MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 5348–5363 doi: 10.1111/mec.13401

# Elucidating the spatio-temporal dynamics of an emerging wildlife pathogen using approximate Bayesian computation

OLIVIER REY,*[1] LISA FOURTUNE,*[1] IVAN PAZ-VINAS,†‡§[1] GÉRALDINE LOOT,*‡ CHARLOTTE VEYSSIÈRE,* BENJAMIN ROCHE¶ and SIMON BLANCHET*†

**Station d'Écologie Expérimentale du CNRS à Moulis, USR 2936, 09200 Moulis, France, †CNRS, UPS, ENFA, Évolution & Diversité Biologique (EDB) UMR 5174, 118 Route de Narbonne, 31062 Toulouse, Cedex 9, France, ‡Université de Toulouse, UPS, UMR-5174 (EDB), 118 route de Narbonne, 31062 Toulouse, Cedex 9, France, §Aix-Marseille Université, CNRS, IRD, Université d'Avignon et des Pays de Vaucluse, UMR 7263 – IMBE, Équipe EGE, Centre Saint-Charles, Case 36, 3 place Victor Hugo, 13331 Marseille, Cedex 3, France, ¶IRD, UPMC, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes (UMMISCO), 32 avenue Varagnat, 93143 Bondy, Cedex, France*

## Abstract

**Emerging pathogens constitute a severe threat for human health and biodiversity. Determining the status (native or non-native) of emerging pathogens, and tracing back their spatio-temporal dynamics, is crucial to understand the eco-evolutionary factors promoting their emergence, to control their spread and mitigate their impacts. However, tracing back the spatio-temporal dynamics of emerging wildlife pathogens is challenging because (i) they are often neglected until they become sufficiently abundant and pose socio-economical concerns and (ii) their geographical range is often little known. Here, we combined classical population genetics tools and approximate Bayesian computation (i.e. ABC) to retrace the dynamics of *Tracheliastes polycolpus*, a poorly documented pathogenic ectoparasite emerging in Western Europe that threatens several freshwater fish species. Our results strongly suggest that populations of *T. polycolpus* in France emerged from individuals originating from a unique genetic pool that were most likely introduced in the 1920s in central France. From this initial population, three waves of colonization occurred into peripheral watersheds within the next two decades. We further demonstrated that populations remained at low densities, and hence undetectable, during 10 years before a major demographic expansion occurred, and before its official detection in France. These findings corroborate and expand the few historical records available for this emerging pathogen. More generally, our study demonstrates how ABC can be used to determine the status, reconstruct the colonization history and infer key evolutionary parameters of emerging wildlife pathogens with low data availability, and for which samples from the putative native area are inaccessible.**

*Keywords*: approximate Bayesian computation, Bayesian clustering, pathogen, population genetics, spatio-temporal dynamic

*Received 30 January 2014; revision received 23 September 2015; accepted 24 September 2015*

## Introduction

Emerging pathogens (i.e. newly identified or evolved pathogens increasing rapidly in incidence and/or expanding their geographical, host or vector ranges) constitute a threat to human health and biodiversity (Daszak *et al.* 2000). So far, an important effort has been devoted to identify the natural and anthropogenic drivers facilitating the emergence of pathogens (Morse 1995; Kilpatrick 2011; Jones *et al.* 2013). Accumulating evidence suggests that the impact of humans on climate, landscapes and biodiversity plays a large role in

Correspondence: Olivier Rey, Fax: +33 5 61 96 08 51; E-mail: olivier.rey@ecoex-moulis.cnrs.fr and Simon Blanchet, Fax: +33 5 61 04 03 60; E-mail:simon.blanchet@ecoex-moulis.cnrs.fr

[1] These authors contributed equally to this work

the emergence of pathogens (Epstein 2001; Altizer *et al.* 2013).

Of particular concern is the intensification of the global human transportation network, which provides new opportunities for pathogens to disperse worldwide, establish and eventually propagate (i.e. 'spill-over') into remote naïve ecosystems (Gozlan *et al.* 2005; Lebarbenchon *et al.* 2008). Alternatively, introduced free-living organisms that have benefited from human transportation may act as competent hosts for native cryptic pathogens (Poulin *et al.* 2011); these introduced hosts may promote and increase pathogen transmission back to local hosts, thus extending the distribution area of native pathogens (i.e. 'spill-back'; Hartigan *et al.* 2011; Poulin *et al.* 2011). Thus, whether the emergence of a newly identified pathogen results either from the introduction of non-native populations (i.e. non-native pathogens) or from the recent expansion of cryptic local populations (i.e. native pathogen) is far from being trivial (Poulin 2014). Nonetheless, tracing out the spatio-temporal dynamics of emerging pathogens is crucial for unravelling ecological and evolutionary factors underlying their emergence and spread (Rachowicz *et al.* 2005).

A critical aspect for reconstructing the spatio-temporal dynamics of emerging pathogens is to have extensive spatial and temporal data on prevalence and occurrences, ideally including data from all putative native areas so as to efficiently (i) test whether or not the pathogen is native and (ii) estimate demographic parameters associated with the colonization/expansion of the pathogen (e.g. Staubach *et al.* 2011; Pullan *et al.* 2012). However, incomplete data sets are a recurrent issue in epidemiology (Woolhouse 2002; de Meeûs *et al.* 2007). This is particularly true for wildlife emerging pathogens, because they are often neglected until they pose socio-economical concerns, and because their actual geographical distributions are often poorly informed (Poulin 2014). Such a lack of precise data greatly hampers our ability to retrace the history of wildlife pathogens back to their emergence.

The recent advent of the approximate Bayesian computation framework (i.e. ABC) in population genetics has greatly improved our ability to retrace the evolutionary history of organisms with limited data availability (Csilléry *et al.* 2010; Lymbery & Thompson 2012). Briefly, ABC applied to population genetics consists in comparing observed empirical data to simulated genetic data sets generated under a range of complex demographic and/or evolutionary scenarios defined a priori and that are likely to explain observed pattern (Beaumont *et al.* 2002). Comparisons between simulated and observed data sets are based on statistics that resume the genetic diversity between and within populations (Beaumont *et al.* 2002; Csilléry *et al.* 2010). This approach allows (i) determining which scenario among a set of likely evolutionary scenarios best explains the observed data (i.e. 'model-choice' procedure) and (ii) inferring key demographic and evolutionary parameters from this most likely scenario (Beaumont *et al.* 2002; Bertorelle *et al.* 2010; Csilléry *et al.* 2010; Estoup & Guillemaud 2010). The ABC framework has been successfully used to identify the geographical origin and to precisely reconstruct the spatio-temporal dynamics of well-monitored pathogens causing socio-economic concerns (e.g. such as the mildew pathogen *Plasmopara viticola* in Europe; Fontaine *et al.* 2013; or *Plasmodium falciparum*, the pathogen that causes Malaria in South America; Yalcindag *et al.* 2012). However, there are still few case studies using the full potential of ABC to unravel the spatio-temporal dynamics of wildlife emerging pathogens (or even non-native free-living species) for which spatial and/or temporal data (from the native or non-native areas) are scarce, a characteristic of most data sets concerning emerging pathogens (e.g. for *Pseudogymnoascus destructans* – the species causing the white-nose disease – when it was first discovered; Warnecke *et al.* 2010).

In this study, we demonstrate the utility of the ABC framework for unravelling the spatio-temporal dynamics of 'information-lacking' emerging pathogens, and notably those for which data on the supposedly native origin are lacking. We specifically aimed at retracing the spatio-temporal dynamics of *Tracheliastes polycolpus* (von Norman 1832), an emerging and virulent ectoparasite of freshwater fishes recently detected in Western Europe (Lootvoet *et al.* 2013; Blanchet *et al.* 2009a; Figs 1 and S1, Supporting information). The case of *T. polycolpus* is intriguing because it is suspected to be non-native in Western Europe (i.e. France, United Kingdom and Spain; Fryer 1982), where it can be highly abundant and virulent, whereas it is less abundant (and hence hard to sample) in its supposedly native area (i.e. Eastern Europe; see Fig. S1, Supporting information). This observation lead to an alternative hypothesis stipulating that the emergence of *T. polycolpus* in Western Europe could result from the expansion of local cryptic populations following biotic (e.g. introduction of a new host species) and/or abiotic (e.g. climatic) changes. Here, the main challenge was to tease apart these two scenarios (i.e. native vs. non-native origin) for a species for which it is difficult to get data from the supposedly native areas. We hence focused exclusively on French watersheds that are supposed to be part of the invaded area (Fig. S1, Supporting information; Fryer 1982). We first determined the genetic diversity and the genetic structure of *T. polycolpus* populations using two clustering approaches. Based on the genetic structure identified and on the historical knowledge available for

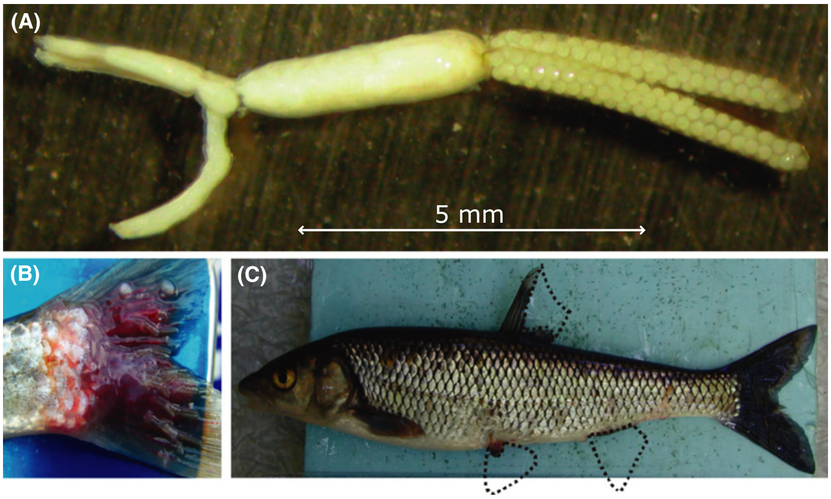

**Fig. 1** Pictures showing (A) a *Tracheliastes polycolpus* parasitic adult female; (B) the inflamed caudal fin of an infected host individual (*Leuciscus burdigalensis*) in response to the attachment of *T. polycolpus*; and (C) a heavily parasitized host individual with the pelvic and anal fins partially or totally destructed and more than 20 *T. polycolpus* attached on the caudal fin. In (C), dotted areas are a rough representation of the fin area destroyed by *T. polycolpus*.

*T. polycolpus*, we formulated an evolutionary model from which we designed six competing scenarios that may explain such a structure, accounting for both the native and non-native hypotheses. Using an original computational pipeline, we identified the most likely scenario using an ABC model-choice procedure. We finally used ABC inference methods based on the most likely scenario to estimate key parameters that characterize the emergence and spatio-temporal colonization history of *T. polycolpus* in French watersheds. Beyond providing fresh insights into the emergence history of *T. polycolpus*, our study illustrates how ABC can be used to (i) determine the status (native or non-native), (ii) reconstruct the spatio-temporal history and (iii) infer key evolutionary parameters of emerging wildlife pathogens with low data availability, and for which samples on the putative native area are scarce and/or inaccessible.

## Material and methods

### Biological model

*Tracheliastes polycolpus* is a freshwater ectoparasite copepod (Crustacea) that displays a direct life cycle (i.e. development involving a single individual host). Only females are parasitic; they attach to the fins of the individual host and feed on its mucus and epithelial cells, hence causing severe infections (Blanchet *et al.* 2009a,b; Loot *et al.* 2004; Fig. 1). Parasitic females of *T. polycolpus* reproduce with free-living males and develop eggs after approximately 3 months (Piasecki 1989). Once matured, the eggs hatch and free-living larvae are released into the water column.

The distribution of *T. polycolpus* is not well documented. *Tracheliastes polycolpus* has been recorded across the most of Eurasia (Fig. S1, Supporting information), including Northern and Eastern Europe (Ponyi & Molnar 1969; Silfverberg 1999), Middle-East (Pazooki & Masoumian 2012) and Northern Asia (Popiolek *et al.* 2011). In Western Europe, *T. polycolpus* was first recorded in the United Kingdom in the 1920s (Aubrook & Fryer 1965) and in France in the 1960s (Tuffery 1967). *Tracheliastes polycolpus* is mainly (but not exclusively) associated with species from the *Leuciscus* complex (Fish: Cyprinidae) including *Leuciscus idus, Leuciscus burdigalensis* and *Leuciscus leuciscus* (Fryer 1982). *Leuciscus idus* is a non-native species in French and English watersheds and has been proposed as the main vector of *T. polycolpus* in Western Europe, where the latter is hence suspected to be a non-native species (i.e. 'the non-native pathogen hypothesis'; Fryer 1982). In French watersheds, *L. idus* is relatively uncommon (Keith *et al.* 2011), whereas *T. polycolpus* is distributed in most watersheds (Fig. S2, Supporting information), where it generally infects *L. leuciscus* and *L. burdigalensis* and, to a lesser extent, other Cyprinid species (*Phoxinus phoxinus*, *Rutilus rutilus*, *Parachondrostoma toxostoma*, *Gobio gobio*, *Barbus barbus* and *Squalius cephalus*; Lootvoet *et al.* 2013). This discrepancy between *L. idus* and *T. polycolpus* distributions in French watersheds suggests an alternative hypothesis whereby *T. polycolpus* is a native species that has recently expanded its geographical range (and/or abundance) following environmental (biotic and/or abiotic) changes (i.e. 'the native pathogen hypothesis').

### Sampling design and microsatellite genotyping

We focus on French watersheds, an area in which *T. polycolpus* has recently emerged (Tuffery 1967) and where it is abundant. A total of 130 sites distributed across all French watersheds were investigated between 2009 and 2011. *Tracheliastes polycolpus* was found in 74 – of the 130 – sites distributed among 53 rivers (Fig. S2, Supporting information), and a total of 663 parasites were collected. Our sampling hence covers the whole known current geographical distribution of *T. polycolpus* in French watersheds (Fig. S2, Supporting information). All parasites were collected exclusively from *L. leuciscus* and *L. burdigalensis*. Fish were captured by electric fishing following standard protocols defined by French environmental agencies (i.e. 'Office National de l'Eau et des Milieux Aquatiques' ONEMA and 'Fédérations Départementales pour la Pêche et la Protection des Milieux Aquatiques' FDAAPPMA; Poulet *et al.* 2011). Parasites were collected from their host individuals using forceps and were directly stored in 70% ethanol for subsequent genetic analyses. All host individuals were returned to their original sampling sites.

Individual DNA was extracted from the cephalothorax of parasites following a salt extraction protocol (Aljanabi & Martinez 1997). Individual multilocus genotypes were obtained at 16 polymorphic microsatellite loci using primers specifically designed for *T. polycolpus* using high-throughput sequencing methods (Loot and Blanchet, unpublished data). Microsatellite loci were amplified using classic polymerase chain reactions (see Appendix S1, Supporting information for more details). Amplified fragments were separated on an ABI PRISM™ 3730 automated capillary sequencer (Applied Biosystems, Foster City, California). Allelic sizes were ultimately scored using GENEMAPPER™ v.4.0. (Applied Biosystems).

### Population structure and genetic diversity

We first determined the uppermost level of genetic structure of *T. polycolpus* populations over French watersheds using the Bayesian clustering approach implemented in STRUCTURE v.2.3.4 (Pritchard *et al.* 2000). This method uses a Markov chain Monte Carlo (MCMC) algorithm to assign each individual into a number of genetic clusters ($K$) so as to maximize Hardy–Weinberg equilibrium and minimize linkage disequilibrium within each cluster. We performed five independent runs for $K$ values ranging from $K = 1$ to $K = 21$. Each run consisted of a burn-in period of $2.5 \times 10^5$ iterations followed by $10^6$ iterations. We assumed correlation of allele frequencies and an admixture model (Hubisz *et al.* 2009). The best model among the 21 models tested was selected on the basis of the second-order rate of change in likelihood ($\Delta K$), according to Evanno *et al.* (2005), using the R package 'CorrSieve' (Campana *et al.* 2011).

Additionally, we conducted the discriminant analysis of principal components (i.e. DAPC) clustering procedure implemented in the R package 'adegenet' (Jombart 2008; Jombart *et al.* 2010). This procedure consists in partitioning individuals into $K$ groups so as to minimize the sum of squares of distances between individuals and the assigned cluster centroids. This complementary approach was used to confirm (or infirm) the number of genetic clusters detected among *T. polycolpus* sampled populations inferred using STRUCTURE.

Finally, for the uppermost hierarchical level of structure identified (i.e. $K = 4$ for both approaches; see Results), we achieved ten additional independent STRUCTURE runs using the same settings as previously to refine the cluster assignation probabilities (Qs) for each individual. The individual Qs were averaged over the ten independent runs using the greedy algorithm implemented in the CLUMPP software (Jakobsson & Rosenberg 2007). We used the individual averaged Qs over the ten runs to generate summary barplots with DISTRUCT v.1.1 (Rosenberg 2004).

Several indices of genetic diversity were estimated for each cluster. Unbiased expected heterozygosities ($H_e$) and fixation indexes ($F_{IS}$) were estimated using GENETIX (Belkhir *et al.* 1996–2004). Allelic richness ($A_r$) and private allele richness ($A_p$) were estimated with the rarefaction procedure implemented in ADZE v1.0 (Szpiech *et al.* 2008) based on the minimal sampling size ($n = 94$; see Results). Pairwise population differentiation values (i.e. pairwise $F_{ST}$) were also estimated between each pair of populations using the GENEPOP software (Rousset 2008).

### Spatio-temporal dynamics of T. polycolpus in French watersheds

We used an ABC approach to unravel the spatio-temporal dynamics of *T. polycolpus* in France. We designed six competing evolutionary scenarios that could potentially explain the current structure of genetic diversity observed for *T. polycolpus* in France, accounting for both the native and the non-native origin hypotheses. We then determined the evolutionary scenario that best explains observed data by applying ABC model-choice procedures (Beaumont *et al.* 2002). Finally, we applied ABC inference procedures based on the most probable evolutionary scenario to estimate key demographic and evolutionary parameters characterizing the emergence of *T. polycolpus* in France. We detail each of these steps below.

*Definition of scenarios.* Scenarios were built from the genetic structure of *T. polycolpus* populations identified using the clustering approaches, that is by considering four main genetic clusters: Central, Northern, Pyrenean and Southern clusters (see Results and Fig. 2A–B). Regarding the 'native pathogen hypothesis', three competing scenarios (namely Sc1 to Sc3) were designed, whereas three additional scenarios (namely Sc4 to Sc6) were designed for the 'non-native pathogen hypothesis'. All scenarios are depicted graphically in Fig. 3, and

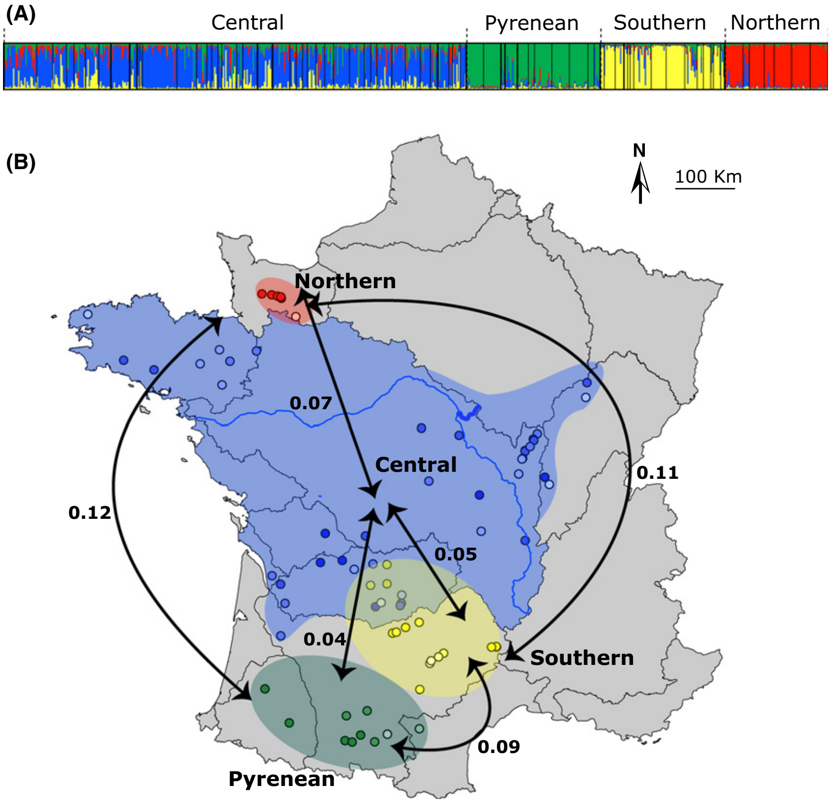

**Fig. 2** Genetic clustering of *Tracheliastes polycolpus* populations in France. (A) Summary barplot indicating the averaged probability for each individual (vertical line) to belong to each of the four clusters identified. (B) Spatial structure of the four *T. polycolpus* clusters determined via the Bayesian clustering approach implemented in the program STRUCTURE. Blue, red, green and yellow dots are sampling sites characterized by individuals belonging to the Central, Northern, Pyrenean and Southern genetic clusters, respectively. The colour intensity of the dots indicates the average probability of individuals from a given site to belong to their assigned genetic clusters. Light to dark colours indicate low to high average probabilities. Delimited regions (black lines) are the main watersheds. Values indicated on arrows are the $F_{ST}$ estimates between clusters.

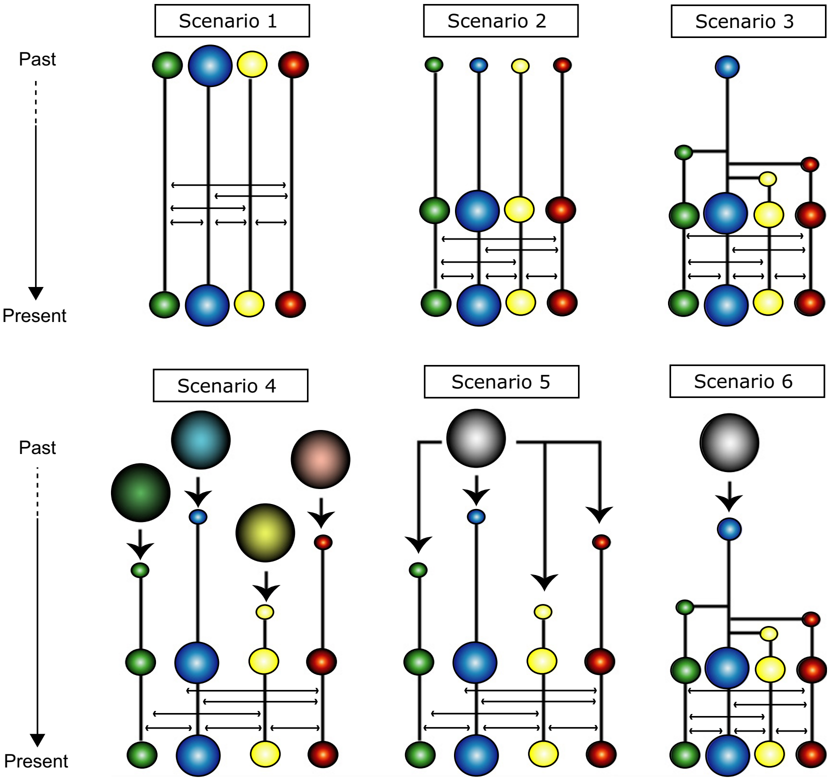

**Fig. 3** Schematic representation of each scenario tested by the approximate Bayesian computation approach. Scenarios are detailed in the main text (Material and methods section). Colours of each population (circles) correspond to the colours of each genetic cluster in Fig. 2 (blue = Central; red = Northern; green = Pyrenean; and yellow = Southern). The sizes of the circles correspond to the effective population sizes. Horizontal arrows represent migration between clusters.

associated parameters (together with their prior values) are described in Table 1. We will first describe parameters that are common to all scenarios and will then fully describe (in a forward-in-time manner, i.e. from past to present) each scenario together with their specific parameters.

For all scenarios, we defined four parameters describing the current effective population size of each cluster, namely $N_{Central}$, $N_{Northern}$, $N_{Pyrenean}$ and $N_{Southern}$. As previous analyses indicated that the Central cluster clearly exhibits a higher effective population size than peripheral clusters (i.e. $N_{Northern}$, $N_{Pyrenean}$ and $N_{Southern}$; data not shown), we constrained $N_{Northern}$, $N_{Pyrenean}$ and $N_{Southern}$ to be smaller than $N_{Central}$ for each scenario. Moreover, the date of first detection of the parasite in Western Europe, that is 1920s in the UK (Aubrook & Fryer 1965) and 1960s in France (Tuffery 1967), was used to calibrate and to bound a parameter called $T_{detect}$ (Table 1). This parameter corresponds to the time (from present) at which clusters experienced a demographic expansion, reaching 'detectable' effective population sizes (i.e. their current population effective sizes $N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$). In all scenarios but Sc1, we further assumed that at $T_{detect}$, clusters started exchanging migrants at a rate $M_{X\text{-}Y}$ (where X and Y indicate the pair of clusters involved in the exchange of migrants, Table 1). For all scenarios, we geographically constrained migration between clusters by assuming that (i) migration rates between non-adjacent clusters are smaller than between adjacent clusters and (ii) migration rates between the Central cluster and the three peripheral clusters are identical ($M_{Central\text{-}Northern} = M_{Central\text{-}Pyrenean} = M_{Central\text{-}Southern}$). Hereafter, we will consider that all scenarios but Sc1 follow the same rule from $T_{detect}$ to present. We now describe in a forward-in-time manner the three scenarios (Sc1 to Sc3) related to the 'native pathogen hypothesis'.

*Scenario 1 (Sc1).* This scenario corresponds to a situation for which *T. polycolpus* is not an emerging parasite, but rather a wildlife parasite that has long been overlooked and detected only recently in France (i.e. in the 1960s). We assumed that the Northern, Central, Pyrenean and Southern clusters have been present since the last glacial event (*c.* 10 000–12 000 years ago) at constant effective population sizes $N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$, respectively, and that clusters exchanged migrants across generations at rates defined by $M_{X\text{-}Y}$ (Sc1 in Fig. 3; Table 1 for details on parameters). In this scenario, we thus did not account for the parameter $T_{detect}$.

*Scenario 2 (Sc2).* In this scenario, the four clusters were historically present, but at low (i.e. undetectable) effective population sizes ($N'_{Northern}$, $N'_{Central}$, $N'_{Pyrenean}$, $N'_{Southern}$) until $T_{detect}$, the time at which (i) effective population sizes reach a detectable size (i.e. their current sizes $N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$) and (ii) migration between clusters begins (Sc2 in Fig. 3; Table 1).

*Scenario 3 (Sc3).* Here, we considered a unique original cluster (i.e. the Central cluster, where *T. polycolpus* was first described in France; Tuffery 1967) of low effective population size $N'_{Central}$ (Sc3 in Fig. 3). At times $T_{Northern}$, $T_{Pyrenean}$ and $T_{Southern}$, individuals from the Central cluster spread and founded the Northern, Pyrenean and Southern clusters, respectively. $T_{Northern}$, $T_{Pyrenean}$ and $T_{Southern}$ were allowed to occur asynchronously. Clusters remained at low effective population sizes ($N'_{Northern}$, $N'_{Central}$, $N'_{Pyrenean}$, $N'_{Southern}$) without exchanging migrants until $T_{detect}$, the time at which (i) effective population sizes reach their current sizes ($N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$) and (ii) migration between clusters begins (Fig. 3; Table 1).

We now describe the three scenarios (Sc4 to Sc6) related to the 'non-native pathogen hypothesis'. In these scenarios, we assumed a latency time of $L_X$ generations (where X indicates the name of the cluster) that follows each introduction event from an original unknown source (for which we assume Ne = 20 000). During $L_X$, the size of the initial founding population remains equal to the number of introduced founder individuals $N''_X$. After $L_X$, populations experienced a first demographic expansion from $N''_X$ to $N'_X$. Latency time between introduction and first expansion (also called time lag) is expected in introduced populations, in particular when the introduced individuals are not (yet) well adapted to the invaded environment or because of purely demographic processes (Facon *et al.* 2006).

*Scenario 4 (Sc4).* Here, we considered that each cluster originated from independent unknown source populations (Sc4 in Fig. 3). At times $T'_{Central}$, $T'_{Northern}$, $T'_{Pyrenean}$ and $T'_{Southern}$, $N''_{Central}$, $N''_{Northern}$, $N''_{Pyrenean}$ and $N''_{Southern}$ individuals originating from independent source populations colonized and founded each cluster. After $L_{Central}$, $L_{Northern}$, $L_{Pyrenean}$ and $L_{Southern}$ generations (i.e. at $T_{Central}$, $T_{Northern}$, $T_{Pyrenean}$ and $T_{Southern}$, respectively), clusters expanded and remained at $N'_{Central}$, $N'_{Northern}$, $N'_{Pyrenean}$ and $N'_{Southern}$ without exchanging migrants until $T_{detect}$, the time at which (i) effective population sizes reach their current sizes ($N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$) and (ii) migration between clusters begins (Fig. 3; Table 1).

*Scenario 5 (Sc5).* Here we considered that all clusters originated from a single unknown source population

(Sc5 in Fig. 3). Founder individuals were introduced asynchronously and founded the different clusters at times $T'_{Central}$, $T'_{Northern}$, $T'_{Pyrenean}$ and $T'_{Southern}$ for the Central, Northern, Pyrenean and Southern clusters, respectively. After $L_{Central}$, $L_{Northern}$, $L_{Pyrenean}$ and $L_{Southern}$ generations (i.e. at $T_{Central}$, $T_{Northern}$, $T_{Pyrenean}$ and $T_{Southern}$, respectively), clusters expanded and remained at $N'_{Central}$, $N'_{Northern}$, $N'_{Pyrenean}$ and $N'_{Southern}$ without exchanging migrants until $T_{detect}$, the time at which (i) effective population sizes reach their current sizes ($N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$) and (ii) migration between clusters begins (Fig. 3; Table 1).

*Scenario 6 (Sc6).* This scenario assumes that a first introduction event from a single unknown source into a single locality occurred at $T'_{Central}$ and was composed of N $''_{Central}$ individuals (Sc6 in Fig. 3). After $L_{Central}$ generations (i.e. at $T_{Central}$), this pool of founder individuals experienced a first demographic expansion and reached $N'_{Central}$ individuals. At times $T_{Northern}$, $T_{Pyrenean}$ and $T_{Southern}$, individuals from the Central cluster spread and founded the Northern, Pyrenean and Southern clusters, respectively. Clusters remained at low densities ($N'_{Central}$, $N'_{Northern}$, $N'_{Pyrenean}$, $N'_{Southern}$) without exchanging migrants until $T_{detect}$, the time at which (i) effective population sizes reach their current sizes ($N_{Northern}$, $N_{Central}$, $N_{Pyrenean}$ and $N_{Southern}$) and (ii) migration between clusters begins (Fig. 3; Table 1).

*Simulation procedure.* We simulated genetic data sets given each scenario by following four major steps: (i) sampling of a vector of parameter values ($\varphi_x$) for a given scenario from prior parameter distributions (defined in Table 1), (ii) simulation of a genetic data set given $\varphi_x$, (iii) calculation of a vector of statistics $s_x$ that summarizes the simulated data set and (iv) repeat steps (i) to (iii) a large number of times. We implemented this procedure in a handmade computational pipeline that combines multiple population genetics and statistical programs (see Appendix S2, Supporting information for more details). This pipeline was developed to allow us (i) calculating a wide set of informative summary statistics that cannot be obtained from only a single population genetics program and (ii) using a genetic data simulator that allows accounting for gene flow between populations, a requisite that is not implemented in some user-friendly ABC softwares (Cornuet *et al.* 2008) and that was necessary in our study, given the small spatio-temporal extent we investigated. The program ABCsampler (Wegmann *et al.* 2010) was used to manage the pipeline (see Appendix S2, Supporting information for more details), and the coalescent-based program SIMCOAL v2.0 (Laval & Excoffier 2004) was used to simulate 1 × $10^6$ microsatellite data sets under each demographic scenario. Sixteen independent microsatellite loci *per* individual were generated *per* simulation, assuming a stepwise mutation model (SMM; Ohta & Kimura 1973) and a neutral mutation rate of $5 \times 10^{-4}$ over loci (Estoup & Angers 1998). We assumed the SMM rather than other mutation models to be conservative, as the most variable microsatellite loci used in this study exhibit allelic distributions expected under an SMM (i.e. uniform and symmetric distributions; data not shown). We sampled a number of diploid individuals *per* cluster *per* simulated data set identical to the number of individuals in the empirical genetic data set (i.e. 321, 96, 132 and 114 for the Central, Northern, Pyrenean and Southern clusters, respectively). From each simulated data set, we computed a series of genetic variation indexes. We used ADZE (Szpiech *et al.* 2008) for estimating mean allelic richness ($A_r$) and mean private allelic richness ($A_p$) over loci *per* cluster. The program arlsumstat (Excoffier & Lischer 2010) was used for estimating, at the cluster level and over loci, expected heterozygosities ($H_e$), their standard distributions $sd(H_e)$, mean allelic ranges ($R$), Garza–Williamson's indexes (GW; Garza & Williamson 2001), along with global $F_{IS}$ and $F_{ST}$ values, and among cluster pairwise $F_{ST}$ values.

*Model choice.* Before applying ABC model-choice procedures, we assessed whether the six demographic scenarios were able to provide a good fit to the observed data by running a principal component analysis (PCA) in the space of 16 summary statistics calculated at different levels: (i) 8 summary statistics calculated within clusters (i.e. $A_r$ and GW for each cluster), (ii) 2 global summary statistics calculated at the whole scenario level (i.e. global $F_{IS}$ and $F_{ST}$) and (iii) 6 summary statistics calculated among clusters (i.e. pairwise $F_{ST}$ values). The use of summary statistics calculated at different levels is recommended in ABC, as it allows to better capture the properties of the modelled demographic scenarios (Cornuet *et al.* 2010). The PCA approach was used to graphically show the position of the observed (real) data set vs. the 10 000 simulations the closest to the observed data set (assuming a Euclidean distance) under each demographic scenario (Cornuet *et al.* 2010; Marino *et al.* 2013).

We then applied ABC model-choice procedures to determine the scenario that best explains the observed data (Beaumont *et al.* 2002) using the subset of 16 summary statistics described above. The posterior support for each scenario given the observed data was assessed using a neural network algorithm with a tolerance rate of 0.0001 (Csilléry *et al.* 2012). The neural network algorithm was chosen in this step to reduce the large

**Table 1** Prior parameter values explored in simulations of genetic data sets under all scenarios tested in the approximate Bayesian computation procedure. Prior values were drawn from uniform distributions for all parameters

| Parameter | Description | Prior values | Scenarios | Unit |
|---|---|---|---|---|
| $N_{Central}$ | Current effective size of the Central cluster | 10–300 | All scenarios | 2n individuals |
| $N_{Southern}$ | Current effective size of the Southern cluster | 25–150 | All scenarios | 2n individuals |
| $N_{Pyrenean}$ | Current effective size of the Pyrenean cluster | 25–150 | All scenarios | 2n individuals |
| $N_{Northern}$ | Current effective size of the Northern cluster | 10–40 | All scenarios | 2n individuals |
| $T_{detect}$ | Time since the last expansion of the different genetic clusters | 16–300 | All but scenario 1 | Generations |
| $T_{Southern}$ | Split time of the peripheral Southern cluster | 161–350 | Scenarios 3–6 | Generations |
| $T_{Pyrenean}$ | Split time of the peripheral Pyrenean cluster | 161–350 | Scenarios 3–6 | Generations |
| $T_{Northern}$ | Split time of the peripheral Northern cluster | 161–350 | Scenarios 3–6 | Generations |
| $T_{Central}$ | Time at which the initial invasive population established from the founder individuals first expanded | 250–400 | Scenarios 6 | Generations |
| $L_{Southern}$ | Latency time between the introduction of $N''_{Southern}$ individuals and first expansion | 1–50 | Scenarios 4–5 | Generations |
| $L_{Pyrenean}$ | Latency time between the introduction of $N''_{Pyrenean}$ individuals and first expansion | 1–50 | Scenarios 4–5 | Generations |
| $L_{Northern}$ | Latency time between the introduction of $N''_{Northern}$ individuals and first expansion | 1–50 | Scenarios 4–5 | Generations |
| $L_{Central}$ | Latency time between the introduction of $N''_{Central}$ individuals and first expansion | 1–50 | Scenarios 4–6 | Generations |
| $T'_{Pyrenean}$ | Time since the introduction of founder individuals at the origin of the Pyrenean cluster | $(T_{Pyrenean} + L_{Pyrenean})$ | Scenarios 4–5 | Generations |
| $T'_{Southern}$ | Time since the introduction of founder individuals at the origin of the Southern cluster | $(T_{Southern} + L_{Southern})$ | Scenarios 4–5 | Generations |
| $T'_{Northern}$ | Time since the introduction of founder individuals at the origin of the Northern cluster | $(T_{Northern} + L_{Northern})$ | Scenarios 4–5 | Generations |
| $T'_{Central}$ | Time since the introduction of founder individuals at the origin of the Central cluster | $(T_{Central} + L_{Central})$ | Scenarios 4–6 | Generations |
| $N'_{Pyrenean}$ | Size of the ancestral Pyrenean cluster relative to $N_{Pyrenean}$ before $T_{detect}$ | 0.1–0.4 | All but scenario 1 | Relative ratio |
| $N'_{Northern}$ | Size of the ancestral Northern cluster relative to $N_{Northern}$ before $T_{detect}$ | 0.1–0.4 | All but scenario 1 | Relative ratio |
| $N'_{Central}$ | Size of the ancestral Central cluster relative to $N_{Central}$ before $T_{detect}$ | 0.1–0.4 | All but scenario 1 | Relative ratio |
| $N'_{Southern}$ | Size of the ancestral Southern cluster relative to $N_{Southern}$ before $T_{detect}$ | 0.1–0.4 | All but scenario 1 | Relative ratio |
| $N''_{Northern}$ | Number of founder individuals of the Northern cluster from an unknown source population | 0.2–0.4 | Scenarios 4–5 | Relative ratio |
| $N''_{Central}$ | Number of founder individuals of the Central cluster from an unknown source population | 0.2–0.4 | Scenarios 4–6 | Relative ratio |
| $N''_{Pyrenean}$ | Number of founder individuals of the Pyrenean cluster from an unknown source population | 0.2–0.4 | Scenarios 4–5 | Relative ratio |
| $N''_{Southern}$ | Number of founder individuals of the Southern cluster from an unknown source population | 0.2–0.4 | Scenarios 4–5 | Relative ratio |
| $M_{Central\text{-}Southern}$ | Migration between the Central and the Southern cluster | 0.001–0.04 | All scenarios | Rate |
| $M_{Central\text{-}Pyrenean}$ | Migration between the Central and the Pyrenean cluster | 0.001–0.04 | All scenarios | Rate |
| $M_{Central\text{-}Northern}$ | Migration between the Central and the Northern cluster | 0.001–0.04 | All scenarios | Rate |
| $M_{Northern\text{-}Southern}$ | Migration between the Northern and the Southern clusters | 0.001–0.04 | All scenarios | Rate |
| $M_{Northern\text{-}Pyrenean}$ | Migration between the Northern and the Pyrenean clusters | 0.001–0.04 | All scenarios | Rate |
| $M_{Southern\text{-}Pyrenean}$ | Migration between the Southern and the Pyrenean clusters | 0.001–0.04 | All scenarios | Rate |

number of selected summary statistics into a smaller number of dimensions (Blum & François 2010). Additionally, we conducted a Bayesian model comparison by computing Bayes factors (BFs) between each pair of competing scenarios.

Finally, the robustness of the model-choice procedure was evaluated by estimating type I (i.e. how frequently a simulation generated under scenario X has been excluded when it has been actually generated under scenario X) and type II (i.e. how frequently a simulation has been assigned to a scenario X when it has not been generated under scenario X) error rates. To do so, we used the leave-one-out cross-validation procedure implemented in the R package 'abc' v.2.0 (Csilléry *et al.* 2012). In this procedure, a given number of simulations are randomly selected as 'validation' simulations, as opposed to 'training' simulations (i.e. the unselected simulations). Each validation simulation is therefore successively submitted to an ABC model-choice algorithm. We used 100 validation simulations *per* scenario, assuming a neural network algorithm, and a tolerance rate of 0.0001.

*Estimation of demographic parameters.* Based on the best-supported scenario, we first assessed the accuracy of ABC and the sensitivity of parameter estimates to the tolerance rate for three alternative estimation methods (i.e. rejection, local linear regression and neural network). This was achieved using the leave-one-out cross-validation procedure implemented in the R package 'abc' v2.0 (Csilléry *et al.* 2012). The procedure was based on 100 cross-validation simulations generated under the best-supported scenario for each estimation method and assuming four different tolerance rates (i.e. 0.001, 0.005, 0.01 and 0.05). For each couple of estimation method/tolerance rate, we then calculated the sum of all standardized estimation errors obtained for each parameter of the best-supported scenario.

Demographic parameters for the best-supported scenario were estimated by considering the couple of estimation method/tolerance rate that showed the smallest sum of standardized error rates during the cross-validation procedure (i.e. neural network method and a tolerance rate of 0.001; see Results section), and using the same set of 16 summary statistics used for the model-choice procedure (i.e. $A_r$, GW, global $F_{IS}$, global $F_{ST}$ and pairwise $F_{ST}$ values). The spatio-temporal dynamics of *T. polycolpus* was then reconstructed by using the mode of the posterior parameter distributions obtained by ABC for each parameter as point estimates, and 2.5–97.5% percentiles as confidence intervals. Times in years were calculated from times inferred by ABC in generations, by assuming that four generations correspond to a year (Piasecki 1989).

We further checked the goodness of fit of the best-supported scenario to the observed data using two alternative methods: (i) a principal component analysis (PCA) in the space of summary statistics and (ii) a formal hypothesis testing procedure where the test statistic is the mean of the distance between simulations' summary statistics and observed summary statistics. These two methods were based on 1000 posterior simulations generated under the best-supported scenario by sampling parameter values from their posterior distributions, and on an alternative set of 16 summary statistics (i.e. $H_e$, $sd(H_e)$, $A_r$ and $A_p$ *per* cluster). Using an alternative set of summary statistics for posterior checks allows avoiding overfitting (Cornuet *et al.* 2010; Marino *et al.* 2013). We finally conducted posterior predictive checks (Gelman *et al.* 2003) to graphically confirm that the best-supported scenario provides a good fit to the *T. polycolpus* real data, by plotting the distribution of the 16 alternative summary statistics computed for the 1000 posterior simulations described above, along with the observed values.

## Results

### Population structure and genetic diversity

Results from the Bayesian clustering approach and the $\Delta K$ test revealed that the uppermost hierarchical level of structure of *T. polycolpus* populations was $K = 4$ (Figs S3 and S4, Supporting information). Individuals within watersheds were generally assigned with high probability (i.e. high $Q$ value; Fig. 2A) to a single genetic cluster. Four similar main genetic clusters were detected using DAPC, although the latter analysis also revealed finer genetic substructure within each main cluster (Fig. S5, Supporting information). The two clustering methods converged towards the identification of a major cluster covering a large central area (hereafter 'Central' cluster), and three peripheral clusters that are more restricted geographically: a cluster was found in Northern France (hereafter 'Northern' cluster), another in South-Western France (hereafter 'Southern' cluster), and the third cluster covered the whole Pyrenean Mountains (hereafter 'Pyrenean' cluster) (Fig. 2B). Genetic clusters covered at least two adjacent watersheds, except the Northern cluster, which was restricted to a single watershed (Fig. 2B).

Estimates of genetic differentiation between pairs of clusters (i.e. $F_{ST}$) were all significant ($P$-values < 0.01) ranging from 0.04 (between the Central and the Pyrenean clusters) to 0.12 (between the Northern and the Pyrenean clusters). Nonadjacent clusters were more genetically differentiated than adjacent clusters ($t$-test, $t = -4.3$, d.f. = 3, $P$-value = 0.01; Fig. 2B). The Central cluster displayed the highest genetic diversity compared to peripheral clusters, although this was much more evident when considering allelic richness rather than heterozygosity (Table 2). Similarly, the Central cluster contained more private alleles ($A_p$ = 0.40) than the other clusters (Table 2).

### *Spatio-temporal dynamics of T. polycolpus in French watersheds*

*Best-supported scenario.* The cross-validation procedure for ABC model-choice displayed an overall misclassification rate of 36% (Table S1, Fig. S6, Supporting information). Type I errors were moderate to high for all scenarios (with rates ranging from 24% for Sc2 to 56% for Sc1) except for Sc4, for which the type I error rate was very low (4%). Two pairs of scenarios displayed high rates of cross-misclassifications: Sc1/Sc3 (32–56%) and Sc2/Sc5 (20–46%). The scenarios related to the 'native pathogen hypothesis' (i.e. Sc1–3) displayed the highest rates of type II errors (9–17.6%), whereas the lowest rates of type II errors were found for the 'non-native' scenarios Sc4 and Sc6 (0% and 1.6%, respectively; Table S1, Fig. S6, Supporting information).

The adequacy of the different scenarios to generate a set of summary statistics similar to that of the empirical data set was confirmed visually by the PCA for all scenarios but the Sc1 and Sc3 (Fig. S7, Supporting information). For these two scenarios, the 99% envelopes of the two principal components of the PCA did not comprehend the observed data set, suggesting that these two models are unlikely to have a good fit with the real data.

Among the six scenarios, Sc6 was the best supported according to the ABC model-choice procedure (posterior probability = 0.7625; Table S2, Supporting information). This result was confirmed by the very high pairwise BFs obtained between Sc6 and all others scenarios, compared to BFs computed between all other pairs of scenarios (Table S3, Supporting information). The five other scenarios were clearly rejected compared to Sc6 (i.e. posterior probabilities <0.1239; Table S2, Supporting information). Under scenario Sc6, *T. polycolpus* populations in French watersheds originate from a single genetic entity from which the Central cluster emerged. The three peripheral clusters emerged from the establishment of few founder individuals originating from the Central cluster into remote watersheds.

**Table 2** Estimates of genetic diversity within each of the four genetic clusters identified in the STRUCTURE analysis

| Cluster name | $N_{ind}$ | $A_r$ | SE ($A_r$) | $A_p$ | SE ($A_p$) | $H_e$ | SE ($H_e$) | $F_{IS}$ |
|---|---|---|---|---|---|---|---|---|
| Northern | 96 | 3.73 | 0.36 | 0.06 | 0.03 | 0.52 | 0.15 | 0.16 |
| Central | 321 | 5.03 | 0.45 | 0.40 | 0.11 | 0.57 | 0.18 | 0.12 |
| Pyrenean | 132 | 4.40 | 0.42 | 0.23 | 0.08 | 0.56 | 0.14 | 0.16 |
| Southern | 114 | 4.40 | 0.39 | 0.15 | 0.05 | 0.54 | 0.21 | 0.14 |

$N_{ind}$, number of *T. polycolpus* individuals assigned to each cluster; $A_r$, allelic richness; $A_p$, private allele richness; $H_e$, unbiased expected heterozygosity; SE, standard error.

*Estimation of demographic parameters.* Our cross-validation procedure indicated that the neural network method and a tolerance rate of 0.001 was the most adequate combination to minimize errors in the estimates of parameters for Sc6 (Table S4, Supporting information). Using these settings, we found low standardized estimation errors for the current effective size of all but the Northern cluster, and low to moderate standardized estimation errors for time since the last expansion of the different genetic clusters and split time of the peripheral clusters. The highest standardized estimation errors concerned the migration rates between clusters (Table S4, Supporting information). This suggests that the estimates of these parameters may be less accurate than the others and should be carefully interpreted.

Both the PCA and the hypothesis testing procedure based on 1000 posterior simulations and the alternative set of summary statistics confirmed the goodness of fit of Sc6 to the observed data set (Figs S8 and S9, Supporting information). The adequacy of Sc6 for generating simulated data sets akin to the real data was further graphically assessed through the comparison between the distribution of the alternative summary statistics computed for the posterior simulations and the observed values (Fig. S10, Supporting information).

The inference of key parameter values from Sc6 with ABC (Fig. 4; Table S5, Supporting information) revealed that 38 founder individuals (i.e. $N''_{Central}$; 2.5–97.5% percentile: [2–44]) were most likely introduced in France at $T'_{Central}$ = 368 [271–434] generations ago, which is about 92 years before sampling (i.e. in 1918), considering a generation time for *T. polycolpus* of four generations *per* year (see Material and methods). After a latency time of about 6 years (at $T_{Central}$ = 343 [260–407] generations), this set of founder individuals established an initial effective population $N'_{Central}$ of 100 individuals [19–111]. Three subsequent colonization events dispersed from this initial population to peripheral watersheds 69 (Northern), 74 (Southern) and 75 (Pyrenean) years before the sampling (i.e. at $T_{Northern}$ = 282 [199–339], $T_{Southern}$ = 301 [202–344] and $T_{Pyrenean}$ = 303 [201–339] generations, respectively; Fig. 4A). These colonization events involved 13 ($N'_{Northern}$; [2–19]), 29 ($N'_{Southern}$; [6–60]) and 31 ($N'_{Pyrenean}$; [5–59]) individuals for the establishment of the Northern, Southern and Pyrenean

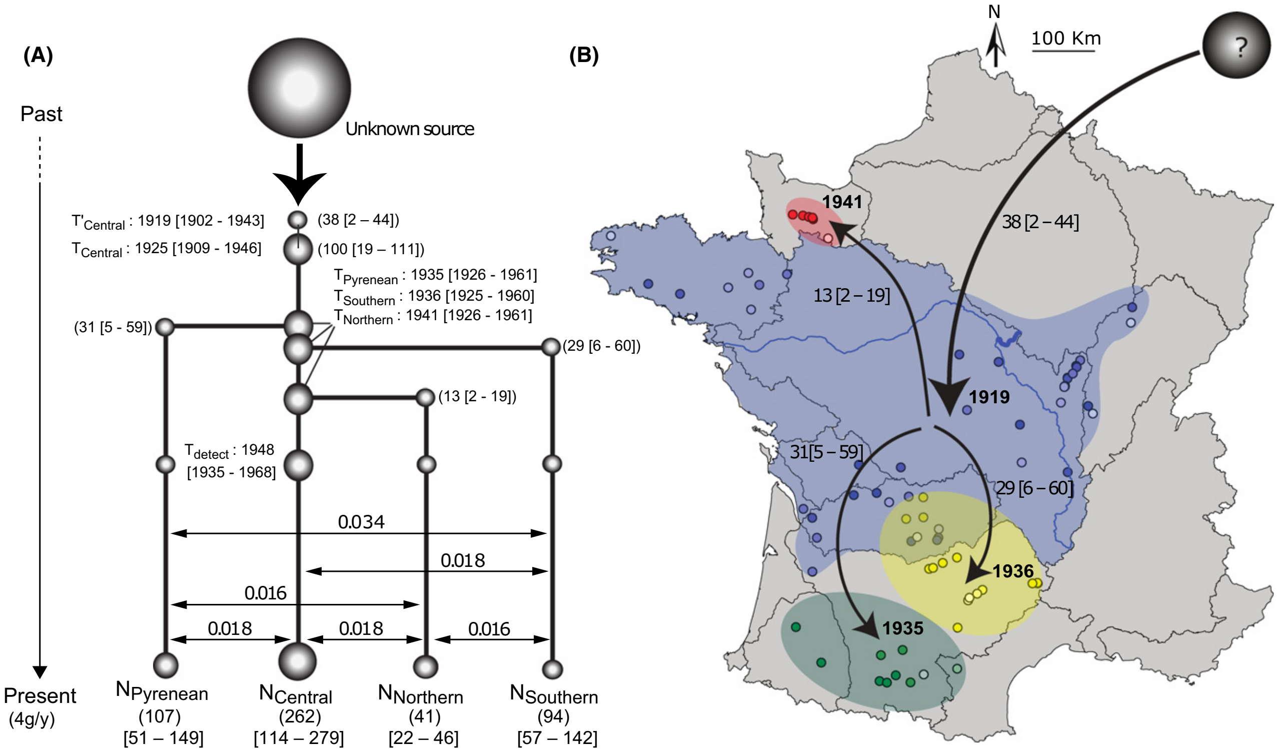
(A)
Past
Unknown source
T'Central : 1919 [1902 - 1943]
(38 [2 – 44])
TCentral : 1925 [1909 - 1946]
(100 [19 – 111])
TPyrenean : 1935 [1926 - 1961]
TSouthern : 1936 [1925 - 1960]
TNorthern : 1941 [1926 - 1961]
(31 [5 - 59])
(29 [6 - 60])
(13 [2 - 19])
Tdetect : 1948 [1935 - 1968]
0.034
0.018
0.016
0.018
0.018
0.016
Present (4g/y)
NPyrenean (107) [51 – 149]
NCentral (262) [114 – 279]
NNorthern (41) [22 – 46]
NSouthern (94) [57 – 142]
(B)
N
100 Km
?
1941
38 [2 – 44]
13 [2 – 19]
1919
31 [5 – 59]
29 [6 – 60]
1936
1935

**Fig. 4** Spatio-temporal dynamics of *Tracheliastes polycolpus* in France, based on the best-supported scenario among the six scenarios tested using approximate Bayesian computation. (A) Schematic representation of the best-supported scenario (i.e. Scenario 6). The numbers in parentheses are the effective population sizes of populations or the number of founding individuals at the time of their introduction. The chronological axis on the left is determined assuming four generations of *T. polycolpus per* year. (B) Map illustrating the propagation of *T. polycolpus* in French watersheds since its introduction, according to the best-supported scenario. In A and B, the 95% confidence interval of each estimate is in square brackets.

clusters, respectively. The four clusters were maintained at low effective sizes until 1948 ($T_{detect}$ = 252 [172–303] generations), when they attained their current effective sizes ($N_{Central}$ = 262 [114–279], $N_{Northern}$ = 41 [22–46], $N_{Southern}$ = 94 [57–142] and $N_{Pyrenean}$ = 107 [51–149] individuals; Fig. 4A; Table S5, Supporting information). Current migration rates between clusters varied significantly, from 0.016 [0.013–0.028] between the Northern/Southern and Northern/Pyrenean clusters, to 0.034 [0.025–0.049] between the Pyrenean and the Southern clusters (Fig. 4A; Table S5, Supporting information).

## Discussion

In this study, we combined population genetics tools and ABC model-based inference to unravel the spatio-temporal dynamics of a poorly documented emerging wildlife pathogen with scarcely available data. Our results provide novel information regarding the non-native status and the colonization history of *T. polycolpus* since its emergence in French watersheds. More generally, this case study illustrates the value of ABC for inferring key evolutionary and demographic parameters in 'information-lacking' wildlife emerging pathogens.

### *Tracheliastes polycolpus in France: a non-native pathogen?*

The main challenge of this study was to tease apart the native vs. non-native origin of *T. polycolpus* in French watersheds without genetic data sampled over the whole distribution of this species. Altogether, our results strongly support that – according to previous expectations (Fryer 1982) – *T. polycolpus* is a non-native species in France. First, among the six scenarios tested, the two statistical approaches used to select the best scenario (posterior probabilities and BFs) clearly favoured one of the non-native scenarios (Sc6) as the most likely to explain observed genetic data. Moreover, all validation tests (Bertorelle *et al.* 2010) confirm our model-choice procedure. Overall, both type I and type II error rates inversely related to the confidence in the choice of a scenario were higher in the 'native scenarios' than in the 'non-native scenarios' (although differences were less pronounced for type I errors). We further tested the goodness of fit between each of the six scenarios and observed data prior to running the model-choice procedure. The three scenarios related to the 'non-native pathogen hypothesis' generated data similar to that observed, whereas two of the three scenarios related to the 'native pathogen hypothesis' failed to do so, hence discrediting the 'native hypothesis' at the expense of the 'non-native hypothesis'. It is, however, noteworthy that the only 'native' scenario that generated data similar to the observed (i.e. Sc2) was also the second best-supported model (Table S2, Supporting information). Further, both the second and the third best-supported scenarios (Sc2 and Sc4) showed non-negligible support (12.39% and 10.13% for Sc2 and Sc4, respectively; Table S2, Supporting information), indicating that these two scenarios were able to generate data akin to that generated under the Sc6 scenario for at least some specific combination of parameters (Table S5, Supporting information). For instance, inferences of parameters under scenarios Sc2 and Sc4 provided similar small ancestral populations effective sizes ($N'_X$ parameters, see Table S5, Supporting information) than those observed for Sc6, a characteristic that may similarly shape genetic diversity generated under the three scenarios (i.e. by imposing strong genetic bottlenecks due to founder effects, see below). Thus, even if the model-choice procedure clearly highlights Sc6 as the best-supported scenario, we cannot completely reject Sc2 and Sc4.

Second, according to the best-supported scenario (Sc6; Fig. 4), the Central population of *T. polycolpus*, initially composed of a few individuals having immigrated from an unknown source population, underwent a strong bottleneck before establishing a stable, initial population. Genetic bottlenecks due to founder effects are a classical demographic imprint of introduced populations (Sakai *et al.* 2001; Blanchet 2012), and the existence of a temporal gap between the first event of introduction and the first detection of non-native organisms is also common (Watari *et al.* 2011). This is especially the case for small-bodied and neglected nonmodel organisms (Litchman 2010). Moreover, we estimated that the first genetic bottleneck event underwent by the *T. polycolpus* Central population occurred in the 1920s, which coincides with the initial introduction of *Leuciscus idus* in France as an ornamental fish in the early 1900s (Keith *et al.* 2011). *Leuciscus idus* is the principal host of *T. polycolpus* in the United Kingdom and Eastern Europe (Fryer 1982), and established wild populations of *L. idus* in France have been almost exclusively reported in the central Loire River watershed (Fig. S2, Supporting information; Keith *et al.* 2011). Altogether, these evidences strongly suggest that *T. polycolpus* was introduced together with *L. idus* in Western Europe before shifting on native host species, and expanding its demographic and geographical range.

Despite all the accumulated evidence that *T. polycolpus* is non-native, the hypothetical native origin of *T. polycolpus* in French watersheds will not be definitely ruled out unless further analyses including samples from the UK, and, ideally, from the supposedly native area (i.e. Eastern Europe) are performed. Unravelling the phylogenetic relationships between populations across most of the geographical distribution of *T. polycolpus* would hence provide new insights into the evolutionary history of this species and shed light on the origin of the populations established in Western Europe. More specifically, the ABC framework would be particularly useful to revaluate the likelihood of non-native vs. native scenarios, by including complex scenarios that imply known potential sources, such as, for instance, the hypothetical two-step scenario in which *T. polycolpus* was first introduced in the United Kingdom before being introduced in French watersheds (Fryer 1982).

### *Invasion dynamics of T. polycolpus in French watersheds*

We were able to retrace the invasion dynamics of *T. polycolpus* among French watersheds since its probable introduction in Central watersheds of France in the 1920s. In particular, we found that the range expansion of *T. polycolpus* through almost all French riverine systems has involved three main dispersal events of few founder individuals (i.e. from 13 to 31 individuals). These three dispersal events have occurred almost synchronously in the late 1930s from the initial Central watersheds population to three distinct remote watersheds. Interestingly, the mean parasite intensity in individual hosts can reach up to 80 individuals on daces (*L. burdigalensis*) and two to five individuals on gudgeons (*G. gobio*) and minnows (*P. phoxinus;* Lootvoet et al. 2013). Thus, we can hypothesize that the stocking of a few infected host individuals (i.e. 1–15 host individuals; depending on the host species stocked) would have been sufficient for *T. polycolpus* to disperse, establish and expand over the main French watersheds. Moreover, our results also indicate that once established, these populations were maintained at a small effective population size before expanding demographically around 1948, that is two decades before *T. polycolpus* was first identified in France (Tuffery 1967). Finally, ongoing gene flow between the Central and peripheral clusters and, to a lower extent, between peripheral clusters was also detected, which indicates that *T. polycolpus* colonization in French watersheds is still an ongoing process. It is noteworthy that large credibility intervals were found for most parameter estimates. Wide credibility intervals for parameter estimates can sometimes be obtained from ABC algorithms mainly because of the loss of information during the ABC process (Csilléry *et al.* 2010; Benazzo *et al.* 2015). In our case, the low level of genetic diversity observed in *T. polycolpus* over all genetic clusters, which is most likely due to its recent evolutionary history in Western Europe, might also have hampered our ability to obtain more accurate estimates of the demographic parameters. Thus, although the parameter estimates obtained are consistent with the limited historical data available for *T. polycolpus*, they should be interpreted carefully.

The apparent fast expansion of *T. polycolpus* since its emergence in the Central watersheds raises important questions regarding the factors that allow this parasite to propagate so rapidly. The direct life cycle of *T. polycolpus* and its low specificity (i.e. the propensity to use a large host spectrum) constitute a real asset for its spread over French watersheds (Criscione *et al.* 2005). Indeed, *T. polycolpus* can infect at least eight common host species that display different dispersal capacities (Lootvoet *et al.* 2013). According to the consensual view that host movement is a major determinant in parasite dispersion (Criscione *et al.* 2005; Blasco-Costa *et al.* 2012), the broad host species range of *T. polycolpus* may allow this parasite to cover long-distance dispersal within watersheds and increase its probability to establish locally as soon as one of these host species is present. Moreover, some alternative host species used by *T. polycolpus* are affiliated to human activities such as aquaculture or angling (e.g. *Gobio gobio*, *Rutilus rutilus* and *Phoxinus phoxinus*) and are hence prone to be stocked from one watershed to another (Lewin *et al.* 2006). This may allow *T. polycolpus* to disperse rapidly within and among watersheds and may explain the non-negligible among-watershed migration rates inferred in this study.

## Conclusion

Understanding the anthropogenic, ecological and evolutionary factors that favour the emergence of wildlife pathogens is a major current challenge (Morens *et al.* 2004). A key step in breaching this challenge is to develop and propose analytical frameworks that help clarifying the spatio-temporal dynamics of pathogens over short-time scales, especially during emergence events. ABC is one of such frameworks. The ABC framework has been mainly used to retrace the eco-evolutionary dynamics of populations at large spatial scales (e.g. populations from different continents; Guillemaud *et al.* 2009) but rarely at low temporal and spatial extents. We show here that by using a large set of informative summary statistics and by explicitly accounting for gene flow among populations, an ABC framework can be used to distinguish between closely similar scenarios at low spatial and temporal resolutions. Moreover, despite the lack of precise data on the expansion and emergence of *T. polycolpus* and the scarcity of samples from the putative native geographical range (i.e. Eastern Europe), we additionally show that ABC can successfully infer key parameters related to the recent colonization of information-lacking organisms. Specifically, our results strongly support the non-native origin of *T. polycolpus* in France and have provided information on the size of the founder populations and the timing of colonization among watersheds. Our results constitute a foundation for future researchers focusing on the evolutionary potential of *T. polycolpus*, and more particularly on the evolution of *T. polycolpus* virulence on naïve host populations or the rise of local adaptation patterns in novel host–pathogen interactions (Dunn 2009).

More generally, we believe that *T. polycolpus* is far from being an isolated case study, since historical records and scientific knowledge of most wildlife pathogens are limited until some associated public health or veterinarian issues become apparent (Magurran *et al.* 2010; but see Staubach *et al.* 2011). The ongoing advent of molecular biology together with the development of computational resources will improve our ability to test increasingly complex and realistic demographic scenarios and improve the accuracy of estimates of a wider range of demographic parameters associated with the emergence and spread of emerging wildlife pathogens.

## Acknowledgements

We thank our colleagues from the 'Office National de l'Eau et des Milieux Aquatiques' and the 'Fédérations Départementales pour la Pêche et la Protection des Milieux Aquatiques' for their invaluable help with the collection of *T. polycolpus* samples. We also thank our international colleagues (more particularly Wojciech Piasecki and Amy Reading) that spent much time in searching some *T. polycolpus* samples in the supposedly native area. We are also grateful to Roman Biek and four anonymous reviewers for their insightful comments and suggestions to improve the final version of the manuscript. This work is part of the projects INCLIMPAR (ANR-11-JSV7-0010, Agence National de la Recherche) and PROBIS (Biodiversa). O.R. is also grateful to the 'Region Midi-Pyrénées' for financial support. We also thank the Fisheries Society of British Islands that has participated financially to this research. The genetic data were generated at the molecular genetic technical facilities of the Génopole Toulouse (Toulouse, France). We thank the CALMIP Group, Boris Dintrans, Nicolas Renon and Pierrette Barbaresco in particular. This work was performed using HPC resources from CALMIP (allocation P1003). This work has been performed in two research units (SEEM & EDB) that are part of the 'Laboratoire d'Excellence' (LABEX) entitled TULIP (ANR-10-LABX-41).

## References

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, **25**, 4692–4693.

Altizer S, Ostfeld RS, Johnson PTJ, Kutz S, Harvell CD (2013) Climate change and infectious diseases: from evidence to a predictive framework. *Science*, **341**, 514–519.

Aubrook EW, Fryer G (1965) The parasitic copepod *Tracheliastes polycolpus* Nordmann in some Yorkshire rivers: the first British records. *Naturalist London*, **893**, 51–56.

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

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F, 1996-2004. *GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations*. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université Montpellier II, Montpellier, France.

Benazzo A, Ghirotto S, Vilaca ST, Hoban S (2015) Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source. *Heredity*, **115**, 262–272.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Blanchet S (2012) The use of molecular tools in invasion biology: an emphasis on freshwater ecosystems. *Fisheries Management and Ecology*, **19**, 120–132.

Blanchet S, Méjean L, Bourque J-F *et al.* (2009a) Why do parasitized hosts look different? Resolving the "chicken-egg" dilemma. *Oecologia*, **160**, 37–47.

Blanchet S, Rey O, Berthier P, Lek S, Loot G (2009b) Evidence of parasite-mediated disruptive selection on genetic diversity in a wild fish population. *Molecular Ecology*, **18**, 1112–1123.

Blasco-Costa I, Waters JM, Poulin R (2012) Swimming against the current: genetic structure, host mobility and the drift paradox in trematode parasites. *Molecular Ecology*, **21**, 207–217.

Blum MGB, François O (2010) Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, **20**, 63–73.

Campana MG, Hunt HV, Jones H, White J (2011) CorrSieve: software for summarizing and evaluating structure output. *Molecular Ecology Resources*, **11**, 349–352.

Cornuet JM, Santos F, Beaumont MA *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

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

Criscione CD, Poulin R, Blouin MS (2005) Molecular ecology of parasites: elucidating ecological and microevolutionary processes. *Molecular Ecology*, **14**, 2247–2257.

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

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

Daszak P, Cunningham AA, Hyatt AD (2000) Emerging infectious diseases of wildlife – threats to biodiversity and human health. *Science*, **287**, 443–449.

Dunn AM (2009) Parasites and biological invasions. *Advances in Parasitology*, **68**, 161–184.

Epstein PR (2001) Climate change and emerging infectious diseases. *Microbes and Infection*, **3**, 747–754.

Estoup A, Angers B (1998) Microsatellite and minisatellites for molecular ecology: theoretical and empirical considerations. In: *Microsatellites: Evolution and Applications* (ed. Carvalho G), pp. 55–86. NATO Press, Amsterdam.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, **19**, 4113–4130.

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

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.

Facon B, Genton BJ, Shykoff J, Jarne P, Estoup A, David P (2006) A general eco-evolutionary framework for understanding bioinvasions. *Trends in Ecology & Evolution*, **21**, 130–135.

Fontaine MC, Austerlitz F, Giraud T *et al.* (2013) Genetic signature of a range expansion and leap-frog event after the recent invasion of Europe by the grapevine downy mildew pathogen *Plasmopara viticola*. *Molecular Ecology*, **22**, 2771–2786.

Fryer G (1982) *The Parasitic Copepoda and Branchiura of British Freshwater Fishes – A Handbook and Key*. Scientific Publications of the Freshwater Biological Association, Ambleside.

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

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2003) *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida.

Gozlan RE, St-Hilaire S, Feist SW, Martin P, Kent ML (2005) Biodiversity – disease threat to European fish. *Nature*, **435**, 1046.

Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2009) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.

Hartigan A, Fiala I, Dykova I *et al.* (2011) A suspected parasite spill-back of two novel *Myxidium spp.* (Myxosporea) causing disease in Australian endemic frogs found in the invasive cane toad. *PLoS ONE*, **6**, e18871.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Jones BA, Grace D, Kock R *et al.* (2013) Zoonosis emergence linked to agricultural intensification and environmental change. *Proceedings of the National Academy of Sciences*, **110**, 8399–8404.

Keith P, Persat H, Feunteun E, Allardi J (2011) *Les Poissons D'eau Douce de France*. Muséum nationale d'histoire naturelle, Paris, Mèze.

Kilpatrick AM (2011) Globalization, land use, and the invasion of West Nile virus. *Science*, **334**, 323–327.

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.

Lebarbenchon C, Brown SP, Poulin R, Gauthier-Clerc M, Thomas F (2008) Evolution of pathogens in a man-made world. *Molecular Ecology*, **17**, 475–484.

Lewin W-C, Arlinghaus R, Mehner T (2006) Documented and potential biological impacts of recreational fishing: insights for management and conservation. *Reviews in Fisheries Science*, **14**, 305–367.

Litchman E (2010) Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. *Ecology Letters*, **13**, 1560–1572.

Loot G, Poulet N, Reyjol Y, Blanchet S, Lek S (2004) The effects of the ectoparasite *Tracheliastes polycolpus* (Copepoda: Lernaeopodidae) on the fins of rostrum dace (*Leuciscus leuciscus burdigalensis*). *Parasitology Research*, **94**, 16–23.

Lootvoet A, Blanchet S, Gevrey M, Buisson L, Tudesque L, Loot G (2013) Patterns and processes of alternative host use in a generalist parasite: insights from a natural host–parasite interaction. *Functional Ecology*, **27**, 1403–1414.

Lymbery AJ, Thompson RCA (2012) The molecular epidemiology of parasite infections: tools and applications. *Molecular and Biochemical Parasitology*, **181**, 102–116.

Magurran AE, Baillie SR, Buckland ST *et al.* (2010) Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution*, **25**, 574–582.

Marino IAM, Benazzo A, Agostini C *et al.* (2013) Evidence for past and present hybridization in three Antarctic icefish species provides new perspectives on an evolutionary radiation. *Molecular Ecology*, **22**, 5148–5161.

de Meeûs T, McCoy KD, Prugnolle F *et al.* (2007) Population genetics and molecular epidemiology or how to "débusquer la bête". *Infection, Genetics and Evolution*, **7**, 308–332.

Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. *Nature*, **430**, 242–249.

Morse SS (1995) Factors in the emergence of infectious diseases. *Emerging Infectious Diseases*, **1**, 7–15.

Ohta T, Kimura M (1973) A model mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201–204.

Pazooki J, Masoumian M (2012) Synopsis of the parasites in Iranian freshwater fishes. *Iranian Journal of Fisheries Sciences*, **11**, 570–589.

Piasecki W (1989) Life cycle of *Tracheliastes maculatus* Kollar, 1835 (Copepoda, Siphonostomatoida, Lernaeopodidae). *Wiadomosci Parazytologiczne*, **35**, 187–245.

Ponyi J, Molnar L (1969) Studies on the parasite fauna of fish in Hungary V. Parasitic Copepods. *Parasitologia Hungarica*, **2**, 137–148.

Popiolek M, Kubizna J, Wolnicki J, Kusznierz J (2011) Parasites of lake minnow, *Eupallasella percnurus* (Pall.): the state of knowledge and threats. *Archives of Polish Fisheries*, **19**, 133–226.

Poulet N, Beaulaton L, Dembski S (2011) Time trends in fish populations in metropolitan France: insights from national monitoring data. *Journal of Fish Biology*, **79**, 1436–1452.

Poulin R (2014) Parasite biodiversity revisited: frontiers and constraints. *International Journal for Parasitology*, **44**, 581–589.

Poulin R, Paterson RA, Townsend CR, Tompkins DM, Kelly DW (2011) Biological invasions and the dynamics of endemic diseases in freshwater ecosystems. *Freshwater Biology*, **56**, 676–688.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pullan RL, Sturrock HJW, Magalhaes RJS, Clements ACA, Brooker SJ (2012) Spatial parasite ecology and epidemiology: a review of methods and applications. *Parasitology*, **139**, 1870–1887.

Rachowicz LJ, Hero JM, Alford RA *et al.* (2005) The novel and endemic pathogen hypotheses: competing explanations for the origin of emerging infectious diseases of wildlife. *Conservation Biology*, **19**, 1441–1448.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Rousset F (2008) GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Sakai AK, Allendorf FW, Holt JS *et al.* (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics*, **32**, 305–332.

Silfverberg H (1999) A provisional list of Finnish Crustacea. *Memoranda-Societas Pro Fauna et Flora Fennica*, **75**, 15–37.

Staubach C, Hoffmann L, Schmid VJ, Ziller M, Tackmann K, Conraths FJ (2011) Bayesian space-time analysis of *Echinococcus multilocularis* infections in foxes. *Veterinary Parasitology*, **179**, 77–83.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics*, **24**, 2498–2504.

Tuffery G (1967) Importance des considérations topographiques, biologiques, écologiques, lors de l'aménagement ou du classement d'un bassin hydrographique. *Bulletin Français du Pisciculture*, **226**, 5–21.

Warnecke L, Turner JM, Bollinger TK *et al.* (2010) Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. *Proceedings of the National Academy of Sciences*, **109**, 6999–7003.

Watari Y, Nagata J, Funakoshi K (2011) New detection of a 30-year-old population of introduced mongoose *Herpestes auropunctatus* on Kyushu Island, Japan. *Biological Invasions*, **13**, 269–276.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, **11**, 116.

Woolhouse MEJ (2002) Population biology of emerging and re-emerging pathogens. *Trends in Microbiology*, **10**, s3–s7.

Yalcindag E, Elguero E, Arnathau C *et al.* (2012) Multiple independent introductions of *Plasmodium falciparum* in South America. *Proceedings of the National Academy of Sciences*, **109**, 511–516.

The study was designed by S.B. and G.L. O.R. and I.P.V. wrote the manuscript with the help of L.F., B.R., G.L. and S.B. S.B., G.L. and C.V. collected the samples; G.L. and C.V. produced the genetic data; I.P.V., O.R. and L.F. conducted the statistical analyses with the help of S.B. and B.R.

## Data accessibility

Detailed information on sampling sites, on samples and on individual microsatellite genotypes: Dryad doi:10.5061/dryad.vn536

## Supporting information

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

**Fig. S1** Worldwide distribution of *Tracheliastes polycolpus* based on all records found in the literature.

**Fig. S2** Sampling map of *Tracheliastes polycolpus* (circles).

**Fig. S3** Plot of mean posterior probability (LnP(D)) values (open circles) *per* cluster (*K*) generated by the STRUCTURE program (Pritchard *et al.* 2000).

**Fig. S4** Magnitude of Δ*K* as a function of the number of *K* clusters (mean over five replicates) according to Evanno *et al.* (2005) using the R package 'CorrSieve'.

**Fig. S5** Scatterplot of the two first principal components of the DAPC used to identify the number of *T. polycolpus* genetically differentiated clusters present among French watersheds.

**Fig. S6** Type I (i.e. how frequently a simulation generated under scenario X has been excluded when it has been actually generated under scenario X; blue bars), type II (i.e. how frequently a simulation has been assigned to a scenario X when it has not been generated under scenario X; red bars) error rates and rate of correct classifications (green bars) estimated from a leave-one-out cross-validation procedure based on 100 validation simulations *per* demographic scenario (Sc1 to Sc6), 16 summary statistics (see Material and methods section), a neural network algorithm and a tolerance rate of 0.0001.

**Fig. S7** Ninety-nine percent envelopes of the two principal components (i.e. PC1 and PC2) of a principal component analysis on the space of 16 summary statistics calculated for the 10 000 simulations closest (using an Euclidean distance) to the observed (i.e. real) dataset under each demographic scenario.

**Fig. S8** Ninety-nine percent envelope of the two principal components (i.e. PC1 and PC2) of a principal component analysis on the space of 16 alternative summary statistics (see Material and methods section) calculated for 1000 simulations generated under the best-supported scenario (Sc6) with parameter values extracted from the posterior distribution of parameters.

**Fig. S9** Histogram of the null distribution of the goodness-of-fit test statistic (i.e. the mean of the distance between accepted and observed summary statistics) along with the test *P*-value, estimated from 1000 simulations generated under the best-supported scenario (Sc6) with parameter values extracted from the posterior distribution of parameters.

**Fig. S10** Posterior predictive checks for *Tracheliastes polycolpus* in France under the best demographic scenario (Sc6), estimated from 1000 simulations generated under Sc6 assuming parameter values extracted from the posterior distribution of parameters.

**Table S1** Confusion matrix obtained using a leave-one-out cross-validation procedure based on 100 validation simulations *per* scenario, a neural network algorithm and a tolerance rate of 0.00001.

**Table S2** Posterior probabilities obtained for the six scenarios through the ABC model-choice procedure, using a neural network algorithm with a tolerance rate of 0.0001.

**Table S3** Bayes factors computed between each pair of scenarios (the definition of each scenario is detailed in the Material and methods section of the main text; see also Fig. S2).

**Table S4** Estimation errors (standardized by feature scaling) obtained for different ABC parameter inference methods (rejection, local linear regression and neural network) and tolerance rates (0.001, 0.005, 0.01 and 0.05) through a leave-one-out cross validation procedure based on 100 simulated datasets generated under the scenario Sc6.

**Table S5** Mode value, 2.5 and 97.5% percentiles of the posterior distributions inferred by the ABC inference procedure for each parameter for the first, second, and third ranked scenarios (i.e., Scenarios 6, 2 and 4 respectively; ranked decreasingly based on the posterior probabilities found with the ABC model-choice procedure), using neural networks and tolerance rates of 0.001.

**Appendix S1** Details on the microsatellite multiplex PCR protocols.

**Appendix S2** The computational pipeline used for implementing the simulation procedure.