JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.12919

Genetic distinction between contiguous urban and rural multimammate mice in Tanzania despite gene flow

S. GRYSEELS*, J. GOÜY DE BELLOCQ*†, R. MAKUNDI‡, K. VANMECHELEN§, J. BROECKHOVE§, V. MAZOCH¶, R. ŠUMBERA¶, J. ZIMA JR†¶, H. LEIRS* & S. J. E. BAIRD†

*Evolutionary Ecology Group, Department of Biology, University of Antwerp, Antwerp, Belgium
†Institute of Vertebrate Biology, Research Facility Studenec, Academy of Sciences of the Czech Republic, Brno, Czech Republic
‡Pest Management Centre, Sokoine University of Agriculture, Morogoro, Tanzania
§Computational Modelling and Programming, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
¶Department of Zoology, University of South Bohemia, České Budějovice, Czech Republic

Keywords:

IMa2;
Mastomys natalensis;
population genetics;
spatial genetics;
synanthropy;
Tanzania;
urbanization.

Abstract

Special conditions are required for genetic differentiation to arise at a local geographical scale in the face of gene flow. The Natal multimammate mouse, *Mastomys natalensis*, is the most widely distributed and abundant rodent in sub-Saharan Africa. A notorious agricultural pest and a natural host for many zoonotic diseases, it can live in close proximity to humans and appears to compete with other rodents for the synanthropic niche. We surveyed its population genetic structure across a 180-km transect in central Tanzania along which the landscape varied between agricultural land in a rural setting and natural woody vegetation, rivers, roads and a city (Morogoro). We sampled *M. natalensis* across 10 localities and genotyped 15 microsatellite loci from 515 individuals. Hierarchical STRUCTURE analyses show a *K*-invariant pattern distinguishing Morogoro suburbs (located in the centre of the transect) from nine surrounding rural localities. Landscape connectivity analyses in Circuitscape and comparison of rainfall patterns suggest that neither geographical isolation nor natural breeding asynchrony could explain the genetic differentiation of the urban population. Using the isolation-with-migration model implemented in IMa2, we inferred that a split between suburban and rural populations would have occurred recently (<150 years ago) with higher urban effective population density consistent with an urban source to rural sink of effective migration. The observed genetic differentiation of urban multimammate mice is striking given the uninterrupted distribution of the animal throughout the landscape and the high estimates of effective migration ($2N_eM$ = 3.0 and 29.7), suggesting a strong selection gradient across the urban boundary.

Introduction

In contrast to allopatric divergence, differentiation of gene pools in the face of gene flow requires special conditions such as extreme selection gradients (Grahame *et al.*, 2006; Johannesson *et al.*, 2010; Bird *et al.*, 2012). Organisms living on the edges of human influence may experience such strong differential selection pressures. Although human urbanization of the environment is relatively recent on evolutionary time scales, many organisms thrive in these new synanthropic niches (Francis & Chadwick, 2012). These are usually generalist species, but a number of studies have demonstrated distinct phenotypic adaptations of urban populations to their relatively new environment, including shifts in mouse circadian rhythm and breeding season (Gliwicz, 1980) and changes in bird beak morphology (Badyaev

Correspondence: Sophie Gryseels, University of Antwerp, Evolutionary Ecology Group, Groenenborgerlaan 171, B-2020 Antwerp – Belgium.
E-mail: sophie.gryseels@gmail.com

et al., 2008) and wing length (Evans *et al.*, 2009). White-footed mice in New York city show directional selection at several loci (Harris *et al.*, 2013). Here, we use a spatial population genetic approach to explore whether a relatively recent urban selection gradient may result in genetic differentiation of a widespread generalist rodent despite the presence of gene flow.

The Natal multimammate mouse, *Mastomys natalensis*, is the most widely distributed rodent species in sub-Saharan Africa. Its abundance and reproductive potential make it a notorious agricultural pest and important carrier of zoonoses, most notably Lassa virus in West Africa. In eastern Africa, it is by far the most numerous mammal in most mesic habitats, particularly cultivated land, grasslands, savannah and low shrublands (Leirs, 2013), with densities being regulated by climatic and density-dependent factors resulting in occasional local population explosions (Leirs *et al.*, 1997; Sluydts *et al.*, 2007). It is the main vertebrate agricultural pest in eastern Africa, damaging up to 80% of maize seedlings and harvest (Mwanjabe & Leirs, 1997; Mulungu *et al.*, 2003). In western Africa, *M. natalensis* often occurs synanthropically inside and around rural human dwellings, whereas the savannah and agricultural niches seem to be dominated by other species, especially *Mastomys erythroleucus* (Duplantier *et al.*, 1997; Brouat *et al.*, 2007). In the extreme western limit of its distribution in Senegal, *M. natalensis* even appears absent from savannah and cultivated land, being found only in association with human habitation (Duplantier *et al.*, 1990). Although in eastern and southern Africa *M. natalensis* may also occasionally be encountered inside houses (Christensen, 1995; Monadjem *et al.*, 2011; Katakweba *et al.*, 2012), *Rattus* spp. are the main synanthropic rodents, relegating *M. natalensis* to the peridomestic area and food stores (Monadjem *et al.*, 2011).

Mastomys natalensis is characterized by a 2N = 32 karyotype (Granjon *et al.*, 1996) and can be divided into two main clades based on mitochondrial DNA sequence information (roughly encompassing western and eastern Africa, as divided by the Rift Valley) (Colangelo *et al.*, 2013). The eastern African *M. natalensis* mitochondrial clade consists of three lineages which were estimated to coalesce more or less simultaneously about 1.4 million years ago, their divergence probably driven by the constrained distribution of suitable habitats during the wet phases of the climatic oscillations in the Pleistocene (Colangelo *et al.*, 2013). At the western, synanthropic, edge of its distribution range in Senegal, the genetic structure of *M. natalensis* also reflects geographic patterns of habitat suitability, with the fine-scale spatial structure of *M. natalensis* being governed by its association with scattered rural settlements (Brouat *et al.*, 2007).

We investigate the population genetic structure of *M. natalensis* in central lowland Tanzania, centred on an area where the animal's biology and population ecology have previously been studied intensively (Leirs *et al.*, 1997; Mohr *et al.*, 2007; Sluydts *et al.*, 2007). We sampled across a 180-km transect, along which the landscape includes agricultural land in a rural setting, natural woodland, rivers, roads and a central city: Morogoro, where the human population has expanded by almost a factor of 30 since 1948 (8000–228 000 in 2002) (Mtatifikolo, 1997; National Bureau of Statistics, 2006).

Because *M. natalensis* is a common widespread generalist, gene flow might be expected to homogenize genetic structure across the study localities. However, the field area is environmentally heterogeneous, so we investigate whether the effect of geographical habitat quality variation is powerful enough for signals of genetic differentiation to be picked up in the face of gene flow. We first define and analyse the genetic structure of *M. natalensis* sampled at ten localities using cytochrome b sequences, a suite of 15 microsatellite loci and a genetic equilibrium maximizing clustering algorithm. We then explore the evolutionary history of this structure further in the light of an isolation-with-migration model (IM), where we co-estimate splitting time, effective population sizes and bidirectional migration rates. We discuss scenarios that might explain the observed structure, how the IM discretization can be interpreted over a continuous field area and the implications of genetic divergence in the face of gene flow driven by synanthropic selection gradients.

Materials and methods

Rodent trapping and sample collection

In order to study the potential effect of environmental heterogeneity in structuring *M. natalensis* genepools, we sampled a 180-km-long transect at regular intervals while crossing various land covers. Ten field localities were selected approximately 20 km apart along the road from Dodoma to Dar es Salaam: Berega, Maguha, Dumila, Dakawa, Mkundi, Morogoro, Mikese, Bwawani, Ubena and Chalinze (Fig. 1). The first six localities were sampled in November and December of 2009 and the last four sites in December 2010 and January 2011. The central locality Morogoro was sampled in both periods.

Small mammals were captured in Sherman live traps baited with a mixture of peanut butter and maize flour. These traps were laid out evenly in a 1-ha square grid of 10 × 10 traps. At every locality, at least four such grids were constructed a minimum of 500 m and maximum of 2.5 km apart from each other, and each grid was sampled for a single night. Depending on the trapping success in that locality, additional grids were set. For each sampling grid, a maximum of 20 *M. natalensis* individuals were euthanized by Isoflurane® inhalation.

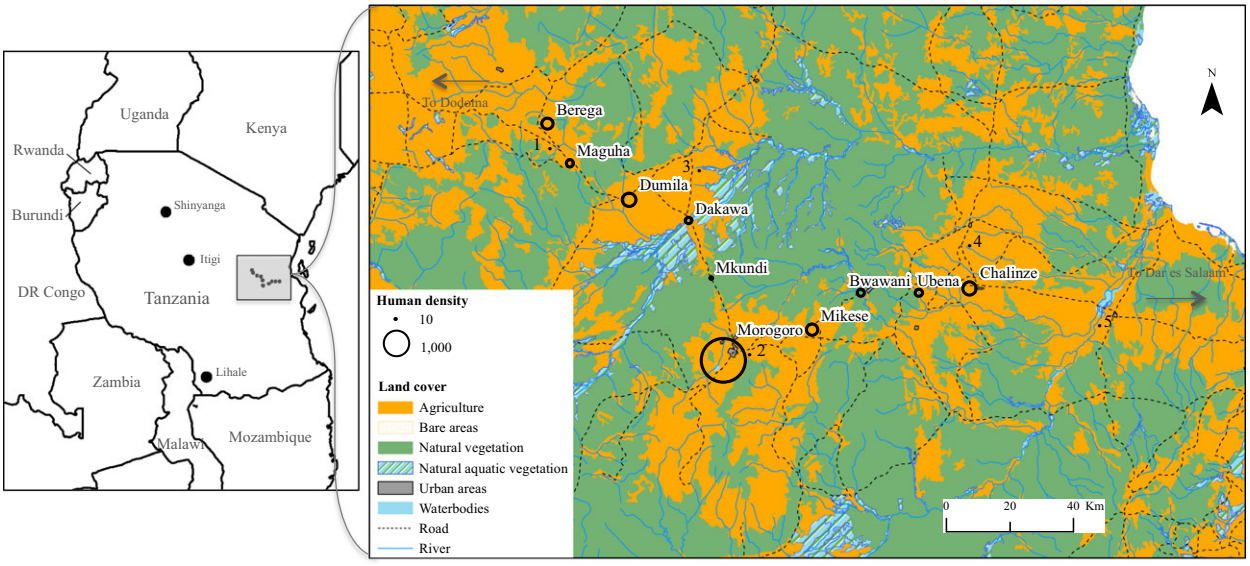

Fig. 1 Sampling area. Left: overview of Tanzania and neighbouring countries including the three wide-scale sampling localities. Right: transect sampling localities. The map background is a simplified version of the land cover layer used in the landscape connectivity analyses. The sizes of the circles represent average human population density (per km^2) in a 5-km radius around each sampling locality. Numbers indicate meteorological stations from which monthly precipitation averages were obtained (see Fig. S3): 1: Berega, 2: Morogoro, 3: Wami prison farm, 4. Lugoba Mission post, 5. Ruvu.

Blood was drawn from the retro-orbital sinus with a capillary and preserved on prepunched filter papers (Serobuvard, LDA 22, Zoopole, France), and organ samples were preserved in RNAlater and 95% ethanol for subsequent studies of viral pathogens. When more than 20 *M. natalensis* were captured in a grid, the 'excess' animals were sedated by Isoflurane® inhalation, and only blood and toe clips were sampled on filter paper and in ethanol, respectively, after weighing the animals. These 'excess' animals were then released at the point of capture.

To put the population genetic structure of our samples into a wider geographical context, we used 'wide-scale' *M. natalensis* samples from Lihale (n = 16), Itigi (n = 16) and Shinyanga Lubaga (n = 8) (Fig. 1). These samples originated from animals captured in Sherman live traps and snap traps placed 10 m apart in two lines for one to two nights per locality, in July–August 2008.

All animal work was approved by the University of Antwerp Ethical Committee for Animal Experimentation (2011–52), and followed regulations of the Research Policy of Sokoine University of Agriculture as stipulated in the 'Code of Conduct for Research Ethics' (revised version of 2012). The import permit into Belgium was obtained from the Federal Agency for Safety of the Food Chain (CONT/IEC/FRT/320373). The export permit from Tanzania was granted by the Directorate of Veterinary Services, Tanzania, based on the fact that the animal specimens are common pest species and are not listed in the IUCN list of endangered species.

Microsatellite and mitochondrial genotyping

DNA was extracted using the Qiagen DNeasy™ Blood & Tissue Kit according to the manufacturer's instructions. From the transect localities, 515 samples were genotyped for 15 microsatellite loci while ensuring at least 20 individuals per locality were genotyped (Table 1). For localities where more than 20 individuals were captured, at least 20 animals for genotyping were randomly chosen across all trapping grids of the locality, giving priority to individuals for which organ samples were available (itself a random selection), and with the number >20 being proportional to the total of samples available for that locality. All 40 available samples from the wide-scale localities were genotyped (Table 1).

Fifteen microsatellite markers for *M. natalensis* were amplified in two multiplex PCRs following Galan *et al.* (2004) and Loiseau *et al.* (2007); however, the first PCR ran loci MH52, MH216, MH30, MH133, MH206, MH141, MH141, MH28, MH60 and MH174, whereas the second ran MH5, MH80, MH105 and MH101, with each primer in 0.125 μM concentration. Automated scoring of the fluorescent peaks in GeneMapper v3.7 (Applied Biosystems, Foster City, CA, USA) was manually verified twice for all samples. If peaks were absent for more than two loci, DNA extraction and fragment analyses were repeated. However, the large collection of samples from Morogoro and Dumila consisted in part of single toe clips, which were entirely used during the first DNA extraction round so that re-extraction was

Table 1 Per locality overview of *Mastomys natalensis* trapping success, number of individuals genotyped, microsatellite cluster assignment, mitochondrial lineage assignment and relevant environmental characteristics of the localities. See Figure 1 for the locations of 'transect localities', which were sampled at regular intervals across a 180-km transect in a standardized manner, and the locations of 'wide-scale localities'.

	Locality	Nr of *M. natalensis* captures/trap nights (trapping success in %)	Nr of *M. natalensis* genotyped (microsatellite | cytochrome b)	Mean microsatellite genotype assignment to yellow | blue | red cluster at $K = 3$ (Fig. 2)	Frequency of mitochondrial lineage B-IV | B-V in %	% Land cover agriculture	% Land cover natural vegetation (grassland, shrubland, woodland and forests)	Human population density (people km^{-2}) in 2002
Transect localities	Berega	23/700 (3.3)	23 | 23	0.41 | 0.37 | 0.22	78.3 | 21.7	19	81	84
	Maguha	23/800 (2.9)	23 | 23	0.13 | 0.68 | 0.19	21.7 | 78.3	23	77	40
	Dumila	152/400 (38.0)	151 | 97	0.06 | 0.72 | 0.22	6.2 | 93.8	81	19	127
	Dakawa	36/400 (9.0)	36 | 34	0.03 | 0.82 | 0.15	17.6 | 82.4	25	75	33
	Mkundi	21/400 (5.3)	20 | 20	0.01 | 0.84 | 0.15	5 | 95	0	100	7
	Morogoro	410/800 (51.3)	138 | 86	0 | 0.27 | 0.72	3.5 | 96.5	49	44	1273
	Mikese	98/500 (19.6)	34 | 28	0.05 | 0.75 | 0.2	3.6 | 96.4	56	44	81
	Bwawani	61/800 (7.6)	25 | 24	0.02 | 0.72 | 0.25	4.2 | 95.8	2	97	35
	Ubena	54/900 (6.0)	30 | 31	0.08 | 0.81 | 0.11	3.2 | 96.8	20	80	41
	Chalinze	78/1100 (7.1)	35 | 33	0.03 | 0.82 | 0.15	9.1 | 90.9	39	60	122
Wide-scale localities	Shinyanga	8	8 | 1	0.79 | 0.1 | 0.11	1 | 0	36	64	120
	Itigi	16	16 | 12	0.97 | 0.02 | 0.01	12 | 0	76	24	<1
	Lihale	16	16 | 15	0 | 0.95 | 0.05	0 | 15	98	0	2

not possible. In those cases, only PCR and fragment analysis were repeated. Although microsatellite genotypes derived from toe clip samples still often had missing data (26.3% of data), from every locality at least 20 genotypes were derived from liver samples that contained on average much less missing data (5.2%). Presence of null alleles was gauged with the program Microchecker (Van Oosterhout *et al.*, 2004), which suggested that 11 of the 15 loci probably contained some null alleles. Although inference of genotype assignment to population genetic clusters (our main goal) has been shown to be only weakly affected by the presence of null alleles, which mainly result in a simple loss of assignment power (Carlsson, 2008), we assessed their potential impact on our results in two ways. Using MicroDrop (Wang *et al.*, 2012), we determined that across individuals, there was a significant correlation between homozygotes and missing genotypes ($\rho = 0.23$, $P < 0.001$), whereas across loci no significant correlation between homozygotes and missing genotypes could be observed ($\rho = 0.28$, $P = 0.157$). This implies that sample-specific factors, such as low DNA concentration or poor sample quality, were in fact the dominant factors for allelic dropout in this data set, and not locus-specific factors, such as null alleles. Secondly, following the null allele protocol recommended by the STRUCTURE manual, we added unobserved recessive alleles for each locus (i.e. the unknown null alleles) to the data set, and subsequently ran STRUCTURE for $K = 2$ to 10 using the same settings as below.

To determine which of the three mitochondrial lineages known to occur in Tanzania were harboured by our samples (Colangelo *et al.*, 2013), we sequenced a portion of the cytochrome b gene after a PCR using primers H15915 and L14723 (Lecompte *et al.*, 2002) in a 0.02 μM concentration, among 1× buffer (GoTaq®Flexi Buffer; Promega, Fitchburg, WI, USA), 0.1 μM $MgCl_2$, 0.02 μM dNTP (Fermentas, Waltham, MA, USA) and 1.35 units GoTaq® DNA Polymerase (Promega). Sequencing was performed using the L14723 primer at the Genetic Sequencing Facility of the Vlaams Instituut voor Biotechnologie (Belgium).

Genetic structure

We analysed the population genetic structure of *M. natalensis* samples using the Bayesian clustering algorithm implemented in the program STRUCTURE v2.3.2 (Pritchard *et al.*, 2000). For a given number of clusters K, this co-estimates the nature of clusters and the assignment of genotypes to those clusters while minimizing genetic disequilibrium within each cluster. The optional prior to include information on which individuals were sampled from the same locality was selected (Hubisz *et al.*, 2009). We used the admixture model of individual genotypes such that individuals are allowed to have genotypes originating from more than one

cluster, and alpha, the Dirichlet parameter for degree of admixture, was allowed to vary between clusters. The algorithm was run 25 times for the scenarios of 1–10 clusters (K = 1–10) in our data set, with run lengths of 1 000 000 iterations after a burn-in of 200 000 iterations. Run lengths were determined based on visual inspection of the convergence of the Ln P(D) values for exploratory runs.

The STRUCTURE analysis outputs 10 × 25 Q-matrices (K range × N replicates) in which the proportional assignment of each genotype to each of K clusters is estimated. The columns in the Q-matrices correspond to the (arbitrarily chosen) colours in standard STRUCTURE plots. To automatically colour our plots using similar colours for correlated clusters found in different runs, we used the software CLUMPAK (Kopelman *et al.*, 2015). For each K level, CLUMPAK groups runs into distinct modes. STRUCTURE can find different posterior estimates of the Q-matrices from run to run: when all K clusters of a run have correlates in a second run, then these runs will be grouped in the same mode in CLUMPAK.

Genetic diversity and relatedness

Cytochrome b sequences were aligned and trimmed, unique haplotypes identified and nucleotide diversity per locality calculated, all using Geneious 5.0 (Biomatters, 2014). Basic genetic descriptive statistics (Weir and Cockerham Fst per locus – across all localities – allelic richness adjusted for sample size, observed and expected heterozygosity, r^2 as a measure of linkage disequilibrium) were calculated for the total set of microsatellite genotypes in the R package hierfstat (Goudet, 2005). The significance of pairwise differences in allelic richness between localities was tested in R (R Development Core Team, 2011) using an ANOVA and a Tukey 'honest significant difference' (HSD) test. To evaluate whether by chance we sampled more related animals in some localities than others, we calculated Li's relationship coefficient (Li *et al.*, 1993) between pairs of microsatellite genotypes sampled across the transect localities in SPAGeDi (Hardy & Vekemans, 2002). We excluded those genotypes with missing data for more than 5 of the 15 loci (leaving a total of 441 genotypes; at least 20 per locality). Li's relationship coefficient r is a summation over all loci of the average proportion of alleles shared by a pair of individuals at a given locus, including a correction for sample size and a weighting of each locus. Deviation from 0 means that individuals within a given sample group are significantly more (positive values) or less (negative values) related than at random. Then, using only the relatedness coefficients within a locality, a linear model was constructed in R relating these pairwise estimates of individuals within a single locality to locality memberships. The pairwise differences between localities in average relatedness of individuals within a locality were tested using an ANOVA and a Tukey HSD test.

Estimating splitting time, effective population sizes and migration using an isolation-with-migration model

Our data revealed a distinct cluster of *M. natalensis* microsatellite genotypes in Morogoro that are geographically circumscribed by a larger cluster that unites all other individuals from the surrounding transect localities. One explanation for this pattern of genetic structure is a splitting history of rural–urban gene pool isolation with ongoing migration between the units thus formed. We investigated this possibility under the assumption of panmictic population units using the software IMa2 (Hey, 2010), which implements an IM coalescent model in the Bayesian inference framework. We distinguish between the *M. natalensis* sampled within Morogoro urban district (U) and those from eight rural localities outside Morogoro (R) (excluding the Berega locality whose individuals are involved in a deeper genetic distinction – see Results). The IM model has six parameters: the time since the split of an ancestral population (t_0, in years) with effective size N_{eA} (in number of individuals), the effective population sizes of the two current populations N_{eU} and N_{eR} (in number of individuals) and two migration rates $M_{U\rightarrow R}$ and $M_{R\rightarrow U}$ (in migrants per generation), summarizing movements between the current populations since the split. We assumed a microsatellite mutation rate μ of 4×10^{-4} mutations per generation, an average across studies and loci (Ellegren, 2000, 2004; Whittaker *et al.*, 2003), assuming stepwise mutation. Mutation rate variation across loci was accounted for by a uniform prior on locus-specific mutation rates with a mean of 4×10^{-4} mutations per generation within a range of (1×10^{-5} – 1×10^{-4}). We assumed an average generation time (G_T) of two generations per year, as estimated for *M. natalensis* from Morogoro (Leirs *et al.*, 1993). During IM calculations, all parameters are arranged as a product or division of the parameter and the mutation rate (e.g. t_0/G_T, M/μ and $4N_e\mu$) so that the final parameter estimates can all be rescaled given mutation rate estimates.

Because of the computational intensity of IMa2, a subset of 105 individuals was chosen for analysis, and to maximize statistical power, only samples without any missing microsatellite data were considered: there were 55 such individuals in the Morogoro samples. There were many data-complete individuals from the rural cluster so 50 were stochastically selected with probability 1 - W, where W is the size of the confidence interval of the assignment of the genotype by STRUCTURE to the rural cluster. We verified with STRUCTURE that this reduced data set showed the same signal of genetic structure as the complete data set

(results not shown). Pinho and Hey (2010) suggest a bias may arise when the same genetic data is used to select individuals for IM analysis as for the analysis itself. We note that while some choice has to be made, our stochastic preferential choice of 'good' members of the rural cluster may reduce estimates of migration from urban to rural.

We augmented the IMa2 code with state checkpointing logic (Ansel *et al.*, 2009), enabling the restart of a computation at the last available checkpoint in the event of a server failure.

We first performed exploratory IMa2 runs with 100 Markov chains to optimize chain mixing (i.e. swapping of the multidimensional parameter states between heated and unheated chains). We used the geometric heating model, with parameters h_a, the degree of non-linearity with which each successive chain is increased in temperature (β), and h_b, the lowest value that β can take. Values of $h_a = 0.99$ and $h_b = 0.5$ led to 'favourable' swapping rates between 0.4 and 0.6 in >80% of chains.

Priors for all parameters were set up relative to an initial estimate of the population mutation rate $\Theta = 4N_e\mu$. We used Arlequin v3.5 (Excoffier & Lischer, 2010) to calculate the average genotypic differences between all pairs of genotypes, $\prod_n$, and a measure of the genetic distance between pairs of sampled localities $(\delta\gamma)^2$, based on the actual number of repeats of the microsatellites. These are estimators for Θ and are given by:

$$\prod_n = \frac{\sum_{i=1}^{k}\sum_{j<i} p_i p_j d_{ij}}{L} \text{ and } (\delta\gamma)^2 = (\gamma_U - \gamma_R)^2$$

respectively, where d_{ij} is an estimate of the number of mutations that have occurred between genotypes i and j, k is the number of haplotypes, p_i is the frequency of genotype i, and L is the number of loci. γ_U and γ_R are the average number of allelic size differences within the sample of genotypes from urban Morogoro (U) and the genotypes sampled in the eight rural localities (R), computed over all loci. The $\prod_n$ estimators for the two current populations R and U were 0.88 and 0.89, and $(\delta\gamma)^2 = 0.90$, so we used $\Theta = 0.89$ as our estimate of Θ.

Exploratory IMa2 runs were used to evaluate the recommendations for bounds on the priors (in terms of Θ) to ensure they were appropriate while being computationally efficient. Although the IMa2 manual recommends the upper bound for the prior on the mutation-rate-scaled migration rate (M/μ) to be approximately $\Theta/2$, our exploratory runs showed significant posterior density abutting this limit, so we used a uniform prior on migration/mutation with much higher (more conservative) upper bounds of 11.2Θ for (backward-time) migration from the urban population to the rural and 44.9Θ for (backward-time) migration from the rural population to the urban population. Similarly, we set conservative upper bounds on the priors for the mutation-rate-scaled current effective population sizes ($4N_{eU}\mu$ and $4N_{eR}\mu$) to 45Θ, and for the ancestral effective population size ($4N_{eA}\mu$) to 112Θ, in contrast to the default recommendation for upper bounds of 5Θ. The upper bound of the uniform prior on the mutation-rate-scaled population splitting time ($t_0\mu/G_T$) was set at 2Θ as recommended in the manual.

Given these settings, we set ten independent runs (with different starting seeds) each with 100 Markov chains to sample the IM parameter space. After burn-in, which was based on visual inspection of convergence in the trend plots for the overall Log[P(G)+P(D|G)] and $t_0\mu/G_T$ parameter estimates, genealogies were saved every 1000 chain steps until the marginal distributions of each parameter from each run visually stabilized through time, and the marginal distributions originating from the different independent runs indicated convergence towards a single distribution. We finally saved between 39 480 and 46 867 genealogies per independent run, giving a total of 431 807 genealogies. The parameter estimates from each of these runs were integrated to calculate their full posterior distributions using the 'L mode' of IMa2. The likelihood of the full model (in which all parameters are unconstrained) was compared to two nested models using likelihood ratio tests: the first nested model constrained the two migration rates to be equal (symmetric migration) and the second constrained the two descendant population sizes to be equal (symmetric split).

Approximate geographic extent of clusters

Isolation-with-migration analyses provide estimates of effective population sizes for genetic clusters, but measures of the geographic extent covered by each cluster would be necessary to understand how these counts are spread as effective population densities over the field area. Our linear transect sampling is unsuited to estimate the 2D extent of the urban and rural populations; we can however consider simple upper and lower bounds. For the urban population, we define the lower bound assuming it occupies at least the area of the minimum bounding polygon around the urban-clustered sample grids (i.e. 1.5 km^2), and an upper bound, which is the disk of which the radius is the distance between the centre of the urban samples and the nearest sampling grid of which the genotypes belong to the rural cluster (Mikese). This disk covers 1470 km^2. We consider the minimum geographic extent for the rural population as the minimum bounding polygon around transect localities Maguha, Dumila, Dakawa, Mkundi, Mikese, Bwawani, Ubena and Chalinze and the wide-scale locality Lihale (i.e. localities where genotypes clustered together in STRUCTURE). This polygon covers 38 900 km^2.

Landscape connectivity analyses

We analysed the land cover of our study area in the Morogoro region (a rectangle with sides being a minimum distance of 40 km from each outermost locality) to estimate the relative potential connectivity between sampled localities in terms of *M. natalensis* movement through generations. We used the Africover Spatially Aggregated Multipurpose Landcover database of Tanzania from the Food and Agriculture Organization of the UN as input maps (available at http://www.africover.org). The land cover map was produced from visual interpretation of digitally enhanced LANDSAT TM images acquired mainly in the year 1997, as well as ground data and expert knowledge. Details on the land cover classification used can be found in Di Gregorio & Jansen (2000). We further categorized these land cover elements into agriculture, natural vegetation and urban cover and determined the proportion of these landscape categories in 5 km radius around the sampling localities (Table 1).

To estimate the relative movement potential (across generations) of *M. natalensis* between each pair of sampling localities, we estimated the 'landscape resistance' between these localities. First, 10 rodent researchers with extensive field experience with *M. natalensis* independently assigned a *M. natalensis* habitat value to each land cover element by expert opinion, choosing from six exponentially increasing numerical categories, after which a weighted average was taken. When linear landscape elements such as roads and rivers crossed a cell of the raster map, the habitat value of that cell was increased to the value of the linear element, unless the habitat value was already higher. The resulting raster of *M. natalensis* habitat values in the landscape was used as input in the software Circuitscape (McRae, 2006), which uses analogy with electric circuit theory to quantify movement probabilities through the landscape. The landscape is then represented as a conductive surface and the effective resistance (as in electric circuits) is calculated between each pair of 10 polygons constructed around all sampling sites within each locality.

We used data from the National Bureau of Statistics (NBS) in Tanzania (available at http://www.nbs.go.tz/) to calculate human population densities around each site, by summing GIS density coverage over enumeration areas across a radius of 5 km per sampled site.

Precipitation patterns

Monthly precipitation means (calculated over at least 20 years) from the meteorological stations in the vicinity of our sampling sites were gathered from the Global Historical Climatology Network (GHCN-Monthly) (Peterson & Vose 1997) through the KNMI Climate Explorer database (http://climexp.knmi.nl/). Rainfall patterns are strongly linked to seasonal breeding patterns of *M. natalensis* (Leirs *et al.*, 1997), thus providing a way to evaluate whether breeding would naturally occur synchronously between sampling sites.

Results

Sampling summary

In the transect localities, we captured a total of 1060 small mammals, of which 956 (90.2%) were *M. natalensis* (Tables 1 and S1). Trapping effort, bearing in mind we ensured at least 20 *M. natalensis* individuals were captured per locality, ranged between 400–1100 trap nights (i.e. number of traps × trapping nights) per locality. Average trapping success (a snapshot surrogate for current relative *M. natalensis* population density at localities) varied between 2.9 and 50.5 captures per 100 trap nights. We further used a total of 40 *M. natalensis* samples from the three wide-scale localities (Table 1).

Mitochondrial haplotypes

From the 427 partial cytochrome b sequences (GenBank accession numbers KF779499–KF779925), we could distinguish 73 distinct haplotypes that can be phylogenetically grouped into two main lineages (Table S2). These correspond to the two lineages Colangelo *et al.*, (2013) described in Tanzania (lineage B-IV and B-V). In the two northern wide-scale localities (Itigi and Shinyanga), only lineage B-IV was detected whereas in the southern wide-scale locality only lineage B-V was found (Table 1). Apart from the northernmost transect locality, Berega, where 78% of the haplotypes are lineage B-IV vs. 22% lineage B-V, in all other transect localities (Maguha, Dumila, Dakawa, Mkundi, Morogoro, Mikese, Bwawani, Ubena and Chalinze), the vast majority of the haplotypes belong to lineage B-V (Tables 1 and S2).

Genetic summary statistics: microsatellites

The number of alleles observed per microsatellite locus ranged from 13 (MH5) to 44 (MH80), with an average of 30 alleles across the 15 loci. The allelic richness per locus and per locality ranged between 6.3 and 19.0, the observed heterozygosity (Ho) between 0.33 and 1 and the expected heterozygosity (He) between 0.64 and 0.97 (Table S3). Loci did not show significant deviation from linkage equilibrium (Table S4). The average allelic richness was not significantly different between any pair of localities (overall effect of locality: $P = 0.80$; pairwise comparisons: P-values ranged between 0.63 and 1). For 3 loci, the Weir and Cockerham Fst values (calculated over the whole sample) were lower than 0.001; for 10 loci, Fst was between 0.01 and 0.02; and for 2 loci, Fst was higher than 0.02 (Table S4).

Li's coefficients of relatedness ranged from −0.27 to 0.71. According to the ANOVA and Tukey HSD test comparing the average relatedness of animals (within a locality) between all pairwise localities, the animals from Mkundi were significantly more related to each other than animals from Berega (95% CI = 0.0018, 0.064; $P = 0.03$), Maguha (95% CI = 0.008, 0.070; $P = 0.025$), Dumila (95% CI = 0.003, 0.009; $P < 0.001$), Mikese (95% CI = 0.033, 0.061; $P = 0.0044$), Ubena (95% CI = 0.036, 0.065; $P = 0.004$) and Chalinze (95% CI = 0.047, 0.074; $P < 0.001$). Animals from Morogoro were significantly more related to each other than animals from Dumila (95% CI = 0.012, 0.006; $P < 0.001$) and Chalinze (95% CI = 0.027, 0.041; $P < 0.001$). All other pairwise comparisons of average within-locality relatedness were not significantly different.

Hierarchical population structure

The pattern of clustering of microsatellite genotypes across the *M. natalensis* transect localities and the three wide-scale localities further north and south in Tanzania is hierarchically structured. That is, increasing K consistently subdivides existing clusters in the same localities of occurrence rather than creating novel ones with a different distribution pattern (Fig. 2). In Fig. 2, we show the most common modes identified by CLUMPAK for each level of K.

When forcing assignment of the individuals into two clusters ($K = 2$), all individuals from the wide-scale northern Tanzanian localities (Itigi and Shinyanga) are majority assigned to one 'northern' yellow cluster, and from the wide-scale southern Lihale to a 'southern' blue cluster. Animals from Berega (the most northern transect locality) are admixed between these two clusters, whereas the genotypes of the animals from all other transect localities are majority assigned to the 'southern' blue cluster. The distribution of the northern and southern superclusters thus matches the distribution of the two mitochondrial lineages (Table 1). The distinction of animals from the wide-scale northern localities is invariant across all investigated levels of K, whereas the southern blue cluster shows strong substructure from $K = 3$ onwards. At $K = 3$, a large 'red' subcluster splits off from the blue cluster, with only animals from Morogoro majority assigned to the red cluster. Figure 2 shows how this general pattern, in which most animals from Morogoro are majority assigned to different sets of clusters than most other animals from all other localities, is invariant across all levels of K. The only other locality-specific clustering of individuals at K levels >3 is from Berega, where animals specifically admixed between northern and southern superclusters seem to occur, and from Lihale, geographically distant from all sampled other localities, from which six (out of 16) are majority assigned to a distinct cluster in all levels of K >4 (Fig. 2). The analysis allowing for null alleles showed an essentially identical clustering pattern (Fig. S1).

Geographic variation

Considering the estimated degree of landscape resistance to dispersal (arbitrary units Ω) (Table 2) between pairs of neighbouring localities, the resistance between Bwawani and its two neighbours was relatively highest (15.2 and 16.4 Ω), whereas the resistance between Dumila and Dakawa was relatively lowest (3.0 Ω). The landscape resistance between Morogoro and its two neighbours had intermediate values (7.8 and 7.2 Ω). For each locality pair, a map of the current flow across the landscape is shown in Fig. S2, depicting the relative potential for movement of *M. natalensis* between the localities, given the specified *M. natalensis* habitat raster.

In a 5-km radius around the sampling points, the percentage of the land cover that is taken up by agriculture (preferred *M. natalensis* habitat) and by natural vegetation (less preferred *M. natalensis* habitat) varied considerably between localities. Mkundi had no mapped agricultural fields (although we did observe small subsistence farms there), whereas 80% of the land around Dumila was taken up by agriculture. Morogoro has 49% of its surrounding land covered by agriculture (Table 1).

The population density of humans also varied considerably between localities. In a 5-km radius around the sampling points, the Morogoro city locality had 1273 persons per km^2 in 2002, whereas the human densities in other localities range between 7 and 127 persons per km^2 (Table 1).

Rainfall pattern and breeding season

As an estimate of the timing of the breeding season, average monthly precipitation at meteorological stations near our sampling sites (Fig. 1) is shown in Supplementary Fig. S3. All sites, apart from Berega (meteorological station 1), experience a bimodal rainfall pattern with coinciding rainfall peaks in March–April and November–January. In Berega, precipitation is unimodal and peaks between December and May.

Isolation-with-migration model estimates

As input populations for the IM model, we considered groups of individuals identified by STRUCTURE: animals from Morogoro (an urban area) vs. those of the surrounding sampled localities (rural areas). We did not consider animals from geographically distant Berega, Lihale, Itigi and Shinyanga. For the IM analyses, these urban and rural groups are assumed to be IM populations arising from a vicariance (isolation) event in the past followed by ongoing gene flow (migration) between the vicars. We calculated parameter estimates

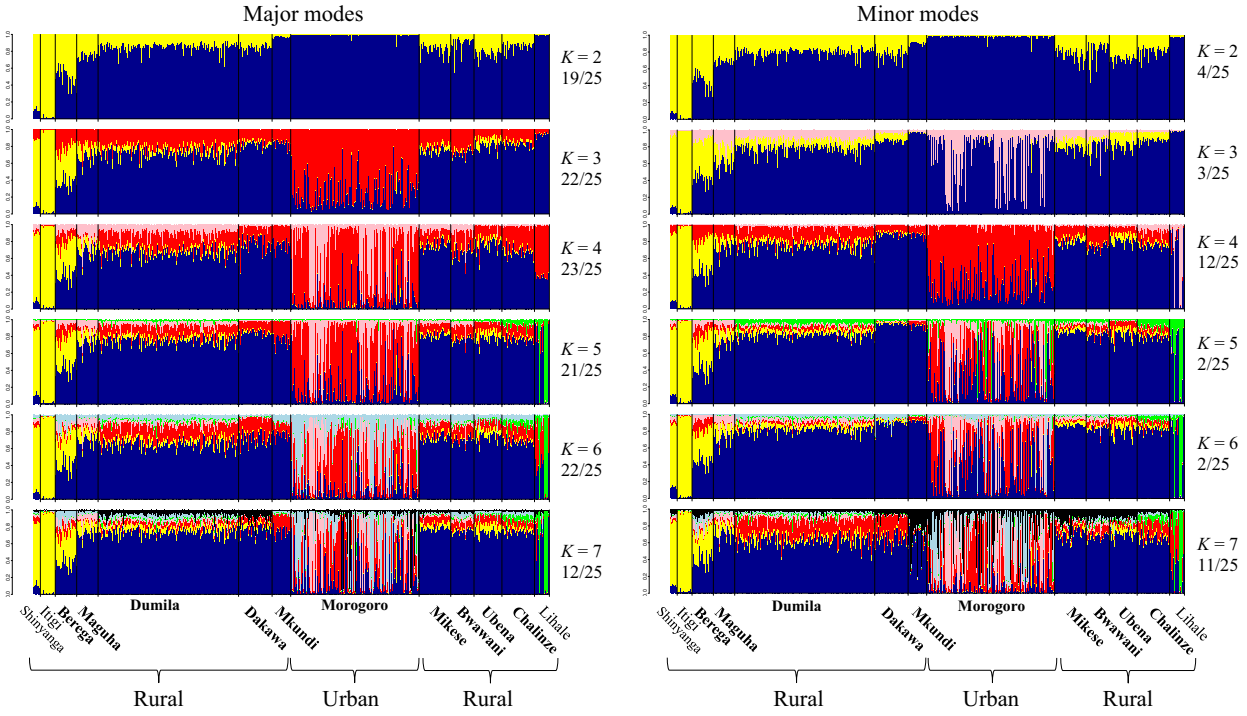

Fig. 2 Mean probabilities of individual cluster memberships of STRUCTURE runs for K = 2–7, with each colour depicting a distinct cluster. Clusters correlated across STRUCTURE runs were labelled using CLUMPAK, and similarity between clusters (i.e. the assignment of the same colour) across different K levels was subsequently determined by eye. Each bar represents the mean probabilities of an individual belonging to the different clusters while allowing for individual admixed genotypes. Left plots represent membership means taken across STRUCTURE run replicates from the most common mode, and right plots represent means taken from minor modes (as identified by CLUMPAK (Kopelman *et al.*, 2015)). Modes with single run representatives are not depicted. Black lines separate individuals from different localities.

Table 2 Geographical distance (km) (upper triangle) and relative resistance distance (in Ω units) (lower triangle) as an inverse measure of landscape connectivity between centroids of sampled localities. In boldface are resistance distances between neighbouring localities.

	Berega	Maguha	Dumila	Dakawa	Mkundi	Morogoro	Mikese	Bwawani	Ubena	Chalinze
Berega	0	11.4	30.7	50.1	64.5	82	95.8	107.4	123.8	142.3
Maguha	**4.4**	0	20.7	40.8	54.1	70.9	85.8	98.4	115.3	134
Dumila	11.7	**12.8**	0	20.2	33.8	52	65.2	77.8	94.9	113.7
Dakawa	13.7	14.9	**3**	0	17	37.5	46	57.7	74.7	93.6
Mkundi	20.4	21.7	11.2	**10.3**	0	20.6	31.9	47.5	65.7	84.6
Morogoro	17.6	18.9	8.7	8.8	**7.8**	0	24.6	44.8	62.6	80.5
Mikese	22	23.2	13	13	12.3	**7.2**	0	20.5	38.1	56.1
Bwawani	27.4	28.7	18.3	18.4	18.9	15	**15.2**	0	18.3	37.1
Ubena	29.7	31	20.6	20.8	21.9	18.2	19.1	**16.4**	0	19
Chalinze	29.5	30.8	20.4	20.6	22	18.3	19.6	18.8	**11.6**	0

of t_0, M and N_e from the mutation-rate-scaled posteriors of the IM model used by IMa2 ($t_0\mu/G_T$, M/μ and $4N_e\mu$). We inferred that the two populations split <130 years ago, with the most probable estimate of t_0 at 37 years ago (95% HPD interval 5.6–129.4; Fig. 3a, Table 3). It must be remembered, however, that our confidence in this estimate relies in turn on our confidence in the estimate of microsatellite mutations rates (Ellegren, 2000, 2004; Whittaker *et al.*, 2003). Thus, if the mutation rate is in reality 10 times slower than the 4×10^{-4} mutations/generation we assumed, the t_0 estimate becomes 10 times larger. The (forward-time) migration rate $M_{R\rightarrow U}$ from the rural population to Morogoro (highest probability (HP) point estimate 7.66×10^{-4} migrators per generation) is significantly lower than the migration rate $M_{U\rightarrow R}$ from the urban

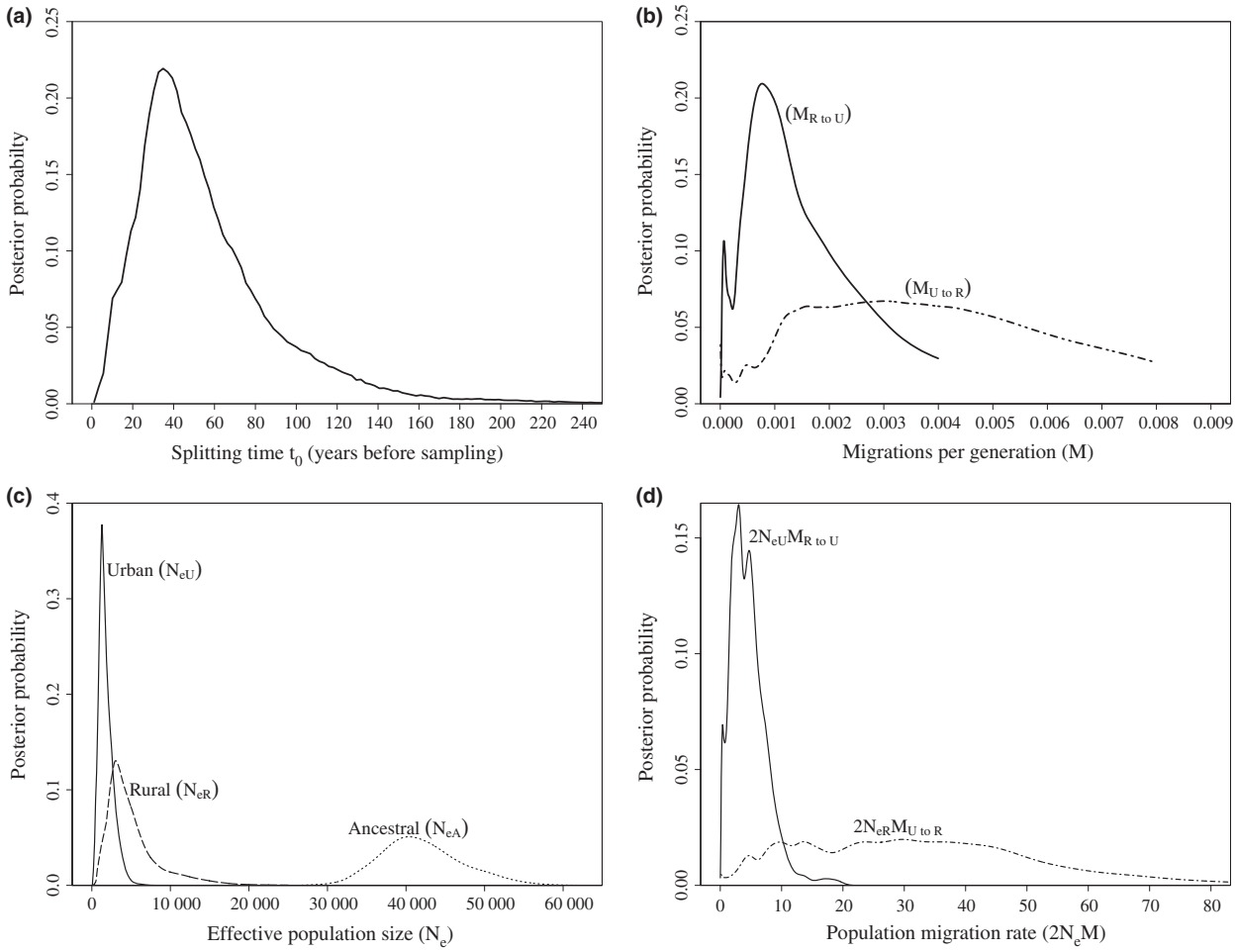

Fig. 3 (a–c) Posterior probability distributions of the six parameters of the isolation-with-migration model: splitting time t_0 (in years), migration rate M (migrators/generation) and three effective populations sizes (effective individuals in the current urban $-N_{eU}-$ and rural populations $-N_{eR}-$ and the ancestral population $-N_{eA}-$ from which both are drawn). (d) Population migration rate $2N_eM$ (individuals times incoming migrations per generation). For each parameter except the last, a microsatellite mutation rate of 4×10^{-4} mutations per generation, following a stepwise mutation model, and a generation time of 6 months were assumed, as IMa2 estimates $t_0\mu/G_T$, M/μ and $4N_e\mu$.

population to rural (HP estimate of 3.012×10^{-3}) ($\chi = 3.816$, d.f. = 1, $P = 0.05$) (Fig. 3b, Table 3); despite the potential inverse bias mentioned previously. The 95% highest posterior density bounds for these estimates could not however be clearly determined as the posterior distribution is to a small extent restricted by the upper bound of the prior (Fig. 3b). The HP estimate for effective population size N_{eU} of the current urban *M. natalensis* samples is, at 1288 effective individuals (95% HPD interval: 313–3688), significantly smaller than that of the modelled rural population (N_{eR}, HP estimate 3038 effective individuals, 95% HPD interval: 588–12 538) ($\chi = 7.204$, d.f. = 1, $P = 0.007$) (Fig. 3c, Table 3).

The population migration rate ($2N_e \times M$) from rural to urban (which is independent of the mutation rate and calculated as the joint posterior density function of N_{eU} and $M_{R \to U}$) is, at 3.001 (95% of the distribution volume: 0.50–11.99), significantly smaller than $2N_{eR}M_{U \to R}$ 29.72 (95% of the distribution volume: 4.10–83.09) from the urban to the rural population, the latter distribution having a much wider confidence interval (Fig. 3d). Being a product of the N_e and M posterior density distributions, the 95% bounds of the HPD interval could not be clearly determined due to the aforementioned restriction by the prior bounds of N_e.

Point estimates of effective population *densities* would require point estimates of the area occupied by the urban and rural clusters. As our linear transect is unsuited for such estimates, instead we place upper and lower bounds on density estimates using convex hull bounds on the urban and rural population ranges, and

Table 3 Posterior estimates of IMa2 parameters. HiPt = the estimate value with the highest posterior probability. HiSmth = smoothed highest posterior probability (only provided for t_0). HPD95Lo = the lower bound of the estimated 95% highest posterior density interval. HPD95Hi = the upper bound of the estimated 95% highest posterior density interval. Between parentheses: the HP distribution is artificially restricted by prior upper bounds.

	HiPt or HiSmth	HPD95Lo	HPD95Hi
t_0 (years)	37.125	5.625	129.375
Number of effective individuals in the urban population (N_{eU})	1287.5	312.5	3687.5
Number of effective individuals in the rural population (N_{eR})	3037.5	587.5	12537.5
Number of effective individuals in ancestral population (N_{eA})	40343.75	12537.5	52906.25
Effective migrations from the urban to the rural population per generation ($M_{U\rightarrow R}$)	3.012×10^{-3}	5.2×10^{-4}	(7.6×10^{-3})
Effective migrations from the rural to the urban population per generation ($M_{R\rightarrow U}$)	7.66×10^{-4}	1.3×10^{-4}	(3.6×10^{-3})
Effective population migration from the urban to the rural population, per generation ($2N_{eR}M_{U\rightarrow R}$)	29.72	4.1	−83.09
Effective population migration from the rural to the urban population, per generation ($2N_{eU}M_{R\rightarrow U}$)	3.001	0.5	−11.99

the appropriate bounds on the posterior distribution of IM's N_e estimates (Table 3). The minimum effective density of the urban population is about 0.21 effective individuals km^{-2} (312.5/1470 km^2), whereas the maximum is 2381 (3687.5/1.5 km^2). The rural population's maximum density is 0.32 effective individuals km^{-2} (12537.5/38 900 km^2), whereas the minimum density is difficult to define: although the rural population's range extends at least as far as Lihale (>600 km to the south of the transect), it seems likely to extend even further. Given these very conservative calculations, it seems likely urban effective densities are an order of magnitude higher than rural.

Discussion

We have shown that over a transect of 180 km through inland Tanzania, the widespread multimammate mouse has a genetic structure setting samples from the city of Morogoro apart from surrounding rural samples. The genetic data is consistent with a recent split of urban vs. rural mice, higher urban mouse densities and a source–sink of migration from urban to rural.

The strongest signal of differentiation ($K = 2$) sets *M. natalensis* from 370 km north of our transect, together with a proportion of genotypes at our most northerly transect locality, apart from all remaining localities. This wide-scale geographic distinction is concordant with the distribution of the two divergent mitochondrial lineages that are present in our data set, strongly supporting a shared mitochondrial and nuclear evolutionary history of *M. natalensis* in this region. The mitochondrial lineages were previously estimated to have allopatrically split about 1 million years ago and are likely in secondary contact around Berega (Colangelo *et al.*, 2013). The Berega admixture of north–south microsatellite genotypes in the current data set is consistent with such a secondary contact, which is the subject of a subsequent study in preparation. Although the infrequent detection of mitochondrial lineage B-IV south of Berega suggests that mitochondrial genes have managed to introgress for large distances across the contact zone, the spatial concordance with the distribution of the southern nuclear microsatellite cluster supports the designation of our *M. natalensis* samples south of Berega as a 'super' taxon in which any further genetic subdivision is embedded.

Allowing our nuclear samples to be divided into three or more genetic equilibrium clusters, we could indeed distinguish further substructure within the southern taxon: most *M. natalensis* individuals sampled within the Morogoro urban district were distinct from all other individuals sampled in 10 rural localities (including a southern locality > 600 km south of the transect localities). Taken alone, this relative uniformity of genotype cluster membership over rural sampling would suggest there is effective panmixis of *M. natalensis* over large (>40 000 km^2) geographic scales. We would then not expect samples from the centre of such a uniform area (the samples from urban Morogoro) to be genetically distinct, and so here we explore possible explanations.

Can the observations be explained by the formation of spurious clusters?

It has been suggested that because the cluster model in STRUCTURE assumes uniform relatedness between sampled animals, all but one exemplar of each closely related group within a data set should be discarded before analysis (Schwartz & McKelvey, 2008). Apart from pragmatic issues (inference will depend critically on which exemplars are chosen, and what measure and value are used to define 'close' relatedness), censoring data in this way must be approached with caution. Whereas inclusion of closely related individuals inflates the frequency of allelic states they share, removing closely related individuals also reduces the frequencies of

their *unshared* alleles within the data set, thus reducing the power of cluster detection in ways *orthogonal* to the intended removal of bias (this is also why exemplar choices are nonexchangeable). Instead, we first minimize variation in relatedness through our sampling design: our locality samples were always derived from at least four sites >500 m apart, larger than *M. natalensis'* home range diameter and larger than the average juvenile dispersal distance (Leirs *et al.*, 1996; Borremans *et al.*, 2014). Second, without censoring our data, we ask: Is it likely the *Morogoro* cluster is a spurious cluster arising due to relatedness? The answer is no: individuals in the Morogoro locality are on average no more related to each other than animals in many other localities, and those other localities (with higher relatedness) do not form clusters. A second potential source of spurious clusters is isolation by distance (IBD): Is it likely the Morogoro cluster arises due to IBD? The answer is no: localities surrounding Morogoro cluster with localities about four times more distant than Morogoro. If a distance of, for example, 80 km (between Mkundi and Chalinze) cannot produce a spurious cluster through IBD, then it seems highly unlikely a distance of, for example, 16 km (between Mkundi and Morogoro) would.

No evidence of spatial or temporal isolation

As dispersal is a strong homogenizing force (Slatkin, 1987), the simplest explanation for observed population structure at a local geographic scale is that it has arisen during a period of isolation, or population vicariance. However, it is highly unlikely the Morogoro urban cluster arose through IBD, nor are there obvious geographic barriers to gene flow between urban and rural multimammate mice in our study area. Our landscape connectivity analysis furthermore indicates resistance to dispersal between Morogoro and its neighbouring localities is moderate to low when compared to other neighbour pairs (Table 2), such that Morogoro is no more isolated in terms of estimated landscape connectivity than other localities, and is less isolated than at least four other localities.

An alternative explanation of the observed population structure is a temporal rather than physical barrier to gene flow. Breeding seasons in *M. natalensis* in Tanzania are well delineated in time and are strongly linked to seasonal rainfall patterns (Leirs *et al.*, 1997; Sluydts *et al.*, 2007). Breeding asynchrony caused by local climatic variation may, through reproductive isolation, lead to genetic divergence of animals between neighbouring localities in the absence of obvious landscape barriers (Moore *et al.*, 2005). Apart from Berega, seasonal rainfall patterns throughout the sampling area in Tanzania were quite similar, with a bimodal precipitation pattern and the highest precipitation peak occurring in April (Fig. S3). It is particularly this rainfall season, when plants and crops are plentiful, which determines the onset of breeding in *M. natalensis* (Leirs *et al.*, 1989, 1997). Asynchronous breeding of multimammate mice caused by natural differences between sampling sites therefore seems an unlikely explanation for the observed pattern of genetic structure. A prolonged breeding period in the Morogoro urban area does not seem unlikely due to increased resource availability, as has been reported for several urbanized species (Francis & Chadwick, 2012). However, prolonged urban breeding does not reduce existing overlap with rural breeders, and so seems unlikely to reduce migration below the threshold required to allow divergence by drift (Slatkin, 1987).

Isolation-with-migration's relative estimates of recent splitting time and large effective population sizes suggest no demographic bottleneck

The IM estimate of the splitting time between the urban and the rural populations, with a HP of 37 years (95% interval of 5.6–129 years), falls within the period of Morogoro's rapid urbanization, and is thus consistent with the hypothesis of differentiation across the rural–urban divide. Morogoro has been an important trade centre since the late 19th century, but human densities only started to grow rapidly since the 1960s (with about 97 persons km^{-2} in 1967 and 234 persons km^{-2} in 1978) (Mtatifikolo, 1997; National Bureau of Statistics, 2006). Currently, there are over 1200 persons km^{-2} living in Morogoro, compared with 7–127 persons km^{-2} in the sampled rural localities (Table 1). Although the IM estimates of absolute splitting time and N_e suffer from uncertainty in both the mutation rate and generation time estimates, their values *relative* to other IM estimates are more robustly informative because they allow some of the uncertainties of the model to be cancelled out. For example, in comparison with the recent estimated splitting time of 37 years (74 generations), the estimates of effective population sizes ($N_{eUrban} = 1287.5$, $N_{eRural} = 3037.5$) are clearly large, and would remain so despite any rescaling of the mutation rate. Therefore, it is highly unlikely either population went through a demographic bottleneck in the time elapsed since the ancestral population split into the urban and rural IM populations.

Estimates of migration and effective densities suggest a hotspot of density within the urban perimeter

Isolation-with-migration analyses indicate a significant asymmetry in gene flow between the urban and the rural populations, but this asymmetry is in the opposite direction than we might expect from the field conformation: across the border of a circle (such as the boundary round a city), one can expect a net gene flow

into the circle simply due to its curvature, there being more individuals contacting the outside of a given arc than contacting the inside. The importance of the IM finding that urban *M. natalensis* are net exporters of genes is therefore strengthened, as the 'null' expectation would, if anything, be the opposite: Morogoro swamped by the all-surrounding rural *M. natalensis*. A higher migration from urban multimammate mice to the surrounding range is consistent with the urban environment being able to sustain higher densities of *M. natalensis* compared with the rural environment in a source–sink relationship. Taking into account upper and lower limits on the geographic extents of the (small) urban and (large) the rural populations, the discrete IM-model-based estimates of a small urban N_e vs. large rural N_e seem very likely to map in geographic reality to much higher effective *M. natalensis* densities within the urban perimeter than in the rural surroundings. We suggest therefore the field area does not have homogeneous effective density, but instead a hotspot of density within the urban perimeter. This is consistent with greater resource availability for *M. natalensis* in the urban environment, a pattern demonstrated for the urban–rural contrast in several birds and mammals (Francis & Chadwick, 2012).

Distinguishing between alternative historical and selective scenarios

Our analyses suggest that the distinctive genetic signal of the urban *Mastomys* samples did not arise due to isolation by either a physical or temporal barrier, it is not associated with a strong bottleneck, and there has been little drift during the development of the divergence from the surrounding rural gene pool. Relative densities and migration rates are consistent with a source–sink relationship between urban and rural *M. natalensis*. An explanation for these observations could therefore be strong natural selection acting on those mice that take advantage of the rapidly growing urban and suburban resources available in Morogoro. Strong selection in a hard selective sweep would have left a bottleneck signal, which was not observed, so selection is unlikely to have acted on a new mutation, but on standing variation already present in the *M. natalensis* genepool (Barrett & Schluter, 2008). This standing variation need not have arisen in Morogoro itself. Although it is unlikely individuals pre-adapted to synanthropic life styles arrived en masse through human transport, it is possible that some of the gene pool contributing to the distinctive Morogoro genotypes arrived in individuals from elsewhere in Tanzania, Morogoro being a regional transport hub. However, the ultimate source of the genes that have prospered is in some sense irrelevant: the suggested proximal process is that despite being surrounded by the rural gene pool, *M. natalensis* genes situated in the urban hub of Morogoro have prospered to such extent that a high-density genetically distinct cluster has been established and, rather than being swamped by gene flow from the surrounding range (Lenormand, 2002), it is actually a net exporter of genes. Moreover, any selective advantage gained in the urban environment does not seem to carry over into the rural environment: the urban cluster is exporting genes, whereas the neighbouring rural multimammate mice remain clear members of the rural cluster.

A number of previous studies have also identified genetic differentiation of rodents living in urban environments in comparison with neighbouring rural sites or in other sites within the same city. However, in these cases the distinction was attributed to intense physical isolation of urban rodents rather than selection (Munshi-South & Kharchenko, 2010; Chiappero *et al.*, 2011; Harris *et al.*, 2013). For example, white-footed mice sampled in New York city are deeply genetically structured according to the degree of intervening impervious surface cover (Munshi-South & Kharchenko, 2010; Munshi-South, 2012; Harris *et al.*, 2015), and genetic drift has led to reduced genetic diversity compared with populations in rural areas (Munshi-South *et al.*, 2016). Our samples were not taken at isolated vegetated spots in the city centre of Morogoro where physical isolation might be expected, but rather from fallow maize fields in the suburbs of the city with no apparent physical barriers to the rural surrounds. The genetic differentiation of urban individuals relative to their rural counterparts, without a loss of diversity and in the face of gene flow, suggests the selection gradient itself is maintaining the genetic differentiation. Strong directional selection on urban rodents has previously been detected at multiple loci (Harris *et al.*, 2013), and cranial capacity has been shown to be repeatedly greater in the urban than in the rural populations (Snell-Rood & Wick, 2013). Future studies might focus on whether there are any behavioural or morphological distinctions associated with the observed genetic distinction between rural and urban *M. natalensis*.

Acknowledgments

We are very grateful to Alexis Ribas (University of Barcelona at the time of fieldwork) and all members of the Pest Management Centre (Sokoine University of Agriculture, Morogoro), most especially to Khalid Kibwana, Shabani Lutea, Kamil Lutea, Ramadhani Idd and Geofrey Sabuni, for their help and support during fieldwork. We thank Christopher Sabuni, Vincent Sluydts, Anne Laudisoit, Hana Patzenhauerová, Elisabeth Fichet-Calvet and Benny Borremans for their expert opinions on *M. natalensis* habitat values. This work was supported by the University of Antwerp (grant no. 24323), the Fund for Scientific Research–Flanders (FWO grant no. 1.5.264.12 and K.2.209.10.N.01), the

Czech Science Foundation (GACR grant P502/11/J070 and P506/10/0983) and NextGen grant CZ.1.07/2.3./20.0303. During most of this study, SG and JGB were PhD and postdoctoral fellows with FWO.

Data accessibility

All collected samples are deposited in the tissue collection of the Evolutionary Ecology group of the University of Antwerp and are available upon request.

Partial cytochrome B DNA sequences: GenBank accession numbers KF779499 - KF779925 (see Table S2 for more details). Data available from the Dryad Digital Repository: doi: 10.5061/dryad.n5v84

- Field data (sample locations, dates and species identification of captured animals)
- Microsatellite data (including location origin information used in STRUCTURE)
- Parameter settings for STRUCTURE
- IMa2 input files (input data, commands for parameter settings)
- Landscape costs (the costs assigned to landscape elements for connectivity analyses in Circuitscape)

References

Ansel, J., Arya, K. & Cooperman, G. 2009. *DMTCP: Transparent Checkpointing for Cluster Computations and the Desktop*. The 23rd IEEE International Parallel and Distributed Processing Symposium, pp. 1–12. Rome, Italy.

Badyaev, A.V., Young, R.L., Oh, K.P. & Addison, C. 2008. Evolution on a local scale: developmental, functional, and genetic bases of divergence in bill form and associated changes in song structure between adjacent habitats. *Evolution* **62**: 1951–1964.

Barrett, R.D.H. & Schluter, D. 2008. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**: 38–44.

Biomatters 2014. Geneious version 6.1.7.

Bird, C.E., Fernandez-Silva, I., Skillings, D.J. & Toonen, R.J. 2012. Sympatric speciation in the post "modern synthesis" era of evolutionary biology. *Evol. Biol.* **39**: 158–180.

Borremans, B., Hughes, N.K., Reijniers, J., Sluydts, V., Katakweba, A.A.S., Mulungu, L.S. *et al*. 2014. Happily together forever: temporal variation in spatial patterns and complete lack of territoriality in a promiscuous rodent. *Popul. Ecol.* **56**: 109–118.

Brouat, C., Loiseau, A., Kane, M., Ba, K. & Duplantier, J.M. 2007. Population genetic structure of two ecologically distinct multimammate rats: the commensal *Mastomys natalensis* and the wild *Mastomys erythroleucus* in southeastern Senegal. *Mol. Ecol.* **16**: 2985–2997.

Carlsson, J. 2008. Effects of microsatellite null alleles on assignment testing. *J. Hered.* **99**: 616–623.

Chiappero, M.B., Panzetta-Dutari, G.M., Gomez, D., Castillo, E., Polop, J.J. & Gardenal, C.N. 2011. Contrasting genetic structure of urban and rural populations of the wild rodent *Calomys musculinus* (Cricetidae, Sigmodontinae). *Mamm. Biol.* **76**: 41–50.

Christensen, J.T. 1995. On the ecology of the multimammate rat, *Mastomys natalensis* (A. Smith 1834), with reference to its role as an agricultural pest. PhD thesis, Aarhus University, Aarhus, Denmark.

Colangelo, P., Verheyen, E., Leirs, H., Tatard, C., Denys, C., Dobigny, G. *et al.* 2013. A mitochondrial phylogeographic scenario for the most widespread African rodent, *Mastomys natalensis*. *Biol. J. Linn. Soc.* **108**: 901–916.

Di Gregorio, A. & Jansen, L. 2000. *Land Cover Classification System, Classification Concepts and User Manual*. Food and Agriculture Organization of the United Nations, Rome.

Duplantier, J.M., Brittondavidian, J. & Granjon, L. 1990. Chromosomal characterization of 3 species of the genus *Mastomys* in Senegal. *Z. Zool. Syst. Evol.* **28**: 289–298.

Duplantier, J.M., Granjon, L. & Ba, K. 1997. Répartition biogéographique des petits rongeurs au Sénégal. *J. Afr. Zool.* **111**: 17–26.

Ellegren, H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**: 551–558.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**: 435–445.

Evans, K.L., Gaston, K.J., Sharp, S.P., McGowan, A. & Hatchwell, B.J. 2009. The effect of urbanisation on avian morphology and latitudinal gradients in body size. *Oikos* **118**: 251–259.

Excoffier, L. & Lischer, H.E.L. 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.

Francis, R.A. & Chadwick, M.A. 2012. What makes a species synurbic? *Appl. Geogr.* **32**: 514–521.

Galan, M., Van Hooft, W.F., Legrand, D., Berthier, K., Loiseau, A., Granjon, L. *et al.* 2004. A multiplex panel of microsatellite markers for widespread sub-Saharan rodents of the genus *Mastomys*. *Mol. Ecol. Notes* **4**: 321–323.

Gliwicz, J. 1980. Ecological aspect of synurbanization of the striped field mouse, *Apodemus agrarius*. *Wiadomosci Ekologiczne* **26**: 117–124.

Goudet, J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**: 184–186.

Grahame, J.W., Wilding, C.S. & Butlin, R.K. 2006. Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution* **60**: 268–278.

Granjon, L., Duplantier, J.M., Catalan, J., Britton-Davidian, J. & Bronner, G.N. 1996. Conspecificity of *Mastomys natalensis* (Rodentia:Muridae) from Senegal and South Africa: evidence from experimental crosses, karyology and biometry. *Mammalia* **60**: 697–706.

Hardy, O.J. & Vekemans, X. 2002. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**: 618–620.

Harris, S.E., Munshi-South, J., Obergfell, C. & O'Neill, R. 2013. Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York metropolitan area. *PLoS ONE* **8**: e74938.

Harris, S.E., O'Neill, R.J. & Munshi-South, J. 2015. Transcriptome resources for the white-footed mouse (*Peromyscus leucopus*): new genomic tools for investigating ecologically

divergent urban and rural populations. *Mol. Ecol. Resour.* **15**: 382–394.

Hey, J. 2010. Isolation with Migration models for more than two populations. *Mol. Biol. Evol.* **27**: 905–920.

Hubisz, M.J., Falush, D., Stephens, M. & Pritchard, J.K. 2009. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**: 1322–1332.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E. & Butlin, R.K. 2010. Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philos. Trans. R. Soc. Lond. Biol. Sci.* **365**: 1735–1747.

Katakweba, A.A.S., Mulungu, L.S., Eiseb, S.J., Mahlaba, T.A., Makundi, R.H., Massawe, A.W. *et al.* 2012. Prevalence of haemoparasites, leptospires and coccobacilli with potential for human infection in the blood of rodents and shrews from selected localities in Tanzania, Namibia and Swaziland. *Afr. Zool.* **47**: 119–127.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15**: 1179–1191.

Lecompte, É., Granjon, L., Peterhans, J.K. & Denys, C. 2002. Cytochrome b-based phylogeny of the Praomys group (Rodentia, Murinae): a new African radiation? *C. R. Biol.* **325**: 827–840.

Leirs, H. 2013. Mastomys. In: *Mammals of Africa: Volume III*, 1st ed edn. Vol. 3 (J. Kingdon, D. Happold, T. Butynski, M. Hoffmann, M. Happold & J. Kalina, eds), pp. 460–472. Bloomsbury Publishing, London, UK.

Leirs, H., Verheyen, W., Michiels, M., Verhagen, R. & Stuyck, J. 1989. The relation between rainfall and the breeding season of *Mastomys natalensis* (Smith, 1834) in Morogoro, Tanzania. *Ann. Soc. R. Zool. Belge.* **119**: 59–64.

Leirs, H., Verhagen, R. & Verheyen, W. 1993. Productivity of different generations in a population of *Mastomys natalensis* rats in Tanzania. *Oikos* **68**: 53–60.

Leirs, H., Verheyen, W. & Verhagen, R. 1996. Spatial patterns in *Mastomys natalensis* in Tanzania (Rodentia, Muridae). *Mammalia* **60**: 545–556.

Leirs, H., Stenseth, N.C., Nichols, J.D., Hinds, J.E., Verhagen, R. & Verheyen, W. 1997. Stochastic seasonality and nonlinear density-dependent factors regulate population size in an African rodent. *Nature* **389**: 176–180.

Lenormand, T. 2002. Gene flow and the limits to natural selection. *Trends Ecol. Evol.* **17**: 183–189.

Li, C.C., Weeks, D.E. & Chakravarti, A. 1993. Similarity of DNA fingerprints due to chance and relatedness. *Hum. Hered.* **43**: 45–52.

Loiseau, A., Konečný, A., Galan, M., Bryja, J., Cosson, J.F. & Brouat, C. 2007. New polymorphic microsatellite loci for rodents of the genus *Mastomys* using PCR multiplexing, and cross-species amplification in *Myomys* and *Praomys*. *Mol. Ecol. Notes* **7**: 684–687.

McRae, B.H. 2006. Isolation by resistance. *Evolution* **60**: 1551–1561.

Mohr, K., Leirs, H., Katakweba, A. & Machang'u, R. 2007. Monitoring rodents movements with a biomarker around introduction and feeding foci in an urban environment in Tanzania. *Afr. Zool.* **42**: 294–298.

Monadjem, A., Mahlaba, T.a.A., Dlamini, N., Eiseb, S.J., Belmain, S.R., Mulungu, L.S. *et al.* 2011. Impact of crop cycle on movement patterns of pest rodent species between fields and houses in Africa. *Wildlife Res.* **38**: 603–609.

Moore, I.T., Bonier, F. & Wingfield, J.C. 2005. Reproductive asynchrony and population divergence between two tropical bird populations. *Behav. Ecol.* **16**: 755–762.

Mtatifikolo, F.P. 1997. The dynamics of the urbanization forces in Tanzania and related policy and research issues. *Tanzan. J. Popul. Stud. Dev.* **4**: 67–83.

Mulungu, L., Makundi, R., Leirs, H., Massawe, A., Vibe Petersen, S. & Stenseth, N. 2003. The rodent density-damage function in maize fields at an early growth stage. In: *Rats, Mice and People: Rodent Biology and Management* (G.R. Singleton, L.A. Hinds, C.J. Krebs & D.M. Spratt, eds), pp. 301–303. Australian Centre for International Agricultural Research, Canberra.

Munshi-South, J. 2012. Urban landscape genetics: canopy cover predicts gene flow between white-footed mouse (*Peromyscus leucopus*) populations in New York City. *Mol. Ecol.* **21**: 1360–1378.

Munshi-South, J. & Kharchenko, K. 2010. Rapid, pervasive genetic differentiation of urban white-footed mouse (*Peromyscus leucopus*) populations in New York City. *Mol. Ecol.* **19**: 4242–4254.

Munshi-South, J., Zolnik, C.P. & Harris, S.E. 2016. Population genomics of the Anthropocene: urbanization is negatively associated with genome-wide variation in white-footed mouse populations. *Evol. Appl.* **9**: 546–64.

Mwanjabe, P.S. & Leirs, H. 1997. An early warning system for IPM-based rodent control in smallholder farming systems in Tanzania. *Belg. J. Zool.* **127**: 49–58.

National Bureau of Statistics. 2006. *Ministry of Planning, Economy and Empowerment*. Vol. 10. Dar es Salaam, Tanzania.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Peterson, T.C. & Vose, R.S. 1997. An overview of the Global Historical Climatology Network temperature database. *Bull. Am. Meteorol. Soc.* **78**: 2837–2849.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwartz, M.K. & McKelvey, K.S. 2008. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conserv. Genet.* **10**: 441–452.

Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science* **236**: 787–792.

Sluydts, V., Crespin, L., Davis, S., Lima, M. & Leirs, H. 2007. Survival and maturation rates of the African rodent, *Mastomys natalensis*: density-dependence and rainfall. *Integr. Zool.* **2**: 220–232.

Snell-Rood, E.C. & Wick, N. 2013. Anthropogenic environments exert variable selection on cranial capacity in mammals. *Proc. Biol. Sci. R. Soc.* **280**: 20131384.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M. & Shipley, P. 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**: 535–538.

Wang, C.L., Schroeder, K.B. & Rosenberg, O.N.A. 2012. A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics* **192**: 651–669.

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G. & Sibly, R. 2003. Likelihood-based estimation of microsatellite mutation rate. *Genetics* **164**: 781–787.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Figure S1 Output from STRUCTURE clustering of the microsatellite data (Q matrices labelled using CLUMPAK (Kopelman *et al.*, 2015)) when allowing for null-alleles using the 'recessive allele method' of STRUCTURE.

Figure S2 The flow of 'electric' current between two localities, presented separately for each pair of sampled localities.

Figure S3 Monthly mean precipitation at meteorological stations near the sampling locations (see Fig. 1).

Table S1 Numbers of small mammals captured at each sampling grid.

Table S2 GenBank accessions numbers for cytochrome b sequences and the corresponding mitochondrial lineage name, following (Colangelo *et al.*, 2013).

Table S3 Basic genetic statistics per sampling locality and per microsatellite locus as calculated in the R package Hierfstat: Weir and Cockrham *Fst* per locus (over all localities), allelic richness, observed heterozygosity, expected heterozygosity, and the *P*-values of the χ^2 test for deviation of Hardy-Weinberg equilibrium.

Table S4 The pairwise correlation among allele frequencies of loci (measure of linkage disequilibrium), r^2.

Data deposited at Dryad: doi: 10.5061/dryad.n5v84

Received 27 May 2016; accepted 14 June 2016