEST – Université de Rennes 1

Directeur de thèse : Guillaume Evanno – Chargé de Recherche INRA – AGROCAMPUS OUEST

Co- encadrement : Sophie Launey – Chargé de Recherche INRA – AGROCAMPUS OUEST

Résumé : Cette thèse étudie le processus de spéciation entre la lamproie fluviatile (*Lampetra fluviatilis*) et la lamproie de Planer (*L. planeri*). Les deux espèces présentent des stratégies d'histoire de vie extrêmement différentes : *L. fluviatilis* est parasite et anadrome alors que *L. planeri* n'est pas parasite et reste strictement dulcicole. Toutefois, leur degré d'isolement reproducteur et leur histoire de divergence demeurent méconnus. Ces questions ont été abordées par des approches expérimentales, de génomique de populations et de simulations démographiques. Des croisements expérimentaux ont révélé un faible isolement reproducteur, confirmé par des degrés variables de flux géniques dans les populations naturelles. Les analyses génétiques ont montré que les deux taxons représentaient probablement des écotypes avec un isolement reproducteur partiel suggérant que les barrières reproductives endogènes ne réduisaient que partiellement la migration efficace entre écotypes. L'importance du contexte géographique actuel et passé dans l'étude de la spéciation a aussi été mise en évidence par des analyses à l'échelle du génome. Ainsi, les populations isolées de *L. planeri* évoluent principalement sous l'effet de la dérive génétique et ont une diversité réduite. Les inférences démographiques ont suggéré que la divergence a été initiée en allopatrie puis suivie de contacts secondaires résultant en un parallélisme génomique partiel entre réplicas de paires de populations. Une hétérogénéité de la divergence génomique a démontré que les ilots génomiques de différenciation ne résultaient pas de l'action récente de la divergence écologique. En outre, nos résultats suggèrent un impact faible de la fragmentation anthropique des cours d'eau sur la diversité génétique des populations de *L. planeri.* Les populations résidentes possèdent une diversité génétique plus grande lorsque le flux de gènes avec *L. fluviatilis* dans les parties aval des cours d'eau est possible. Globalement cette thèse a démontré que les paires d'écotypes parasites et non-parasites de lamproies représentent un excellent modèle d'étude de la spéciation et notamment de l'architecture génomique de la divergence.

Mots clés : Spéciation, flux de gènes, parallélisme, modélisation, biogéographie, histoire évolutive, *Lampetra*

Evolution of divergence between the river lamprey (*Lampetra fluviatilis*) and the brook lamprey (*L. planeri*) inferred by experimental approaches and population genomics

Summary: This thesis investigates the process of speciation between the European lampreys *Lampetra fluviatilis* and *L. planeri.* The two species have drastically different life history strategies: *L. fluviatilis* is parasitic and anadromous while *L. planeri* is non-parasitic and strictly freshwater resident. Yet their level of reproductive isolation and history of divergence remain poorly understood. A multidisciplinary approach including experiments, population genomics analyses and historical reconstruction was undertaken to address these issues. Experimental crosses revealed a very low level of reproductive isolation, partially mirrored by variable levels of gene flow in wild populations. Genetic analyses revealed that the two taxa were best described as partially reproductively isolated ecotypes suggesting that endogenous genetic barriers partially reduced effective migration between ecotypes. Genome wide analyses showed the importance of the current and ancient geographical context of speciation. In particular, parapatric *L. planeri* populations diverged mostly through drift and displayed a reduced genetic diversity. Demographic inferences suggested that divergence have likely emerged in allopatry and then secondary contacts resulted in partial parallelism between replicate population pairs. A strong heterogeneity of divergence across the genome was revealed by sympatric populations suggesting that genomic islands of differentiation were not linked to ongoing ecological divergence. Further investigations showed that the genetic diversity of *L. planeri* populations was weakly affected by human-induced river fragmentation. Resident populations displayed a higher diversity when gene flow was possible with L. *fluviatilis* populations in downstream sections of rivers. Overall this thesis showed that parasitic and non-parasitic lamprey ecotypes represent a promising model for studying speciation and notably the genomic architecture of divergence.

Keywords: Speciation, Gene flow, parallelism, modelling, biogeography, evolutionary history, *Lampetra*