Heredity (2017) 119, 64–75

www.nature.com/hdy

ORIGINAL ARTICLE

Genetic structure and invasion history of the house mouse (*Mus musculus domesticus*) in Senegal, West Africa: a legacy of colonial and contemporary times

C Lippens[1,7], A Estoup[2,3,6], MK Hima[1,8], A Loiseau[2], C Tatard[2], A Dalecky[1,9], K Bâ[4], M Kane[4], M Diallo[4], A Sow[4], Y Niang[4], S Piry[2], K Berthier[5], R Leblois[2,3], J-M Duplantier[1] and C Brouat[1,6]

Knowledge of the genetic make-up and demographic history of invasive populations is critical to understand invasion mechanisms. Commensal rodents are ideal models to study whether complex invasion histories are typical of introductions involving human activities. The house mouse *Mus musculus domesticus* is a major invasive synanthropic rodent originating from South-West Asia. It has been largely studied in Europe and on several remote islands, but the genetic structure and invasion history of this taxon have been little investigated in several continental areas, including West Africa. In this study, we focussed on invasive populations of *M. m. domesticus* in Senegal. In this focal area for European settlers, the distribution area and invasion spread of the house mouse is documented by decades of data on commensal rodent communities. Genetic variation at one mitochondrial locus and 16 nuclear microsatellite markers was analysed from individuals sampled in 36 sites distributed across the country. A combination of phylogeographic and population genetics methods showed that there was a single introduction event on the northern coast of Senegal, from an exogenous (probably West European) source, followed by a secondary introduction from northern Senegal into a coastal site further south. The geographic locations of these introduction sites were consistent with the colonial history of Senegal. Overall, the marked microsatellite genetic structure observed in Senegal, even between sites located close together, revealed a complex interplay of different demographic processes occurring during house mouse spatial expansion, including sequential founder effects and stratified dispersal due to human transport along major roads.

Heredity (2017) **119,** 64–75; doi:10.1038/hdy.2017.18; published online 29 March 2017

INTRODUCTION

The reconstruction of invasion histories is crucial to understand the ecological and evolutionary processes underlying invasions (Estoup and Guillemaud, 2010). One of the main features to have emerged from several well-documented examples is that invasion histories involving human activities are often far more complex than initially thought, with multiple introductions, bridgehead effects and stochastic processes leading to the development of a genetic structure within invaded areas that is difficult to predict (see, for example, Lombaert *et al.*, 2014). Although human travels and trade have always facilitated the dispersal of other organisms, most of these case studies have concerned recent introductions (Dlugosch and Parker, 2008).

Commensal rodents are ideal models to study the complexity of invasion histories over different timescales, as these animals have been dispersing with humans since Neolithic times (Jones *et al.*, 2013). The house mouse, *Mus musculus domesticus*, in particular, is recognized as a major invasive taxon (http://www.issg.org/database/) having dramatic impacts on biodiversity, human health and human activities (Singleton *et al.*, 2003). This subspecies from the *Mus musculus* complex originates from South-West Asia (Suzuki *et al.*, 2013), and became commensal during the initial settlements of humans in the Middle East at ~10 000 BC (Cucchi *et al.*, 2012). The distribution range of *M. m. domesticus* then expanded, probably thanks to increasing human trade, around the Mediterranean Sea during the Iron Age (Cucchi *et al.*, 2012). The subspecies subsequently spread to North-West Europe during the Viking era, and then to much of the rest of the world following the Age of Discovery (Jones *et al.*, 2013).

Recent phylogeographic studies have described the past and recent colonization histories of *M. m. domesticus* in Europe (see, for example, Bonhomme *et al.*, 2011), and in some islands (see, for example, Gray *et al.*, 2014). However, only a few historical records (Dalecky *et al.*, 2015 and references therein) and a few genetic data (a unique

[1] Ird, CBGP (UMR INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, Montferrier sur Lez, France; [2] Inra, CBGP (UMR INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, Montferrier sur Lez, France; [3] Institut de Biologie Computationnelle, Montpellier, France; [4] Ird, CBGP (UMR INRA/IRD/Cirad/Montpellier SupAgro), Campus ISRA/IRD de Bel Air, Dakar, Senegal and [5] INRA, UR0407 Pathologie Végétale, Montfavet, France
[6] These two authors contributed equally to this work.
[7] Current address: Université de Bourgogne, Laboratoire BioGeoSciences UMR 6282, Dijon, France.
[8] Current address: Université Abdou Moumouni, Faculté des Sciences et Techniques, Département de Biologie, Niamey, Niger.
[9] Current address: Ird, Aix Marseille Univ, LPED, Marseille, France.
Correspondence: Dr C Brouat, CBGP, IRD, Campus International Baillarguet, CS 30016, Montferrier sur Lez, 34988, France.
E-mail: carine.brouat@ird.fr
Received 3 October 2016; revised 24 February 2017; accepted 25 February 2017; published online 29 March 2017

population sample from Cameroon (Ihle *et al.*, 2006); 12 individuals from Senegal (Bonhomme *et al.*, 2011)) were available to document the evolutionary history of *M. m. domesticus* in Africa. The house mouse may have been present in West Africa since the arrival of Portuguese sailors in fifteenth century (Rosevear, 1969). In Senegal, which was a focal area for European settlers, large and stable populations of house mice have been described in the colonial cities along the Atlantic coast since the middle of the nineteenth century (Dalecky *et al.*, 2015 and references therein). Following the development of human transport, the subspecies has spread further inland since the twentieth century. Its range now covers the northern half of the country, and is still expanding (Granjon and Duplantier, 2009; Dalecky *et al.*, 2015; BPM Database: http://vminfotron-dev.mpl.ird.fr/bdrss/index.php).

The aim of this study was to decipher the invasion history and spatial demographic dynamics of *M. m. domesticus* in Senegal, and to assess the consequences of human history in shaping neutral genetic variation of this subspecies in its expanding range. We used two different types of genetic markers to characterize the genetic variation of the house mouse: sequences from the mitochondrial DNA control region (D-loop) and 16 nuclear microsatellites. The D-loop is the only molecular marker for which substantial data are available over the entire distribution of the house mouse (see, for example, Bonhomme *et al.*, 2011). It is therefore a useful marker for investigations of the exogenous origin of this subspecies. Microsatellites provide more detail about introduction history and the spatial expansion processes at work within the invaded area. We first carried out classic phylogenetic and population genetics analyses on an extensive sample set covering the entire distribution area in Senegal. We placed the D-loop data in a wider context, by including a large set of previously published sequences in the phylogenetic analyses. Approximate Bayesian computation methods (ABC) were then applied to the microsatellite data in order to compare different introduction scenarios and estimate several parameters of interest, such as introduction time.

MATERIALS AND METHODS

Sample collection and laboratory analyses

In Senegal, the distribution of the house mouse is restricted to villages and towns along four main roads (Dalecky *et al.*, 2015) in the north (the northern road), the centre of the country (the central road), within the Ferlo region (the Ferlo road) and along the coast (the coastal road; Figure 1). Between 2011 and 2013, house mice (target sample size: 20 individuals) were sampled by live trapping in 36 human settlements (villages or cities, hereafter referenced as sites) along these main roads (Figure 1), according to a standardized protocol described by Dalecky *et al.* (2015) (but see also Supplementary Table S1). Fieldwork was carried out under the framework agreement established between the Institut de Recherche pour le Développement and the Republic of Senegal, as well as with the Senegalese Head Office of Waters and Forests. Handling procedures were performed under our lab agreement for experiments on wild animals (no. 34-169-1), and followed the official guidelines of the American Society of Mammalogists (Sikes *et al.*, 2011). Trapping campaigns within houses were systematically performed with prior explicit agreement from relevant local authorities. Each captured mouse was killed by cervical dislocation, weighed, measured and necropsied. A piece of liver was stored in 95% ethanol for molecular analyses.

The complete D-loop sequence was amplified for 119 mice sampled from distant houses within each of the studied sites (from 2 to 10 mice per site, Table 1), with the PCR primers and conditions described in Rajabi-Maham *et al.* (2008). PCR products were sequenced in both directions by Eurofins MWG (Ebersberg, Germany).

We genotyped 16 nuclear microsatellite loci (D1Mit291, D2Mit456, D3Mit246, D4Mit17, D4Mit241, D6Mit373, D7Mit176, D8Mit13, D9Mit51, D10Mit186, D11Mit236, D14Mit66, D16Mit8, D17Mit101, D18Mit8 and D19Mit30: available from the MMDBJ database: http://www.shigen.nig.ac.jp/mouse/mmdbj/top.jsp) for the total of 763 mice sampled (including the 119 individuals for which the D-loop was sequenced). The selected loci had perfect dinucleotide motifs, flanking sequences suitable for primer binding and were located on different chromosomes (except D4Mit17 and D4Mit241). They were amplified in three multiplex PCRs (Supplementary Table S2). PCR products were separated and detected with an ABI 3130 automated sequencer (Applied Biosystems, Foster City, CA, USA) and analysed with GeneMapper v.3.7. For each mouse successfully genotyped at some loci but not at others, each failed locus was reamplified by simplex PCR (to prevent primer competition).

Sequence analyses

All 119 D-loop sequences from Senegal were aligned with 1673 sequences retrieved from GenBank (1313 published in Bonhomme *et al.* (2011) and 361 obtained from house mice sampled in Europe, Asia, Oceania and Africa: see references in Figure 3). We used the Multiple Alignment of the Fast Fourier Transform algorithm (MAFFT v.7; Katoh and Standley, 2013). Haplotypes were identified with FaBox v1.41 (Villesen, 2007). Bayesian phylogenetic reconstruction (four chains, burn-in $= 2 \times 10^3$ iterations; chain length $= 2 \times 10^7$ iterations) was performed on all haplotypes with MRBAYES v.3.2 (Ronquist and Huelsenbeck, 2003), with a Hasegawa–Kishino–Yano (Hasegawa *et al.*, 1985) mutational model previously identified as the best model by JmodelTest v.2.1.3 (Darriba *et al.*, 2012) using the Akaike information criterion. A phenogram was then constructed with NETWORK v.4.6.11 (Bandelt *et al.*, 1999) on haplotypes from Senegal in order to illustrate their relative frequencies within the invaded area.

Microsatellite population diversity and structure

Deviations from Hardy–Weinberg equilibrium within loci and sites, and genotypic linkage disequilibrium between pairs of loci, were assessed using GENEPOP v.4 (Rousset, 2008). We corrected for multiple testing by the false discovery rate approach (Benjamini and Hochberg, 1995) implemented in the QVALUE package (Dabney *et al.*, 2011) of R. Heterozygote deficiencies are often found in house mouse populations and are classically attributed to their social system, resulting in subpopulation structuring (Ihle *et al.*, 2006). We analysed the subpopulation structure by calculating the kinship coefficient (ρ) of Loiselle *et al.* (1995) between all pairs of individuals at each site, with SPAGeDI v.1.4 (Hardy and Vekemans, 2002), using the genotype data for each site as the reference for allelic frequencies.

Genetic diversity at each site was estimated with FSTAT v.2.9.3 (Goudet, 2001) by calculating the allelic richness a_r (rarefaction procedure; minimum sample size of 14 diploid individuals), and Nei's unbiased genetic diversity (H_s, Nei, 1987). The mean *M* index (Garza and Williamson, 2001), an indicator of past demographic changes, was calculated across loci for each site with DIYABC v.2.1 (Cornet *et al.*, 2014). Genetic differentiation between sites was summarized by calculating pairwise F_{ST} estimates (Weir and Cockerham, 1984) with FSTAT.

We characterized the spatial genetic structure of mice in Senegal using two approaches. First, we used the clustering approach implemented in STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000) in order to estimate the number of homogeneous genetic groups (*K*) in the data set. The analyses were performed with a model including admixture and correlated allele frequencies (Falush *et al.*, 2003). We performed 20 independent runs for each *K* value (from $K = 1$ to 20). Each run included 500 000 burn-in iterations followed by 1 000 000 iterations. The number of genetic groups was inferred by the delta*K* method applied to the log probabilities of data (Evanno *et al.*, 2005). We checked that a single mode was obtained in the results of the 20 runs for all *K*-values explored, using the Greedy Algorithm implemented in CLUMPP v.1.2.2 (Jakobsson and Rosenberg, 2007). Barplots were finally generated with DISTRUCT v.1.1 (Rosenberg, 2004). Second, we used the spatial Bayesian clustering method implemented in TESS v.2.3.1 (Chen *et al.*, 2007). In TESS, the spatial information considered is a neighbourhood network of the sample sites, obtained from a Dirichlet tessellation of their coordinates. As allowed in TESS, the network was modified in order to delete unrealistic neighbourhood relationships between individuals sampled in sites that are not directly

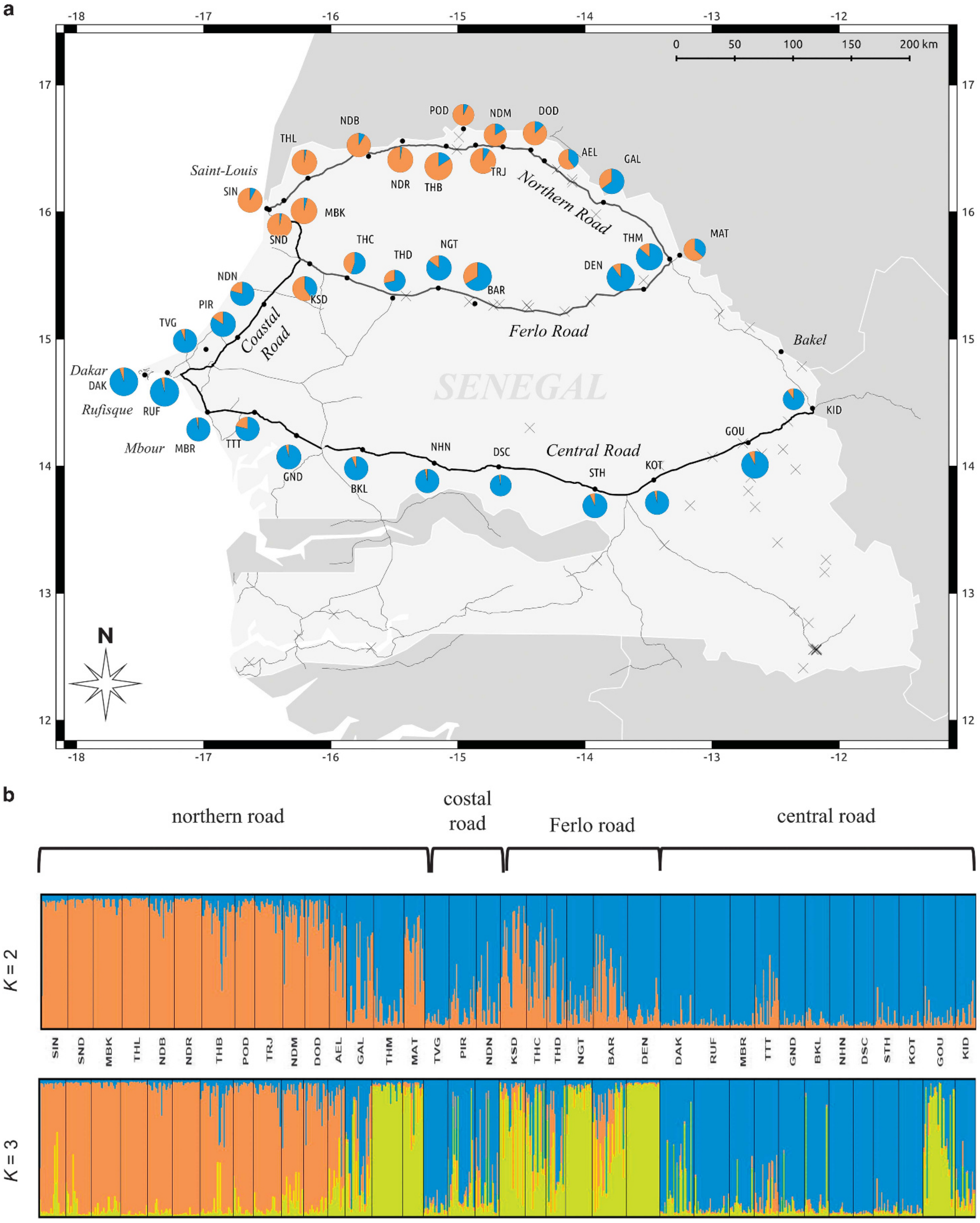

Figure 1 Geographic origin and genetic clustering of sampling sites (code names in Table 1) for *Mus musculus domesticus* in Senegal. (**a**) Geographic distribution of the two main genetic groups ($K=2$) obtained using STRUCTURE. For each site, colours in pie charts indicated the proportions of house mice that were assigned to each genetic group. (**b**) Individual ancestry estimates assuming two or three genetic groups in STRUCTURE. Each vertical line represents an individual, and each colour represents a genetic group. Individuals are grouped by site and sites are ordered along each sampled road according to a west–east gradient. Clustering patterns obtained with TESS were similar (Supplementary Figure S4).

Table 1 Genetic estimates of key statistics within sampling sites of *Mus musculus domesticus* in Senegal

Site	*Code*	n_{mic}	F_{IS}	ρ	a_r	H_S	M	$n_{D\text{-}loop}$	$H_{D\text{-}loop}$
Aere Lao	AEL	14	0.20*	−0.016	3.8	0.59	0.46	2	H1
Barkedji	BAR	28	0.15*	−0.002	3.9	0.57	0.45	3	H5/H8/H13
Mbirkilane	BKL	20	0.15*	−0.015	4.4	0.63	0.47	2	H2
Dakar	DAK	28	0.31*	−0.033	4.5	0.59	0.45	5	H1 /H2
Dendoudi	DEN	27	0.05	−0.008	3.2	0.49	0.46	3	H1
Dodel	DOD	20	0.21*	−0.028	3.9	0.61	0.47	2	H1
Ida Seco	DSC	16	−0.03	−0.024	4.3	0.63	0.46	2	H2
Galoya	GAL	22	0.08	−0.007	4.7	0.61	0.50	4	H1/H4
Gandiaye	GND	21	0.20*	−0.009	4.4	0.61	0.46	2	H1/H2
Goudiri	GOU	26	0.19*	−0.018	4.5	0.61	0.45	7	H3/H9/H14/H15
Kidira	KID	16	0.18*	−0.026	4.4	0.60	0.50	10	H1
Kothiari	KOT	20	0.00	−0.011	3.3	0.48	0.48	2	H1
Keur Seyni Dieng	KSD	21	0.25*	−0.039	4.0	0.59	0.50	3	H1/H5
Matam	MAT	17	0.14	−0.014	4.9	0.67	0.42	3	H1
Mbakhana	MBK	24	0.10	−0.017	4.8	0.65	0.47	3	H1
Mbour	MBR	20	0.18*	−0.017	5.0	0.62	0.42	3	H1
Ndombo	NDB	20	0.19*	−0.018	3.9	0.55	0.47	2	H1
Ndioum	NDM	18	0.02	−0.004	5.0	0.68	0.46	2	H1
Ndande	NDN	20	0.10	−0.005	4.3	0.59	0.49	3	H1/H3
Ndiareme	NDR	23	0.12	−0.009	3.8	0.59	0.56	2	H2
Nguith	NGT	22	0.07	−0.011	3.7	0.53	0.47	2	H1
Niahene	NHN	20	0.10	−0.012	4.3	0.58	0.47	2	H2
Pire	PIR	22	0.11	−0.001	4.8	0.60	0.46	3	H1/H10
Podor	POD	16	0.05	−0.014	4.6	0.66	0.45	4	H1/H4/H6/H7
Rufisque	RUF	29	0.19*	−0.017	4.8	0.59	0.46	6	H1/H2
St Louis Ile Nord	SIN	21	0.22*	−0.022	6.1	0.74	0.47	4	H1/H12
St Louis Sor	SND	21	0.25*	−0.048	5.1	0.67	0.46	3	H1
Sinthiou Maleme	STH	21	0.08	−0.008	4.5	0.63	0.45	2	H1/H2
Thille Boubacar	THB	27	0.13*	−0.005	4.5	0.61	0.47	6	H1/H2
Thiamene Cayor	THC	17	0.21*	−0.029	4.4	0.59	0.45	3	H1
Thiamene Djolof	THD	16	0.16*	−0.016	4.9	0.68	0.46	3	H1
Thilene	THL	22	0.21*	−0.015	4.0	0.57	0.47	3	H1/H2
Thiambe	THM	25	0.19*	−0.006	4.7	0.67	0.46	5	H1/H2/H11
Taredji	TRJ	23	0.10	−0.008	4.7	0.66	0.50	2	H1/2
Tattaguine	TTT	20	0.18*	−0.005	4.7	0.64	0.46	3	H1
Tvine Tangor	TVG	20	0.15*	−0.009	4.7	0.69	0.45	3	H1

The numbers of house mice genotyped for microsatellite markers (n_{mic}) and sequenced for the mitochondrial D-loop ($n_{D\text{-loop}}$) are also indicated. For microsatellite data, the table includes the genetic estimates of F_{IS} (* indicated values significant after correction for multiple testing), the median pairwise kinship coefficients ρ (Loiselle *et al.*, 1995), mean allelic richness for a sample size of 14 individuals (a_r), Nei's unbiased genetic diversity (H_S) and mean *M* index (Garza and Williamson, 2001). For mitochondrial data, the table includes D-loop haplotype names ($H_{D\text{-loop}}$). D-loop haplotype frequencies per site and the GenBank accession numbers of haplotypes are given in Supplementary Table S4 and Supplementary Figure S1.

connected by roads, but similar results were obtained considering the unmodified network (results not shown). We performed 20 independent runs for each *K*-value ranging from 2 to 20, using the admixture model CAR, a burn-in period of 10 000 sweeps followed by 30 000 sweeps and the interaction parameter set to 0.6. The number of genetic groups was inferred using the delta*K* method applied to deviance information criterion. We also used CLUMP to check that a single mode was obtained in the results for each *K*.

Both STRUCTURE and TESS results may be biased because of deviations from Hardy–Weinberg equilibrium. We validated the clustering solution obtained using Discriminant Analysis on Principal Components (DAPC) that is not based on a predefined population genetics model and is thus free from Hardy–Weinberg equilibrium assumptions (Jombart *et al.*, 2010). DAPC was performed using the adegenet package (Jombart, 2008) in R. The consistency of the results was assessed through 10 independent DAPC runs.

Regular loss of genetic diversity along colonization routes is often expected because of the occurrence of successive bottlenecks during the expansion of the range of the colonizing species (Ramachandran *et al.*, 2005). In the context of an invasion, geographic gradients of genetic diversity may thus provide insight into the source populations that were initially introduced. STRUCTURE and TESS analyses identified two main genetic groups (see the Results section). We tested the hypothesis suggested by historical data that the source populations of these two groups were initially introduced into the main colonial cities of Senegal located on the Atlantic coast (Dalecky *et al.*, 2015). To this aim, we performed Spearman's rank correlation analyses between genetic diversity estimates (a_r, H_S) and the longitude (that is closely related to distance from the coast) of the sampled sites for each genetic group.

If each of the genetic groups of Senegalese house mice had an independent origin, we could expect a greater genetic diversity at sites of admixture. We defined the admixture rate as the mean proportion of membership (between 0 and 50%) of each site to the alternate genetic group given by STRUCTURE. We then applied Spearman's rank correlation analysis to the entire data set to assess the relationship between such admixture rates and genetic diversity estimates (a_r, H_S).

Inference of dispersal from microsatellite data

The dispersal of offspring over limited distances from their parents results in an increase in genetic differentiation with geographic distance through a process known as isolation by distance (IBD; Rousset, 1997). We characterized the dispersal patterns of Senegalese house mice by conducting IBD analyses at two different spatial scales: (1) at a local scale, that is, within sites (as in Verdu *et al.*, 2010), to estimate the spatial restriction of dispersal between houses within villages; and (2) at a larger scale, along presumed expansion roads, to evaluate the contribution of long-distance dispersal to the genetic structure.

We used two different inference methods for this purpose. First, for both the local and large scales, we used the regression method based on the expected linear relationship between genetic and geographic distances (Rousset, 1997, 2000). These analyses were run with GENEPOP, using the pairwise genetic differentiation estimator e_r calculated between individuals (Watts *et al.*, 2007) and Euclidean geographic distances between individuals or their logarithms, depending on whether dispersal occurred principally in one dimension (along roads) or in two dimensions (within sites). The minimum distance between sites (3 km) was used as a threshold to exclude pairs of individuals from the same site or from different sites in the analyses performed along roads and within sites, respectively. Mantel tests with 10 000 permutations were performed to assess the correlation between matrices of genetic and geographic distances, with a home-made R script that modified the Mantel test to calculate rank correlation coefficients and to permute the pairwise distances within sites or between individuals from different sites only (script available upon request).

Second, IBD was explored at the large scale by the maximum likelihood method implemented in MIGRAINE that infers model parameters using importance sampling algorithms (de Iorio *et al.*, 2005) extended to consider linear IBD as a model for population structure (Rousset and Leblois, 2007). A geometric distribution is considered for dispersal and a *K* allele model for mutation (Rousset and Leblois, 2007, 2012). MIGRAINE provides point estimates, 95% coverage confidence intervals (CIs) and two-dimensional parameter likelihood profiles for several parameters: the scaled local population size ($\theta = 2 \times N_{\text{genes}} \times \mu$, where N_{genes} is the local population size expressed in number of genes and μ the mutation rate per locus per generation), the scaled emigration rate (number of emigrant per generation: $\gamma = 2 \times N_{\text{genes}} \times m$, where m is the total emigration rate per generation for a local population), the geometric dispersal distribution parameter (g) and neighbourhood size ($N_s = 2 \times D \times \sigma^2$, where D is the density of individuals and σ^2 the mean squared parent–offspring dispersal distance). All MIGRAINE runs were performed under a linear model of IBD (that is, 1D IBD) on the 16 microsatellites with the following computing parameters: 1000 trees, 600 points and 2 iterations. We translated the parameters inferred from MIGRAINE into effective population size (N_{genes}) using the mutation rate commonly used for microsatellites: 5×10^{-4} (Sun *et al.*, 2012).

Inference about introduction scenarios from ABC on microsatellites

ABC analyses were performed on microsatellite data only, for which we had population samples (see Table 1). The small size and large geographical scale of the sampling of mitochondrial DNA variation (that is, a subset of individuals from all sites sampled in Senegal) was clearly not appropriate for ABC analyses that assume Hardy–Weinberg population units. The common practice of pooling differentiated site samples may give misleading results in ABC analyses (Lombaert *et al.*, 2014). Hence, ABC analyses were conducted on sites chosen to be representative of each genetic group identified by the clustering analyses, and known on the basis of rodent community data (collected from the nineteenth century to the present days: see references in Dalecky *et al.*, 2015) to be in the most likely areas of introductions. We chose to test a small number of competing scenarios rather than an exhaustive list to focus computational efforts on well-founded introduction hypotheses. The first and second scenarios involved two introduction events in Senegal, one in the north and the other further south, from two different unsampled ancestral populations (scenario 1, Figure 2a) or from a single unsampled ancestral population (scenario 2, Figure 2b). The third and fourth scenarios involved a single introduction event from a single unsampled ancestral population in a southern (scenario 3, Figure 2c) or in a northern (scenario 4, Figure 2d) coastal site, with a subsequent secondary introduction event from the first introduced population into a northern or southern coastal site, for scenarios 3 and 4, respectively.

Significant genetic population substructure was observed locally within many sampled sites (Table 1). We hence evaluated the potential effect of local substructure within our sampled sites on scenario choice when using ABC treatment in which the absence of local population substructure is assumed within the analysed samples. To this aim we analysed different sets of simulated pseudo-observed data sets characterized by the absence or presence of genetic substructure within samples (Supplementary Appendix S1).

ABC analyses were performed with DIYABC v.2.1 (Cornuet *et al.*, 2014). The prior distributions of the historical, demographic and mutational parameters are described in Supplementary Table S3 (prior distribution set 1, including only uniform priors). Wild house mice are generally thought to have a generation time of 3 months (Nachman and Searle, 1995). Priors for introduction and divergence times were thus defined within the last 2000 generations to encompass the period during which Europeans initially arrived in Senegal (fifteenth century: Sinou, 1993) within the possible values. A second set of prior distributions was used to evaluate the robustness of the ABC inferences to prior choice. It included (1) normal distributions with the same mean and bounds as in prior set 1 for demographic parameters; and (2) logUniform distributions with the same bounds as in prior set 1 for mutation parameters (see Supplementary Table S3, prior distribution set 2).

We summarized the genetic information within and between populations using all single-sample and two-sample summary statistics (that is, 16 summary statistics) available in DIYABC (see p. 16 in the DIYABC user manual, available from http://www1.montpellier.inra.fr/CBGP/diyabc/). In a preliminary study, we evaluated the confidence in the choice of scenario and accuracy of parameter estimation under a given scenario for different sets of summary statistics using DIYABC simulated pseudo-observed data sets (pods) drawn randomly from prior distributions for both the scenario ID and the parameter values. We showed that the use of all summary statistics provided a better discrimination among the tested scenarios without degrading the estimation of parameter values under a given scenario than the more or less arbitrary choice of a subset of statistics.

We simulated 10^6 data sets per scenario, and the posterior probability of each competing scenario was estimated by a polychotomous logistic regression on the 1% of simulated data sets closest to the observed data set. We carried out a linear discriminant analysis transformation of the 16 summary statistics before calculating the logistic regression (Estoup *et al.*, 2012). We then estimated the posterior distributions of demographic parameters under the selected scenario by local linear regression on the 1% of simulated data sets closest to the observed data set (Cornuet *et al.*, 2008). We used raw (that is, non-linear discriminant analysis transformed) summary statistics for this analysis (see, for example, Lombaert *et al.*, 2014).

We evaluated confidence in the choice of scenario and the accuracy of parameter estimation under a given scenario, using simulated pseudo-observed data sets (pods), for which the true scenario identity (ID) and parameter values are known. Pods were simulated from posterior distributions to focus around the observed data set as error and accuracy indicators conditional to the observed data set (that is, from posterior distributions) are clearly more relevant than indicators blindly calculated over the whole prior data space. We used the new option proposed by DIYABC v.2.1 to compute posterior error rates for model choice and posterior accuracy indicators for parameter estimation from sets of 5000 pods (see DIYABC manual p. 5 and sections 3.5.2 and 3.5.5 for details).

Finally, we evaluated a Bayesian equivalent of goodness of fit for the selected scenario using the model checking option of DIYABC. From the 10^6 data sets simulated under the selected scenario, we obtained a posterior sample of 10^4 values from the posterior distributions of parameters through a rejection step based on Euclidean distances and linear regression post treatment (as previously described). We then simulated 10^4 data sets and corresponding summary statistics with parameter values drawn with replacement from this posterior sample. Finally, we ranked the summary statistics for the observed data against those for the simulated data sets. For the model fit to be considered good, the number of observed statistics falling in the margins of the distributions of simulated statistics (that is, statistics with a Proportion (simulated < observed values) < 5 or > 95%) has to be low (that is, < 10% of the 16 summary statistics used here as test statistics).

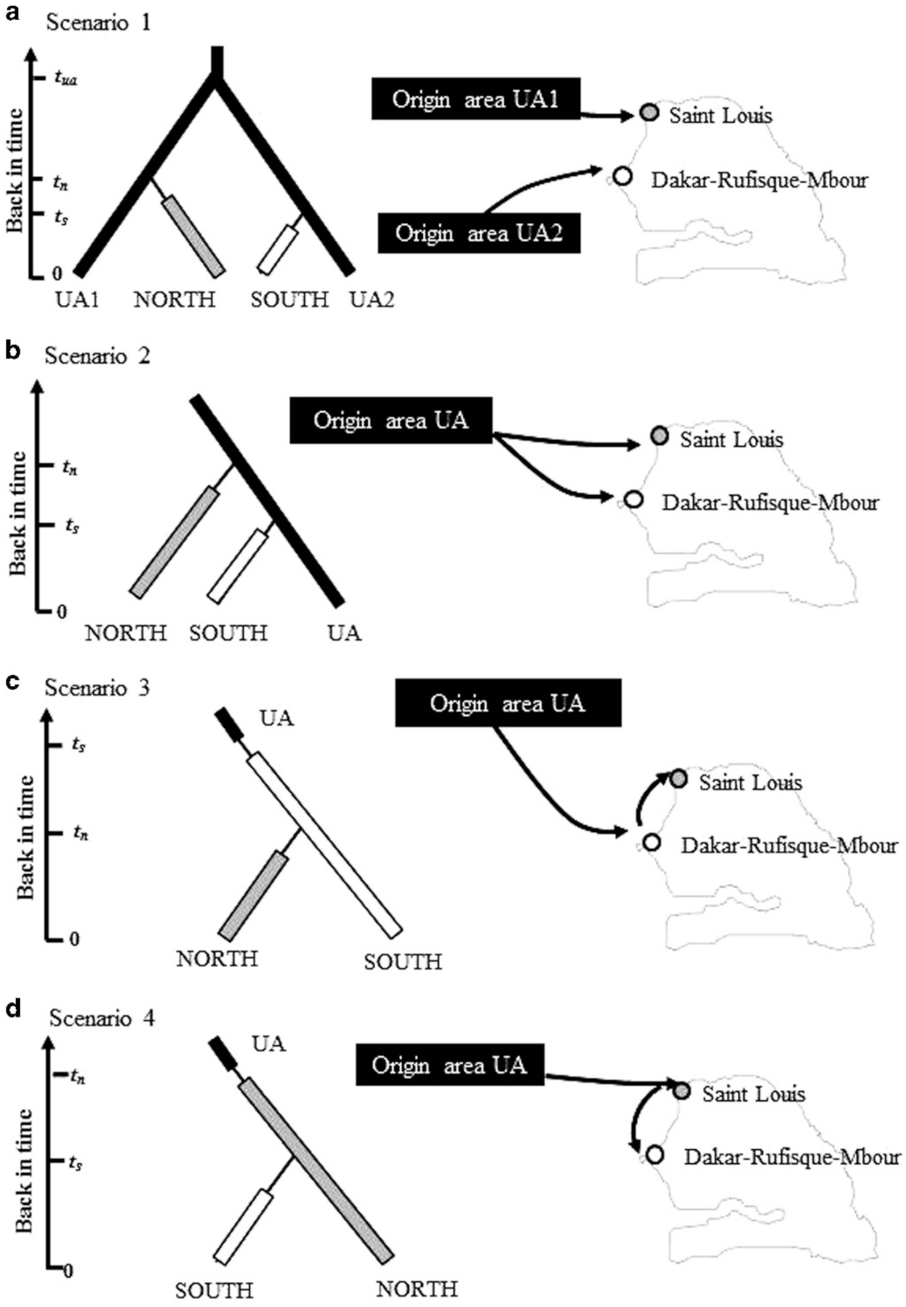

Figure 2 Graphical representation of the four competing introduction scenarios for *Mus musculus domesticus* in Senegal compared by ABC. UA, unsampled ancestral population. Time 0 is the sampling date. The main historical events are represented on the timescale to the left of each scenario. Black, grey and white bars represent different stable effective population sizes in the ancestral populations and in Senegal, and thin lines represent bottleneck events characterized by their own effective number of founders and duration. All parameters and their associated prior distributions are described in Supplementary Table S3. (**a**) In scenario 1, there are two independent introduction events in Senegal from two unsampled ancestral populations UA1 and UA2 that diverged t_a generations ago, one (t_n generations ago) giving rise to the NORTH group, the other (t_s generations ago) giving rise to the SOUTH group; $t_n < t_a$ and $t_s < t_a$. Graphically, t_s is represented as more recent than t_n, but no assumption is actually made about the chronological order of these parameters. (**b**) In scenario 2, there are two independent introduction events in Senegal as in scenario 1, but the populations introduced are considered to originate from a single unsampled population UA. (**c**) In scenario 3, there is a single primary introduction event in the south of Senegal from a single unsampled population UA (t_s generations ago), followed by a secondary introduction event further north t_n generations ago; $t_s > t_n$. (**d**) Scenario 4 also involves a single primary introduction event from a single unsampled population UA, this time in the north of Senegal, followed by a secondary introduction event further south; $t_n > t_s$.

RESULTS

Mitochondrial sequence analysis

In the 119 D-loop sequences obtained, there were only 11 variable sites, defining 15 haplotypes (Supplementary Table S4; mean haplotype diversity $h = 0.51 \pm 0.04$; mean nucleotide diversity $\pi = 0.003 \pm 0.0003$). Two major haplotypes (H1 and H2) were found in 79 and 21 individuals, respectively (Figure 3a). The other haplotypes were separated from H1 or H2 by only a few mutational steps (only

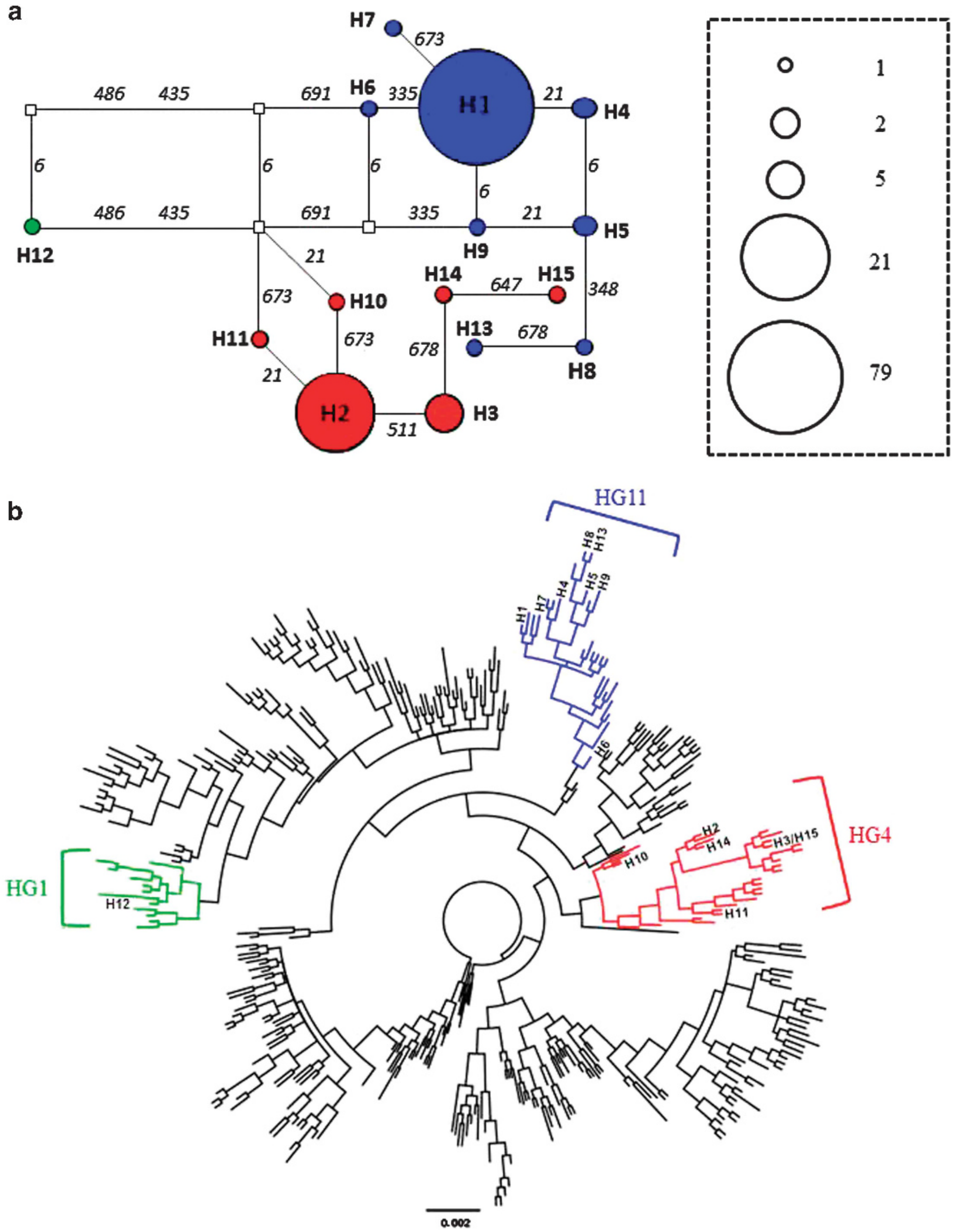

Figure 3 Mitochondrial D-loop haplotypes (701 bp) in Senegalese *Mus musculus domesticus*. (**a**) Median joining network of the 15 D-loop *M. m. domesticus* haplotypes found in Senegal: white squares correspond to nonobserved haplotypes, and blue, red and green symbols correspond to haplotypes from three haplogroups (HG11, HG4 and HG1, respectively) identified by Bonhomme *et al.* (2011). The positions of mutational steps are indicated by the numbers in italics, and symbol size scales are proportional to the number of house mice, as indicated in the legend to the right. (**b**) Phylogenetic tree for the 367 D-loop haplotypes sequences found in *M. m. domesticus.* Haplotypes were identified in a data set containing the 119 sequences from this study, 1313 sequences from Bonhomme *et al.* (2011) and 361 sequences from other studies (Prager *et al.*, 1996, 1998; Gündüz *et al.* 2000, 2001, 2005; Ihle *et al.*, 2006; Searle *et al.*, 2009a, b; Jones *et al.*, 2010; Linnenbrinck *et al.*, 2013; Suzuki *et al.*, 2013; Gabriel *et al.*, 2015; Jones and Searle, 2015). The 15 labelled haplotypes (from H1 to H15) are those found in Senegal. They belong to three haplogroups (in blue: HG11; in red: HG4; in green: HG1) identified by Bonhomme *et al.* (2011). Although haplogroups appear as reasonably cohesive, they are not statistically supported in phylogenetic analyses, as it can be expected from a recent expansion phenomenon (Bonhomme *et al.*, 2011).

1 for 12 of the 19 remaining individuals), except for the more distantly related haplotype H12 that was found in one mouse from Saint-Louis (SIN; Figure 3a). The observed distribution of D-loop haplotypes in Senegal followed no clear geographic pattern (Supplementary Figure S1). Under a HKY85 mutational model, Bayesian reconstruction showed that haplotypes H1, H2 (and the haplotypes derived from them) and H12 belonged to the haplogroups HG11 (or clade E), HG4 (or clade F) and HG1 (or clade C1), respectively, in the nomenclature defined by Bonhomme *et al.* (2011) (or by Jones *et al.*, 2011) (Figure 3b).

Microsatellite genetic diversity and structure

Linkage disequilibrium was significant for 27 of the 4320 tests performed, and hence the 16 loci were considered to be genetically independent. Only three loci (D4Mit241, D11Mit236 and D16Mit8) were at Hardy–Weinberg equilibrium at all sites. All others displayed significant heterozygote deficiencies at most sites. Null alleles were unlikely to explain heterozygote deficiencies, because only a small number of null genotypes were observed (0–5) per locus. Overall, positive F_{IS} values were obtained at 21 sites (Table 1). Within sites, the median kinship coefficient ρ ranged from -0.048 to -0.001 (Table 1). Very high ρ values ($\rho > 0.5$) were obtained for only a few pairs of individuals at some sites (Supplementary Figure S2), indicating the occurrence of some full siblings. Analysis of the restricted data set corresponding to house mice captured in different buildings only (540 individuals) yielded similar results for deviations from Hardy–Weinberg equilibrium, F_{IS} and ρ values (results not shown), suggesting that buildings were not the relevant units for defining genetic subgroups within sites.

Allelic richness (a_r) ranged from 3.2 to 6.1 alleles (mean 4.4 ± 0.5) and H_S from 0.48 to 0.74 (mean 0.61 ± 0.05). Mean values of the M index (between 0.42 and 0.56) were all consistent with a bottleneck signal (<0.68; Garza and Williamson, 2001). Pairwise F_{ST} values (Supplementary Table S5) ranged from 0.05 to 0.34, with a global mean F_{ST} value of 0.19 (95% CI = 0.17–0.21). Substantial genetic structure was observed even between sites that were geographically close together (Supplementary Table S5).

Spatial genetic structure was first characterized with STRUCTURE. The highest delta$K$ value was that for $K = 2$ (Supplementary Figure S3a). At $K \geqslant 4$, there was no congruence between the 20 runs for each K. At $K = 2$, sites along the northern road between SIN and AEL were largely assigned to a first group, whereas those along the central and coastal roads were largely assigned to a second group (Figure 1). House mice from the Ferlo road and from eastern sites (GAL, THM, MAT) had a variable mixed inferred ancestry (Figure 1). At $K = 3$, the genetic groups corresponding to the northern route between SIN and AEL, on the one hand, and the central and coastal roads, on the other hand, remained mostly unchanged. The third group corresponded to individuals from the GOU site, from the Ferlo road (between KSD and DEN) and from the eastern sites of GAL, THM and MAT that were admixed at $K = 2$ (Figure 1b).

Using TESS, the highest delta$K$ value was that for $K = 3$ (Supplementary Figure S3b). Note that the delta$K$ value cannot be calculated for $K = 2$, as it is not possible to run a TESS analysis for $K = 1$. At $K = 2$ and 3, the clustering patterns were identical among runs and similar to those obtained using STRUCTURE (Supplementary Figure S4). For $K \geqslant 4$, there was no congruence between the different runs for each K value, as observed with STRUCTURE.

The DAPC clustering pattern at $K = 2$ was similar to those obtained with STRUCTURE and TESS (Supplementary Figure S4). Some differences concerned THM, MAT and DEN (eastern sites), and sites between KSD and BAR (along the Ferlo road) that had a variable mixed inferred ancestry in STRUCTURE and TESS (see Supplementary Figure S4 in the Supplementary Material). These inconsistencies may result from admixture effects that cannot be accounted for in the DAPC.

In summary, clustering analyses identified two main genetic groups: the NORTH group, mostly located along the northern road between SIN and AEL, and the SOUTH group, mostly distributed along the central road (Figure 1). Other sites (along the coastal and Ferlo roads, and the eastern-most sites along the northern road) displayed variable levels of admixture between the two groups. There was a tendency for allelic richness a_r to decrease with increasing longitude for sites along the northern road between SIN and AEL (Spearman's rank correlation coefficient: $r_s = -0.51$, $P = 0.09$), and along the central road ($r_s = -0.56$, $P = 0.06$). No significant correlation was observed between longitude and H_S for sites between SIN and AEL ($P = 0.59$), and along the central road ($P = 0.89$). No relationship was found between admixture rate and a_r ($P = 0.90$) or H_S ($P = 0.65$) calculated for all sites.

Two-dimensional IBD was significant within sites (Mantel test: $P < 0.0001$; slope $b = 0.038$, 95% CI = 0.034–0.049). The slope of the IBD regression line provides a robust estimator of $1/4\pi D\sigma^2$, the inverse of neighbourhood size (Rousset, 1997,2000). From the inferred slope, we calculated that $D\sigma^2 = 6.5$ (5.0–7.4). Using a rough estimate of $D = 100$ house mice per km^2 (based on the mean number of households occupied by house mice and the mean surface area of the sampled sites: data not shown), we obtained an estimate of $\sigma = 255$ m.

In contrast, linear IBD patterns were very weak and were globally nonsignificant for between sites analyses along the northern road (Mantel test: $P = 1$; slope $b = 9.5 \times 10^{-8}$, 95% CI $= 5.2 \times 10^{-8}$ to 1.5×10^{-7}), and along the central road (Mantel test: $P = 1$; slope $b = 2.0 \times 10^{-8}$, 95% CI $= -1.2 \times 10^{-8}$ to 5.3×10^{-8}). Slope values (the inverse of the neighbourhood size estimates) gave σ values of 16 and 35 km for the north and central roads, respectively.

Similar inferences emerged from MIGRAINE between sites along the northern and central roads (Table 2). Very high neighbourhood size values (N_s) were inferred by MIGRAINE, indicating weak IBD patterns and, therefore, frequent long-distance dispersal events. In addition, the island model (corresponding to $g = 1$) was not rejected for either the northern or the central road, consistent with a lack of spatial restriction of dispersal. The numbers of mice per village were calculated from estimates of scaled population size (θ) (126 (104–152) and 108 (86–128) mice per village for the northern and central roads, respectively) and were hence close to our rough estimate of 100 mice per village.

ABC inferences about introduction scenarios

The introduction history of the NORTH and SOUTH genetic groups was studied using ABC. Six ABC analyses were processed independently with pairs of sites corresponding to the major colonial cities of the coast, from the NORTH (St Louis: SIN or SND) and SOUTH (Dakar: DAK; Rufisque: RUF, or Mbour: MBR) groups (Table 3).

For all six sample pairs considered, scenario 4 consistently had the highest posterior probability (for example, $P = 0.89$ for the SIN-RUF sample pair, 95% CI = 0.888–0.897; Table 3). This scenario involves a primary introduction event on the northern part of the coast from a single unsampled population, and subsequent divergence due to a secondary introduction event from Northern Senegal to a coastal site

Table 2 Isolation-by-distance (IBD) parameters estimated using MIGRAINE for house mice sampled along their main dispersal axes in Senegal

Parameter	*Northern road*	*Central road*
θ	0.125 (0.10 to 0.15)	0.107 (0.09 to 0.13)
γ	3.84 (3.5 to 4.2)	5.24 (4.66 to 6.16)
g	1 (0.98 to 1)	1 (0.98 to 1)
N_s	5.1×10^{12} (2.5×10^{8} to 5.5×10^{12})	9.3×10^{12} (1.3×10^{8} to 1.3×10^{13})

Analyses were performed for sites distributed along the northern and central routes (see Figure 1). Point estimates and 95% confidence intervals are provided for θ (the scaled local population size), γ (the scaled emigration rate), g (the parameter of the geometric dispersal distribution) and N_s (the neighborhood size).

Table 3 ABC model choice results for the introduction history of *Mus musculus domesticus* in Senegal

Sample pair	P *(S4) (95% CI)*	P *(S3+S4) (95% CI)*	*Posterior error rate (among S1, S2, S3, S4)*	*Posterior error rate (S1+S2 vs S3+S4)*
SIN+RUF	0.893 (0.888; 0.898)	0.959 (0.951; 0.967)	0.232	0.081
SIN+MBR	0.687 (0.637; 0.699)	0.777 (0.756; 0.796)	0.201	0.074
SIN+DAK	0.766 (0.756; 0.776)	0.898 (0.881; 0.914)	0.221	0.074
SND+RUF	0.593 (0.580; 0.605)	0.756 (0.736; 0.775)	0.366	0.169
SND+MBR	0.628 (0.619; 0.637)	0.849 (0.832; 0.865)	0.371	0.168
SND+DAK	0.583 (0.570; 0.596)	0.749 (0.728; 0.770)	0.296	0.089
Mean	0.691	0.831	0.281	0.109

Abbreviations: ABC, approximate Bayesian computation; CI, confidence interval.
For all six pairs of samples considered, scenarios including a single primary introduction event followed by a secondary introduction event (scenarios 3 and 4) were clearly more supported (cf., *P* (S3+S4)) than scenarios including two primary introduction events (scenarios 1 and 2). More specifically, scenario 4 (primary introduction in the North) was in all cases the best supported scenario (cf., *P* (S4)) with 95% CIs of probabilities that did not overlap with those of other scenarios. Posterior error rates are presented when choosing among each four scenarios (among S1, S2, S3, S4) or between scenarios 1+2 and 3+4 (S1+S2 *vs* S3+S4). The different pairs of sampled sites (codes in Table 1 and Figure 1) are representative of the introduction areas in the North (SIN or SND) and in the South (RUF, MBR or DAK). All analyses have been processed assuming the prior set 1. Similar probabilities (*P*) and error rates were obtained when using the prior set 2 (results not shown).

further south. The second-best scenario was scenario 3 (for example, $P=0.066$ for the SIN-RUF sample pair, 95% CI = 0.062–0.069) that also involved a single introduction event but occurring on the southern part of the coast. The data provided weaker support for scenarios involving two independent introduction events (for example, scenario 1: $P=0.002$ (0.001–0.002); scenario 2: $P=0.04$ (0.036–0.042), for the SIN-RUF sample pair). Simulation-based analyses showed that the conclusion about the most likely scenario among the four compared scenarios is not challenged by the level of local population substructure observed in the present study (Supplementary Appendix S1).

Posterior error rates are presented in Table 3 for the choice among the four scenarios considered individually or between scenarios 1+2 (scenarios including a single primary introduction event) and scenarios 3+4 (scenarios including two primary introduction events). Posterior error rates were relatively low (that is, ~10%) for the choice between scenarios 1+2 and 3+4, but were substantially higher (that is, ~30%) for the choice among the four scenarios considered independently. Thus, confidence in the choice between scenarios 1, 2, 3 and 4 in the vicinity of the observed data set is rather poor, whereas simply choosing between histories involving a single primary introduction event versus histories involving two independent primary introduction events has more statistical support.

When the model checking option of DIYABC was applied with the selected scenario 4 and associated parameter posterior probabilities, we found that none of the 16 summary statistics used as test quantities had a low tail-probability value (that is, $0.05<P<0.95$ for all test quantities; Supplementary Table S6). The inferred scenario–posterior combination therefore provides a good fit to the observed data set. Accordingly, the projections of the simulated data sets onto the principal component axes for the tested scenario–posterior combination were relatively well grouped and centred on the target point corresponding to the observed data set (Supplementary Figure S5).

ABC inference about demographic and historical parameters

We inferred the posterior distributions of demographic parameters under scenario 4. The ABC analyses reported below concerned the SIN (for NORTH) and RUF (for SOUTH) sites, but the other sample pairs provided similar results (data not shown). For most parameters, the estimated posterior distributions were not much more informative than the priors (Supplementary Tables S3 and S7 and Supplementary Figure S6 for prior set 1; data not shown for prior set 2). Consistent with this finding, the RMedAD values obtained from pods were similar to those calculated as base level from prior information only (that is, without genetic information) for all parameters, including the introduction times for the NORTH and SOUTH groups (Supplementary Table S8). More information was obtained for composite parameters (Supplementary Table S8), but it remains difficult to interpret these estimates biologically. Finally, we found that each introduction event was followed by a demographic bottleneck that was less intense for the primary introduction in the north (median bottleneck intensity $tb_n/Nb_n=0.26$) than for the secondary introduction in the south ($tb_s/Nb_s=0.54$; Supplementary Table S7 and Supplementary Figure S6).

DISCUSSION

In this study, we aimed to investigate the invasion history of *M. m. domesticus* in Senegal by characterizing its genetic structure with both mitochondrial sequences and microsatellite markers. We wanted (1) to evaluate whether the introduction history of the subspecies and its spatial demographic dynamics are consistent with human history at colonial and contemporary times and (3) to give insights into the evolutionary processes that may underlie the invasions of commensal rodents.

Introduction history

We found some evidence from D-loop data that a small group of mice was introduced in Senegal in a single main introduction event. Only two major haplotypes (H1 and H2) were found in Senegal, and mean haplotype and nucleotide diversities in Senegal ($h=0.51$ and $\pi=0.003$) were substantially lower than those for the house mice of Western Europe (mean *h* from 0.82 to 0.95; mean π from 0.002 to 0.008; see, for example, Rajabi-Maham *et al.*, 2008; Searle *et al.*, 2009b; Jones *et al.*, 2011; Gabriel *et al.*, 2015) or from invaded areas after multiple introductions (mean *h* from 0.66 to 0.91; mean π from 0.004 to 0.01; see, for example, Searle *et al.*, 2009a; Gabriel *et al.*, 2015). In addition, both major haplogroups were found at coastal sites, and no geographic pattern was observed in the distribution of haplotypes across Senegal. These features are consistent with the presence of ancestral polymorphism in a single initial introduction area, with a subsequent spatial spread inland.

Microsatellite data also suggested that there had been a single primary introduction event in Senegal. Consistent with the notion that one of the two main genetic groups spreading in Senegal originated from the other, we found no relationship between admixture levels and genetic diversity within sites. ABC analyses provided more statistical support for scenarios involving a single primary introduction event than those involving two independent introduction events

(Table 3). More specifically, the best scenario selected by ABC (scenario 4 in Figure 2), which involves a single primary introduction event in northern Senegal, was repeatedly selected in each of the six ABC analyses carried out, despite substantial differentiation between the sites chosen as representative of each genetic group. This suggests that we can be confident in the selection of scenario 4, despite high posterior error rates associated with this choice.

It remains a challenge to finely identify the origin of the first introduced house mouse population in Senegal. Both microsatellite and historical data suggest that Saint Louis, the first colonial port to be developed in Senegal (Sinou, 1993) and a major colonial city involved in the trading of slaves and Arabic gum during the eighteenth century (Bonnardel, 1992), might be the putative area of introduction. France was involved in the establishment of Saint Louis, and the British controlled the city for 80 years during the eighteenth century (Sinou, 1993). Unfortunately, the lack of precise information about introduction times provided by ABC makes it impossible to evaluate the consistency of these times with historical data. All the mitochondrial haplogroups found in Senegal are typical from Western Europe (Bonhomme *et al.*, 2011). Both major mitochondrial haplotypes (H1 and H2) and their closely related haplotypes have been reported at relatively high frequencies (>10%) in not only Western France and Great Britain, but also Germany (H1), Norway (H2) or Morocco (H1) (Bonhomme *et al.*, 2011; Linnenbrinck *et al.*, 2013; Supplementary Table S9). The unique haplotype H12 from haplogroup HG1 was found at high frequency (>20%) in Southern France and Portugal (Bonhomme *et al.*, 2011). Nevertheless, D-loop data for Europe are sparse and concern sites with no particular connection to colonial history. D-loop and microsatellite data from house mouse populations located close to major harbours historically involved in trade with Senegal (such as Nantes or Bordeaux in France, Liverpool in Britain) may facilitate identification of the precise Western European source of the mouse populations of Senegal.

At first glance, the known occurrence of H1 in Morocco (Bonhomme *et al.*, 2011) might suggest another scenario of colonization by a continental route from North-West Africa to Senegal. This scenario would be unlikely, however, to explain the primary distribution area of the house mouse in Senegal that was shown to be restricted to coastal villages and towns (Dalecky *et al.*, 2015). Indeed, several hundred km separated Senegalese mouse populations from the nearest populations further north (Granjon and Duplantier, 2009), and the trade between Senegal and North-West Africa did not historically occur via the Atlantic coast, but via inland sites (Miège, 1981).

Spatial expansion

Historical data and longitudinal surveys of commensal rodent communities in Senegal have suggested that the spread of house mice in Senegal is recent (twentieth century) and related to the development of road traffic (see Dalecky *et al.*, 2015 and references therein). Indeed, mouse populations would first have become established in villages and towns on the coast, possibly because of the development of railway trade between St Louis and Dakar at the end of the nineteenth century (Bonnardel, 1992). Genetic admixture between the NORTH and SOUTH groups would have occurred in this area before expansion to the east with the development of asphalt roads inside the country.

In the context of biological invasions, spatial expansion is often linked to high levels of gene flow that may minimize population structure and IBD patterns (Marrs *et al.*, 2008). It may also be characterized by sequential founder events, leading to strong genetic structure and spatial decrease of allelic diversity along the colonization axis (Clegg *et al.*, 2002). Founder events may strongly limit or at least delay the rise of the IBD pattern because of independent changes in allele frequencies at each introduction. In Senegalese house mice, substantial genetic structure was observed in the analysis of microsatellites, even within the main genetic groups identified by Structure and TESS, indicating that founder events may have occurred repeatedly during the expansion process. Mean M values are consistent with bottleneck signals and decreases in allelic richness along the main expansion road of each genetic group from the coast further suggested serial founder events during expansion (Ramachandran *et al.*, 2005).

At the local geographic scale (that is, within sites), IBD was significant and associated with low estimates of neighbourhood size, reflecting the spatial limitation of dispersal. These results are consistent with the scarce estimates of home ranges of a few tens of metres reported to date for commensal house mice (Pocock *et al.*, 2005). However, the occurrence of long-distance dispersal events over a larger spatial scale is clearly suggested by IBD analyses between sites along the northern and central roads, showing large neighbourhood size estimates. Genetic signatures involving both local diffusion and long distance dispersal are often observed in invasive species with a limited capacity of autonomous dispersal but with many opportunities for passive dispersal by humans (Marrs *et al.*, 2008). This seems to be the case for the house mouse that is generally thought to display active dispersal over only short distances (Pocock *et al.*, 2005). Anthropogenic dispersal probably occurs both between neighbouring villages and over large distances, as mice can take advantage of even small vehicles to disperse.

Estimated values of σ given by IBD regression analyses are compatible with the size of the attraction area of villages having weekly rural markets in Senegal (about 10–20 km: Ninot, 2003) that may be viewed as ‘invasion hubs’ for the mouse towards geographically close villages (Dalecky *et al.*, 2015). The occurrence of long-distance dispersal events in Senegal is highlighted by the assignment of eastern sites along the northern road to the SOUTH genetic group. The distinguishing feature of these eastern sites is to be inhabited by families of human emigrants sending sufficient financial resources to pay for large amount of goods to be brought in directly from Dakar (Bredeloup, 1997), creating opportunities for long-distance transport of mice. Another example is provided by the genetic grouping of individuals from GOU, THM and MAT at $K=3$ in STRUCTURE. This grouping could be explained by the past transport of goods between these sites, before the construction of an asphalt road between Bakel and Kidira (KID) (Kayser and Tricart, 1957).

Evolutionary processes underlying invasions

Multiple introductions leading to genetic admixture in the introduced populations may play an important role in invasion success (Kolbe *et al.*, 2004). We did not formally test the hypothesis that the first population of house mice introduced in Senegal was a pool of individuals from multiple differentiated European sites, as we wished to focus on a limited number of competing scenarios. The predominance of two mitochondrial haplogroups in Senegal (HG4 and HG11) suggests that two maternal lineages were introduced, but these two lineages may have originated from the same site in Western Europe. Little evidence of multiple introductions is generally found in house mouse populations from remote islands (see, for example, Gabriel *et al.*, 2015). This supports behavioural studies suggesting that once established, populations of mice are substantially closed to immigration of conspecifics (Palanza *et al.*, 1996). This may also account for the marked microsatellite genetic structure observed in Senegal, even between sites located close together.

A similar pattern involving a small number of successful introduction events was found for the tropical fire ant that invaded the Old World as a result of Spanish colonial trade (Gotzek *et al*., 2015). These (and others) undoubtedly successful invasions provide support for the notion that multiple introductions are not key events explaining the expansion of introduced populations (Dlugosch and Parker, 2008). As suggested by Dlugosch *et al*. (2015), further research is needed to identify the genetic basis of adaptation allowing spread into new areas, even in the presence of close competitors.

DATA ARCHIVING

DNA sequences: GenBank accession nos KY686322–KY686440. Microsatellite genotypes and final sequence assemblies for D-loop haplotypes: data available from the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.n0n60

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all the people in Senegal who welcomed us into their homes for the purposes of small mammal trapping, and the Senegalese Head Office of Waters and Forests for allowing us to do fieldwork and to export samples. Molecular data were generated at the molecular biology platform of the CBGP and the technical facilities of the LabEx CEMEB. Part of this work was carried out with the computing grids of the INRA MIGALE (http://migale.jouy.inra.fr/), GENOTOUL (Toulouse Midi-Pyrénées), Montpellier Bioinformatics Biodiversity (MBB) and CBGP HPC (with technical assistance from A Dehne-Garcia) platform services. We thank J-M Cornuet, J-M Marin, C Robert, P Pudlo, E Jousselin and F Rousset for fruitful discussions about the analyses. Funding was provided by the Institut de Recherche pour le Developpement (IRD) and the ANR ENEMI project (ANR-11-JSV7-0006).

Bandelt H-J, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc B* **57**: 289–300.

Bonhomme F, Orth A, Cucchi T, Rajabi-Maham H, Catalan J, Boursot P *et al.* (2011). Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. *Proc Roy Soc Lond B Biol* **278**: 1034–1043.

Bonnardel R (1992). *Saint Louis du Sénégal: mort ou naissance?*. L'Harmattan: Paris.

Bredeloup S (1997). Migrants et politiciens à Ouro Sogui (moyenne vallée du fleuve Sénégal): pour quelle dynamique urbaine?. In: Bertrand M, Dubresson A (eds). *Petites et moyennes villes d'Afrique noire.* Karthala: Paris, pp 279–304.

Chen C, Durand E, Forbes F, François O (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol Ecol Notes* **7**: 747–756.

Clegg SM, Degnan SMK, Kikkawa J, Moritz C, Estoup A, Owens APF (2002). Genetic consequences of sequential founder events by an island-colonizing bird. *Proc Natl Acad Sci USA* **99**: 8127–8132.

Cornet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gauthier M, Leblois R *et al*. (2014). DIYABC v2.0: a software to make approximate bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30**: 1187–1189.

Cornuet JM, Santos F, Beaumont MA, Robert C, Marin JM, Balding DJ *et al.* (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate bayesian computation. *Bioinformatics* **24**: 2713–2829.

Cucchi T, Auffray JC, Vigne J-D (2012). On the origin of the house mouse synanthropy and dispersal in the Near East and Europe: zooarchaeological review and perspectives. In: Macholán M, Baird SJE, Munclinger P, Piálek J (eds). *Evolution of the House Mouse.* Cambridge University Press: Cambridge, pp 65–93.

Dabney A, Storey JD, Warnes GR (2011). Qvalue: Q-value Estimation for False Discovery Rate Control. R package, R Foundation for Statistical Computing: Vienna, Austria.

Dalecky A, Bâ K, Piry S, Lippens C, Diagne CA, Kane M *et al.* (2015). Range expansion of the invasive house mouse *Mus musculus domesticus* in Senegal, Western Africa: a synthesis of trapping data over three decades, 1983-2014. *Mammal Rev* **45**: 176–190.

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**: 772.

de Iorio M, Griffiths RC, Leblois R, Rousset F (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor Popul Biol* **68**: 41–53.

Dlugosch KM, Parker IM (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol Ecol* **17**: 431–449.

Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD (2015). The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Mol Ecol* **24**: 2095–2111.

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

Estoup A, Lombaert E, Marin J-M, Guillemaud T, Pudlo P, Robert CP *et al.* (2012). Estimation of demo-genetic model probabilities with approximate bayesian computation using linear discriminant analysis on summary statistics. *Mol Ecol Res* **12**: 846–855.

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

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Gabriel SI, Mathias ML, Searle JB (2015). Of mice and the 'Age of Discovery': the complex history of colonization of the Azorean archipelago by the house mouse (*Mus musculus*) as revealed by mitochondrial DNA variation. *J Evol Biol* **28**: 130–145.

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

Gotzek D, Axen HJ, Suarez AV, Helms Cahan S, Shoemaker D (2015). Global invasion history of the tropical fire ant: a stowaway on the first global trade routes. *Mol Ecol* **24**: 374–388.

Goudet J (2001). *FSTAT: A Program to Estimate and Test Gene Diversities and Fixation Indices* (Version 2.9.3). Available from http://www2.unil.ch/popgen/softwares/fstat.htm.

Granjon L, Duplantier JM (2009). *Les rongeurs de l'Afrique sahélo-soudanienne*. IRD Editions/Publications scientifiques du Muséum: Marseille.

Gray MM, Wegmann D, Haasl RJ, White MA, Gabriel SI, Searle JB *et al.* (2014). Demographic history of a recent invasion of house mice on the isolated Island of Gough. *Mol Ecol* **23**: 1923–1939.

Gündüz I, Tez C, Malikov V, Vaziri A, Polyakov AV, Searle JB (2000). Mitochondrial DNA and chromosomal studies of wild mice (*Mus*) from Turkey and Iran. *Heredity* **84**: 458–467.

Gündüz I, Auffray J-C, Britton-Davidian J, Catalan J, Ganem G, Ramalhinho MG *et al.* (2001). Molecular studies on the colonization of the Madeiran archipelago by house mice. *Mol Ecol* **10**: 2023–2029.

Gündüz I, Rambau RV, Tez C, Searle JB (2005). Mitochondrial DNA variation in the western house mouse (*Mus musculus domesticus*). close to its site of origin: studies in Turkey. *Biol J Linn Soc* **84**: 473–485.

Hardy OJ (2002). Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Mol Ecol* **12**: 1577–1588.

Hasegawa M, Kishino H, Yano T-A (1985). Dating of the human-ape splitting by molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D (2006). An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mole Biol Evol* **23**: 790–797.

Jakobsson M, Rosenberg NA (2007). CLUMPP: a cluster matching and permuting 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 EP, van der Kooi J, Solheim R, Searle JB (2010). Norwegian house mice (*Mus musculus musculus/domesticus*): distributions, routes of colonization and patterns of hybridization. *Mol Ecol* **19**: 5252–5264.

Jones EP, Johannesdottir F, Gündüz I, Richards MB, Searle JB (2011). The expansion of the house mouse into north-western Europe. *J Zool* **283**: 257–268.

Jones EP, Eager HM, Gabriel SI, Johannesdottir F, Searle JB (2013). Genetic tracking of mice and other bioproxies to infer human history. *Trends Genet* **29**: 298–308.

Jones EP, Searle JB (2015). Differing Y chromosome versus mitochondrial DNA ancestry, phylogeography, and introgression in the house mouse. *Biol J Linn Soc* **115**: 348–361.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kayser B, Tricart J (1957). Rail et route au Sénégal. *Ann Géog* **66**: 328–350.

Kolbe JJ, Glor RE, Schettino LR, Lara AC, Larson A, Losos JB (2004). Genetic variation increases during biological invasion by a Cuban lizard. *Nature* **431**: 177–181.

Linnenbrinck M, Wang J, Hardouin EA, Künzel S, Metler D, Baines JF (2013). The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol Ecol* **22**: 1904–1916.

Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* **82**: 1420–1425.

Lombaert E, Guillemaud T, Lundgren J, Koch R, Facon B, Grey A *et al.* (2014). Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the case of the Asian ladybird *Harmonia axyridis*. *Mol Ecol* **23**: 5979–5997.

Marrs RA, Sforza R, Hufbauer RA (2008). When invasion increases population genetic structure: a study with *Centaurea diffusa*. *Biol Invasions* **10**: 561–572.

Miège J-L (1981). Le commerce transsaharien au XIXe siècle. *Rev. de l'Occ. Musulman et de la Médit* **32**: 93–119.

Nachman MW, Searle JB (1995). Why is the house mouse karyotype so variable? *Trends Ecol Evol* **10**: 397–402.

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

Ninot O (2003). *Vie de relations, organisation de l'espace et développement en Afrique de l'Ouest: la région de Tambacounda au Sénégal*. PhD thesis, Université de Rouen.

Palanza P, Re L, Mainardi D, Brain PF, Parmigiani S (1996). Male and female competitive strategies of wild house mice pairs (*Mus musculus domesticus*) confronted with intruders of different sex and age in artificial territories. *Behaviour* **133**: 863–882.

Pocock MJO, Hauffe HC, Searle JB (2005). Dispersal in house mice. *Biol J Linn Soc* **84**: 565–583.

Prager EM, Tichy H, Sage RD (1996). Mitochondrial DNA sequence variation in the eastern house mouse, *Mus musculus*: comparison with other house mice and report of a 75-bp tandem repeat. *Heredity* **143**: 427–446.

Prager EM, Orrego C, Sage RD (1998). Genetic variation and phylogeography of Central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150**: 835–861.

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

Rajabi-Maham H, Orth A, Bonhomme F (2008). Phylogeography and postglacial expansion of *Mus musculus domesticus* inferred from mitochondrial DNA coalescent, from Iran to Europe. *Mol Ecol* **17**: 627–641.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102**: 15942–15947.

Ronquist F, Huelsenbeck J (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

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

Rosevear DR (1969). *Rodents of West Africa*. British Museum of Natural History: London.

Rousset F (1997). Genetic differentiation and estimation of gene flow from *F*-Statistics under isolation by distance. *Genetics* **145**: 1219–1228.

Rousset F (2000). Genetic differentiation between individuals. *J Evol Biol* **13**: 58–62.

Rousset F, Leblois R (2007). Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Mol Biol Evol* **24**: 2730–2745.

Rousset F (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol Ecol Res* **8**: 103–106.

Rousset F, Leblois R (2012). Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. *Mol Biol Evol* **29**: 957–973.

Searle JB, Jamieson PM, Gündüz I, Stevens MI, Jones EP, Gemmill CEC *et al*. (2009a). The diverse origins of New Zealand house mice. *Proc Roy Soc Lond B Biol* **276**: 209–217.

Searle JB, Jones CS, Gündüz I, Scascitelli M, Jones EP, Herman JS *et al.* (2009b). Of mice and (Viking?) men: phylogeography of British and Irish house mice. *Proc Roy Soc Lond B Biol* **276**: 201–207.

Sikes RS, Gannon WI, the Animal Care and Use Committee of the American Society of Mammalogists (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *J. Mammalogy* **92**: 235–253.

Singleton GR, Hinds LA, Krebs CJ, Spratt DM (2003). *Rats, Mice and People: Rodent Biology and Management*. ACIAR Monograph 96 Australian Centre for International Agricultural Research: Canberra.

Sinou A (1993). *Comptoirs et villes coloniales du Sénégal: Saint-Louis, Gorée, Dakar*. Editions Khartala-Orstom: Paris.

Sun JX, Helgason A, Mason G, Ebenesersdottir SS, Li H, Mallick S *et al.* (2012). A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161–1165.

Suzuki H, Nunome M, Koinoshita G, Aplin KP, Vogel P, Kryuskov AP *et al.* (2013). Evolutionary and dispersal history of Eurasian house mice *Mus musculus* clarified by more extensive geographic sampling of mitochondrial DNA. *Heredity* **111**: 375–390.

Verdu P, Leblois R, Froment A, Théry S, Bahuchet S, Rousset F *et al.* (2010). Limited dispersal in mobile hunter–gatherer Baka Pygmies. *Biol Letters* **6**: 858–861.

Villesen P (2007). FaBox: an online toolbox for fasta sequences. *Mol Ecol Notes* **7**: 965–968.

Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ *et al.* (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera). populations: analysis of 'neighbourhood size' using a more precise estimator. *Mol Ecol* **16**: 737–751.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

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