ORIGINAL ARTICLE

On the role played by the carrying capacity and the ancestral population size during a range expansion

S Mona^{1,2}

Most species are structured and various population genetics models have been proposed to investigate their history. For mathematical tractability, most of these models make the simplifying assumption of equilibrium. Here we focus on the properties of a nonequilibrium spatial explicit model, range expansions (REs). Despite their abundance, many details of their genetic consequences need yet to be fully investigated. The model we studied is characterized by four main parameters: the effective population size of each deme (N), the migration rate per generation per deme (m), the time of the expansion (T_{exp}) and the effective size of the deme from which the expansion started (N_{anc}). By means of extensive coalescent simulations, we focused on two aspects of range expansions for fixed Nm : (1) the separate influence of N and m and (2) the role of N_{anc} . We compared our results with an equilibrium stepping stone model and found two main features typical of REs: an excess of rare variants for larger N and a complex interaction between N , T_{exp} and N_{anc} in shaping the degree of population differentiation (which depends only on Nm in the stepping stone model). Finally, we developed an approximate Bayesian computation approach to jointly estimate N and m and to infer N_{anc} . When applied to pseudo-observed data sets, we could correctly recover both N and m (but not N_{anc}), provided a large number of demes were sampled. These findings highlight how it will be possible to estimate the dispersal rate in nonequilibrium metapopulations by using population genetics approaches.

Heredity (2017) **118**, 143–153; doi:10.1038/hdy.2016.73; published online 7 September 2016

INTRODUCTION

In recent years, advances in both theoretical and empirical population genetics have highlighted the importance of taking population structure into account when interpreting the variability of a species and inferring demographic parameters. Most, if not all, populations are indeed not isolated, but linked in a network of demes exchanging migrants to some extent (that is, they belong to a metapopulation). Many recent works have shown that failing to take the structure of the metapopulation into account may lead to spurious signatures of effective population size changes through time (Chikhi *et al.*, 2010; Peter *et al.*, 2010; Stadler *et al.*, 2009) and to biased estimates of the mutation rate (Navascues and Emerson, 2009). Moreover, models of population structure can also leave a genetic footprint similar to episodes of ancient hybridization (Eriksson and Manica, 2014).

Of particular importance in this context are the range expansions (REs). Episodes of REs have been postulated for many species (Rendine *et al.*, 1986; Ibrahim *et al.*, 1996; Cavalli-Sforza and Feldman, 2003; Klopstein *et al.*, 2006; Boessenkool *et al.*, 2009; Brown *et al.*, 2010) but they have been quantitatively tested in a spatially explicit framework only in few studies (Barbujani *et al.*, 1995; Hamilton *et al.*, 2005; Ray *et al.*, 2005; Francois *et al.*, 2008; Neuenschwander *et al.*, 2008; Gaggiotti *et al.*, 2009; Schneider *et al.*, 2010; Mona *et al.*, 2013). REs leave distinct signatures in the genetic diversity of populations (Barbujani *et al.*, 1995; Ray *et al.*, 2003; Edmonds *et al.*, 2004; Travis *et al.*, 2007; Sousa *et al.*, 2014), proceeding as a series of founder events followed by continuous migrations among neighboring demes

(Austerlitz *et al.*, 1997; Ray *et al.*, 2003). In a seminal work on the genetics of metapopulation, Wakeley (1999) and Wakeley and Aliacar (2001) have elegantly separated the coalescent process of genes sampled in a structured population into two separate phases, namely the *scattering* and the *collecting* phase. The *scattering* phase, proceeding backward in time, starts at the present generation, ending when all sampled genes have either coalesced or migrated to different demes. At this point, the *collecting* phase begins: this period is characterized by a large number of migration events and a few coalescent events that can happen only when two lineages end up in the same deme. The coalescent process during the *collecting* phase in a structured population was shown to have the same properties as in an unstructured population, with time scaled by the number of demes, deme size and migration rates (Wakeley, 1999). A similar separation of timescales was also shown to hold in more complex and realistic models taking into account geographic features (Wakeley and Aliacar, 2001) in extinction/recolonization models (Wakeley and Aliacar, 2001; Wakeley, 2004), in continuous populations (Wilkins and Wakeley, 2002; Wilkins, 2004) as well as in range expansions (Ray *et al.*, 2003; Wegmann *et al.*, 2006; Nullmeier and Hallatschek, 2013; Mona *et al.*, 2014).

Despite the growing interest in the genetics properties of REs, many of their features are not yet completely understood. Most of the attention has focused so far on REs in homogeneous environments (Ray *et al.*, 2003), where the driving parameter of the genetic diversity is Nm (effective population size times the migration rate). This result

¹EPHE, PSL Research University, Paris, France and ²Institut de Systématique, Évolution, Biodiversité (ISYEB), UMR 7205 - CNRS, MNHN, UPMC, EPHE, Ecole Pratique des Hautes Etudes, Paris Sorbonne Universités, Paris, France

Correspondence: Dr S Mona, EPHE, PSL Research University, UMR 7205, 16 Rue Buffon, F-75005 Paris, France.

E-mail: mona@mnhn.fr

Received 3 July 2015; revised 19 July 2016; accepted 26 July 2016; published online 7 September 2016

mirrors Wright's finite island model where Nm determines not only the diversity of the metapopulation (Wakeley, 1998), but also its genetic structure (Wright, 1931). However, REs are spatial explicit nonequilibrium models where the colonization process starts from an ancestral deme. So far, the role played by the effective population size of the ancestral deme (hereafter N_{anc}) on shaping the genetic variation of the metapopulation has not been thoroughly investigated. Moreover, both Ray *et al.* (2003) and Wegmann *et al.* (2006) noted that, for low Nm , demes with large carrying capacity have a higher number of haplotypes compared with demes with low carrying capacity. They explained this finding suggesting that gene lineages may co-migrate in the same deme and coalesce quickly with low N . This increases the probability of identity and decreases the number of haplotypes compared with demes where N is larger. Nevertheless, no further analyses were conducted to test this hypothesis. To better characterize the effect of different N values given a fixed Nm on the pattern of genetic diversity as well as to define the influence of N_{anc} , we simulated REs using SPLATCHE2 (Ray *et al.*, 2010). We found that for a fixed Nm : (1) intra-deme genetic diversity depends on N (that is, for increasing N values we found a monotonic increase in the number of haplotypes and in the number of segregating sites) and (2) increasing N leads to a skew in the site frequency spectrum toward an excess of rare variants that leads to more significantly negative values of Tajima's D . We compared these results with an equilibrium stepping stone model: fixing Nm we observed an increase in the genetic diversity for higher N as in the REs but we found no influence on the frequency of rare variants. We also identified the role played by the ancestral population size of the founding deme particularly on the degree of population structure of the whole metapopulation. We further demonstrated how it is indeed possible to correctly jointly estimate N and m in a RE using an approximate Bayesian computation framework (ABC), provided we have sampled many demes (at least more than one) and we have typed at least few independent loci.

MATERIALS AND METHODS

We simulated a range expansion of haploid individuals over a two-dimensional array of demes arranged on a lattice using the software SPLATCHE2 (Ray *et al.*, 2010). Each deme exchanges migrants with its four nearest neighbors as in a two-dimensional stepping stone model (Kimura and Weiss, 1964) with bouncing edges. Simulations are done in two consecutive steps, namely the forward (demographic) and the backward (coalescent) steps. The forward simulation starts from a founding (or ancestral) deme that sends migrants to its neighboring demes. The founding deme is characterized by an effective population size N_{anc} , a parameter that does not play a role in the forward process but it will determine the rate of coalescence in the backward phase (see below). Migrations to empty demes represent new colonization events. Each deme has an intrinsic growth rate r (the same value for all demes) and its density is logistically regulated by its carrying capacity (Ray *et al.*, 2003). We note that for this reason the effective population size in each deme varies slightly each generation, but the effect of such oscillations on the genetic variability are negligible. For this reason and for the sake of comparison with previous work and with equilibrium models, we called hereafter the carrying capacity N . After the regulation step within each deme, migrants are sent to the four neighboring demes at rate m . The process is repeated generation by generation for each non-empty deme, resulting in a wave of advance of the whole population. The second phase of the algorithm then starts at the present generation, proceeding backward in time. The actual number of individuals present in a deme is used to compute the probability of a coalescent event, and the migration rates determine the probability of each sampled genes to emigrate, backward in time, to the surrounding demes. The coalescent process stops after all genes have coalesced. The lineages that did not coalesce in the

scattering or in the collecting phase will coalesce in the ancestral deme, with a rate determined by N_{anc} .

The size of the array of stepping stones was fixed to 50×50 demes and the origin of the expansion was set to position $<1;1>$ of the square lattice (North-West corner). We tested six carrying capacities ($N=20, 50, 100, 200, 500$ and 1000), with migration rates varying accordingly to get a fixed Nm . To cover observed population differentiation in real data, we tested four Nm values ($Nm=1, 10, 50$ and 100). The resulting F_{st} value would range from ~ 0.005 to ~ 0.33 in a finite island model. Expansion times (T_{exp}) were set to 50 000, 15 000, 3000, 1000 and 500 generations ago and the ancestral population size to $N_{anc}=100\ 000$, and 10 000. These parameter combinations were chosen to represent a post-glacial expansion of species with different generation time and to explore a broad range of plausible biological demographic scenarios. We simulated DNA sequences of 1000 bp with a mutation rate of 0.001 per generation for the whole locus and performed 1000 coalescent simulations for each demographic scenario. We sampled 9 fixed demes in the array, as displayed in Figure 1. From each deme we sampled 20 lineages. All these simulations were repeated under an instantaneous expansion model as in Hamilton *et al.* (2005), where all the demes are colonized at the same time (hereafter, IE model). Finally, we simulated an equilibrium stepping stone models for the four Nm and the same combinations of N and m as above using fastsimcoal2 v2.5.1 (Excoffier *et al.*, 2013). Simulations of the equilibrium stepping stone were performed with a mutation rate of one order of magnitude lower than the one used in REs in order to obtain comparable levels of variability. We computed several statistics to characterize the genetic diversity at the intra-deme level: the number of haplotypes H (that is, the number of different sequences present in the sample), the number of segregating sites S (that is, the number of polymorphic sites in the data) and the mean pairwise difference π (that is, the mean number of differences between all pairs of sequences in the sample). To detect departure from the Wright-Fisher model (that is, an isolated population of constant effective size) because of the range expansion as well as the presence of metapopulation structure, we computed two summary statistics and then tested their departure from mutation-drift equilibrium: Tajima's D that is based on the comparison of two estimators of the scaled mutation rate θ (namely θ_S and θ_π) and Fu's F_s that evaluates the probability of having a number of haplotypes similar or smaller than the observed value given the observed number of the mean pairwise difference. To determine the extent of genetic differentiation between demes we computed the ϕ_{st} (based on molecular distance metrics; Excoffier *et al.*, 1992) and F_{st} (based on allele frequency). All summary statistics were calculated using Arlequin 3.5 (Excoffier and Lischer, 2010). Intra-population statistics were computed on the central deme of the array as well as on the nine pooled demes. ϕ_{st} and F_{st} were

Figure 1 The simulated array of 50×50 demes. Black dots: sampled demes. Gray dot: founding deme (origin of the range expansion). Gray circles: demes sampled to compute ϕ_{st} and F_{st} presented in the text.

computed among the central and the bottom left deme (Figure 1), as well as among all demes.

To test the ability to estimate the four parameters of our RE model: N , m , N_{anc} and T_{exp} , as well as the Nm product, we developed an ABC framework (Beaumont *et al.*, 2002; Beaumont, 2010; Bertorelle *et al.*, 2010) in four steps. *Step 1*: we created our reference tables for the later estimation process by performing 100 000 simulations sampling from uniform prior distributions for N (U: 5–2000), m (U: 0.004–1), T_{exp} (U: 500–20 000) and N_{anc} (U: 50–15 000). The resulting compound prior for Nm is very close to an exponential distribution with rate 0.002 bounded within 0–2000. We sampled 9 demes and 20 lineages per deme (Figure 1) and simulated 1000 independent loci, each of 1000 bp length with a mutation rate of 0.001 per generation for the whole

locus as above. *Step 2*: we computed two sets of summary statistics for the ABC analysis: (1) SET1: the intra-population statistics (H , S , π , D and F_s) in the 9 demes pooled and the ϕ_{st} (for a total of six summary statistics) and (2) SET2: the same five intra-population statistic computed in the central deme only. Each statistic was averaged over the number of loci. *Step 3*: we generated pseudo observed data sets (*pods*) by randomly extracting 1000 observations from the reference table, recording the corresponding vector of prior values. *Step 4*: we analyzed *pods* using the ABC algorithm, with a weighted local linear regression (Beaumont *et al.*, 2002) performed on the 5% of the closest simulations. For each analyzed *pods* we retained the mode and the median estimate of the posterior distribution and plotted them against the ‘real’ value recorded in step 3. We also computed the scaled root mean square error (SRMSE) (Walther and Moore, 2005) both for the median and the mode and the 95% coverage. A power analysis was carried out by subsampling from the total number of loci (1, 10 and 100) for both sets of summary statistics and sampling scheme. To determine which statistics is more informative and the importance of pooling demes we define three other sets of summary statistics: SET3 (summary statistics computed on all demes separately plus the ϕ_{st} for a total of 46 statistics), SET4 (the same as SET1 without ϕ_{st}) and SET5 (the same as SET1 without the neutrality tests). Summary statistics used in each set are listed in Table 1.

RESULTS

Unless specified, we present and discuss results in REs relative to: (1) the nine pooled demes concerning the statistics computed at the

Table 1 Summary statistics used in the approximate Bayesian computation (ABC) cross-validation procedure

SET	Summary statistics
SET1	H , S , π , D and F_s in the 9 demes pooled and the ϕ_{st} (6 summary statistics)
SET2	H , S , π , D and F_s computed in the central deme only (5 summary statistics)
SET3	H , S , π , D and F_s in each of the nine demes and ϕ_{st} (46 summary statistics)
SET4	H , S , π , D and F_s in the 9 demes pooled (5 summary statistics)
SET5	H , S , π in the 9 demes pooled and the ϕ_{st} (4 summary statistics)

Figure 2 Distribution of the number of haplotypes (H) for $T_{exp}=3000$ and $N_{anc}=100$ for four Nm values computed on the pooled demes. N values on the x axis.

intra-population level (Figure 1); (2) the central deme concerning the neutrality tests; and (3) the central and the bottom left deme, concerning F_{st} and ϕ_{st} (Figure 1). The other sampling schemes displayed a similar pattern. We note that the grid of demes was not fully colonized for $Nm=1$ and $T_{exp} \leq 1000$, and therefore we did not report results of these simulations.

For fixed Nm we observe a monotonic increase in genetic diversity (as measured by H , S and π) for growing N (Figure 2). Both H and S showed a dependency on N for all the T_{exp} tested as well as for the equilibrium stepping stone model (Figure 2 and Supplementary Figures S1 and S2). Conversely, we observed a different pattern for π that is insensitive to N for low T_{exp} but depends on it for higher expansion time ($T_{exp}=50\,000$) and in the equilibrium stepping stone model (Supplementary Figure S3). The excess of private haplotypes for larger N leads to an increase of H and S that is reflected in the more negative D and F_s (Figure 3 and Supplementary Tables S1–S4). The influence of N on the neutrality tests decreases for high T_{exp} and it is

lost when the process arrives at equilibrium (Figures 3 and 4). This result can be explained as followed. For fixed Nm more mutations will occur for higher N . The length of the internal branches of the genealogy is constrained by T_{exp} , implying that this excess of mutations will be concentrated in the terminal branches. These new mutations are therefore at low frequencies having almost no influence on π . Conversely, for large T_{exp} and in the stepping stone the internal branches will be longer so that mutations will have time to occur. In this case, these mutations will be at intermediate frequencies, having an impact on π (and, as a consequence, on the neutrality tests). The percentage of significant outcomes (as calculated by Arlequin 3.5), both positive and negative, is also related to N for all the four Nm tested (Figure 4 and Supplementary Tables S1–S4). Low Nm can produce false signature of bottleneck in equilibrium model at the deme level (Chikhi *et al.*, 2010). We found a similar result in RE model, where for $Nm=1$ many data sets display a significant signature of reduction of the effective population size (Supplementary Table S3).

Figure 3 Distribution of Tajima's D for $Nm=10$ in a RE for various values of T_{exp} and the equilibrium stepping stone model in the 9 pooled demes.

However, when N grows from 20 to 1000, the number of significant positive Tajima's D drop considerably (Supplementary Table S3). Similarly, the power to detect a RE depends on the T_{exp} (Ray *et al.*, 2003): here we show that even for large T_{exp} the probability of detecting an expansion increases substantially for larger N (Figure 4), and the probability of erroneously inferring a bottleneck drops (Supplementary Tables S1–S4).

We present the ϕ_{st} and F_{st} distribution in REs for $Nm=10$ and $N_{\text{anc}}=100$ in Figure 5 and compare the results with the equilibrium stepping stone model. We observed a negative correlation between both measures of population differentiation and N when $T_{\text{exp}} \geq 3000$ (Figure 5). This result is consistent with the monotonic increase of diversity for larger N : higher heterozygosity tends to lower the degree of population differentiation, in particular when it is measured from allele frequencies (Hedrick, 2005). However, ϕ_{st} does not depend on N in the equilibrium stepping stone, whereas the behavior of F_{st} is consistent to that of nonequilibrium models. We found a similar trend for all the Nm values we tested, ranging from 1 to 100 (Supplementary Figures S4–S6). However, the decay of population differentiation for large N it is particularly strong for $Nm=1$ where the median value of F_{st} is ~ 0.8 for $N=20$, falling to ~ 0.35 for $N=1000$. Such difference is weaker for ϕ_{st} (decreasing from ~ 0.8 to ~ 0.6), consistent with the

suggestion that measures based on molecular difference can better correct for the heterozygosity excess (Hedrick, 2005). For $T_{\text{exp}} \leq 1000$ the pattern is more complex and it is not consistent between the two differentiation measures. First of all, we observed a slight increase in the mean and in the variance of the distribution of ϕ_{st} for higher N (Figures 5–7 and Supplementary Figure S5 and S6), whereas no dependency on N was found in the F_{st} . Second, the monotonic increase in ϕ_{st} for larger N is more pronounced when $N_{\text{anc}}=10\,000$. This pattern is more pronounced in IEs than in REs, where we also found similar results for $T_{\text{exp}}=3000$ when $N_{\text{anc}}=10\,000$ (Figure 7 and Supplementary Figure S6). Independently of Nm , no differences related to N were observed for the F_{st} . The genome-wide distribution of measures of population differentiation is therefore dependent on the interaction between N , T_{exp} and N_{anc} in RE model. Detecting outlier loci in genomic scan of selection will require a good knowledge of all these parameters when investigating structured species.

We performed a cross-validation ABC procedure to investigate the power to jointly estimate N and m in RE model. We presented the modes of the estimated N in Figure 8 and Supplementary Figure S7 for all the combination of sets of summary statistics (SET1 and SET2) and number of loci (1, 10, 100 and 1000). The black line represents a perfect match: the closer the points are to the line, the closer the

Figure 4 Proportion of significant negative outcomes of Tajima's D for the four Nm values as a function of N in the nine pooled demes.

Figure 5 Distribution of ϕ_{st} (on the left of the dotted line) and F_{st} (on the right of the dotted line) for $Nm=10$ in a RE for various values of T_{exp} and the equilibrium stepping stone model, N_{anc} fixed to 100.

estimated mode is to the ‘real’ parameter. As expected, increasing the number of loci reduces the variance of the modes around the line. When we compared SET1 and SET2, we found that pooling more demes increases the performance of the ABC procedure, particularly when a small number of loci are used. A consistent pattern was found for all other parameters (Supplementary Figures S8–S10), with the exception of N_{anc} , poorly estimated in all cases (Supplementary Figure S11). We reported the SRMSE averaged over the priors for both the mode and the median for all the four parameters of the RE model and the Nm product in Supplementary Table S5. An SRMSE of one means that mode (or median) is on average one time higher or smaller than the real value. SRMSE decreases with the number of loci and it is always smaller in SET1 compared with SET2. We noted that the accuracy in the estimate of Nm is slightly higher than that of N , whereas in m it is considerably worse than both of them.

Adding more loci did not improve the estimate of N_{anc} that was unexpectedly poor. Conversely, few loci were enough to correctly estimate T_{exp} with any sets of summary statistics considered. We further tested three more combinations of summary statistics. First, we evaluated whether pooling more demes increases the performance in the ABC estimation process. To this end, we used the intra-deme summary statistics calculated separately in the nine demes: adding ϕ_{st} we defined SET3, characterized by 46 summary statistics. SRMSE were similar to SET1 (Supplementary Table S5), suggesting that it is important to sample more demes but not necessarily to pool them. We note however that pooling reduces the number of predictors in the regression step of ABC and it could be particularly helpful when studying large data sets (that is, composed by many demes). We then investigated the role played by ϕ_{st} by removing it from SET1 (thus obtaining the SET4). We obtained results very

Figure 6 Distribution of ϕ_{st} (on the left of the dotted line) and F_{st} (on the right of the dotted line) for $Nm=10$ in a RE for two values of T_{exp} .

similar to SET2 (Supplementary Table S5): migration rate is poorly estimated and, as expected, Nm shows an SRMSE twice as big as the one found in SET1. Finally, we obtained SET5 by removing D and F_s from SET1. We observed a decrease in the performance of the estimation of T_{exp} , N and m , all of which showed an SRMSE two times bigger than SET1. Conversely, the compound parameter Nm was unaffected.

DISCUSSION

Empirical population genetics is now experiencing a revolution. Thanks to the spread of next-generation sequencing technology, an impressive amount of genetic data are now easy to collect in both model and nonmodel organisms. These huge amounts of data carry precious information on the demographic and selective history of populations and species. However, having a lot of data has a potential drawback: the interpretation of genetic variation can be significantly incorrect when an incorrect demographic model is applied. Most (if not all) species are structured, that is, they are composed by many demes exchanging migrants to various extent. The two structured models that have been explored the most in theoretical population

genetics are the island and the stepping stone model, both of which are at equilibrium. It is therefore urgent to explore in greater details nonequilibrium structured model and to integrate spatial components. It is for this reason that here we explored REs. Previous studies have shown that in homogeneous environment the main driver of genetic diversity after a RE is the compound Nm product (Ray *et al.*, 2003). Here we focus on the influence of N for fixed Nm values and on the role played by N_{anc} in shaping the genetic diversity. Finally, we tested how well we can recover the demographic parameters after a RE using an ABC approach and how many demes it is important to sample to obtain accurate estimates. All those issues have not been investigated before and they are likely to be informative for empirical studies.

To better understand the specificity of REs, we compared our results with an equilibrium stepping stone model, fixing the Nm product and exploring various combinations of N and m . We considered both the variability within deme and the population differentiation computed among demes. At the intra-deme level, we found a monotonic increase in genetic diversity in both RE and stepping stone for increasing N . However, we found two important characteristics of REs: (1) an excess of low frequency haplotypes for

Figure 7 Distribution of ϕ_{st} (on the left of the dotted line) and F_{st} (on the right of the dotted line) for $Nm=10$ in an IE for two values of T_{exp} .

increasing N and (2) a complex interaction between N , N_{anc} and T_{exp} in shaping the degree of population differentiation. It is for these reasons that we could estimate N and m using an ABC approach. Below we provide some coalescent arguments to explain our findings.

Signatures of REs

Coalescent in metapopulation can be described in terms of *scattering* and *collecting* phases (Wakeley, 1999). In RE models the number of emigrants sent out to each generation is Nm , whereas the within-deme probability of coalescent is determined by N only. Therefore, for a fixed Nm the *scattering* phase will be longer for higher N and fewer within-deme coalescences will happen. This has three consequences, as for higher N : (1) the number of lineages entering the *collecting* phase will be higher; (2) there will be more H and more private S as in RE models mutations can also occur during the *scattering* phase; and (3) the ratio of external to internal branch lengths increases. Indeed, we observed a monotonic increase in both the number of haplotypes and the number of segregating sites for larger N when we fixed Nm (Figure 2 and Supplementary Figure S1). This is also true for the

equilibrium stepping stone and it is therefore related to the increase of the scaled mutation rate (Supplementary Figure S2). However, the distribution of mean pairwise differences follows a different pattern: we observed a monotonic increase for larger N only when $T_{exp} \geq 50\,000$ and at equilibrium. Both θ_π and θ_s are estimators of the scaled mutation rate θ and Tajima's D is based on their difference that is centered around zero under neutrality (Tajima, 1989). For a fixed Nm , if S increases monotonically with N but π remains constant, the difference between the two estimators will become negative, a signature typical of population expansion. For this reason, the power to detect an RE for fixed Nm depends heavily on N (Figure 4). This has practical implications: the signature of a RE will be lost more quickly in species characterized by small-sized demes. This is clearer when we compare the results of $Nm=10$ between RE and the equilibrium stepping stone, where N does not play a role in determining the shape of the genealogy (Figures 3 and 4). Moreover, in equilibrium model, low Nm gives rise to false signature of population bottleneck (Chikhi *et al.*, 2010). This is also true in RE (Ray *et al.*, 2003) but the number of positive outcomes of Tajima's D

Figure 8 Cross-validation of N . Modes (y axis) are estimated under the ABC framework using the SET1 and plotted against the real values (x axis). The analyses are performed with 1, 10, 100 and 1000 loci (see text for details).

drops when the deme size increases (Supplementary Table S3). We argue that in empirical studies, particularly when concerning endangered species, N and m must be estimated separately to fully recover the underlying demography.

Population differentiation

The *collecting* phase dominates the coalescent process in equilibrium model (Wakeley, 1999), whereas its length in RE/IE will depend both on N and T_{exp} . In RE/IE, backward in time lineages will contract toward the founding/ancestral deme: for a fixed Nm there will be less coalescence with higher N and more lineages entering the *ancestral* phase of the coalescent process. The number of lineages surviving until the *ancestral* phase is also influenced by T_{exp} . Indeed, when T_{exp} is low, there is less opportunity for the sampled lineages to coalesce in the array of demes. The surviving lineages will coalesce in the *ancestral* deme according to a Kingman coalescent dependent on N_{anc} only, accumulating mutations proportionally to N_{anc} and the mutation rate. Therefore, for fixed Nm larger N will increase the heterozygosity and the number of lineages reaching the founding deme, the two phenomena having opposite effect on the degree of population differentiation. The increased within-deme heterozygosity tends to lower both ϕ_{st} and F_{st} (Hedrick, 2005; Meirmans, 2006), causing a negative correlation between N and the genetic differentiation between

demes (Figure 5 and Supplementary Figure S4). This is clearer for $T_{\text{exp}} \geq 3000$ because the difference in heterozygosity among the tested N is of higher magnitude. We found a similar behavior in the equilibrium stepping stone (Figure 5), consistently with the observed pattern of intra-deme diversity. Conversely, more lineages coming from different demes will arrive in the *ancestral* phase for larger N and $T_{\text{exp}} < 3000$. Here, the occurring mutations enhance the between deme component of molecular variance, increasing ϕ_{st} but not F_{st} for increasing N (Figures 5–7 and Supplementary Figures S5 and S6). When $T_{\text{exp}} \leq 1000$ generations, under our parameter combination, the influence of N on the ϕ_{st} is remarkable (Figure 4). The larger the N_{anc} , the higher will be the average ϕ_{st} and its variance (Figures 6 and 7 and Supplementary Figures S5 and S6). This effect is of greater intensity for IE (Figure 7 and Supplementary Figure S6), where the number of lineages arriving in the *ancestral* phase is larger because the contraction phase typical of RE, where many coalescent events happen before arriving in the ancestral deme (Excoffier *et al.*, 2009), is absent. These findings have important implications in the analysis of real data. For example, invasive species are typical instances of young REs: the genome-wide distribution of the ϕ_{st} will differ enormously between species characterized by small- vs large-sized demes. Many tests of selection are based on the idea of finding outliers in the distribution of some summary statistics such as the ϕ_{st} . A good knowledge of the

effective deme size is therefore needed to generate a correct null distribution to avoid false positive results in genomic scan of selection.

Estimating parameters of a RE

The influence of the carrying capacity on the genetic variability at both the intra and inter-population levels suggested that it should be possible to jointly estimate N and m using summary statistics. To this end, we developed an ABC framework and check by cross-validation the power to estimate the parameters of a RE. We varied the number of simulated loci, the number of demes sampled and the set of summary statistics chosen. To be consistent with all the previous simulations, we computed inter and intra-population summary statistics in both a single deme and a pool of lineages coming from all nine demes. We measured the accuracy of the estimation of all parameters by computing the SRMSE averaged through the prior distributions. For any combination of loci/demes/summary statistics, we obtained a similar accuracy in the estimation of N and Nm (Supplementary Table S5). This shows unambiguously how N affects the distribution of genetic variability, playing a role that can be easily disentangled from m . Several independent loci need to be typed to take into account the stochastic variability of the coalescent process: however, it seems that going from 100 to 1000 loci only slightly improves the accuracy of our estimate, suggesting that at least under our model a relatively small set of loci should provide consistent results. When using summary statistics computed in pooled lineages (SET1) or in the nine demes independently (SET3) the performance of the ABC procedure increases drastically compared with SET2, where statistics from a single deme are used (Figure 8 and Supplementary Figures S7–S11). The number of demes studied is therefore important to correctly characterize the parameter of REs. Summary statistics convey different information on each parameter. As expected, estimating the migration rate m is crucial to add the ϕ_{st} in the ABC algorithm (Supplementary Table S5). Indeed even when we pooled demes, the SRMSE is higher than one for m if we do not include the ϕ_{st} , whereas the drop in accuracy is less drastic for the compound parameter Nm (SET1 and SET4, Supplementary Table S5). Similarly, neutrality test statistics are also useful to disentangle N and m (Supplementary Table S5, SET5), consistent with the finding that N influences the power to detect a RE for fixed Nm . It should be possible to devise an optimal sampling strategy in terms of number of demes and number of loci, but this is beyond the scope of this paper. We only highlight here the importance of sampling more than just one deme to estimate the parameters of a RE and to compute a measure of population differentiation to estimate the migration rate. Despite the large number of loci studied, we could not correctly estimate the N_{anc} . Contrary to N , which plays a direct role on the intra-population diversity and on the measure of differentiations, N_{anc} interacts with N and T_{exp} to shape F_{st} and ϕ_{st} . We argue that for this reason it may be not identifiable under our setting. At the same time, we note that other summary statistics (for example, combining the variance of both F_{st} and ϕ_{st}) may provide more inferential power on the estimation of this parameter that will be the subject of further investigations.

CONCLUSIONS

We show that for fixed Nm , the carrying capacity of demes play an important role during a RE at both the intra and inter-population levels. Species are unlikely to be at equilibrium. We therefore stress the importance of estimating N and m separately as well as investigating the ancestral phase of the coalescent process. Our results highlight that we can actually disentangle N and m provided many demes and a consistent set of summary statistics is used in the estimation process

(for example, using an ABC algorithm). This would be extremely useful in conservation studies and also when investigating invasive species. Finally, we note that species undergoing recent REs will retain the signature of the ancestral demography that could be uncovered focusing on the genome-wide distribution of population differentiation.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

I thank Laurent Excoffier for stimulating the present study and Oscar Lao for his interesting comments. I thank the anonymous reviewers and the associate editor for greatly improving a previous version of this manuscript. I am grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources. Finally, I thank Erin Koval for revising the English.

-
- Austerlitz F, Jung-Muller B, Godelle B, Gouyon PH (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theor Popul Biol* **51**: 148–164.
- Barbujani G, Sokal RR, Oden NL (1995). Indo-European origins: a computer-simulation test of five hypotheses. *Am J Phys Anthropol* **96**: 109–132.
- Beaumont MA (2010). Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Syst* **41**: 379–406.
- Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19**: 2609–2625.
- Boessenkool S, Austin JJ, Worthy TH, Scofield P, Cooper A, Seddon PJ *et al.* (2009). Relict or colonizer? Extinction and range expansion of penguins in southern New Zealand. *Proc Biol Sci* **276**: 815–821.
- Brown RM, Nichols RA, Faulkes CG, Jones CG, Bugoni L, Tatayah V *et al.* (2010). Range expansion and hybridization in Round Island petrels (*Pterodroma* spp.): evidence from microsatellite genotypes. *Mol Ecol* **19**: 3157–3170.
- Cavalli-Sforza L, Feldman MW (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet* **33** (Suppl): 266–275.
- Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**: 983–995.
- Edmonds CA, Lillie A S, Cavalli-Sforza LL (2004). Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci USA* **101**: 975–979.
- Eriksson A, Manica A (2014). The doubly conditioned frequency spectrum does not distinguish between ancient population structure and hybridization. *Mol Biol Evol* **31**: 1618–1621.
- Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**: e1003905.
- Excoffier L, Foll M, Petit RJ (2009). Genetic consequences of range expansions. *Annu Rev Ecol Syst* **40**: 481–501.
- Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–567.
- Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.
- Francois O, Blum MG, Jakobsson M, Rosenberg NA (2008). Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet* **4**: e1000075.
- Gaggiotti OE, Bekkevold D, Jorgensen HB, Foll M, Carvalho GR, Andre C *et al.* (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution* **63**: 2939–2951.
- Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**: 409–417.
- Hedrick PW (2005). A standardized genetic differentiation measure. *Evolution* **59**: 1633–1638.
- Ibrahim K, Nichols R, Hewitt G (1996). Spatial patterns of genetic variation by different forms of dispersal during range expansion. *Heredity (Edinb)* **77**: 282–291.
- Kimura M, Weiss GH (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561–576.
- Klopfstein S, Currat M, Excoffier L (2006). The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol* **23**: 482–490.

- Meirmans PG (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**: 2399–2402.
- Mona S, Mordret E, Veuille M, Tommaseo-Ponzetta M (2013). Investigating sex-specific dynamics using uniparental markers: West New Guinea as a case study. *Ecol Evol* **3**: 2647–2660.
- Mona S, Ray N, Arenas M, Excoffier L (2014). Genetic consequences of habitat fragmentation during a range expansion. *Heredity (Edinb)* **112**: 291–299.
- Navascues M, Emerson BC (2009). Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. *Mol Ecol* **18**: 4390–4397.
- Neuenschwander S, Lurgiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008). Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**: 757–772.
- Nullmeier J, Hallatschek O (2013). The coalescent in boundary-limited range expansions. *Evolution* **67**: 1307–1320.
- Peter BM, Wegmann D, Excoffier L (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol Ecol* **19**: 4648–4660.
- Ray N, Currat M, Berthier P, Excoffier L (2005). Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* **15**: 1161–1167.
- Ray N, Currat M, Excoffier L (2003). Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* **20**: 76–86.
- Ray N, Currat M, Foll M, Excoffier L (2010). SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* **26**: 2993–2994.
- Rendine S, Piazza A, Cavalli-Sforza LL (1986). Simulation and separation by principal components of multiple demic expansions in Europe. *Am Nat* **128**: 681–706.
- Schneider N, Chikhi L, Currat M, Radespiel U (2010). Signals of recent spatial expansions in the grey mouse lemur (*Microcebus murinus*). *BMC Evol Biol* **10**: 105.
- Sousa V, Peischl S, Excoffier L (2014). Impact of range expansions on current human genomic diversity. *Curr Opin Genet Dev* **29**: 22–30.
- Stadler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Travis JM, Munkemüller T, Burton OJ, Best A, Dytham C, Johst K (2007). Deleterious mutations can surf to high densities on the wave front of an expanding population. *Mol Biol Evol* **24**: 2334–2343.
- Wakeley J (1998). Segregating sites in Wright's island model. *Theor Popul Biol* **53**: 166–174.
- Wakeley J (1999). Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- Wakeley J (2004). Metapopulation models for historical inference. *Mol Ecol* **13**: 865–875.
- Wakeley J, Aliacar N (2001). Gene genealogies in a metapopulation. *Genetics* **159**: 893–905.
- Walther B, Moore J (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28**: 815–829.
- Wegmann D, Currat M, Excoffier L (2006). Molecular diversity after a range expansion in heterogeneous environments. *Genetics* **174**: 2009–2020.
- Wilkins JF (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**: 2227–2244.
- Wilkins JF, Wakeley J (2002). The coalescent in a continuous, finite, linear population. *Genetics* **161**: 873–888.
- Wright S (1931). Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Supplementary Information accompanies this paper on Heredity website (<http://www.nature.com/hdy>)