MOLECULAR ECOLOGY

Molecular Ecology (2015) doi: 10.1111/mec.13345

INVITED REVIEWS AND SYNTHESES

Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems

I. PAZ-VINAS,*† ‡ G. LOOT,† § V. M. STEVENS§ and S. BLANCHET*§

*Centre National de la Recherche Scientifique (CNRS), École Nationale de Formation Agronomique (ENFA), UMR 5174 EDB (Laboratoire Évolution & Diversité Biologique), Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France, †UPS, UMR 5174 (EDB), Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 4, France, ‡UMR 7263 – IMBE, Équipe EGE, Centre Saint-Charles, Aix-Marseille Université, CNRS, IRD, Université d'Avignon et des Pays de Vaucluse, Case 36, 3 place Victor Hugo, 13331 Marseille Cedex 3, France, §Station d'Écologie Expérimentale du CNRS à Moulis, USR 2936, Centre National de la Recherche Scientifique (CNRS), 2 route du CNRS, 09200 Moulis, France

Abstract

Describing, understanding and predicting the spatial distribution of genetic diversity is a central issue in biological sciences. In river landscapes, it is generally predicted that neutral genetic diversity should increase downstream, but there have been few attempts to test and validate this assumption across taxonomic groups. Moreover, it is still unclear what are the evolutionary processes that may generate this apparent spatial pattern of diversity. Here, we quantitatively synthesized published results from diverse taxa living in river ecosystems, and we performed a meta-analysis to show that a downstream increase in intraspecific genetic diversity (DIGD) actually constitutes a general spatial pattern of biodiversity that is repeatable across taxa. We further demonstrated that DIGD was stronger for strictly waterborne dispersing than for overland dispersing species. However, for a restricted data set focusing on fishes, there was no evidence that DIGD was related to particular species traits. We then searched for general processes underlying DIGD by simulating genetic data in dendritic-like river systems. Simulations revealed that the three processes we considered (downstream-biased dispersal, increase in habitat availability downstream and upstream-directed colonization) might generate DIGD. Using random forest models, we identified from simulations a set of highly informative summary statistics allowing discriminating among the processes causing DIGD. Finally, combining these discriminant statistics and approximate Bayesian computations on a set of twelve empirical case studies, we hypothesized that DIGD were most likely due to the interaction of two of these three processes and that contrary to expectation, they were not solely caused by downstream-biased dispersal.

Keywords: approximate Bayesian computation, asymmetric gene flow, colonization, dendritic ecological networks, genetic diversity, meta-analysis, random forest, river network, simulated genetic data, spatial patterns of biodiversity

Received 15 October 2014; revision received 30 July 2015; accepted 13 August 2015

Introduction

Spatial patterns of biological diversity are defined as repeatable gradients of biodiversity along geographic descriptors (e.g. latitude, longitude or altitude; Levin 1992; Lawton 1996; Hillebrand 2004). Describing and understanding spatial patterns of biodiversity is a central and critical topic of ecological, evolutionary and conservation sciences (Gotelli *et al.* 2009; Chave 2013). The coupling of empirical and theoretical works has been incredibly helpful in improving our understanding of spatial patterns for numerous facets of biodiversity

Correspondence: Ivan Paz-Vinas, Fax: +33 4 13 55 07 86; E-mail: ivanpaz23@gmail.com and Simon Blanchet, Fax: +33 5 61 04 03 60; E-mail: simon.blanchet@ecoex-moulis.cnrs.fr

(Chave 2013). However, despite its ecological and evolutionary importance (Hughes *et al.* 2008; Caballero & García-Dorado 2013), intraspecific genetic diversity remains an aspect of biodiversity for which many spatial patterns remain to be revealed and understood.

Several general patterns have been described at the intraspecific genetic level such as patterns of isolation by distance (Wright 1943; Sexton *et al.* 2014), the decrease in genetic diversity along routes of colonization (Taberlet *et al.* 1998), reduced genetic diversity at range boundaries (Kirkpatrick & Barton 1997; Eckert *et al.* 2008; Liggins *et al.* 2015) and more recently patterns of isolation by adaptation (Nosil 2009; Sexton *et al.* 2014). Beyond these widely acknowledged patterns, the increasing availability of genetic data sets has generated evidences for some additional patterns in specific ecosystems. For instance, in river ecosystems several studies reported a *downstream increase in genetic diversity* (hereafter DIGD) (e.g. Hänfling & Weetman 2006; Kikuchi *et al.* 2009; Alp *et al.* 2012; Torterotot *et al.* 2014), which may constitute an additional pattern of intraspecific genetic diversity. Unravelling whether or not DIGD constitutes a *spatial pattern* or a casual observation is of major importance for conservation purposes, as this may imply that the spatial structure of genetic diversity might be predictable, and hence that freshwater-protected areas may theoretically benefit several species at a time. However, albeit there have been some attempts to synthesize observations across taxonomic groups regarding the spatial organization of intraspecific genetic diversity in river systems (Hughes 2007; Finn *et al.* 2011; Hughes *et al.* 2013), studies aiming at thoroughly explore whether or not DIGD constitutes a general spatial pattern of genetic diversity in rivers remain scarce (but see Honnay *et al.* 2010 for a meta-analysis on riparian plants).

River ecosystems offer unique opportunity to make such generalization across taxa because population genetic studies have accumulated for a wide variety of organisms living in rivers or along their banks (Pauls *et al.* 2014), and simple hypotheses can be generated regarding the processes generating a spatial organization of genetic diversity in those systems. River ecosystems are specific cases of dendritic ecological networks, characterized by their tree-like geometric branching pattern (Benda *et al.* 2004; Campbell Grant *et al.* 2007), and strongly structured by elevation, making water flow unidirectional. These two characteristics (i.e. branching geometry and unidirectional water flow) strongly constrain movements of individuals and hence dispersal. Dispersal in turn facilitates gene flow (Ronce 2007) and colonization, which might generate spatial patterns of intraspecific genetic diversity in riverscapes (Altermatt 2013). Some theoretical studies have investigated the genetic consequences of network geometry (Labonne *et al.* 2008), dendritic connectivity *per se* (Paz-Vinas & Blanchet 2015) and particular dispersal modalities in simulated river systems (Chaput-Bardy *et al.* 2009; Morrissey & de Kerckhove 2009). Nevertheless, there have been no or few attempts, to broadly explore the processes generating DIGD in river ecosystems.

In theory, DIGD may be the result of three main processes affecting neutral genetic diversity at the deme level, by moving the balance between the forces increasing diversity (mutation and immigration), and those that reduce it (emigration and genetic drift). The first process is *downstream-biased gene flow*, which may result from asymmetric dispersal costs due to unidirectional water flow (Morrissey & de Kerckhove 2009; Paz-Vinas *et al.* 2013). In consequence, upstream demes would entail higher loss of alleles by emigration and genetic drift, while immigration would compensate for drift in downstream demes (Ritland 1989; see also Müller 1954 for a closely related hypothesis called 'the drift paradox hypothesis'). This hypothesis is generally the first (and often the only one) to be invoked to explain empirical DIGD. Second, DIGD may be the ultimate result of *variation in habitat availability*, which typically increases from sources to river mouth because of the downstream increase in river width and hence habitat availability, at least for weakly specialized species (Muneepeerakul *et al.* 2007; Raeymaekers *et al.* 2008; Carrara *et al.* 2014). This hypothesis is based on the observation that higher genetic diversity can be reached in populations with higher effective sizes (N_e): under the assumption that N_e covaries with the abundance of individuals (which in turn is positively correlated with habitat availability), the amount of available habitat may positively correlate with genetic diversity (Nei 1987; Frankham 1996). Consequently, species may be more genetically diverse in downstream sections than in upstream sections (i.e. conform to a DIGD). Finally, although less acknowledged, DIGD can be a particular case of declining genetic diversity along colonization routes (Cyr & Angers 2012). Assuming that the remnant (or founding) populations are located downstream, the progressive *upstream-directed colonization* would create a succession of founding events, typically accompanied by a loss of genetic variation. This can be expected, for instance, after a glacial event for which the glacial refugees were situated in the downstream section of a river basin, during an introduction or a biological invasion (Hewitt 1996), or during an upstream range shift due to the breakdown of natural and/or anthropogenic barriers (e.g. during a range shift due to climate change; Conti *et al.* 2015). Ultimately, this may generate the observed DIGD. To our knowledge, no study has simultaneously tested which of these three

processes – or which combination of processes – is more likely to generate DIGD.

The first general objective of this study was to provide a quantitative and exhaustive synthesis of previously published studies describing spatial gradients of genetic diversity in river networks, so as to test for the generality of a spatial pattern in intraspecific diversity (namely DIGD) across taxonomic groups. To that aim, we used neutral allelic richness calculated using microsatellite markers as a surrogate of intraspecific genetic diversity, and we used meta-analytical tools to test the generality of DIGD across a variety of taxa including plants, arthropods, mollusks and vertebrates. This database was also used to explore potential species traits explaining why the strength of DIGD may vary among species. We first test the hypothesis that contrasting spatial patterns of genetic diversity should be observed for organisms displaying two contrasting dispersal modes: exclusively waterborne dispersal (e.g. fish) vs. overland dispersal (e.g. riparian plants, amphibians and arthropods). We expected that waterborne dispersers should display a DIGD stronger than overland dispersers (Alp *et al.* 2012). Second, we restrict the data set to fish species (the most abundant group in the meta-analysis) to test whether or not life history, ecological or morphometric species traits can explain variation in the strength of DIGD. Finally, on a secondary perspective, we used this database to test another spatial pattern of intraspecific genetic diversity that has been hypothesized in river networks; that is, that upstream demes are more genetically differentiated than downstream demes (Finn *et al.* 2011; Paz-Vinas & Blanchet 2015). This is an important pattern to explore for conservation purposes because – if verified – it would mean that upstream demes are genetically unique and might largely contribute to the whole network-scale genetic diversity (Finn *et al.* 2011).

Our second general objective was to go beyond the description of spatial patterns by identifying processes that may generate DIGD. To that aim, we used pattern-oriented simulations (Hoban 2014; Pauls *et al.* 2014) to theoretically explore which process of downstream-biased gene flow, variation in habitat availability or upstream-directed colonization (or which combination of these three processes) is more likely to generate DIGD. We hypothesize that the most likely process to generate DIGD is downstream-biased gene flow – alone and/or in combination with the two other processes – as it is the process that is generally invoked to explain empirical DIGD (Hänfling & Weetman 2006; Kikuchi *et al.* 2009; Paz-Vinas *et al.* 2013). We further used the output of these simulations and random forest classification models (Breiman 2001) to highlight the most informative summary statistics that best capture the genetic signature of these processes, and that may help identifying the most likely process generating DIGD when it is actually observed in natural populations. Finally, we came back to empirical observations by applying approximate Bayesian computations (ABC) on the most discriminant statistics to hypothesize which (or which combination) of the three processes considered in this study was likely to have generated DIGD observed in a subset of contrasted empirical data sets.

Materials and methods

Patterns of genetic diversity

We described spatial patterns of genetic diversity as the Pearson's correlation coefficient between the distance of each sampled deme from the river mouth (or the distance of each deme from the first downstream confluence shared by all demes when the study did not cover the whole river basin) and the mean allelic richness (hereafter COR_{AR}) for both empirical and simulated data sets (see below). Mean allelic richness is the mean number of alleles *per* sampling deme, averaged over loci and corrected for the number of individuals genotyped in a deme. DIGD translates into negative COR_{AR}. Although we mainly focused on DIGD, we secondarily described (both for simulations and for a subset of case studies selected in the meta-analysis) spatial patterns of genetic differentiation as the Pearson's correlation coefficient between the distance of each sampled deme from the river mouth and within-deme F_{ST} values (i.e. a measure of the genetic uniqueness of a deme at the meta-population level, calculated as the average of the pairwise F_{ST} values observed between a deme and all other demes; Coleman *et al.* 2013). We will hereafter refer to this statistic as $COR_{F_{ST}}$.

Meta-analysis of empirical data

We conducted a literature survey of scientific papers published during the period 2004–2013 that report genetic diversity data for freshwater organisms sampled in river networks. The search was done using the ISI Web of Knowledge® platform (last accessed the 18th of July 2013) and combining the following keywords: 'river', 'genetic diversity' and 'microsatellites'. We restricted our search to papers that (i) used microsatellite markers (to fit our simulations and reduce the variation due to the use of other marker types) and (ii) directly reported COR_{AR} or (if this was not the case) reported values of allelic richness for each sampled deme as well as a map representing sampling locations. The maps were used in these cases to calculate topological distances of each deme from the river mouth (or from the most down-

stream confluence shared by all demes when maps did not cover the whole river basin) using Inkscape v.0.48.2, hence allowing the calculation of COR_{AR}. We retained studies for which the sampling was done only in a single stream, or over a dendritic network. This led to 79 case studies from six major taxonomic groups (plants, mollusks, arthropods, amphibians, agnates and fish, see Table S1, Supporting information).

Simulations

To determine the propensity of downstream-biased gene flow, variation in habitat availability and upstream-directed colonization to generate DIGD, we simulated neutral genetic (microsatellite-like) data under eight distinct dendritic models using a coalescent-based approach (Kingman 1982). All models had the same spatial configuration but varied in specific parameters (see below and Fig. 1; Table 1). Three models were ruled by one of the three processes independently (hereafter 'geneflow model', 'habitat availability model' and 'colonization model', respectively, for the processes generated by downstream-biased gene flow, variation in habitat availability or upstream-directed colonization; Fig. 1). Additionally, we ran simulations under five complementary models: (i) three models ruled by two-way interactions between these processes (i.e. geneflow/habitat, geneflow/colonization and habitat/colonization models), (ii) one model ruled by a three-way interaction between the three processes (i.e. geneflow/habitat/colonization model) and (iii) a dendritic model where none of the processes mentioned above were simulated (i.e. the null model). This last model will reflect the solely effects of dendritic connectivity *per se* on spatial patterns of genetic diversity (see Paz-Vinas & Blanchet 2015).

Spatial configuration. All eight models were composed of 33 demes arranged in a dendritic fashion: there were eight upstream branches that ultimately feed a downstream section arranged in a linear stepping-stone chain composed of five demes (Fig. 1). The deme at the bottom of the linear chain was considered as the most downstream deme (i.e. 'river mouth deme'), whereas the upper demes were considered as the most upstream demes (i.e. 'headwater demes'). The range of the explored parameter values for each model is defined in Table 1. Hereafter, we thoroughly present each single-process model (geneflow model, habitat availability model and colonization model) and then explain briefly how the null and the interacting models were built.

Model for downstream-biased gene flow (geneflow model). We assumed that all demes had the same effective population size (N_{DEMES}), which was constant across generations (Fig. 1A; Table 1). We considered two different dispersal rates: a downstream-directed dispersal rate ($D_{DOWNSTREAM}$) and an upstream-directed dispersal rate, $D_{UPSTREAM} = D_{DOWNSTREAM}/P_{ASYM}$, where P_{ASYM} is a parameter representing the level of asymmetry in gene flow (Fig. 1A). A P_{ASYM} of 1 means that gene flow is symmetric, whereas values >1 indicate downstream-biased gene flow.

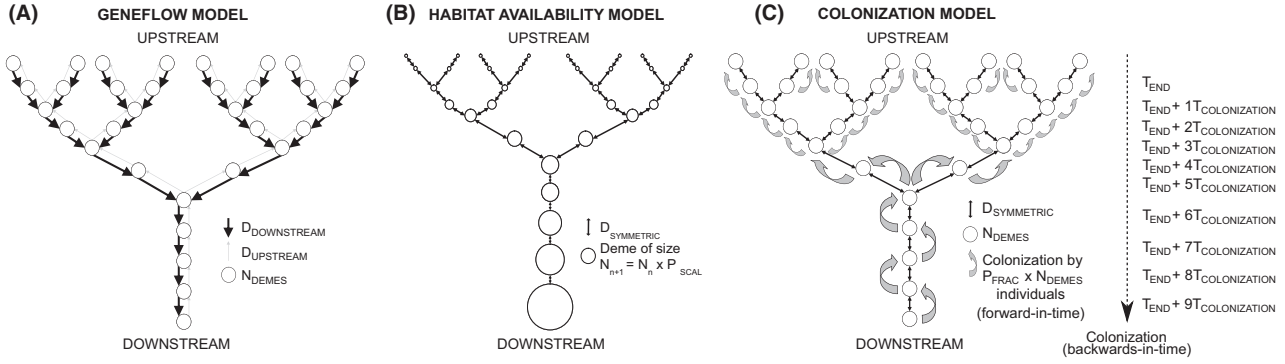

Fig. 1 Figure representing the three independent dendritic population models considered for simulating genetic data. (A) illustrates the downstream-biased geneflow model (geneflow model), with demes of equal size N_{DEMES} and downstream-directed dispersal equal to or higher than upstream-directed migration (i.e. $D_{DOWNSTREAM} \geq D_{UPSTREAM}$). (B) represents the model with variation in habitat availability (habitat availability model), in which symmetric dispersal is assumed ($D_{SYMMETRIC}$) and the size of the demes increases along the upstream–downstream gradient by the recurrence relationship $N_{n+1} = N_n \times P_{SCAL}$. (C) illustrates the model with upstream-directed colonization (colonization model), in which the entire population is progressively colonized from downstream-to-upstream by successive colonization steps of length $T_{COLONIZATION}$ and $P_{FRAC} \times N_{DEMES}$ colonizing individuals *per* colonization step. The colonization process stops when headwater populations are colonized at time T_{END}. Although not represented in this figure, the null model and the two- and three-way interacting models (gene flow/habitat, gene flow/colonization, habitat/colonization and gene flow/habitat/colonization) share the same spatial structure than the single-process models.

Table 1 Prior parameter values considered for simulating genetic data under the eight dendritic models (gene flow, habitat availability, colonization, gene flow/habitat, gene flow/colonization, habitat/colonization, gene flow/habitat/colonization and null)

Model	Parameter	Description	Prior parameter values	Number of parameter values tested
Gene flow	N_{DEMES}	Size of the demes (diploid individuals)	50; 1000 or 10 000	3
	P_{ASYM}	Asymmetry in dispersal rate	1 to 20 by 0.1	191
	$D_{DOWNSTREAM}$	Downstream-directed dispersal rate	0.01 to 0.3 by 0.01	30
Habitat availability	$N_{HEADWATER}$	Size of the most upstream demes (diploid individuals)	50 to 500 by 5	100
	P_{SCAL}	Scaling parameter for calculating downstream demes sizes	1.0 to 1.5 by 0.1	6
	$D_{SYMMETRIC}$	Symmetric dispersal rate	0.01 to 0.3 by 0.01	30
Colonization	N_{DEMES}	Size of the demes (diploid individuals)	50; 1000 or 10 000	3
	$D_{SYMMETRIC}$	Symmetric dispersal rate	0.01 to 0.3 by 0.01	30
	T_{END}	Time of the ending of the stepwise colonization (generations)	10 to 500 by 10	50
	$T_{COLONIZATION}$	Time elapsed between each colonization step (generations)	5 to 100 by 5	20
	P_{FRAC}	Fraction of individuals colonizing a new deme	0.1 to 0.3 by 0.1	3
Gene flow/habitat	$N_{HEADWATER}$	Size of the most upstream demes (diploid individuals)	50 to 500 by 5	100
	P_{ASYM}	Asymmetry in dispersal rate	1 to 20 by 0.1	191
	$D_{DOWNSTREAM}$	Downstream-directed dispersal rate	0.01 to 0.3 by 0.01	30
	P_{SCAL}	Scaling parameter for calculating downstream demes sizes	1.0 to 1.5 by 0.1	6
Gene flow/ colonization	N_{DEMES}	Size of the demes (diploid individuals)	50; 1000 or 10 000	3
	P_{ASYM}	Asymmetry in dispersal rate	1 to 20 by 0.1	191
	$D_{DOWNSTREAM}$	Downstream-directed dispersal rate	0.01 to 0.3 by 0.01	30
	T_{END}	Time of the ending of the stepwise colonization (generations)	10 to 500 by 10	50
	$T_{COLONIZATION}$	Time elapsed between each colonization step (generations)	5 to 100 by 5	20
	P_{FRAC}	Fraction of individuals colonizing a new deme	0.1 to 0.3 by 0.1	3
Habitat/colonization	$N_{HEADWATER}$	Size of the most upstream demes (diploid individuals)	50 to 500 by 5	100
	P_{SCAL}	Scaling parameter for calculating downstream demes sizes	1.0 to 1.5 by 0.1	6
	$D_{SYMMETRIC}$	Symmetric dispersal rate	0.01 to 0.3 by 0.01	30
	T_{END}	Time of the ending of the stepwise colonization (generations)	10 to 500 by 10	50
	$T_{COLONIZATION}$	Time elapsed between each colonization step (generations)	5 to 100 by 5	20
	P_{FRAC}	Fraction of individuals colonizing a new deme	0.1 to 0.3 by 0.1	3
Gene flow/ habitat/ colonization	$N_{HEADWATER}$	Size of the most upstream demes (diploid individuals)	50 to 500 by 5	100
	P_{ASYM}	Asymmetry in dispersal rate	1 to 20 by 0.1	191

Table 1 *Continued*

Model	Parameter	Description	Prior parameter values	Number of parameter values tested
	$D_{DOWNSTREAM}$	Downstream-directed dispersal rate	0.01 to 0.3 by 0.01	30
	P_{SCAL}	Scaling parameter for calculating downstream demes sizes	1.0 to 1.5 by 0.1	6
	T_{END}	Time of the ending of the stepwise colonization (generations)	10 to 500 by 10	50
	$T_{COLONIZATION}$	Time elapsed between each colonization step (generations)	5 to 100 by 5	20
	P_{FRAC}	Fraction of individuals colonizing a new deme	0.1 to 0.3 by 0.1	3
Null model	N_{DEMES}	Size of the demes (diploid individuals)	50; 1000 or 10 000	3
	$D_{SYMMETRIC}$	Symmetric dispersal rate	0.01 to 0.3 by 0.01	30

Model for variation in habitat availability model (habitat availability model). We assumed a single symmetric dispersal rate, $D_{SYMMETRIC}$ (Fig. 1B; Table 1), and we considered that demes increased in effective population size (N_e) along the upstream–downstream gradient. This was modelled by (i) setting a parameter defining the N_e of all headwater demes, $N_{HEADWATER}$, and (ii) by determining the size of the following demes by the recurrence relationship $N_{n+1} = N_n \times P_{SCAL}$. P_{SCAL} represents a positive scaling parameter, and the starting value for N_n was $N_0 = N_{HEADWATER}$. Populations in this model were characterized by larger N_e in downstream than in upstream demes (Fig. 1B).

Model for upstream-directed colonization (colonization model). We considered that populations experienced a stepwise colonization that started from the river mouth deme and ended in the headwater demes at a time T_{END} (Fig. 1C). The end of the colonization process can be recent or ancient, depending on the value of T_{END} (Table 1). The colonization speed was determined by the parameter $T_{COLONIZATION}$, which defines the time (in generations) separating two colonization steps (Fig. 1C). As we used a backwards-in-time simulation framework, the simulation process begins at a 'current' situation where all demes are colonized, and then traces back the stepwise colonization starting at time T_{END} (Fig. 1C). This procedure ensures that all demes have been colonized during the simulation process. The fraction of colonizing individuals at each step was determined by the parameter $P_{FRAC.}$ All demes had the same effective population size, N_{DEMES}, and a unique symmetric dispersal rate, $D_{SYMMETRIC.}$

Two- and three-way interacting models (geneflow/habitat, geneflow/colonization, habitat/colonization models and geneflow/habitat/colonization models). These models were arranged in the same spatial fashion than geneflow, habitat availability and colonization models, but differed from them in that they were characterized by parameters relative to the two or three processes implied in each of these models. The parameters used in each interacting model are those described in Table 1.

Null model. This model was arranged in the same spatial fashion than the others and assumed the same parameter values, but migration among demes was symmetric ($D_{UPSTREAM} = D_{DOWNSTREAM}$), N_e among demes were equal across generations, and no colonization process was modelled for each simulation (Table 1). This model hence reflects the solely effects of dendritic connectivity on spatial patterns of genetic diversity, which is predicted to generate bell-shaped patterns of allelic richness along the upstream–downstream gradient (Paz-Vinas & Blanchet 2015).

Simulation procedure. The simulation procedure comprised four major steps: (i) sampling of a vector of parameter values (φ_x) for a specific model from prior parameter distributions (defined in Table 1); (ii) simulation of a genetic data set D_x, given φ_x; (iii) calculation of a vector of statistics s_x that summarizes the simulated data D_x (see the Summary statistics section below); and (iv) repeat steps (i) to (iii) many times (i.e. 300 000 simulations *per* model in our case).

To implement the simulation procedure, we set a computational pipeline based on the program ABCSAMPLER (Wegmann *et al.* 2010) that integrates several additional population genetics and statistical programs (see Appendix S1, Supporting information for details). The coalescent-based genetic data simulator SIMCOAL v2.1.2

(Laval & Excoffier 2004) was used to simulate microsatellite data under the eight models described above, given φ_x. We simulated fifteen independent microsatellite loci *per* individual, assuming a stepwise mutation model (SMM) and a mutation rate of 5×10^{-4} over loci. The calculation of summary statistics was based on 25 diploid individuals sampled from each deme, irrespective of the model considered. These conditions correspond to typical sampling schemes in population genetics.

Summary statistics. We used the software ARLSUMSTAT (Excoffier & Lischer 2010) to calculate over loci and for each deme the expected heterozygosity (H_E) and the Garza–Williamson's statistic (*GW*; i.e. the mean ratio between the number of alleles observed on microsatellite loci and the range in allele size at this loci; Garza & Williamson 2001). The GW statistic was calculated because it informs on the demographic history of populations (Garza & Williamson 2001), and can thus be potentially informative for discriminating models experiencing upstream-directed colonization (a process that implies a succession of founder events) from the rest of the models. ARLSUMSTAT was also used to calculate statistics at the landscape level, including global F_{ST}, F_{IS} and F_{IT} values, as well as pairwise F_{ST} values between all possible pairs of demes. We used the software ADZE v1.0 (Szpiech *et al.* 2008) to estimate allelic richness (*AR*) and mean private allelic richness (*PA*) at the deme level. We then averaged pairwise F_{ST} values observed between a site and all the other sites in the network to obtain within-deme F_{ST} estimates. Within-deme F_{ST} indicates how unique is a deme in terms of genetic differentiation compared to the other demes in the network (Coleman *et al.* 2013). Finally, we used the R software v.2.13 to characterize spatial patterns of genetic diversity for each D_x and for each model by calculating COR_{AR} (as described above). In addition, we calculated other potential discriminant statistics that were representative of the spatial distribution of different genetic diversity and differentiation indices along the upstream–downstream gradient of dendritic networks. Specifically, we computed Pearson's correlation coefficients between the distance of each deme to the putative river mouth and H_E, *PA*, within-deme F_{ST} and *GW* (hereafter COR_{HE}, COR_{PA}, $COR_{F_{ST}}$ and COR_{GW}, respectively). We finally performed, for each D_x, multiple regressions on distance matrices (Lichstein 2006) in which the matrix of genetic differentiation calculated for all demes (i.e. pairwise F_{ST}) was the dependent variable, and the two independent variables were (i) a matrix informing of the geographic distances between all demes (in number of demes) and (ii) a binary matrix informing of the level of flow connectivity between all pairs of demes (two demes were considered flow-connected when water can flow from the upstream deme to the downstream deme; flow-unconnected demes are two demes that share a common confluence downstream but do not share flow; see fig. 3 in Peterson *et al.* 2013). This analysis produced two standardized regression coefficients: one informing of the effect of geographic distances on genetic differentiation (i.e. 'isolation by distance', with slope COR_{IBD}) and the other informing of the effect of among-demes flow connectivity on genetic differentiation ('isolation by flow', with slope COR_{IBF}). Multiple regressions on distance matrices were performed with the R package 'ecodist' (Goslee & Urban 2007).

We generated 300 000 D_x *per* model, which covered the entire parameter space defined in Table 1 for the geneflow, habitat availability, colonization and null models, and a substantial proportion of the parameter space defined for the two- and three-way interacting models (geneflow/habitat, geneflow/colonization, habitat/colonization and geneflow/habitat/colonization models). Simulations were performed on an ALTIX ICE 8200 EX (Silicon Graphics International, Fremont, CA, USA) and on a BULLx DLC cluster (Bull SAS, Les Clayes-sous-Bois, Yvelines, France) hosted by the CALMIP group (UMS 3667, University Paul Sabatier, Toulouse, France).

Statistical analyses

Spatial patterns of genetic diversity in freshwater organisms. Based on the meta-analysis, we first asked whether or not there was an overall significant negative COR_{AR} over all case studies. After having transformed each COR_{AR} into standardized effect sizes (Fisher's z; Nakagawa & Cuthill 2007), we used a meta-regression approach (based on Bayesian mixed-effects meta-analysis, BMM; Hadfield 2010; Nakagawa & Santos 2012) to estimate the mean effect size (MES) over all case studies, which was then back-transformed into a global correlation coefficient for the entire meta-analysis (hereafter meta-COR_{AR}). We included 'study identity' as a random factor in the BMM, and we estimated the MES (with its 95% confidence intervals, 95% CI) as the intercept of the null model (i.e. no fixed effect). In BMM, each standardized effect size was weighted by the inverse of the asymptotic variance (v_z) using the following formula: $v_z = (n-3)$, where n is the number of sampled populations.

We then ran a first additional BMM in which the 'taxonomic group' was included as a categorical fixed effect to test whether or not the MES varied significantly among major taxonomic groups. In a second additional BMM, we tested whether COR_{AR} differs

when dispersal is only possible through water channels (i.e. fish, mollusks) from when overland dispersal is also possible (through wind, air or terrestrial dispersal, i.e. plants, one amphibian and some arthropods). The 'dispersal mode' was included as a categorical fixed effect to test the working hypothesis that organisms being able to use overland dispersal should display COR_{AR} closer to zero. The deviance information criteria (DIC) of these two additional models were compared to the DIC of the null model to evaluate the support of models including the taxonomic identity or the dispersal mode of each species. We considered that models including the taxonomic identity or the dispersal mode of each species were best supported by the data if their DIC was lower than four units compared to the DIC of the null model ($\Delta DIC > 4$). In a third additional model, we restricted the data set to fish species (which was the most represented taxonomic group, Table S1, Supporting information) to test whether or not COR_{AR} can be predicted from simple life history, ecological and/or morphometric traits. We focused on traits that may influence COR_{AR} because directly or indirectly related to dispersal propensity and/or levels of genetic diversity. For each fish species, we gathered information on (i) migratory type (whether the species is anadromous, potadromous, catadromous or nonmigratory), (ii) habitat use (whether the species is a benthic or a pelagic feeder), (iii) the generation time (in years), (iv) the maximum body length, (v) the shape factor (ratio of maximum body length to maximum body height) and (vi) the swimming factor (ratio of minimum caudal peduncle height to the area of the caudal fin). Life history and ecological data were gathered from FishBase (Froese & Pauly 2015), whereas the two morphometric traits were quantified from a database built on pictures available on the web (S. Brosse & S. Villéger, unpublished data set). We built a full BMM including these six traits as fixed effects and the 'species identity' as a random factor. Then, we built all possible models resulting from the combination of these six variables (i.e. 64 models) and calculated the DIC for each of them. From this set of models, we considered and selected the model(s) with $\Delta DIC < 4$ as the most parsimonious model(s) to explain the data.

Finally, to test the hypothesis that genetic differentiation should be greater in upstream than in downstream demes (Finn *et al.* 2011), we used a subset of the full data set that met the following criteria: (i) observed COR_{AR} is equal to or lower than meta-COR_{AR} (so as to select empirical studies that are likely to match the simulated conditions for further model-based inferences; see *Processes inferred from empirical data sets* subsection), (ii) the sampling was carried out over an entire dendritic network rather than on a single river stretch, and (iii) authors calculated pairwise F_{ST} so that within-deme F_{ST} was calculable (see above). This led to a set of 12 studies (see Table 2) from which we calculated $COR_{F_{ST}}$ for each study. Using the same meta-regression approach than the one described above for COR_{AR}, we estimated the MES meta-$COR_{F_{ST}}$ (with its 95% confidence intervals, 95% CI) over all 12 case studies as the intercept of the null model. The BMM analyses were performed with the R package 'MCMCglmm' (Hadfield 2010).

The propensity of downstream-biased gene flow, variation in habitat availability and upstream-directed colonization to generate DIGD. We used nonparametric probability density functions to visually inspect the frequency of

Table 2 Values of COR_{IBF}, global F_{ST}, COR_{AR}, COR_{IBD} and $COR_{F_{ST}}$ calculated for a subset of populations extracted from the full meta-analysis that met the following criteria: (i) observed COR_{AR} is equal to or lower than meta-COR_{AR}, (ii) the sampling was carried out over an entire dendritic network instead of on a single river, and (iii) within-deme F_{ST} was reported or calculable from pairwise F_{ST} matrices

Species name and Reference	Taxonomic group	COR_{IBF}	F_{ST}	COR_{AR}	COR_{IBD}	$COR_{F_{ST}}$
Gammarus fossarum (Alp *et al.* 2012)	Arthropods	0.26	0.237	−0.85	0.47	−0.281
Semotilus atromaculatus (Boizard *et al.* 2009)	Fishes	0.51	0.25	−0.62	0.39	−0.614
Poecilia reticulata (Crispo *et al.* 2005)	Fishes	0.36	0.302	−0.78	0.41	0.837
Telestes souffia (Dubut *et al.* 2012)	Fishes	0.517	0.056	−0.72	0.523	0.582
Cottus gobio (Hänfling & Weetman 2006)	Fishes	0.29	0.27	−0.57	0.54	0.304
Salmo trutta (Horreo *et al.* 2011) – Nive	Fishes	−0.114	0.21	−0.53	0.561	0.538
Cottus gobio (Junker *et al.* 2012)	Fishes	0.44	0.053	−0.66	0.657	0.387
Salix hukaoana (Kikuchi *et al.* 2009) – Tadami	Plants	0.571	0.173	−0.49	0.571	0.293
Anodonta californiensis (Mock *et al.* 2010)	Mollusks	0.218	0.089	−0.81	0.409	0.368
Gasterosteus aculeatus (Raeymaekers *et al.* 2008)	Fishes	−0.26	0.15	−0.62	0.53	0.152
Gasterosteus aculeatus (Raeymaekers *et al.* 2009)	Fishes	0.403	0.07	−0.87	0.67	0.784
Oncorhynchus clarkii (Wofford *et al.* 2005)	Fishes	0.19	0.125	−0.63	0.184	0.048

COR_{AR} and $COR_{F_{ST}}$ values observed for each model. In addition, we independently calculated for each model (i) the proportion of simulations generating a COR_{AR} lower than the meta-COR_{AR} and (ii) the proportion of simulations generating COR_{AR} falling within the 95% CI of meta-COR_{AR}. Given the low sample size for $COR_{F_{ST}}$ ($n = 12$), this later calculation was not performed for $COR_{F_{ST}}$.

Summary statistics discriminating among processes generating DIGD. To identify summary statistics that may discriminate among the three processes (and their interactions) considered in this study when they actually generate DIGD, we built a random forest classification model (RF; Breiman 2001) in which the response variable was categorical (here, the identity of the eight models under which simulations were generated) and the predictor variables were the summary statistics calculated from simulations (i.e. COR_{AR}, COR_{H_E}, COR_{PA}, COR_{GW}, $COR_{F_{ST}}$, COR_{IBD}, COR_{IBF}, F_{ST}, F_{IS}, F_{IT}). RFs are a type of machine-learning algorithm combining multiple decision trees to obtain averaged predictions based on all the trees in the forest (the forest being the ensemble of all the generated trees; Breiman 2001; Cutler *et al.* 2007). Each decision tree in the forest is built using a random subset of simulations and a number p of randomly chosen predictor variables (here, summary statistics) for each split of the tree (Liaw & Wiener 2002). Here, our forest was composed of 500 classification trees (Breiman *et al.* 1984), and p was determined by calculating the square root of the maximum number of predictor variables (here, $p = 3$ for a total number of 10 summary statistics; Breiman 2001; Liaw & Wiener 2002). Simulations not used for building the trees at each bootstrap step (i.e. out-of-bag simulations, OOB) were then used to estimate an OOB error rate (Liaw & Wiener 2002; Cutler *et al.* 2007). We then used the RF to rank summary statistics through a predictor importance measure: the mean decrease (in percentage) in accuracy of the trees in the forest when the observed values of a predictor are randomly permuted in the OOB simulations (higher values indicate higher predictor importance). Only simulations generating realistic DIGD (i.e. COR_{AR} equal to or lower than the meta-COR_{AR}) were considered for building the RF. We used the R package 'randomForest' (Liaw & Wiener 2002) to build the RF.

Processes inferred from empirical data sets. We applied ABC model-choice procedures (Beaumont *et al.* 2002) to assess which of the three processes – or their interactions – considered in this study most likely generated DIGD observed in a subset of studies taken from the meta-analysis. Independent ABC analyses were specifically performed for the twelve case studies that were used to estimate meta-$COR_{F_{ST}}$ (see section Spatial patterns of genetic diversity in freshwater organisms above; Table 2). In ABC, summary statistics (s_x) calculated from many simulated data sets (D_x) generated under competing models are compared with those calculated from observed data (s_{obs}). D_x producing the closest s_x to s_{obs} are retained to subsequently approximate the posterior probabilities of the models. The choice of s in ABC is critical: increasing the number of s increases the amount of information injected in the procedure, but also drastically increases the number of simulations necessary to ensure the efficiency of the procedure (Aeschbacher *et al.* 2012; Blum *et al.* 2013). To integrate a maximum of information while keeping a reasonable number of summary statistics, we here considered for the ABC model-choice procedures a restricted set of four summary statistics (i.e. COR_{AR}, COR_{IBF}, F_{ST} and COR_{IBD}) that (i) were shown to be highly powerful to discriminate among processes by the RF (i.e. that displayed high mean decreases in accuracy; see Results section) and (ii) were available from published data. We did not include COR_{GW} and F_{IT} in the ABC even though they were highlighted as highly discriminant statistics by the RF (see Results section), because these statistics were not always reported in all the empirical studies investigated here. However, the increase in OOB classification error rate between a RF built on the full set of summary statistics as predictors (OOB error rate of 11.94%; Table 3) and a RF built only on the restricted set of summary statistics (OOB error rate of 18.58%; Table S2, Supporting information) was very low (6.64%), suggesting that proper inferences are achievable using this restricted set of summary statistics. Pairwise F_{ST} were used to calculate COR_{IBD} and COR_{IBF} by gathering topological distances and between-demes flow connection matrices determined from published maps (Table 2).

Model-choice procedures were performed using multinomial logistic regressions and considering a tolerance rate of 0.0001 with the R package 'abc' (Csilléry *et al.* 2012). As a comparison basis, we tested whether or not predictions made from the RF classification model built on the restricted number of summary statistics provide results similar to those obtained through the ABC approach.

Results

Spatial patterns of intraspecific diversity in river ecosystems

Overall, the distribution of COR_{AR} was skewed to the right (skewness = 0.762) with 77.22% negative COR_{AR}

Table 3 Out-of-bag confusion matrix obtained for a random forest classification model composed of 500 classification trees where the identity of the models having generated simulation (i.e. gene flow, habitat availability, colonization, gene flow/habitat, gene flow/colonization, habitat/colonization, gene flow/habitat/colonization and null) is the response variable, and 10 summary statistics (i.e. COR_{AR}, $COR_{F_{ST}}$, COR_{H_E}, COR_{PA}, COR_{GW}, F_{ST}, F_{IS}, F_{IT}, COR_{IBD}, COR_{IBF}) were the predictor variables

Model under which OOB simulations have been generated:	Percentage of OOB simulations assigned to								Classification error (%)
	Gene flow	Habitat availability	Colonization	Gene flow/ habitat	Gene flow/ colonization	Habitat/ colonization	Gene flow/ habitat/ colonization	Null model	
Gene flow	**97.56**	0.58	0.11	0.71	0.58	0.33	0.14	0.00	2.45
Habitat availability	0.07	**63.34**	0.34	0.22	0.05	35.70	0.27	0.00	36.65
Colonization	0.02	0.53	**89.40**	0.00	2.04	8.00	0.00	0.00	10.59
Gene flow/habitat	1.11	1.29	0	**95.48**	0.08	1.13	0.91	0.00	4.52
Gene flow/ colonization	0.43	0.03	0.71	0.07	**97.22**	0.14	1.40	0.00	2.78
Habitat/ colonization	0.06	24.19	1.79	0.19	0.22	**73.14**	0.41	0.00	26.86
Gene flow/habitat/ colonization	0.14	0.28	0.00	1.06	2.88	0.59	**95.05**	0.00	4.95
Null model	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00

The last column reports the OOB classification error *per* model. Percentages of correct assignments are reported in bold.

(Fig. 2A). Some populations exhibited very strong DIGD patterns (i.e. COR_{AR} lower than −0.9, Table S1, Supporting information) such as, for instance, for the fish species *Xyphophorus helleri* (Tatarenkov *et al.* 2010) and *Xyrauchen texanus* (Dowling *et al.* 2012), or for the Agnate *Lethenteron sp* (Yamazaki *et al.* 2011; Table S1, Supporting information). We calculated a global coefficient of correlation (i.e. meta-COR_{AR}) of −0.41 (95% CI: −0.55 to −0.27), indicating that – overall – DIGD is actually a significant and general spatial pattern of genetic diversity in rivers (Fig. 2A). However, there were notable exceptions, with some strongly positive relationships between allelic richness and distance to river mouth (e.g. COR_{AR} of 0.92 for the fish *Squalius torgalensis*, Henriques *et al.* 2010; Table S1, Supporting information, Fig. 2A), and nine case studies for which a better fit was obtained with a quadratic, rather than a linear relationship between allelic richness and distance to the river mouth (e.g. bell-shaped relationships for *Salmo salar*, Primmer *et al.* 2006; and U-shaped relationships for *Perca flavescens*, Leclerc *et al.* 2008; Table S1, Supporting information).

We found significant differences in effect sizes among taxonomic groups (DIC comparison between the null model and a model including the taxonomic group as a categorical factor: ΔDIC = 5.59): the mean effect size was not significantly different from zero for plants and amphibians, whereas it was significantly negative for agnates, arthropods, fish and mollusks (Fig. 3). It is noteworthy that this later result should be interpreted with care, as agnates and amphibians were represented by a single species each (Table S1, Supporting information). As expected, organisms able to use overland dispersal displayed lower (and non-significant) mean effect size (mean effect size = 0.025, 95% CI: −1.20 to 0.64; mean COR_{AR} = −0.158, 95% CI: −0.342 to 0.026; see Fig. 2A) than organisms with dispersal restricted to water corridors (mean effect size = −0.47, 95% CI: −0.61 to −0.28; mean COR_{AR} = −0.365, 95% CI: −0.475 to −0.255, see Fig. 2A). Regarding fish species traits, we found no single model that best explained the data, given that 60 of the 64 possible models had ΔDIC < 4. The null model (i.e. no fixed effects) was included in this set of selected models, meaning that this null model explains the data as well as any other models. There was hence no clear tendency between species traits and COR_{AR}, which was confirmed by visually inspecting the relationships between each trait and COR_{AR} (not shown).

Over the subset of 12 case studies, positive $COR_{F_{ST}}$ values were detected in 9 case studies (Table 2). However, the global coefficient of correlation was weak (meta-$COR_{F_{ST}}$ = 0.139) and the 95% CI included 0 (−0.257 to 0.531), indicating that an increase in genetic differentiation in upstream demes cannot be considered as a general spatial pattern.

Processes generating DIGD

Using nonparametric probability density functions, we showed that the frequency of DIGD (i.e. negative COR_{AR}) was very high for all models except the colonization and the null models (Fig. 2A). More than

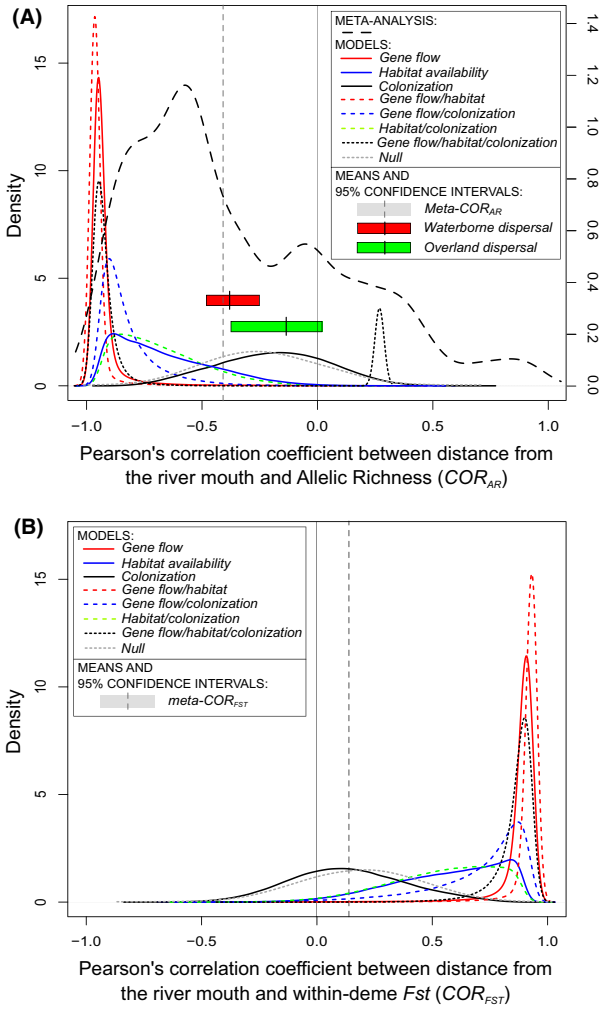

Fig. 2 (A) Figure representing probability density estimations of Pearson's correlation coefficient values between distance from the river mouth and allelic richness (i.e. COR_{AR} values) (i) for simulations from each eight models (coloured lines, left *y*-axis) and (ii) for populations included in the meta-analysis (black dotted line, right *y*-axis). The grey vertical dotted line represents the mean effect size calculated over all case studies from the meta-analysis (meta-COR_{AR}). The red and green boxes represent the 95% confidence intervals for COR_{AR} values observed for species displaying exclusively waterborne dispersal (red) or overland dispersal (green), along with their means (vertical black dashes). (B) Probability density estimations of Pearson's correlation coefficient values between distance from the river mouth and within-deme F_{ST} (i.e. $COR_{F_{ST}}$ values) for simulations from each eight models (coloured lines). The grey vertical dotted line represents the mean effect size calculated over twelve case studies from the meta-analysis (meta-$COR_{F_{ST}}$), along with its confidence interval.

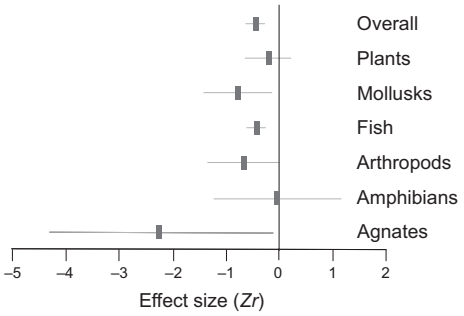

Fig. 3 Comparison of the mean effect sizes (*Zr*) obtained with Bayesian mixed-effects meta-analysis using the 'taxonomic group' as a categorical fixed effect. Grey lines represent 95% confidence intervals (95% CI). Mean effect sizes whose 95% CI did not overlap zero are significant.

80.83% of the simulations had COR_{AR} equal to or lower than the meta-COR_{AR} for all models but the null and colonization models, which generated COR_{AR} below the observed meta-COR_{AR} in only 21.01% and 19.39% of simulations, respectively (Fig. 2A). Variable proportions of simulations generated COR_{AR} comprised in the 95% CI bounding meta-COR_{AR} depending on the assumed model: 0.15% for the geneflow/habitat/colonization model, 0.21% for the geneflow/habitat model, 0.85% for the geneflow model, 3.77% for the geneflow/colonization model, 19.65% for the habitat/colonization model, 21.64% for the habitat availability model, 29.43% for the colonization model and 37.73% for the null model. As expected (Paz-Vinas & Blanchet 2015), simulations generated under the null model also revealed that, when dispersal is symmetric when, there is no variation in habitat availability nor upstream-directed colonization, then the relationship between allelic richness and distance from the river mouth was generally bell-shaped (see Fig. S1A, Supporting information). Most surprisingly, the distribution of COR_{AR} values obtained from simulations generated under the colonization model was very similar to that observed for the null model (Fig. 2A), hence suggesting that the effects of upstream-directed colonization on genetic diversity patterns may be confounded with the unique effect of dendritic connectivity *per se*.

Globally, the distribution of $COR_{F_{ST}}$ was inverse to that found for COR_{AR} for almost all models (Fig. 2B). Indeed, $COR_{F_{ST}}$ values were positive for almost all simulations, irrespective of the model considered. Only the null model and the colonization models displayed non-negligible proportions of negative $COR_{F_{ST}}$ values (21.76 and 30.79%, respectively).

The signature of the three processes generating DIGD

All the processes we considered in this study obviously had the potential to generate DIGD. Our idea is that they probably leave distinct genetic signatures that can be captured by other summary statistics than COR_{AR}, and that these summary statistics may then be used to

discriminate among competing processes. Using a random forest classification model (Breiman 2001), we found that the most important statistics for discriminating among processes – and their two- and three-way interactions – generating DIGD were COR_{IBF}, F_{ST} and COR_{GW}, followed by F_{IT}, COR_{AR} and COR_{IBD} (Figs 4 and S2, Supporting information). Other statistics exhibited low mean decreases in accuracy, indicating that they have a low predictive importance (Fig. 4).

The predictive performance of the random forest was very high, with an OOB classification error of only 11.94% (Table 3). Most misclassifications occurred between the habitat availability vs. the habitat/colonization models, indicating a lower statistical power for efficiently discriminating simulations from these two models (Table 3). It is noteworthy that all simulations generated under the null model were correctly classified and that no simulations generated with the other models were assigned to the null model (Table 3).

Processes inferred from empirical data sets

ABC model-choice procedures unambiguously identified the most likely processes (among the three processes we considered in the study) underlying DIGD in natural populations for 9 (of 12) case studies, with posterior mode probabilities >0.5 (Table 4). Six of these nine DIGD were likely to be generated by two interacting processes (Table 4), whereas three other were predicted to rely on a single process (two from upstream-directed colonization and one from variation in habitat availability; Table 4). The upstream-directed colonization process was implied (either individually or in interaction with another process) in 8 of the 9 case studies, indicating that it may be a major process influencing DIGD in the wild. For the three DIGD for which process inference was more challenging, more than 74% of the total posterior model probabilities were equally shared by two concurrent hypotheses (i.e. gene flow/colonization and colonization alone for Wofford *et al.* 2005; habitat/colonization and habitat availability alone for Junker *et al.* 2012; and habitat/colonization and gene flow/colonization for Alp *et al.* 2012; Table 4).

When processes were predicted from the alternative random forest classification model built on the same summary statistics than those for the ABC procedure (i.e. COR_{AR}, COR_{IBF}, COR_{IBD} and F_{ST}), we found a congruency between the two approaches for nine of the twelve case studies (Table 4). Two of these three incongruencies between methods occurred for case studies for which the ABC failed to clearly identify the process underlying DIGD (i.e. Wofford *et al.* 2005, Alp *et al.* 2012). For these two cases, the random forest identified the process with the second highest posterior probability as the most likely process for explaining DIGD (Table 4).

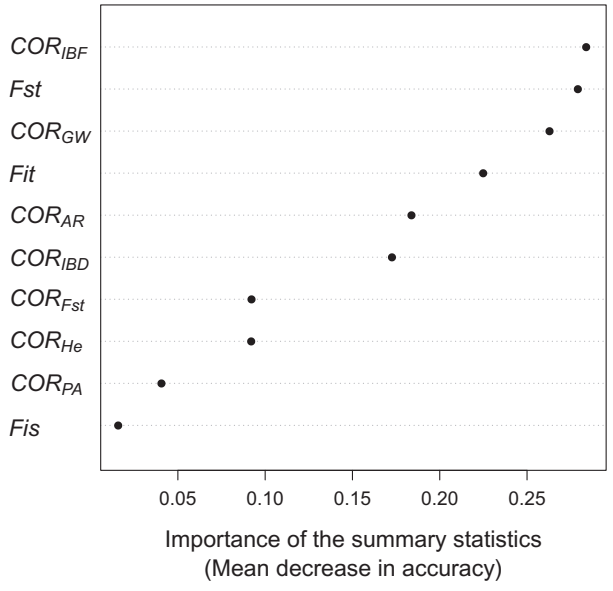

Fig. 4 Variable importance estimated for 10 summary statistics (i.e. COR_{AR}, COR_{H_E} , COR_{PA}, COR_{GW}, $COR_{F_{ST}}$, COR_{IBD}, $COR_{IBF,\ FST}$, F_{IS}, F_{IT}) from a random forest classification model built for predicting the model having generated a simulation (i.e. model identity was the response variable). The importance of each summary statistic has been assessed by measuring the mean decrease in accuracy of a tree in the forest when the observed values of the summary statistic are randomly permuted in the out-of-bag simulations. Higher values denote higher importance of the summary for predicting the correct model.

Discussion

This study illustrates the usefulness of coupling meta-analysis, pattern-oriented simulations, variable selection approaches and model-choice procedures to characterize spatial patterns of biodiversity in the wild, and to identify the foremost processes generating these patterns at the intraspecific genetic level, an ecologically and evolutionary important facet of biodiversity (Hughes *et al.* 2008). The meta-analysis revealed that neutral genetic diversity increases as one moves downstream in river ecosystems when characterized over all case studies. There was, however, strong variation among taxonomic groups in the strength of the relationship between genetic diversity and distance from the river mouth. We theoretically showed that the three processes we considered (downstream-biased gene flow, variation in habitat availability and upstream-directed colonization) can act independently or in interaction to generate such a spatial pattern. Combining model-choice procedures (ABC) and a restricted – but informative – set of discriminant statistics determined

Table 4 Posterior model probabilities obtained with ABC model-choice procedures (based on multinomial logistic regression) for a subset of twelve populations exhibiting significant DIGD extracted from the meta-analysis

	Posterior model probabilities							
Species name and Reference	Gene flow	Habitat availability	Colonization	Gene flow/ habitat	Gene flow/ colonization	Habitat/ colonization	Gene flow/ habitat/ colonization	Null model
Gammarus fossarum (Alp *et al.* 2012)	0.0000	0.0029	0.0077	0.0000	0.3388*	**0.4468**	0.2038	0.0000
Semotilus atromaculatus (Boizard *et al.* 2009)	0.0000	0.0000	0.0024	0.0000	**0.9408***	0.0000	0.0569	0.0000
Poecilia reticulata (Crispo *et al.* 2005)	0.0000	0.0000	0.0008	0.0000	**0.9847***	0.0076	0.0069	0.0000
Telestes souffia (Dubut *et al.* 2012)	0.0863	0.2697	0.0000	0.0052	0.0000	**0.5762***	0.0626	0.0000
Cottus gobio (Hänfling & Weetman 2006)	0.0008	0.0000	0.1333	0.0000	**0.8380***	0.0271	0.0008	0.0000
Salmo trutta (Horreo *et al.* 2011) – Nive	0.0000	0.0000	**0.8698***	0.0000	0.1027	0.0275	0.0000	0.0000
Cottus gobio (Junker *et al.* 2012)	0.0565	0.4138	0.0000	0.1049	0.0000	**0.4200***	0.0049	0.0000
Salix hukaoana (Kikuchi *et al.* 2009) – Tadami	0.0040	0.0000	0.0531	0.0000	**0.6730***	0.0000	0.0000	0.2699
Anodonta californiensis (Mock *et al.* 2010)	0.0042	0.0688	0.1131*	0.0000	0.0915	**0.6631**	0.0592	0.0000
Gasterosteus aculeatus (Raeymaekers *et al.* 2008)	0.0000	0.0000	**0.9237***	0.0000	0.0478	0.0285	0.0000	0.0000
Gasterosteus aculeatus (Raeymaekers *et al.* 2009)	0.0242	**0.5275***	0.0000	0.0182	0.0000	0.4293	0.0008	0.0000
Oncorhynchus clarkii (Wofford *et al.* 2005)	0.0083	0.0000	**0.4154**	0.0000	0.3295*	0.0049	0.0000	0.2418

Bold numbers highlight the highest posterior model probabilities found for each population with the ABC procedure.
*The model predicted by the random forest classification model.

through a machine-learning approach (i.e. a random forest classification model), we finally hypothesized what are the most likely processes (among the processes we considered) underlying DIGD observed in twelve empirical case studies. In parallel, our results suggest that the hypothesis that genetic differentiation should increase as one moves upstream in river networks was not verified.

Downstream increase in genetic diversity: a general spatial pattern

Kermit Ritland (1989) was the first to propose the idea of DIGD in riverscapes by predicting DIGD in riparian and aquatic plants due to the downstream-biased dispersal of seeds transported by water. Nevertheless, several studies failed at identifying DIGD in natural populations (e.g. Tero *et al.* 2003), making it difficult to generalize the relationship between neutral genetic diversity and distance from the river mouth. Here, we explored a broad taxonomic spectrum at a worldwide spatial extent, and we quantitatively demonstrate that DIGD can actually be considered as a general spatial pattern of biodiversity (sensu Lawton 1996).

As expected, the negative relationship between allelic richness and distance from the river mouth was stronger for organisms with purely aquatic dispersal such as fishes or mollusks than organisms with overland dispersal (e.g. riparian plants, a winged insect and an amphibian). This suggests that overland dispersal may counteract processes such as waterborne asymmetric dispersal, hence allowing the maintenance of genetic diversity in upstream reaches and homogenizing genetic diversity over the whole river catchment (Chaput-Bardy *et al.* 2008; Campbell Grant *et al.* 2010; Geismar *et al.* 2015). Overland dispersal may also explain why – as previously demonstrated by Honnay *et al.* (2010) – we found that riparian plants do not meet the DIGD. This is actually surprising given the strong expected role of – downstream-directed – waterborne seeds dispersal (Ritland

1989). This absence of relationships between genetic diversity and distance from the river mouth in riparian plants may result from overland pollen transfer (transported either by insects or by air flows), which is probably independent from river network configuration and which participates to gene flow. It would not be surprising that overland pollen diffusion can impede the establishment of DIGD. These later results confirm that the ability to disperse outside the river network is probably a trait of major importance for predicting the strength and shape of empirical DIGD.

When focusing on fish species, we did not identify biological traits that were statistically related to the strength of the relationship between allelic richness and distance from the river mouth. Interestingly, this conclusion holds true even for traits that are traditionally believed to underlie dispersal ability in fish such as maximum body length, swimming ability or migratory behaviour. This means that our understanding of the links between biological traits and the spatial distribution of genetic diversity is still scarce, although our results suggest that these links might be better understood if traits are considered in interactions with other structuring forces. For instance, DIGD observed in two river networks for the fish *Gasterosteus aculeatus* (Raeymaekers *et al.* 2008, 2009) were likely caused by two different mechanisms (Table 4), suggesting interactive effects among landscape (network) structure, historical contingency and biological characteristics of the taxa. Such interactive effects also probably explain why in some cases, the spatial pattern of genetic diversity strikingly varies among river networks for a single species. The most extreme example was for the fish *Salmo trutta* (Horreo *et al.* 2011), which displayed four different relationships in four different river systems (i.e. negative, positive, U-shaped and bell-shaped relationships; Table S1, Supporting information). It is worth mentioning that such intraspecific differences could also be in part explained by intraspecific variation in biological traits (e.g. differences in maximum body length between different populations of a single species) and/or differences among the river networks (e.g. size of the river network), two features that were not included in our analyses. We hence anticipate that future empirical and theoretical studies should focus more specifically on such interactive effects, and especially on the interactions between network characteristics and both intra- and interspecific variation in biological traits.

Other spatial patterns than DIGD

It is noteworthy that we found exceptions to the general pattern of increase in genetic diversity downstream, some of them deserving considerations. We noticed that several populations display a positive relationship between distance from the river mouth and genetic diversity. Such patterns may be due to the same, but inverted, processes as those generating DIGD. For instance, it is possible that effective population sizes increase upstream due, for instance, to relaxed competition for resources, to altitudinal shifts of species ranges (e.g. due to climate change; Conti *et al.* 2015), or that a downstream-directed colonization occurs after a mass extinction, an introduction or a glacial event (Cyr & Angers 2012). In these cases, the processes can generate positive relationships between distance from the river mouth and allelic richness, as was confirmed by supplementary simulations (see Appendix S2, Supporting information). Upstream-biased gene flow on the contrary is less biologically likely (although it can theoretically generate increased genetic diversity *upstream*; see Appendix S2, Supporting information), as most species considered disperse in or on moving water.

The detection of nonlinear spatial patterns of allelic richness in 11.4% of our simulations was surprising. In most of them, allelic richness was higher in the middle of the river network. This may empirically arise when population densities follow a similar distribution (Watanabe *et al.* 2008) or when species ranges are limited at boundaries, with peripheral populations (e.g. the most upstream and downstream populations) exhibiting lower local genetic diversities (Kirkpatrick & Barton 1997; Eckert *et al.* 2008; Liggins *et al.* 2015), two situations that were not simulated here. In our simulations, this pattern was mainly due to the unique effect of dendritic connectivity *per se,* as previously demonstrated by Paz-Vinas & Blanchet (2015). Demes situated in intermediate sections of the river network are enriched in genetic diversity, because they act as 'crossroads' between genetically distinct demes (notably when migration rates are low), a result that comes to complement similar findings on taxonomic diversity (Carrara *et al.* 2012). This confirms that new connectivity-based statistics such as network centrality measures (informing how a node is important in terms of its connectivity) need to be considered to gain insights into the role of network connectivity on spatial patterns of biodiversity in rivers (Altermatt 2013).

Additionally, we explored the hypothesis that genetic differentiation (as measured by F_{ST}) should be higher in upstream than in downstream demes in river networks, a spatial pattern of genetic differentiation that has been observed both empirically (Finn *et al.* 2011; Múrria *et al.* 2013) and theoretically (this study, see also Paz-Vinas & Blanchet 2015). Our meta-analysis did not allow us to generalize this pattern, given that the tendency for a positive correlation between genetic differentiation and distance from the outlet was not significant. On the

contrary, our simulations confirmed previous theoretical works demonstrating that dendritic connectivity can generate this pattern (Paz-Vinas & Blanchet 2015), and actually show that this pattern can be reinforced when asymmetric gene flow, differences in effective population sizes and/or upstream-directed colonization occur. This discrepancy between our empirical and theoretical findings might simply be due to the low number of case studies we considered here ($n = 12$). Despite this weakness, our empirical results somewhat corroborate the findings of Finn *et al.* (2011). In a meta-analysis, they indeed showed that genetic differentiation was actually significantly higher in isolated upstream populations than in more connected downstream populations when differentiation was measured using an ecological index of dissimilarity (Sorensen's dissimilarity), but not when measured using the 'classical' F_{ST} measure. It seems therefore that in natural populations, genetic differentiation (when measured using F_{ST}) might not be higher in upstream than in downstream populations in dendritic networks. We believe that further insights on the patterns of population differentiation in dendritic networks could be obtained in the future by developing studies partitioning the turnover and nestedness components of population genetic differentiation (Baselga 2010; Liggins *et al.* 2015).

Different processes can lead to similar patterns, but with different genetic signatures

Our simulations revealed that downstream-biased gene flow, variation in habitat availability between demes and upstream-directed colonization can all generate DIGD, either independently or in interaction. Downstream-biased gene flow generates DIGD in most cases, as even weak asymmetric gene flow breaks down the effect of dendritic connectivity and generates strong DIGD, even in the presence of another interacting process (see Appendix S3 and Fig. S1B, Supporting information). However, this pattern was less often identified by the ABC procedures as the most likely process generating empirical DIGD (i.e. it is supported in 4 of the 12 empirical case studies, always in combination with colonization; Table 4). Similarly, weak differences in effective population sizes along the upstream–downstream gradient of habitat availability were sufficient to generate moderate to strong DIGD, especially when this process acts alone or in interaction with upstream-directed colonization (see Appendix S3 and Fig. S1C, Supporting information). In contrast, upstream-directed colonization generated moderate DIGD only under restricted (and complex) combinations of parameter values when acting alone. Stronger DIGD were most likely to occur in the colonization model in simulations involving demes of small effective sizes connected by low dispersal rates (Appendix S3, Supporting information). These simulations hence revealed that although all the processes we assumed can generate DIGD, the probability of observing DIGD in river networks might be higher for populations experiencing asymmetric gene flow or a gradient in habitat availability (or any other parameter acting similarly on effective population sizes).

This conclusion was, however, challenged by the inference of processes generating DIGD in natural populations with ABC model-choice procedures, which shows the limitation of theoretical works when they are not confronted to empirical patterns. The ABC model-choice procedures revealed that DIGD observed in the wild are probably not primarily generated by a single specific process, but rather by the combination of two interacting processes. In addition, upstream-directed colonization was probably implied in almost all the empirical cases investigated (either independently or in interaction with another process), which suggest that this process has a great importance in natural populations. Unexpectedly, the process that was first proposed by Kermit Ritland as the main generator of DIGD and that is generally suspected as the most likely to generate DIGD (i.e. downstream-directed gene flow) was implied in only four case studies and always in interaction with the colonization process (Table 4). It is noteworthy that one potential risk in model-based statistical inference (and hence, in ABC-based statistical inference) is the sensitivity of the inference methods to misspecification of the models. Here, we simulated a single generic landscape structure with only three putative processes generating DIGD, whereas a specific landscape structure that mimics the real empirical one should ideally be designed for each case study. Notwithstanding this, our results clearly show that competing processes related to the demography, history and dispersal patterns of populations can all underlie DIGD in the wild, and must therefore be considered conjointly when predicting spatial patterns of intraspecific diversity.

We show here that each of these processes leaves distinct genetic footprints that can be apprehended with a small number of simple descriptive statistics, which provides a possible way to discriminate among competing processes when DIGD are observed empirically. More particularly, we demonstrated that the three processes leave strikingly different signatures in terms of genetic differentiation between flow-connected and flow-unconnected demes (isolation by flow). Genetic differentiation was generally higher between flow-unconnected demes than between flow-connected demes, but this pattern was stronger for models ruled by asymmetric gene flow (either independently or in interaction with other processes) than for any other model (Fig. S2, Supporting

information). This illustrates how metrics accounting for network connectivity may improve the characterization of spatial patterns of biodiversity, especially in dendritic ecological networks (Carrara *et al.* 2012; Altermatt 2013; Peterson *et al.* 2013). However, the discriminant ability of our approach was much higher when information on isolation by flow was combined to four other statistics. Three of them (F_{ST}, F_{IT} and the slope of isolation by distance) are routinely used by population geneticists, whereas the fourth one (the correlation between Garza–Williamson's statistic and distance to river mouth) is more rarely reported, although it proved to be a really insightful statistic. Surprisingly, when the random forest classification model built on a restricted set of four discriminant summary statistics was used for predicting processes from empirical patterns of DIGD, we found a high congruency between this approach and results from the ABC (Table 4). Despite this high congruency, random forest classification models should not be used directly for inferring processes from empirical genetic data but only to reduce the number of summary statistics that will be used in ABC, as there is no study – to our knowledge – evaluating and validating the efficiency of random forests vs. ABC procedures for model-choice issues. That being said, our study shows that the combination of few simple, but informative statistics (combined to an ABC model-choice approach) proved here to be powerful for identifying the most likely process underlying empirical DIGD among a set of competing processes, which paves the way towards a new framework for investigating empirical spatial patterns of intraspecific diversity.

Conclusions and implications

Our study provides novel insights into the description and understanding of spatial patterns of biodiversity in river systems. The pattern of downstream increase in genetic diversity highlighted here complements the well-established observation that taxonomic diversity also increases as one moves downstream (Campbell Grant *et al.* 2007; Altermatt 2013). This suggests that biodiversity patterns might often be congruent across biological levels in dendritic ecological networks (e.g. Finn & Poff 2011; Blum *et al.* 2012). Future studies should investigate to which extent the processes generating spatial patterns of genetic and taxonomic diversity correlate in dendritic networks (Vellend & Geber 2005; Pauls *et al.* 2014; Vellend *et al.* 2014) so as to develop management plans accounting for multiple biodiversity facets.

Our study demonstrates the strength of model and empirical coupling to unravel the patterns of intraspecific genetic diversity, even within a meta-analytical framework. We demonstrate that coupling meta-analyses and simulation-based studies is powerful for describing empirical genetic patterns and for discriminating among alternative underlying processes. Intraspecific genetic diversity is the fuel for species to adapt to environmental variability and can also have strong indirect influences on ecosystem functioning (Hughes *et al.* 2008; Caballero & García-Dorado 2013). Although next works should rapidly focus on the spatial distribution of adaptive genetic diversity in river networks (e.g. Nukazawa *et al.* 2014), understanding the distribution of neutral genetic diversity in river networks is a first step for predicting the potential for local adaptation in these ecosystems. There have been very few attempts to describe spatial patterns of local adaptation in rivers, or more generally in dendritic networks. However, the nonrandom distribution of genetic diversity in these systems may have a strong influence on the ability of populations to adapt to local conditions (Kawecki & Holt 2002). We hence believe that the time is ripe to make a valuable use of the large amount of data that has been collected in the last decades in population genetics, so as to transform a collection of independent case studies into general rules (ArchMiller *et al.* 2015). As describing and understanding the patterns of biological diversity is one of the bases of integrative and efficient conservation measures (Chave 2013), this should become a priority for population and conservation geneticists.

Acknowledgements

We thank the CALMIP group, Pierette Barbaresco, Boris Dintrans and Nicolas Renon in particular. This work was performed using HPC resources from CALMIP (allocation P1003). Olivier Rey, Joost Raeymaekers, Vincent Dubut, Camille Pagès, Eric Petit, Louis Bernatchez as well as four anonymous referees are thanked for their constructive and stimulating comments. Radika Michniewicz and Keoni Saint-Pé corrected the English. We thank K.E. Mock, J. Brim-Box and the Confederated Tribes of the Umatilla Indian Reservation for sharing their data set on freshwater mussels. We also thank Maria Alp, Erika Crispo, Satoshi Kikuchi, Vincent Dubut, Bernard Angers, Bernt Hänfling, Joost Raeymaekers and Julien Junker for sharing their data. Ivan Paz-Vinas has been financially supported by a PhD scholarship from the 'Ministère de l'Enseignement Supérieur et de la Recherche'. This study is part of the European project 'IMPACT' and has been carried out with financial support from the Commission of the European Communities, specific RTD programme 'IWRMNET'. This work has been carried out in two research units (EDB & SEEM) that are part of the 'Laboratoire d'Excellence' (LABEX) entitled TULIP (ANR-10-LABX-41).

References

Aeschbacher S, Beaumont MA, Futschik A (2012) A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, **192**, 1027–1047.

Alp M, Keller I, Westram AM, Robinson CT (2012) How river structure and biological traits influence gene flow: a population genetic study of two stream invertebrates with differing dispersal abilities. *Freshwater Biology*, **57**, 969–981.

Altermatt F (2013) Diversity in riverine metacommunities: a network perspective. *Aquatic Ecology*, **47**, 365–377.

ArchMiller AA, Bauer EF, Koch RE *et al.* (2015) Formalizing the definition of meta-analysis in *Molecular Ecology*. *Molecular Ecology*, **24**, 4042–4051.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

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

Benda L, Poff NL, Miller D *et al.* (2004) The network dynamics hypothesis: how channel networks structure riverine habitats. *BioScience*, **54**, 413.

Blum MJ, Bagley MJ, Walters DM *et al.* (2012) Genetic diversity and species diversity of stream fishes covary across a land-use gradient. *Oecologia*, **168**, 83–95.

Blum MGB, Nunes MA, Prangle D, Sisson SA (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, **28**, 189–208.

Boizard J, Magnan P, Angers B (2009) Effects of dynamic landscape elements on fish dispersal: the example of creek chub (*Semotilus atromaculatus*). *Molecular Ecology*, **18**, 430–441.

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

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees*. Wadsworth, Belmont, California.

Caballero A, García-Dorado A (2013) Allelic diversity and its implications for the rate of adaptation. *Genetics*, **195**, 1373–1384.

Campbell Grant EH, Lowe WH, Fagan WF (2007) Living in the branches: population dynamics and ecological processes in dendritic networks. *Ecology Letters*, **10**, 165–175.

Campbell Grant EH, Nichols JD, Lowe WH, Fagan WF (2010) Use of multiple dispersal pathways facilitates amphibian persistence in stream networks. *Proceedings of the National Academy of Sciences*, **107**, 6936–6940.

Carrara F, Altermatt F, Rodriguez-Iturbe I, Rinaldo A (2012) Dendritic connectivity controls biodiversity patterns in experimental metacommunities. *Proceedings of the National Academy of Sciences*, **109**, 5761–5766.

Carrara F, Rinaldo A, Giometto A, Altermatt F (2014) Complex interaction of dendritic connectivity and hierarchical patch size on biodiversity in river-like landscapes. *The American Naturalist*, **183**, 13–25.

Chaput-Bardy A, Lemaire C, Picard D, Secondi J (2008) In-stream and overland dispersal across a river network influences gene flow in a freshwater insect, *Calopteryx splendens*. *Molecular Ecology*, **17**, 3496–3505.

Chaput-Bardy A, Fleurant C, Lemaire C, Secondi J (2009) Modelling the effect of in-stream and overland dispersal on gene flow in river networks. *Ecological Modelling*, **220**, 3589–3598.

Chave J (2013) The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology Letters*, **16**, 4–16.

Coleman RA, Weeks AR, Hoffmann AA (2013) Balancing genetic uniqueness and genetic variation in determining conservation and translocation strategies: a comprehensive case study of threatened dwarf galaxias, *Galaxiella pusilla* (Mack) (Pisces: Galaxiidae). *Molecular Ecology*, **22**, 1820–1835.

Conti L, Comte L, Hugueny B, Grenouillet G (2015) Drivers of freshwater fish colonisations and extirpations under climate change. *Ecography*, **38**, 510–519.

Crispo E, Bentzen P, Reznick DN, Kinnison MT, Hendry AP (2005) The relative influence of natural selection and geography on gene flow in guppies. *Molecular Ecology*, **15**, 49–62.

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.

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

Dowling TE, Saltzgiver MJ, Marsh PC (2012) Genetic structure within and among populations of the endangered razorback sucker (*Xyrauchen texanus*) as determined by analysis of microsatellites. *Conservation Genetics*, **13**, 1073–1083.

Dubut V, Fouquet A, Voisin A *et al.* (2012) From late miocene to holocene: Processes of differentiation within the Telestes genus (Actinopterygii: Cyprinidae). *PLoS ONE*, **7**, e34423.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology*, **17**, 1170–1188.

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.

Finn DS, Poff NL (2011) Examining spatial concordance of genetic and species diversity patterns to evaluate the role of dispersal limitation in structuring headwater metacommunities. *Journal of the North American Benthological Society*, **30**, 273–283.

Finn DS, Bonada N, Múrria C, Hughes JM (2011) Small but mighty: headwaters are vital to stream network biodiversity at two levels of organization. *Journal of the North American Benthological Society*, **30**, 963–980.

Frankham R (1996) Relationship of genetic variation to population size in wildlife. *Conservation Biology*, **10**, 1500–1508.

Froese R, Pauly D (2015) FishBase. World Wide Web electronic publication. URL: http://www.fishbase.org/

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

Geismar J, Haase P, Nowak C, Sauer J, Pauls SU (2015) Local population genetic structure of the montane caddisfly *Drusus discolor* is driven by overland dispersal and spatial scaling. *Freshwater Biology*, **60**, 209–221.

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1–19.

Gotelli NJ, Anderson MJ, Arita HT *et al.* (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, **12**, 873–886.

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, **33**, 1–22.

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.

Henriques R, Sousa V, Coelho MM (2010) Migration patterns counteract seasonal isolation of *Squalius torgalensis*, a criti-

cally endangered freshwater fish inhabiting a typical Circum-Mediterranean small drainage. *Conservation Genetics*, **11**, 1859–1870.

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.

Hillebrand H (2004) On the generality of the latitudinal diversity gradient. *The American Naturalist*, **163**, 192–211.

Hoban S (2014) An overview of the utility of population simulation software in molecular ecology. *Molecular Ecology*, **23**, 2383–2401.

Honnay O, Jacquemyn H, Nackaerts K, Breyne P, Van Looy K (2010) Patterns of population genetic diversity in riparian and aquatic plant species along rivers. *Journal of Biogeography*, **37**, 1730–1739.

Horreo JL, Martinez JL, Ayllon F *et al.* (2011) Impact of habitat fragmentation on the genetics of populations in dendritic landscapes. *Freshwater Biology*, **56**, 2567–2579.

Hughes JM (2007) Constraints on recovery: using molecular methods to study connectivity of aquatic biota in rivers and streams. *Freshwater Biology*, **52**, 616–631.

Hughes AR, Inouye BD, Johnson MTJ, Underwood N, Vellend M (2008) Ecological consequences of genetic diversity. *Ecology Letters*, **11**, 609–623.

Hughes JM, Huey JA, Schmidt DJ (2013) Is realised connectivity among populations of aquatic fauna predictable from potential connectivity? *Freshwater Biology*, **58**, 951–966.

Junker J, Peter A, Wagner CE *et al.* (2012) River fragmentation increases localized population genetic structure and enhances asymmetry of dispersal in bullhead (*Cottus gobio*). *Conservation Genetics*, **13**, 545–556.

Kawecki TJ, Holt RD (2002) Evolutionary consequences of asymmetric dispersal rates. *The American Naturalist*, **160**, 333–347.

Kikuchi S, Suzuki W, Sashimura N (2009) Gene flow in an endangered willow *Salix hukaoana* (Salicaceae) in natural and fragmented riparian landscapes. *Conservation Genetics*, **12**, 79–89.

Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Kirkpatrick M, Barton NH (1997) Evolution of a species' range. *The American Naturalist*, **150**, 1–23.

Labonne J, Ravigne V, Parisi B *et al.* (2008) Linking dendritic network structures to population demogenetics: The downside of connectivity. *Oikos*, **117**, 1479–1490.

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.

Lawton JH (1996) Patterns in ecology. *Oikos*, **75**, 145.

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.

Levin S (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Lichstein JW (2006) Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, **188**, 117–131.

Liggins L, Booth DJ, Figueira WF *et al.* (2015) Latitude-wide genetic patterns reveal historical effects and contrasting patterns of turnover and nestedness at the range peripheries of a tropical marine fish. *Ecography*, doi:10.1111/ecog.01398.

Mock KE, Brim Box JC, Chong JP *et al.* (2010) Genetic structuring in the freshwater mussel *Anodonta* corresponds with major hydrologic basins in the western United States. *Molecular Ecology*, **19**, 569–591.

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.

Müller K (1954) Investigations on the organic drift in north Swedish streams. *Report – Institute of Fresh-water Research, Drottningholm*, **35**, 532–537.

Muneepeerakul R, Weitz JS, Levin SA, Rinaldo A, Rodriguez-Iturbe I (2007) A neutral metapopulation model of biodiversity in river networks. *Journal of Theoretical Biology*, **245**, 351–363.

Múrria C, Bonada N, Arnedo MA, Prat N, Vogler AP (2013) Higher β- and γ-diversity at species and genetic levels in headwaters than in mid-order streams in *Hydropsyche* (Trichoptera). *Freshwater Biology*, **58**, 2226–2236.

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews of the Cambridge Philosophical Society*, **82**, 591–605.

Nakagawa S, Santos ESA (2012) Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, **26**, 1253–1274.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nosil P (2009) Adaptive population divergence in cryptic color-pattern following a reduction in gene flow. *Evolution*, **63**, 1902–1912.

Nukazawa K, Kazama S, Watanabe K (2014) A hydrothermal simulation approach to modelling spatial patterns of adaptive genetic variation in four stream insects. *Journal of Biogeography*, **42**, 103–113.

Pauls SU, Alp M, Bálint M *et al.* (2014) Integrating molecular tools into freshwater ecology: developments and opportunities. *Freshwater Biology*, **59**, 1559–1576.

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, 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.

Peterson EE, Ver Hoef JM, Isaak DJ *et al.* (2013) Modelling dendritic ecological networks in space: an integrated network perspective. *Ecology Letters*, **16**, 707–719.

Primmer CR, Veselov AJ, Zubchenko A *et al.* (2006) Isolation by distance within a river system: genetic population structuring of Atlantic salmon, *Salmo salar*, in tributaries of the Varzuga River in northwest Russia. *Molecular Ecology*, **15**, 653–666.

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, Raeymaekers D, Koizumi I, Geldof S, Volckaert FAM (2009) Guidelines for restoring connectivity around water mills: a population genetic approach to the management of riverine fish. *Journal of Applied Ecology*, **46**, 562–571.

Ritland K (1989) Genetic differentiation, diversity, and inbreeding in the mountain monkeyflower (*Mimulus caespitosus*) of the Washington Cascades. *Canadian Journal of Botany*, **67**, 2017–2024.

Ronce O (2007) How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 231–253.
Sexton JP, Hangartner SB, Hoffmann AA (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, **68**, 1–15.
Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics*, **24**, 2498–2504.
Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.
Tatarenkov A, Healey CIM, Avise JC (2010) Microgeographic population structure of green swordail fish: genetic differentiation despite abundant migration. *Molecular Ecology*, **19**, 257–268.
Tero N, Aspi J, Siikamäki P, Jäkäläniemi A, Tuomi J (2003) Genetic structure and gene flow in a metapopulation of an endangered plant species, *Silene tatarica*. *Molecular ecology*, **12**, 2073–2085.
Torterotot JB, 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.
Vellend M, Geber MA (2005) Connections between species diversity and genetic diversity. *Ecology Letters*, **8**, 767–781.
Vellend M, Lajoie G, Bourret A *et al.* (2014) Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. *Molecular Ecology*, **23**, 2890–2901.
Watanabe K, Monaghan MT, Omura T (2008) Longitudinal patterns of genetic diversity and larval density of the riverine caddisfly *Hydropsyche orientalis* (Trichoptera). *Aquatic Sciences*, **70**, 377–387.
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.
Wofford JEB, Gresswell RE, Banks MA (2005) Influence of barriers to movement on within-watershed genetic variation of coastal cutthroat trout. *Ecological Applications*, **15**, 628–637.
Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.
Yamazaki Y, Yamano A, Oura K (2011) Recent microscale disturbance of gene flow in threatened fluvial lamprey, *Lethenteron sp. N*, living in a paddy water system. *Conservation Genetics*, **12**, 1373–1377.

I.P.-V., G.L. and S.B. designed the study. I.P.-V. implemented the simulations. I.P.-V., G.L., V.M.S. and S.B. analysed the results and wrote the article. All authors read and approved this version of the manuscript.

Data accessibility

All the simulated genetic data sets generated under the eight different models, the scripts that were used to simulate them and data and scripts used for conducting the meta-analyses are available at Dryad Digital Repository doi: 10.5061/dryad.38c1 g.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Pearson's correlation coefficient values between allelic richness and distance to the river mouth (COR_{AR}) obtained or computed for 79 populations from a literature survey, along with publication details and species names for each surveyed population.

Table S2 Out-of-bag confusion matrix obtained for an alternative Random Forest classification model composed of 500 classification trees where the identity of the models having generated simulations (i.e. gene flow, habitat availability, colonization, gene flow/habitat, gene flow/colonization, habitat/colonization, gene flow/habitat/colonization and null) is the response variable, and a restricted set of four summary statistics (i.e. COR_{AR}, $COR_{F_{ST}}$, F_{ST} and F_{IT}) were the predictor variables.

Fig. S1 Mean Allelic Richness *per* deme in function of the distance from the putative river mouth (in demes) for (A) simulations obtained with the null model assuming $D_{SYMMETRIC} = 0.03$ and $N_{DEMES} = 10\,000$, (B) simulations obtained with the gene flow model assuming $D_{DOWNSTREAM} = 0.03$, $N_{DEMES} = 10\,000$ and $P_{ASYM} = 3$, and (C) simulations obtained for the habitat availability model assuming $D_{SYMMETRIC} = 0.03$, $P_{SCAL} = 1.3$ and $N_{HEADWATER} = 500$. Grey vertical lines represent standard errors.

Fig. S2 Probability density estimations of (A) COR_{IBF}, (B) $COR_{IBD,\ (C)\ FST}$, (D) F_{IT} and (E) COR_{GW} for each model.

Appendix S1 The computational pipeline used for implementing the simulation procedure.

Appendix S2 Description of the supplementary simulations considering inverted gene flow asymmetry and decreasing habitat availability along the upstream-downstream gradient.

Appendix S3 Regression trees built for each model.