A young age of subspecific divergence in the desert locust *Schistocerca gregaria*

Marie-Pierre Chapuis[1,2], Louis Raynal[3,4], Christophe Plantamp[5], Laurence Blondin[6], Jean-Michel Marin[3,4] and Arnaud Estoup[4,7]

[1]CIRAD, CBGP, Montpellier, France.

[2]CBGP, CIRAD, INRA, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France.

[3] IMAG, Univ de Montpellier, CNRS, Montpellier, France.

[4]Institut de Biologie Computationnelle (IBC), Montpellier, France.

[5]ANSES, Laboratoire de Lyon, France.

[6]CIRAD, BGPI, Montpellier, France.

[7]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France.

Corresponding author: Marie-Pierre Chapuis, marie-pierre.chapuis@cirad.fr

Keywords: Approximate Bayesian Computation; colonization; desert locust; divergence time; Holocene; Orthoptera; paleo-vegetation; Random Forest; *Schistocerca gregaria*

Short title: Evolutionary history of *Schistocerca gregaria*

Abstract

Dating population divergence within species from molecular data and relating such dating to climatic and biogeographic changes is not trivial. Yet it can help formulating evolutionary hypotheses regarding local adaptation and future responses to changing environments. Key issues include statistical selection of a demographic and historical scenario among a set of possible scenarios and estimation of the parameter(s) of interest under the chosen scenario. Such inferences greatly benefit from new statistical approaches including Approximate Bayesian Computation - Random Forest (ABC-RF), the later providing reliable inference at a low computational cost, with the possibility to take into account prior knowledge on both biogeographical history and genetic markers. Here, we used ABC-RF, including or not independent information on evolutionary rate and pattern at microsatellite markers, to decipher the evolutionary history of the African arid-adapted pest locust, *Schistocerca gregaria*. We found that the evolutionary processes that have shaped the present geographical distribution of the species in two disjoint northern and southern regions of Africa were recent, dating back 2.6 Ky (90% CI: 0.9 – 6.6 Ky). ABC-RF inferences also supported a southern colonization of Africa from a low number of founders of northern origin. The inferred divergence history is better explained by the peculiar biology of *S. gregaria*, which involves a density-dependent swarming phase with some exceptional spectacular migrations, or by a brief fragmentation of the African forest core during the interglacial late Holocene, rather than a continuous colonization resulting from the continental expansion of open vegetation habitats during the past Quaternary glacial episodes.

Introduction

As in other regions of the world, Africa has gone through several major episodes of climate change since the early Pleistocene (deMenocal 1995 and 2004). During glaciation periods, the prevalent climate was colder and drier than nowadays, and became more humid during warmer interglacial periods. These climatic phases resulted in shifts of vegetation (Vivo and Carmignotto 2004) and are most likely at the origin of the current isolation between northern and southern distributions of arid-adapted species (Monod 1971). In Africa, at least fifty-six plant species show disjoint geographical distributions in southern and northern arid areas (Monod 1971; Jurgens 1997; Lebrun 2001). Similarly, a number of animal vertebrate species show meridian disjoint distributions on this continent, including eight mammals and 29 birds (Monod 1971; de Vivo and Carmignotto 2004; Lorenzen *et al.* 2012). The desert locust, *Schistocerca gregaria*, is among the few examples of insect species distributed in two distinct regions along the north-south axis of Africa. Other known disjunctions in insects are interspecific and concern species of the families Charilaidae (Orthoptera) and Mythicomyiidae (Diptera), and of the genus Fidelia (Hymenoptera) (Le Gall *et al.* 2010). Similarities in extant distributions of African arid-adapted species across divergent taxonomic groups point to a common climatic history and an important role of environmental factors. Yet, to our knowledge, studies relating evolutionary history and climatic history have rarely been carried out in this continent (but see mitochondrial studies by Miller *et al.* 2011 on the ostrich, Atickem *et al.* 2018 on the black-backed jackal, and Moodley *et al.* 2018 on the white rhinoceros).

Dating population or subspecies divergence within a species and relating such dating to climatic and biogeographic changes in the species history is not trivial. First, global climate models have been largely calibrated using northern hemisphere drivers and validation datasets. Their quality has therefore been tested less often in Africa, even less so when it comes to hindcasting potential distributions using projections of such climate models into different past temporal windows. Recent comparisons between botanical and climate models have suggested that climate forcing in Africa may operate in a different way, and have therefore shed some doubts regarding the

validity of such projections, in particular into long time periods involving several thousand years into the past (Chase and Meadows 2007; Dupont 2011). Second, finding a reliable calibration to convert measures of genetic divergence into units of absolute time is challenging, especially so for recent evolutionary events (Ho *et al.* 2008). Extra-specific fossil calibration may lead to considerable overestimates of divergence times and internal fossil records are often lacking (Ho *et al.* 2008). A sensible approach when internal calibration is available for a related species is to import an evolutionary rate estimated from sequence data of this species (Ho *et al.* 2008). Unfortunately, on the African continent, fossils, such as radiocarbon-dated ancient samples, remain relatively rare and are often not representative of modern lineages (*e.g.*, Le Gall 2010 for insects). The lack of paleontological and archaeological records is partly due to their fragility under the aridity conditions of the Sahara. The end-result is that the options to relate population divergence to biogeographic events in this region are very limited.

In this context, the use of versatile molecular markers, such as microsatellite loci, for which evolutionary rates can be obtained from direct observation of germline mutations in the species of interest, represents a useful alternative. Microsatellite mutation rates exceed by several orders of magnitude that of point mutation in DNA sequences, ranging from 10^{-6} to 10^{-2} events per locus and per generation (Ellegren 2000). This providing allows both to observe mutation events in parent-offspring segregation data of realistic sample size and work out the recent history of related populations. However, the use of microsatellite loci to estimate divergence times at recent evolutionary time-scales still needs overcoming significant challenges. Since microsatellite allele sizes result from the insertion or deletion of single or multiple repeat units and are tightly constrained, these markers can be characterized by high levels of homoplasy that can obscure inferences about gene history (*e.g.*, Estoup *et al.* 2002). In particular, at large time scales (i.e., for distantly related populations), genetic distance values do not follow anymore a linear relationship with time, reach a plateau and hence provide biased unreliable estimation of divergence time (Takezaki and Nei 1996; Feldman *et al.* 1997; Pollock 1998). Microsatellites remain informative

with respect to divergence time only if the population split occurs within the period of linearity with time (Feldman *et al.* 1997; Pollock 1998). The exact value of the differentiation threshold above which microsatellite markers would no longer accurately reflect divergence times will depend on constraints on allele sizes and population-scaled mutation rates (Feldman *et al.* 1997; Pollock 1998). For any inferential framework, including independent information on microsatellite allele size constraints and mutation rates (for instance into priors when using Bayesian methods) is expected to improve the accuracy of parameter estimation, especially when considering divergence times between populations.

The desert locust, *S. gregaria*, is a generalist herbivore that can be found in arid grasslands and deserts in both northern and southern Africa (Figure 1a). In its northern range, the desert locust is one of the most widespread and harmful agricultural pest species with a huge potential outbreak area, spanning from West Africa to Southwest Asia. The desert locust is also present in the southwestern arid zone (SWA) of Africa, which includes South-Africa, Namibia, Botswana and south-western Angola. The southern populations of the desert locust are termed *S. g. flaviventris* and are geographically separated by nearly 2,500 km from populations of the nominal subspecies from northern Africa, *S. g. gregaria* (Uvarov 1977). The isolation of *S. g. flaviventris* and *S. g. gregaria* lineages was recently supported by highlighting distinctive mitochondrial DNA haplotypes and male genitalia morphologies (Chapuis *et al.* 2016). Yet, the precise history of divergence remains elusive.

The main objective of the present study is to unravel the historical and evolutionary processes that have shaped the present disjoint geographical distribution of the desert locust and the genetic variation observed both within and between populations of its two subspecies. To this aim, we first used paleo-vegetation maps to construct biogeographic scenarios relevant to African species from arid grasslands and deserts. We then used molecular data obtained from microsatellite markers for which we could obtain independent information on evolutionary rates and allele size constraints in the species of interest from direct observation of germline mutations (Chapuis *et al.*

2015). We applied newly available algorithms of the Approximate Bayesian Computation - random forest method (ABC-RF; Pudlo *et al.* 2016; Estoup *et al.* 2018a; Raynal *et al.* 2019) on our microsatellite population genetic data to compare a set of thoroughly formalized and justified evolutionary scenarios and estimate the divergence time between *S. g. gregaria* and *S. g. flaviventris* under the most likely of our scenarios. Finally, we interpret our results in the light of the paleo-vegetation information we compiled and various biological features of the desert locust.

New approaches

Due to its great flexibility, Approximate Bayesian Computation (ABC, Beaumont *et al.* 2002) is an increasingly common statistical approach used to perform model-based inferences in a Bayesian setting, especially when complex models are considered (*e.g.*, Beaumont 2010, Bertorelle *et al.* 2010, Csilléry *et al.* 2010). However, both theoretical arguments and simulation experiments indicate that scenario' posterior probabilities can be poorly evaluated by standard ABC methods, even though the numerical approximations of such probabilities can preserve classification (Robert *et al.* 2011). To overcome this problem, Pudlo *et al.* (2016) recently proposed a novel approach based on a machine learning tool named random forests (RF) (Breiman 2001), hence leading to the ABC-RF methodology. When compared with standard ABC methods, the ABC-RF approach enables efficient discrimination among scenarios and estimation of posterior probability of the best scenario while being computationally less intensive. Building on that success, Raynal *et al.* (2019) recently proposed an extension of the RF methodology applied in a (non-parametric) regression setting to estimate the posterior distributions of parameters of interest under a given scenario. When compared with various ABC solutions, this new RF method offers many advantages: a significant gain in terms of robustness to the choice of the summary statistics; independence from any type of tolerance level; and a good trade-off in term of quality of point estimator precision of parameters and credible interval estimations for a given computing time (Raynal *et al.* 2019). An overview of the ABC-RF methods used in the present paper is provided in Supplementary Material S1. Readers

can consult Pudlo *et al.* (2016), Fraimout *et al.* (2017), Estoup *et al.* (2018a,b) and Marin *et al.* (2018) for scenario choice, and Raynal *et al.* (2019) for parameter estimation to access to further detailed statistical descriptions, testing and applications of ABC-RF algorithms.

To our knowledge, the present study is the first one using recently developed ABC-RF algorithms to carry out inferences about both scenario choice and parameter estimation, on a real multi-locus microsatellite dataset. It includes and illustrates three novelties in statistical analyses that were particularly useful for reconstructing the evolutionary history of the divergence between *S. g. gregaria* and *S. g. flaviventris* subspecies: model grouping analyses based on several key evolutionary events, assessment of the quality of predictions to evaluate the robustness of our inferences, and incorporation of previous information on the mutational setting of the used microsatellite markers.

(1) *Model grouping*. Both the poor knowledge on the species history and the complex climatic history of Africa make it necessary to consider potentially complex evolutionary scenarios. We formalized eight competing scenarios including (or not) three key evolutionary events that we identified as having potentially played a role in setting up the disjoint distribution of the two locust subspecies (for details see the section *Formalization of evolutionary scenarios* in Materials and methods). Following the new approach proposed by Estoup *et al.* (2018a), we processed ABC-RF analyses grouping scenarios based on the presence or absence of each type of evolutionary event, before considering all scenarios separately. Such grouping approach in scenario choice is of great interest to disentangle the level of confidence of our approach to make inferences about each specific evolutionary event of interest.

(2) *Assessing the quality of predictions*. For scenario choice and parameter estimation, we evaluated the robustness of our inferences at both a global (*i.e.*, prior) and a local (*i.e.*, posterior) scale. The global prior error was computed, using the computationally parsimonious out-of-bag prediction method for scenarios identity and parameter values covering the entire prior multidimensional space. Since error levels may differ depending on the location of an observed

dataset in the prior data space, prior-based indicators are poorly relevant, aside from their use to select the best classification method and set of predictors, here our summary statistics. Therefore, in addition to global prior errors, we computed local posterior errors, conditionally to the observed dataset. The latter errors measure prediction quality exactly at the position of the observed dataset. For model choice, we demonstrated that the error measure given the observation can be computed as 1 minus the posterior probability of the selected scenario. For parameter estimation, we propose an innovative way to approximate local posterior errors, again relying partly on out-of-bag predictions. See the section *Local posterior errors* in Supplementary Material S1 for details. These statistical novelties were implemented in a new version of the R library abcrf (version 1.8) available on R CRAN. Finally, for estimation of divergence time between the two subspecies, we evaluated how accurately the divergence time posterior distributions reflected true divergence time values and the threshold above which the divergence time posterior estimates reach a plateau. To do this, we used simulated pseudo-observed datasets to compute error measures conditionally to a subset of fixed divergence time values chosen to cover the entire prior interval.

(3) *Incorporation of previous information into the microsatellite mutational setting*. Our ABC-RF statistical treatments benefited from the incorporation of previous estimations of mutation rates and allele size constraints for the microsatellite loci used in this study. Microsatellite mutation rate and pattern of most eukaryotes remains to a large extent unknown, and, to our knowledge, the present study is a rare one where independent information on mutational features was incorporated into the microsatellite prior distributions. We thoroughly evaluated to which extent the incorporation of such independent information improved the performance of ABC-RF for choosing among evolutionary scenarios and for estimating the time of divergence between the two locust subspecies.

Results

Formalization of evolutionary scenarios

Using a rich corpus of (paleo-)vegetation data, we reconstructed the present time (Fig. 1C) and past time (Figs. 1D-F) distribution ranges of *S. gregaria* in Africa, going back to the Last Glacial Maximum period (LGM, 26 to 14.8 Ky ago). Maps of vegetation cover for glacial arid maximums (Figs. 1E and 1F) showed an expansion of open vegetation habitats sufficient to make the potential range of the species continuous from the Horn of Africa in the north-west to the Cape of Good Hope in the south. Maps of vegetation cover for interglacial humid maximums (Fig. 1D) showed a severe contraction of deserts. These maps helped us formalize eight competing evolutionary scenarios (Figure 2), as well as bounds of prior distributions for various parameters (see the section *Prior setting for divergence parameters* in Materials and methods). The eight competing scenarios included different combinations of three key evolutionary events that we identified as having potentially played a role in setting up the observed disjoint distribution of the two locust subspecies: (i) a long population size contraction in the ancestral population, due to the reduction of open vegetation habitats during the interglacial periods, (ii) a bottleneck in the southern subspecies *S. g. flaviventris* right after divergence, associated to a single long-distance migration event of a small fraction of the ancestral population, and (iii) a secondary contact with an asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*, in order to consider the many climatic transitions of the last Quaternary.

Scenario choice

ABC-RF analyses supported the same best scenario or group of scenarios for all ten replicate analyses (Table 1). The classification votes and posterior probabilities estimated for the observed microsatellite dataset were the highest for the groups of scenarios in which (i) *S. g. flaviventris* experienced a bottleneck event at the time of the split (average of 2890 votes out of 3,000 RF-trees;

posterior probability = 0.965), (ii) the ancestral population experienced a population size contraction (2245 of 3,000 RF-trees; posterior probability = 0.746), and (iii) no admixture event occurred between populations after the split (2370 of 3,000 RF-trees; posterior probability = 0.742). When considering the eight scenarios separately, the highest classification vote was for scenario 4, which congruently excludes secondary contact and includes a population size contraction in the ancestral population and a bottleneck event at the time of divergence in the *S. g. flaviventris* subspecies (1777 of 3,000 RF-trees). The posterior probability of scenario 4 averaged 0.584 over the ten replicate analyses (Table 1).

Table S2.1 (Supplementary Material S2) shows that only two other scenarios obtained at least 5% of the votes: scenario 2 including only a single bottleneck event in *S. g. flaviventris* (mean of 537 votes) and scenario 8 with a bottleneck event in *S. g. flaviventris*, a population size contraction in the ancestral population and a secondary contact with admixture from *S. g. gregaria* into *S. g. flaviventris* (mean of 380 votes). All other scenarios obtained less than 5% of the votes and were hence even more weakly supported. Scenario 4 obtained the highest number of votes also for analyses based on a naive mutational prior setting for microsatellite markers, *i.e.*, when drawing prior values for mean mutation parameters from uniform distributions instead of setting them to a fixed value as in our informed mutational prior setting (Table 1 and Table S3.1, Supplementary Material S3; see also the Materials and methods section *Microsatellite dataset, mutation rate and mutation model* for details about the microsatellite prior distributions for the informed and naive mutational settings). Posterior probability values for scenario 4 and for the best groups of scenarios were slightly lower when using a naive mutational prior setting, except for the group without any admixture event (Table 1).

We found that posterior error rates (*i.e.*, 1 minus the posterior probabilities) were lower than prior error rates for the analyses considering either groups of scenarios based on the presence (or not) of a bottleneck in *S. g. flaviventris* (*i.e.*, 3.5% versus 10.2%) or the scenarios separately (*i.e.*, 41.6% versus 47.9%). For other groups of scenarios, the discrimination power was similar at both

the global (prior error rates) and local (posterior error rates) scales, with values ranging from 23.5% to 25.8% (Table 1). Altogether, these results indicate that the observed dataset belongs to a region of the data space where the power to discriminate among scenarios is higher than the global power computed over the whole prior data space, and that the presence or absence of a bottleneck in *S. g. flaviventris* is the demographic event with the most robust prediction in our ABC-RF treatments. These results hold true when using a naive mutational prior setting (Table 1). They can be visually illustrated by the projection of the reference table datasets and the observed one on a single (when analyzing pairwise groups of scenarios) or on the first two linear discriminant analysis (LDA) axes (when analyzing the eight scenarios considered separately) (Figure S2.1, Supplementary Material S2 and Figure S3.1, Supplementary Material S3).

Figure S2.2, Supplementary Material S2, illustrates how RFs automatically rank the summary statistics according to their level of information. It shows that the set of most informative statistics is different depending on the comparisons (groups of scenarios or individual scenarios). Two sample statistics that measure the amount of genetic variation shared between populations (F_{ST}, LIK and DM2) were among the most informative when discriminating among groups of scenarios including or not an admixture event. For groups of scenarios differing by population size variation events, statistics summarizing variation between the two subspecies samples (F_{ST} and DM2 for the bottleneck event in *S. g. flaviventris*; DAS and LIK for the population size contraction in the ancestral population) and statistics summarizing genetic variation within subspecies samples (mean expected heterozygosity and mean number of alleles for both population size variation events) were among the most discriminative ones. Only eight single sample statistics were not informative (according to their position relatively to the noise statistics added to our treatments) when considering the eight individual scenarios separately. All those non informative statistics were associated to the set of transcribed microsatellites (Figure S2.3, Supplementary Material S2). When using a naive mutational prior setting, twice as many more summary statistics turned out to be non-informative (Figure S3.2, Supplementary Material S3).

Parameter estimation

Figure 3A shows point estimates with 90% credibility intervals of the posterior distribution of the divergence time between the two subspecies under the best supported scenario 4. Our estimations point to a young age of subspecies divergence, with a median divergence time of 2.6 Ky and a 90% credibility interval of 0.9 to 6.6 Ky, when using some informed mutational priors and assuming an average of three generations per year (Table 2 and Table S2.2, Supplementary Material S2). The naive mutational prior setting led to a median estimate of 1.7 Ky with a wider 90% credibility interval of 0.4 to 7.9 Ky (Fig. 3a, Table 2 and Table S3.2, Supplementary Material S3). Accuracy of divergence time estimation was similar at both the global and local scales (*i.e.*, normalized mean absolute errors of 0.369 and 0.359, respectively; Table 3). The incorporation of independent information into prior distributions of mutational parameters allowed a more accurate estimation of the median divergence time (*cf.* NMAE values were 30 % higher when using the naive mutational prior setting; Table 3). This observation holds true for the three other demographic parameters, with NMAE values 4 to 35 % lower when using informed mutational priors (Table 3).

Using the median as a point estimate, we estimated that the population size contraction in the ancestor could have occurred at a time about three fold older than the divergence time between the subspecies (Table 2). Estimations of the ratio of stable effective sizes of the *S. g. gregaria* and *S. g. flaviventris* populations (*i.e.*, N_f / N_g) showed large 90% credibility intervals and include the rate value of 1 (Table 2). Accuracy analysis indicates that our genetic data withhold little information on this composite parameter (Table 3). The bottleneck intensity during the colonization of south-western Africa (*i.e.*, db_f / Nb_f) shows the highest accuracy of estimation (Table 3). The median of 1 and the 90% credibility interval of 0.5 to 2.4 exclude severe and mild bottlenecks and rather sustain a strong to moderate event (Table 2).

The most informative summary statistics were different depending on the parameter of interest (results not shown). For the time since divergence between the two subspecies, the most

informative statistics corresponded to the expected heterozygosity computed within the *S. g. flaviventris* sample and the mean index of classification from *S .g. flaviventris* to *S. g. gregaria* (Figure S2.4, Supplementary Material S2). The addition of noise variables in our treatments showed that most statistics characterizing genetic variation within the *S. g. gregaria* sample were not informative. These results hold true when using a naive mutational prior setting (Figure S3.3, Supplementary Material S3).

Constraints on allele sizes in conjunction with high population-scaled mutation rates potentially strongly affect the linearity of the relationship between mutation accumulation and time of divergence estimated from microsatellite data. We thus evaluated the accuracy of ABC-RF estimation of the population divergence time as a function of the time scale, under scenario 4. Analyses of pseudo-observed datasets using informed mutational priors showed that the ABC-RF median estimate of divergence time reached a plateau for time scales ≥ 100,000 generations (Figure 4). Thus, the divergence time between *S. g. flaviventris* and *S. g. gregaria* estimated on our real microsatellite dataset (~10,000 generations) is positioned within the period of linearity with time, well before reaching a plateau reflecting a saturation of genetic information at microsatellite markers. It is hence expected to represent a sensible estimation of the actual divergence time. Figure 4 also showed that the use of a naive mutational prior setting led to a downward bias of the point estimate and to a lower accuracy of estimations. As a result, the incorporation of independent information into the prior distributions of mutational parameters considerably decreased both the NMAE for median estimates and the relative amplitude for time-scales < 100,000 generations (Figure 5).

Discussion

A young age of subspecific divergence

With a 90% credibility interval of the posterior density distribution of the divergence time at 0.9 to 6.6 Ky, our ABC-RF analyses clearly point to a divergence of the two desert locust subspecies occurring during the present Holocene geological epoch (0 to 11.7 Ky ago; Figure 3A). The posterior median estimate (2.6 Ky) and interquartile range (1.8 to 3.7 Ky) postdated the middle-late Holocene boundary (4.2 Ky). This past time boundary corresponds to the last transition from humid to arid conditions in the African continent (Figure 3B). This increasing aridity was shown to be a progressive change, with a concomitant maximum in northern and southern Africa at around 4 to 4.2 Ky ago, where aridity caused a contraction of the forest at its northern and southern peripheries without affecting its core region (Guo *et al.* 2000; Maley *et al.* 2018). Interestingly, the earliest archeological records of the desert locust found in Tin Hanakaten (Algeria) and Saqqara (Egypt) archaeological sites date back to this period (see Figure 3B and references within). Pollen records also showed that during this period the plant community was dominated by the desert and semi-desert taxa found today, including some species of prime importance for the current ecology of the desert locust (Kröpelin *et al.* 2008, Shi *et al.* 1998, Duranton *et al.* 2012). Then, the past 4 Ky are thought to have been under environmental stability and as dry as at present. One can therefore reasonably assume that, at the inferred divergence time between the two locust subspecies, the connectivity between the two African hemispheres was still limited by the moist equator, in particular at the west, and by the savannahs and woodlands of the eastern coast (Figure 1C). Consequently, contrary to most phylogeographic studies on other African arid-adapted species (Atickem *et al.* 2018, Moodley *et al.* 2018), it is unlikely that the rather ancient Quaternary climatic history explained the Southern range extension of the desert locust; see Supplementary Material S4 for additional points of discussions on the influence of climatic cycles on *S. gregaria*.

Recent geological and palynological research has shown that a brief fragmentation of the African primary forest occurred during the Holocene interglacial from 2.5 Ky to 2.0 Ky ago (reviewed in Maley *et al.* 2018). This forest fragmentation period is characterized by relatively warm temperatures and a lengthening of the dry season rather than an arid climate. Although this

period does not correspond to a phase of general expansion of savannas and grasslands, it led to the opening of the Sangha River Interval (SRI). The SRI corresponds to a 400 km wide (14–18° E) open strip composed of savannas and grasslands dividing the rainforest in a north-south direction. The SRI corridor is thought to have facilitated the southern migration of Bantu-speaking pastoralists, along with cultivation of the semi-arid sub-Saharan cereal, pearl millet, *Pennisetum glaucum* (Schwartz 1992; Bostoen *et al*. 2015). The Bantu expansion took place between approximately 5 and 1.5 Ky ago and reached the southern range of the desert locust, including northern Namibia for the Western Bantu branch and southern Botswana and eastern South Africa for the Eastern Bantu branch (Vansina 1995). We cannot exclude that the recent subspecific distribution of the desert locust has been mediated by this recent climatic disturbance, which included a north-south corridor of open vegetation habitats and the diffusion of agricultural landscapes through the Bantu expansion. The progressive reappearance of forest vegetation 2 Ky ago would have then led to the present-day isolation and subsequent genetic differentiation of the new southern populations from northern parental populations.

Our ABC-RF results indicate that a demographic bottleneck (*i.e.*, a strong transitory reduction of effective population size) occurred in the nascent southern subspecies of the desert locust. The high posterior probability value (96.5%) shows that this evolutionary event could be inferred with strong confidence. This result can be explained by the abovementioned colonization hypothesis if the proportion of suitable habitats for the desert locust in the SRI corridor was low, strongly limiting the carrying capacity during the time for range expansion. Alternatively, the bottleneck event in *S. g. flaviventris* can be explained by a southern colonization of Africa through a long-distance migration event. Long-distance migrations are possible in the gregarious phase of the desert locust, with swarms of winged adults that regularly travel up to 100 km in a day (Roffey and Magor 2003). However, since effective displacements are mostly downwind in this species, the likelihood of a southwestern transport of locusts depends on the dynamics of winds and pressure over Africa (Nicholson 1996, Waloff and Pedgley 1986). Because in southern Africa, winds blow

mostly from the north-east toward the extant south-western distribution of the desert locust (at least in southern winter, *i.e.*, August; Figure 1A), only exceptional conditions of a major plague event may have brought a single or a few swarm(s) in East Africa (see Figure 1B) and sourced the colonization of south-western Africa. In agreement with this, rare southward movements of desert locust have been documented along the eastern coast of Africa, for instance in Mozambique in January 1945 during the peak of the major plague of 1941-1947 (Waloff 1966)

Gain in statistical inferences when incorporating independent information into the mutational prior setting

The mutational rate and spectrum at molecular markers are critical parameters for model-based population genetics inferences (*e.g.*, Estoup *et al.* 2002). We found that the specification into prior distributions of previous estimations of microsatellite mutation rates and allele size constraints substantially improved the accuracy of the divergence time estimation. The using of a naive mutational prior setting, where values for mutational parameters were drawn from uniform distributions allowing for larger uncertainties with respect to mutation rates and allele size constraints, resulted in a larger credibility interval of the divergence time estimated from the observed dataset. The latter credibility interval did not include, however, another transition to a dry climatic period, such as the Younger Dryas (YD, 12.9 to 11.7 Ky) or the Last Glacial Maximum (LGM, 21.1 to 17.2 Ky), two periods with a more continuous potential ecological range for the desert locust. Simulation studies also showed that a naive mutational prior setting resulted in a downward bias in median estimate, which could have altered the historical interpretation of our results. For example, the down-biased estimate of the divergence time obtained when using a naive mutational prior setting (median of 1.7 Ky) agrees less with the timing of the aridity associated with the SRI opening (2.5 Ky to 2 Ky). For scenario choice, the inferential gain in incorporating independent information in mutational prior setting was weaker, with power and error rates decreasing by only a few percent.

It is legitimate to ask the question of whether the observed increases in confidence levels in scenario choice and parameter estimation are worth the substantial efforts required to estimate microsatellite mutation rates from direct observation of germline mutations in non-model species. As food for thought, the use of uniform prior rather than a log-uniform prior for time period parameters led to an absolute bias and increase in credibility interval in divergence time estimate similar to that observed when using a naive rather than an informed mutational prior setting (Supplementary Material S5). Using a log-uniform distribution remains a sensible choice for parameters with ranges of values covering several if not many log-intervals, as doing so allows assigning equal probabilities to each of the log-intervals. The observed effect of prior shape distributions highlights, once again, the well-known potential impacts of the prior settings assumed in Bayesian analyses, and calls for processing various error and accuracy analyses using different prior settings as done in the present study.

Implication for the evolution of phase polyphenism

Interestingly, the southern subspecies *S. g. flaviventris* lacks, at least partly, the capacity to mount some of the phase polyphenism responses associated with swarming observed in the northern subspecies *S. g. gregaria* (reviewed in Chapuis *et al.* 2017). Since the *S. g. flaviventris* lineage arose about 7,700 generations ago, it seems unlikely that a hard selective sweep from *de novo* mutation(s) is responsible for the loss of phase polyphenism, although the large effective population sizes may prevent their loss by genetic drift and increase the efficacy of selection (Kimura 1962). Selection on standing genetic variation may therefore better explain such a rapid evolution, since beneficial alleles are immediately available, less likely to be lost by drift than new mutations, and may have been pre-tested by selection in past environments (Barrett and Schluter 2008). Such a scenario would require that variants associated with the reduction of phase polyphenism in *S. g. flaviventris* were already present in past *S. g. gregaria* environments at relatively high frequencies, which may have occurred through *prior* adaptation. First, temporal heterogeneity in selection between low-

density (solitarious) and high-density (gregarious) environments in the northern range may have contributed to retain a high level of genetic variance on this trait (Siepielski *et al.* 2009; Pélissié *et al.* 2016). Second, the southern colonization was preceded by a prolonged and severe contraction of northern deserts, providing ecological conditions favorable for the evolution of a solitarious phase in the native environment that may have facilitated adaptation in the novel southern range of the species.

Hundreds to thousands of genes have been previously identified as differentially expressed between isolated (solitarious) and crowded (gregarious) phases of the desert locust but the challenge of targeting those relevant to the polyphenetic switch is daunting (Badisco *et al.* 2011, Bakkali and Martín-Blázquez 2018). In this context, a promising investigation axis to identify key genes (or transcripts) is to use population genomics (or transcriptomics) approaches comparing highly polyphenic *S. g. gregaria* populations and less polyphenic *S. g. flaviventris*. In particular, genomics studies based on genome scans (reviewed in Vitti et al. 2013) use population samples to measure genetic diversity and differentiation at many loci, with the goal of detecting loci under divergent selection. Since the variance in differentiation estimates across loci is expected to be lower in poorly differentiated populations (Hoban et al. 2016), the recent divergence between desert locust lineages should ease the detection of signatures of natural selection. Genome scans can lead to misleading signals of selection if the effects of geographical, temporal and demographic factors are not properly accounted for (Li *et al.* 2012; Vitti *et al.* 2013). For example, bottlenecks may create spurious signatures that mimic those left by positive selection. Future genome scan studies will therefore greatly benefit from the historical and demographic parameters inferred in the present study, as they could be explicitly included in the analytical process (*e.g.* Vitalis *et al.* 2001; Nielsen *et al.* 2009).

Materials & Methods

Formalization of evolutionary scenarios

To help formalize the evolutionary scenarios to be compared, we relied on maps of vegetation cover in Africa from the Quaternary Environment Network Atlas (Adams and Faure 1997), considering more specifically the periods representative of arid maximums (LGM and YD; Fig.1E-F, humid maximums (HCO; Fig.1D), and present-day arid conditions (Fig.1C). Desert and xeric shrubland cover fits well with the present-day species range during remission periods. Tropical and Mediterranean grasslands were added separately to the desert locust predicted range since the species inhabits such environments during outbreak periods only. The congruence between present maps of species distribution (Fig.1A) and of open vegetation habitats (Fig.1C) suggests that vegetation maps for more ancient periods could be considered as good approximations of the potential range of the desert locust in the past. Maps of vegetation cover during ice ages (Figs. 1E and 1F) show an expansion of open vegetation habitats (*i.e.*, grasslands in the tropics and deserts in both the North and South of Africa) sufficient to make the potential range of the species continuous from the Horn of Africa in North-West to the Cape of Good Hope in the South.

Based on the above climatic and paleo-vegetation map reconstructions, we considered a set of alternative biogeographic hypotheses formulated into different types of evolutionary scenarios. First, we considered scenarios involving a more or less continuous colonization of southern Africa by the ancestral population from a northern origin. In this type of scenario, effective population sizes were allowed to change after the divergence event, without requiring any bottleneck event (*i.e.*, without any abrupt and strong reduction of population size) right after divergence. Second, we considered the situation where the colonization of Southern Africa occurred through a single (or a few) long-distance migration event(s) of a small fraction of the ancestral population. This situation was formalized through scenarios that differed from the formers by the occurrence of a bottleneck event in the newly founded population. The bottleneck event occurred into *S. g. flaviventris* right after divergence and was modelled through a limited number of founders during a short period.

Because the last Quaternary cycle includes several arid climatic periods, including the intense punctuation of the Younger Dryas (YD) and the last glacial maximum (LGM), we also considered scenarios that incorporated the possibility of secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*. Since previous tests based on simulated data showed a poor power to discriminate between a single versus several admixture events (results not shown), we considered only models including a single admixture event.

Finally, at interglacial humid maximums, the map of vegetation cover showed a severe contraction of deserts, which were nearly completely vegetated with annual grasses and shrubs and supported numerous perennial lakes (Fig.1D; deMenocal *et al.* 2000). We thus envisaged the possibility that climatic-induced contractions of population sizes have pre-dated the separation of the two subspecies. Hence, whereas so far scenarios involved a constant effective population size in the ancestral population, we formalized alternative scenarios in which we assumed that a long population size contraction event occurred into the ancestral population at a time t_{ca}, with an effective population size Nc_a for a duration dc_a.

Combining the presence or absence of the three above-mentioned key evolutionary events (a bottleneck in *S. g. flaviventris*, an asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*, and a population size contraction in the ancestral population) allowed defining a total of eight scenarios, that we compared using ABC-RF. The eight scenarios with their historical and demographic parameters are graphically depicted in Figure 2. All scenarios assumed a northern origin for the common ancestor of the two subspecies and a subsequent southern colonization of Africa. This assumption is supported by recent mitochondrial DNA data showing that *S. g. gregaria* have higher levels of genetic diversity and diagnostic bases shared with outgroup and congeneric species, whereas *S. g. flaviventris* clade was placed at the apical tip within the species tree (Chapuis *et al.* 2016). All scenarios considered three populations of current effective population sizes N_f for *S. g. flaviventris*, N_g for *S. g. gregaria*, and N_a for the ancestral population, with *S. g. flaviventris* and *S. g. gregaria* diverging t_{div} generations ago from the ancestral population. The bottleneck event

which potentially occurred into *S. g. flaviventris* was modelled through a limited number of founders N_{bf} during a short period d_{bf}. The potential asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris* occurred at a time t_{sc}, with an effective population size N_{ca} and a proportion r_g of genes of *S. g. gregaria* origin. The potential population size contraction event occurred into the ancestral population at a time t_{ca}, with an effective population size N_{ca} during a duration d_{ca}.

Prior setting for historical and demographical parameters

Prior values for time periods between sampling and secondary contact, divergence and/or ancestral population size contraction events (t_{ca}, t_{div} and t_{sc}, respectively) were drawn from log-uniform distributions bounded between 100 and 500,000 generations, with $t_{ca} > t_{div} > t_{sc}$. Assuming an average of three generations per year (Roffey and Magor 2003), this prior setting corresponds to a time period that goes back to the second-to-latest glacial maximum (150 Ky ago) (de Vivo and Carmignotto 2004, deMenocal *et al.* 2000). Preliminary analyses showed that assuming a uniform prior shape for all time periods (instead of log-uniform distributions) do not change scenario choice results, with posterior probabilities only moderately affected, and this despite a substantial increase of out-of-bag prior error rates (*e.g.*, + 50% when considering the eight scenarios separately; Table S5.1, Supplementary Material S5). Analyses of simulated pseudo-observed datasets (pods) showed that assuming a uniform prior rather than a log-uniform prior for time period parameters would have also biased positively the median estimate of the divergence time and substantially increased its 90% credibility interval (Figure S5.1 and Table S5.2, Supplementary Material S5).

We used uniform prior distributions bounded between $1x10^4$ and $1x10^6$ diploid individuals for the different stable effective population sizes N_f, N_g *and* N_a (Chapuis *et al.* 2014). The admixture rate (r_g; *i.e.*, the proportion of *S. g. gregaria* genes entering into the *S. g. flaviventris* population), was drawn from a uniform prior distribution bounded between 0.05 and 0.5. We used uniform prior distributions bounded between 2 and 100 for both the numbers of founders (in diploid individuals)

and durations of bottleneck events (in number of generations). For the contraction event, we used uniform prior distributions bounded between 100 and 10,000 for both the population size N_{ca} (in diploid individuals) and duration d_{ca} (in number of generations). Assuming an average of three generations per year (Roffey and Magor 2003), such prior choice allowed a reduction in population size for a short to a relatively long period, similar for instance to the whole duration of the HCO (from 9 to 5.5 Ky ago) which was characterized by a severe contraction of deserts.

Microsatellite dataset, mutation rate and mutational model

We carried out our statistical inference on the microsatellite datasets previously published in Chapuis *et al.* (2016). The 23 microsatellite loci genotyped in such datasets were derived from either genomic DNA (14 loci) or messenger RNA (9 loci) resources, and were hereafter referred to as untranscribed and transcribed microsatellite markers (following Blondin *et al.* 2013). These microsatellites were shown to be genetically independent, free of null alleles and at selective neutrality (Chapuis *et al.* 2016). Previous levels of F_{ST} (Weir 1996) and Bayesian clustering analyses (Pritchard *et al.* 2000) among populations showed a weak genetic structuring within each subspecies (Chapuis *et al.* 2014, 2017). For each subspecies, we selected and pooled three population samples in order to ensure both a large sample size (i.e., 80 and 90 individuals for *S. g. gregaria* and *S. g. flaviventris*, respectively), while ensuring a non-significant genetic structure within each subspecies pooled sample, as indicated by non-significant (i.e. p-value > 0.05; Genepop 4.0; Rousset 2008) (i) Fisher's exact tests of genotypic differentiation among the three initial population samples within subspecies and (ii) exact tests of Hardy-Weinberg equilibrium for each subspecies pooled sample. More precisely, the *S. g. gregaria* sample consisted in pooling the population samples 8, 15 and 22 of Chapuis *et al.* (2014) and the *S. g. flaviventris* sample included the population samples 1, 2 and 6 of Chapuis *et al.* (2017).

Mutations occurring in the repeat region of each microsatellite locus were assumed to follow a symmetric generalized stepwise mutation model (GSM; Zhivotovsky *et al.* 1997; Estoup *et al.*

2002). Prior values for any mutation model settings were drawn independently for untranscribed and transcribed microsatellites in specific distributions. The informed mutational prior setting was defined as follows. Because allele size constraints exist at microsatellite markers, we informed for each microsatellite locus their lower and upper allele size bounds using values estimated in Chapuis et al. (2015), following the approach of Pollock *et al.* (1998) and microsatellite data from several species closely related to *S. gregaria* (Blondin *et al.* (2013). Prior values for the mean mutation rates ($\overline{\mu_R}$) were set to the empirical estimates inferred from observation of germline mutations in Chapuis *et al.* (2015), *i.e.*, 2.8×10^{-4} and 9.1×10^{-5} for untranscribed and transcribed microsatellites, respectively. The parameters for individual microsatellites were then drawn from a Gamma distribution with mean = $\overline{\mu_R}$ and shape = 0.7 (Estoup *et al.* 2001) for both types of microsatellites. We ensured that the chosen value of shape parameter generated the same inter-loci variance as estimated in Sun *et al.* (2012) from direct observations of thousands of human microsatellites. Prior values for the mean parameters of the geometric distributions of the length in number of repeats of mutation events ($\bar{P}$) were set to the proportions of multistep germline mutations observed in Chapuis *et al.* (2015), *i.e.*, 0.14 and 0.67 for untranscribed and transcribed microsatellites, respectively. The *P* parameters for individual loci were then standardly drawn from a Gamma distribution (mean = $\bar{P}$ and shape = 2). We also considered mutations that insert or delete a single nucleotide to the microsatellite sequence. To model this mutational feature, we used the DIYABC default setting values (*i.e.*, a uniform distribution bounded between [10^{-8}, 10^{-5}] for the mean parameter $\overline{\mu_{SNI}}$ and a Gamma distribution (mean = $\overline{\mu_{SNI}}$ and shape = 2) for individual loci parameters; Cornuet *et al.* 2010; see also DIYABC user manual p. 13, http://www1.montpellier.inra.fr/CBGP/diyabc/).

We evaluated how the incorporation of independent information on prior distributions for mutational parameters affected both the posterior probabilities of scenarios and the posterior parameter estimation under our inferential framework. To this aim, we re-processed our inferences using a naive mutational prior setting, often used in many ABC microsatellite studies (*e.g.*, Estoup

et al. 2002). In this case, prior values for mean mutation parameters were drawn from uniform distributions instead of being set to a fixed value as in the informed mutational prior setting. For each set of untranscribed or transcribed microsatellites, all loci were free of allele size constraints (cf. allele size bounds were fixed to very different values such as 2 and 500 for the lower and upper bounds, respectively), prior values for $\overline{\mu_R}$ were drawn from a uniform distribution bounded between 10^{-5} and 10^{-3}, $\bar{P}$ values were drawn in a uniform distribution bounded between 0.1 and 0.3. Finally, the mean rate of single nucleotide indel mutations and all parameters for individual loci were set to the DIYABC default values (Chapuis *et al.* 2014; 2015).

Analyses using ABC Random Forest

We used the software DIYABC v.2.1.0 (Cornuet *et al.* 2014) to simulate datasets constituting the so-called reference tables (i.e. records of a given number of datasets simulated using the scenario ID and the evolutionary parameter values sampled from prior distributions and summarized with a pool of statistics). Random-forest computations were then performed using a new version of the R library ABCRF (version 1.8) available on the CRAN. This version includes all ABC-RF algorithms detailed in Pudlo *et al.* (2016), Raynal *et al.* (2019) and Estoup *et al.* (2018a) for scenario choice and parameter estimation, as well as several statistical novelties allowing to compute error rates in scenario choice and accuracy measures for parameter estimation (see details below).

For scenario choice, the outcome of the first step of the ABC-RF statistical treatment applied to a given target dataset is a classification vote for each scenario which represents the number of times a scenario is selected in a forest of *n* trees. The scenario with the highest classification vote corresponds to the scenario best suited to the target dataset among the set of compared scenarios. This step also provides an error rate relevant to the entire prior sampling space, the global prior error. See the section *Global prior errors* in Supplementary Material S1 for details. The second RF analytical step provides a reliable estimation of the posterior probability of the best supported scenario. One minus such posterior probability yields the local posterior error associated to the

observed dataset (see the section *Local posterior errors* in Supplementary Material S1). In practice, ABC-RF analyses were processed by drawing parameter values into the prior distributions described in the two previous sections and by summarizing microsatellite data using a set of 32 statistics (see Table S6.1, Supplementary Material S6, for details about such summary statistics as well as their values obtained from the observed dataset) and the one LDA axis or seven LDA axes (i.e. number of scenarios minus 1; Pudlo *et al.* 2016) computed when considering pairwise groups of scenarios or individual scenarios, respectively. We processed ABC-RF treatments on reference tables including 100,000 simulated datasets (*i.e.*, 12,500 per scenario). Following Pudlo et al. (2016), we checked that 100,000 datasets was sufficient by evaluating the stability of prior error rates and posterior probabilities estimations of the best scenario on 50,000, 80,000 and 90,000 and 100,000 simulated datasets (Table S6.2, Supplementary Material S6). The number of trees in the constructed random forests was fixed to n = 3,000, as this number turned out to be large enough to ensure a stable estimation of the prior error rate (Figure S6.1, Supplementary Material S6). We predicted the best scenario and estimated its posterior probability and prior error rate over ten replicate analyses based on ten different reference tables.

In order to decipher the main evolutionary events that occurred during the evolutionary history of the two desert locust subspecies, we first conducted ABC-RF treatments on three pairwise groups of scenario (with four scenarios per group): groups of scenarios with *vs.* without a bottleneck in *S. g. flaviventris*, groups with *vs.* without a population size contraction in the ancestral population, and groups with *vs.* without a secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*. We then conducted ABC-RF treatments on the eight scenarios considered separately.

For parameter estimation, we constructed ten independent replicate RF treatments based on ten different reference tables for each parameter of interest (Raynal *et al.* 2019): the time since divergence, the ratio of the time of the contraction event into the ancestral population on the time since divergence, the intensity of the bottleneck event in the sampled *S. g. flaviventris* population

(defined as the ratio of the bottleneck event of duration db_f on the effective population size Nb_f) and the ratio of the stable effective population size of the two sampled populations. For each RF treatment, we simulated a total of 100,000 datasets for the selected scenario (drawing parameter values into the prior distributions described in the two previous sections and using the same 32 summary statistics). Following Raynal *et al.* (2019), we checked that 100,000 datasets was sufficient by evaluating the stability of the measure of accuracy on divergence time estimation using 50,000, 80,000 and 90,000 simulated datasets (Table S6.3, Supplementary Material S6). The number of trees in the constructed random forests was fixed to n = 2,000, as such number turned out to be large enough to ensure a stable estimation of the measure of divergence time estimation accuracy (Figure S6.2, Supplementary Material S6). For each RF treatment, we estimated the median value and the 5% and 95% quantiles of the posterior distributions. It is worth noting that we considered median values as the later provided more accurate estimations (according to out-of-bag predictions) than when considering mean values (results not shown). Accuracy of parameter estimation was measured using out-of-bag predictions and the normalized mean absolute error (NMAE). NMAE corresponds to the mean of the absolute difference between the point estimate (here the median) and the (true) simulated value divided by the simulated value (formula detailed in Supplementary Material S1).

Finally, because microsatellite markers tend to underestimate divergence time for large time scales due to allele size constraints, we evaluated how the accuracy of ABC-RF estimation of the time of divergence between the two subspecies was sensitive to the time scale. To this aim, we used DIYABC to produce pseudo-observed datasets assuming fixed divergence time values chosen to cover the prior interval (100 ; 250 ; 500; 1,000 ; 2,500; 5,000 ; 10,000 ; 25,000 ; 50,000; 100,000; 250,000 generations) and using the best scenario with either the informed or the naive mutational prior setting. We simulated 5,000 of such test datasets for each of the eleven divergence time values. Each of these test dataset was treated using ABC-RF in the same way as the above target

observed dataset. In addition, we computed for each test dataset the relative amplitude of parameter estimation, as the 90% credibility interval divided by the (true) simulated value.

Acknowledgements

This work was supported by research funds from the French Agricultural Research Centre for International Development (CIRAD), the project ANR-16-CE02-0015-01 (SWING), the INRA scientific department SPE (AAP-SPE 2016), and the Labex NUMEV (NUMEV, ANR10-LABX-20). The data used in this work were partly produced through the technical facilities of the Centre Méditerranéen Environnement Biodiversité, Montpellier. We thank Christine N. Meynard for careful English language editing and insightful discussions on past climate models for Africa, Pierre-Emmanuel Gay for assistance with maps, Antoine Foucart, Gauthier Dobigny and Jean-Yves Rasplus for fruitful discussions, and Renaud Vitalis for constructive comments on an earlier version of the manuscript.

References

Adams JM, Faure H (1997) (eds), QEN members. Review and Atlas of Palaeovegetation: Preliminary land ecosystem maps of the world since the Last Glacial Maximum. Oak Ridge National Laboratory, TN, USA.

Atickem A, Stenseth NC, Drouilly M, Bock S, Roos C, Zinner D (2018) Deep divergence among mitochondrial lineages in African jackals. *Zool Scripta*, **47**, 1-8.

Aumassip G (2002) L'algérie des premiers hommes. Ibis Press, 230 p.

Badisco L, Ott SR, Rogers SM, Matheson T, Knapen D, Vergauwen L, Verlinden H, Marchal E, Sheehy MRJ, Burrows M , *et al.* (2011) Microarray-based transcriptomic analysis of differences between long-term gregarious and solitarious desert locusts. *PloS One*, **6**, e28110.

Bakkali M, Martín-Blázquez R (2018) RNA-Seq reveals large quantitative differences between the transcriptomes of outbreak and non-outbreak locusts. *Sci Rep*, **8**, 9207.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol*, **19**(13), 2609–2625.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends Ecol Evol*, **23**, 38–44.

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

Blondin L, Badisco L, Pagès C, Foucart A, Risterucci AM, Bazelet CS, Vanden Broeck J, Song H, Ould Ely S, Chapuis M-P (2013) Characterization and comparison of microsatellite markers derived from genomic and expressed libraries for the desert locust. *J Appl Entomol*, **137**, 673–683.

Bond G *et al.* (1997) A pervasive millenial-scale cycle in North Atlantic Holocene and Glacial climates. Science, **278**, 1257-1265.

Bostoen K, Clist B, Doumenge C, Grollemund R, Hombert JM, Muluwa JK, Maley J (2015) Middle to Late Holocene paleoclimatic change and early Bantu expansion in the rain forests of Western Central Africa. *Curr Anthropol*, **56**, 354–384.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**(1), 5–32.

Chapuis M-P, Bazelet CS, Blondin L, Foucart A, Vitalis R, Samways MJ. (2016) Subspecific taxonomy of the desert locust, *Schistocerca gregaria* (Orthoptera: Acrididae), based on molecular and morphological characters. *Syst Entomol*, **41**, 516-530.

Chapuis M-P, Foucart A, Plantamp P, Blondin L, Leménager N, Benoit L, Gay P-E, Bazelet CS (2017) Genetic and morphological variation in non-polyphenic southern African populations of the desert locust. *Afr Entomol*, **25**, 13-23.

Chapuis M.-P, Plantamp P, Blondin B, Pagès C, Vassal J.-M., Lecoq M (2014) Demographic processes shaping genetic variation of the solitarious phase of the desert locust. *Mol Ecol*, **23**, 1749–1763.

Chapuis M-P, Plantamp C, Streiff R, Blondin L, Piou C (2015) Microsatellite evolutionary rate and pattern in *Schistocerca gregaria* inferred from direct observation of germline mutations. *Mol Ecol*, **24**, 6107-6119.

Chase BM, Meadows ME (2007) Late Quaternary dynamics of southern Africa's winter rainfall zone. *Earth Sci Rev*, **84**, 103-138.

Chase BM *et al.* (2009) A record of rapid Holocene climate change preserved in hyrax middens from SW Africa. *Geology* **37**, 703-706.

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

Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup A (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.

Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol*, **25**(7), 410–418.

deMenocal PB (1995) Plio-Pleistocene African climate, *Science*, 270, 53-59.

deMenocal PB, Ortiz J, Guilderson T, Adkins J, Sarnthein M, Baker L, Yarusinsky M (2000) Abrupt onset and termination of the African Humid Period: Rapid climate response to gradual insolation forcing. *Quat Sci Rev*, **19**, 347-361.

deMenocal PB (2004) African climate change and faunal evolution during the Pliocene-Pleistocene. *Earth and Planetary Science Letters*, **220**, 3–24.

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, *et al.* (1996) A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature*, 380, 152–154

Dupont LM (2011) Orbital scale vegetation change in Africa *Quat Sci Rev*, **30**, 3589-3602.

Duranton JF, Foucart A, Gay P-E (2012) Florule des biotopes du criquet pèlerin en Afrique de l'Ouest et du Nord-Ouest à l'usage des prospecteurs de la lutte antiacridienne. Rome : FAO, 487 p.

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

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol Ecol*, **11**, 1591-1604.

Estoup A, Raynal L, Verdu P, Marin J-M (2018a) Model choice using Approximate Bayesian Computation and Random Forests: analyzes based on model grouping to make inferences about the genetic history of Pygmy human populations. *J Soc Fr Statistique*, **159**, 167-190.

Estoup A, Verdu P, Marin J-M, Robert C, Dehne-Garcia A, Cornuet J-M, Pudlo P (2018b) Application of approximate Bayesian computation to infer the genetic history of Pygmy

hunter-gatherers populations from Western Central Africa. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

Estoup A, Wilson IJ, Sullivan C, Cornuet J-M, Moritz C (2001) Inferring population history from microsatellite et enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671-1687.

Feldman MW, Bergman A, Pollock DD, Goldstein DB (1997) Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics*, **145**, 207–216.

Fraimout A, Debat V, Fellous S, Hufbauer RA, Foucaud J, Pudlo P, Marin J-M, Price DK, Cattel J *et al.* (2017). Deciphering the routes of invasion of *Drosophila suzukii* by means of ABC random forest. *Mol Biol Evol*, **34**(4), 980–996.

Gasse F (2000) Hydrological changes in the African tropics since the Last Glacial Maximum. *Quat Sci Rev*, **19**, 189–211.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463-471.

Guo Z, Petit-Maire N, Kroepelin S (2000) Holocene non-orbital climatic events in present-day arid areas of northern Africa and China. *Global Planet Change* v.26 p.97-103.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol*, **25**, 1965-1978.

Ho SY, Saarma U, Barnett R, Haile J, Shapiro B (2008) The Effect of Inappropriate Calibration: Three Case Studies in Molecular Ecology. *PLoS One*, 3, e1615.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am Nat*, **188**, 379–397.

Jürgens N (1997) Floristic biodiversity and history of African arid regions. *Biodiv Conserv*, **6**, 495-514.

Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–719.

Kröpelin S, Verschuren D, Lézine A-M, Eggermont H, Cocquyt C, Francus P, Cazet JP, Fagot M, Rumes B, Russel JM, *et al.* (2008). Climate-driven ecosystem succession in the Sahara: The past 6000 years. *Science*, **320**, 765–8.

Latinne A, Meynard CN, Herbreteau V, Waengsothorn S, Morand S, Michaux JR (2015) Influence of past and future climate changes on the distribution of three Southeast Asian murine rodents. *J Biogeogr*, **42**, 1714-1726.

Lebrun JP (2001) Introduction à la flore d'Afrique. 156 pp. CIRAD and Ibis Press.

Le Gall P, Silvain JF, Nel A, Lachaise D (2010) Les insectes actuels témoins des passés de l'Afrique : essai sur l'origine et la singularité de l'entomofaune de la région afrotropicale. *Ann Soc Entomol Fr*, **46**, 297-343.

Lorenz MW (2009) Migration and trans-Atlantic flight of locusts. Quaternary International, **196**, 4-12.

Lorenzen ED, Heller R, Siegismund HR (2012) Comparative phylogeography of African savannah ungulates. *Mol Ecol*, **21**, 3656–3670.

Malek J (1997) The locusts on the daggers of Ahmose. In: E. Goring, N. Reeves and J. Riffle (eds.), Chief of Seers: Egyptian Studies in Memory of Cyril Aldred, London, 207-219.

Maley J, Doumenge C, Giresse P, Mahé G, Philippon N, Hubau W, Lokonda MO, Tshibamba JM, Chepstow-Lusty A (2018) Late Holocene forest contraction and fragmentation in central Africa. *Quat Res*, **89**, 43–59.

Marin JM, Pudlo P, Estoup A, Robert CP (2018) Likelihood-free Model Choice. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

Meinzingen WF (1993) A guide to migrant pest management in Africa. FAO, Rome, Italy.

Meynard CN, Gay PE, Lecoq M, Foucart A, Piou C, Chapuis M-P (2017) Climate-driven geographic distribution of the desert locust during recession periods: subspecies' niche

differentiation and relative risks under scenarios of climate change. *Glob Change Biol*, **23**(11), 4739-4749.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. (2010) Widespread genomic divergence during sympatric speciation. *Proc Natl Acad Sci USA*, **107**, 9724–9729.

Miller JM, Hallager S, Monfort S, Newby J, Bishop K, Tidmus S, Black P, Houston B, Matthee C, Robinson J, Fleischer RC (2011). Phylogeographic analysis of nuclear and mtDNA supports subspecies designations in the Ostrich (*Struthio camelus*). *Conserv Genet*, **12**, 423–431.

Monod T (1971) Remarques sur les symetries floristiques des zones sèches nord et sud en Afrique. *Mitteilungen der Botanischen Staatssammlung München*, **10**, 375-423.

Moodley Y, Russo I-RM, Robovsky J, Dalton D, Kotzé A, Smith S, Stejskal J, Ryder OA, Hermes R, Walzer C, Bruford MW (2018) Contrasting evolutionary history, anthropogenic declines and genetic contact in the northern and southern white rhinoceros (*Ceratotherium simum*). *Proc R Soc B*, **285**, 1-9.

Nicholson SE (1996) A review of climate dynamics and climate variability in Eastern Africa. In: Johnson, T.C., Odada, E.O. (Eds.), The Limnology, Climatology and Paleoclimatology of the East African Lakes. Gordon and Breach, Amsterdam, pp. 25-56.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, *et al*. (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Res*, **19**, 838–849.

Pélissié B, Piou C, Jourdan H, Pagès C, Blondin L, Chapuis M-P (2016) Extra molting and selection on larval growth in the desert locust. *PLoS One*, **11**(5), e0155736.

Pollock DD, Bergman A, Feldman MW, Goldstein DB (1998) Microsatellite behaviour with range constraints: parameter estimation and improved distances for use in phylogenetic reconstruction. *Theor Popul Biol*, **53**, 256–271.

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

Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. *Bioinformatics*, **32**, 859–866.

Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A (2019) ABC random forests for Bayesian parameter inference. *Bioinformatics*, **35**, 1720–1728.

Robert CP, Cornuet J-M, Marin J-M, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. *P Natl Acad Sci USA*, **108**, 15112–15117.

Roberts N, Taieb M, Baker P, Damnati B, Icole M, Williamson D (1993) Timing of the Younger Dryas event in East Africa from lake-level changes. *Nature*, **366**,146–8

Roffey J, Magor JI (2003) Desert Locust population parameters. Desert Locust Field Research Stations, Technical Series, 30, 29 p. FAO, Rome, Italy.

Rousset F (2008) GenePop'007: a complete re-implementation of the GenePop software for Windows and Linux. *Mol Ecol Resour*, **8**, 103–106.

Schwartz D (1992) Assèchement climatique vers 3000 B.P. et expansion Bantu en Afrique centrale atlantique: quelques réflexions. *B Soc Geol Fr*, **163**, 353–61

Shi N, Dupont LM, Beug H-J, Schneider R (1998) Vegetation and climate changes during the last 21 000 years in S.W. Africa based on a marine pollen record. *Veg Hist Archaeobot*, **7**, 127-140.

Siepielski AM, DiBattista JD, Carlson SM (2009) It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecol Lett*, **12**(11),1261-76.

Stokes S, Thomas DSG, Washington R (1997) Multiple episodes of aridity in southern Africa since the last interglacial period. *Nature*, **388**, 154–158.

Stute M, Talma AS (1997) Glacial temperatures and moisture transport regimes reconstructed from noble gases and ^{18}O, Stampriet aquifer, Namibia. Proceedings of International Symposium on Isotope Techniques in the Study of Past and Current Environmental Changes in the Hydrosphere and the Atmosphere, Vienna, International Atomic Energy Agency.

Sun JX, Helgason A, Masson G *et al.* (2012) A direct characterization of human mutation based on microsatellites. *Nat Genet*, **44**, 1161–1165.

Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) Approximate Bayesian computation. *PLoS Comput Biol*, **9**, e1002803.

Sword GA, Lecoq M, Simpson SJ (2010). Phase polyphenism and preventative locust management. *J Insect Physiol*, **56**, 949–957.

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, **144**, 389–399.

Talma AS, Vogel JC (1992) Late Quaternary paleotemperatures derived from a speleothem from Cango Caves. Cape Province, South Africa. *Quat. Res.*, **37**, 203-13.

Uvarov BP (1977) Grasshoppers and Locusts, vol. 2. Centre for Overseas Pest Research, London, UK.

Van Andel TH, Tzedakis PC (1996). Palaeolithic landscapes of Europe and environs: 150,000-25,000 years ago: an overview. *Quat Sci Rev*, **15**, 481-500.

Vansina J (1995) New Linguistic Evidence and the Bantu expansion. J Afr Hist, **36**, 173–195.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Vivo M, Carmignotto AP (2004) Holocene vegetation change and the mammal faunas of South America and Africa. *J Biogeogr*, **31**, 943–957.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annu Rev Genet*, **47**, 97–120.

Walker MJC, Berkelhammer M, Björck S, Cwynar LC, Fisher DA, Long AJ, Lowe JJ, Newnham RM, Rasmussen SO, Weiss H (2012) Formal subdivision of the Holocene Series/Epoch: a Discussion Paper by a Working Group of INTIMATE (Integration of ice-core, marine and terrestrial records) and the Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy). *J Quat Sci*, **27**, 649–659

Waloff Z (1966) The upsurges and recessions of the desert locust plague: an historical survey. *Anti-Locust Memoir*, 8, 111 p.

Waloff Z, Pedgley DE (1986) Comparative biogeography and biology of the South American locust, *Schistocerca cancellata* (Serville), and the South African desert locust, *S. gregaria flaviventris* (Burmeister) (Orthoptera: Acrididae): a review. *Bull Entomol Res*, 76**,** 1-20.

Weir BS (1996) Genetic Data Analysis II. Sinauer Associates, Sunderland, Massachusetts.

Zhivotovsky LA, Feldman MW, Grishechkin SA (1997) Biased mutations and microsatellite variation. *Mol Biol Evol*, **14**, 926–933.

Supporting Material

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

Figure legends

Figure 1. **Present time distribution range of *Schistocerca gregaria* in Africa under remission periods with winds in August A) and January B), and vegetation habitats suitable for the species during the present period C), the Holocene Climatic Optimum (HCO, 9 to 6 Ky ago) D), the Younger Dryas (YD, 12.9 to 11.7 Ky ago) E) and the Last Glacial Maximum (LGM, 26 to 14.8 Ky ago) F).**

(A-B) Distribution range and winds are adapted from Sword *et al.* (2010) and Nicholson (1996), respectively. In northern Africa, at least since 2.7Ky, the strong northeast trade winds bring desert locust swarms equatorward in the moist intertropical convergence zone (Kröpelin *et al.* 1998). Most transports are westward, with records of windborne locusts in the Atlantic Ocean during plague events (Waloff 1960), including the exceptional trans-Atlantic crossing from West Africa to the Caribbean in 1988 (Lorenz 2009). Nevertheless, at least in northern winter (January), easterly winds flow more parallel to the eastern coast of Africa. (C-F) Vegetation habitats are adapted from Adams and Faure (1997). Open vegetation habitats suitable for the desert locust correspond to deserts (light orange), xeric shrublands (dark orange) and tropical - Mediterranean grasslands (pink). Other unsuitable habitat classes (white) are forests, woodlands and temperate shrublands and savannas.

Figure 2. Evolutionary scenarios compared using ABC-RF.

The subscripts *g*, *f* and *a* refer to the subspecies *S. g. gregaria*, *S. g. flaviventris* and their unsampled common ancestor, respectively. Eight scenarios are considered and identified by a number (from 1 to 8). Such scenarios differ by the presence or absence of three evolutionary events: a bottleneck in *S. g. flaviventris* (*b*) right after divergence between the two subspecies, a population size contraction

in the ancestral population (c_a) and a secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris* (*sc*). For convenience, only the scenario 8 that includes all three evolutionary events is represented graphically. Looking forward in time, time periods are t_{ca}, the time of ancestral population size contraction, t_{div}, the time of divergence between the two subspecies, and t_{sc}, the time of the secondary contact between subspecies (with $t_{ca} > t_{div} > t_{sc}$). r_g is the admixture rate, *i.e.* the proportion of genes from the *S. g. gregaria* lineage entering the *S. g. flaviventris* population at time t_{sc}. N_g, N_f and N_a are the stable effective population sizes of *S. g. gregaria*, *S. g. flaviventris* and the ancestor, respectively. Nc_a is the effective population size during the contraction event of duration dc_a in the ancestor. Nb_f is the effective population size during the bottleneck event of duration db_f.

Figure 3. Divergence time between *S. g. gregaria* and *S. g. flaviventris* inferred under the best supported scenario (scenario 4) A) in relation to bioclimatic changes in Northern and Southern Africa B).

A) Dashed and solid lines represent the formal subdivision of the Holocene and Pleistocene epochs (Walker at al. 2012). Dotted lines with labels on the right side are the median value and 90% confidence interval of the posterior density distributions of the divergence time estimated using an informed or a naive mutation prior setting (assuming an average of three generations per year; Roffey and Magor 2003). Asterisks refer to earliest archeological records of the desert locust. In the Algerian Sahara, remains of locusts were found in a special oven dating back to about 6Ky ago, in the rock shelter of Tin Hanakaten (Aumassip 2002). In Egypt, locusts were depicted on daggers of the pharaoh Ahmose, founder of the Eighteenth Dynasty (about 3.5 Ky ago) (Malek 1997) and, at Saqqara, on tombs of the Sixth Dynasty (about 4.2 to 4.4 Ky ago) that is thought to have felt with the impact of severe droughts (Meinzingen 1993). B) Climatic episodes include major cycles and additional transitions of aridity (sandy brown) and humidity (steel blue). The grey coloration means that there is debate on the climatic status of the period (arid vs. humid). HCO: Holocene Climatic

Optimum; YD: Younger Dryas; LGM: Last Glacial Maximum; LIG: Last Inter Glacial. Delimitations of climatic periods were based on published paleoclimatic inferences from geological sediment sequences (*e.g.*, eolian deposition, oxygen isotope data) and biological records (*e.g.*, pollen or insect fossils assemblages) from marine cores or terrestrial lakes. References are Bond *et al.* (1997), Guo *et al.*(2000), Kröpelin *et al.* (2008), Roberts *et al.* (1993) and van Andel and Tzedakis (1996) for northern Africa, and Talma and Vogel (1992), Stokes *et al.* (1997), and Shi *et al.* (1998) for southern Africa. See also Gasse (2000) for a review.

Figure 4. Point estimates of posterior distributions A) and differences in accuracy B-C) of ABC-RF estimations of the divergence time obtained using an informed or a naive mutational prior setting under the best supported scenario (scenario 4).

Simulated pseudo-observed datasets (5,000 per divergence time) were generated for fixed divergence time values of 100 ; 250 ; 500; 1,000 ; 2,500; 5,000 ; 10,000 ; 25,000 ; 50,000; 100,000; and 250,000 generations (cf. x-axis with a log-scale). A) The estimated median (plain lines) and 90% credibility interval (dashed lines), averaged over the 5,000 datasets, are represented (y-axis) using the informed (black color) or the naive (grey color) mutational prior setting. B) The difference in accuracy, the latter being measured by the normalized mean absolute error (NMAE) calculated from the estimated median values, is represented as NMAE-informed minus NMAE-naive. The negative values of NMAE differences indicate a higher accuracy of estimations based on the informed mutational prior setting. C) The difference in accuracy, here measured by the relative amplitude of estimation (averaged over the 5,000 datasets), is represented as relative amplitude-informed minus relative amplitude-naïve.

Fig. 1

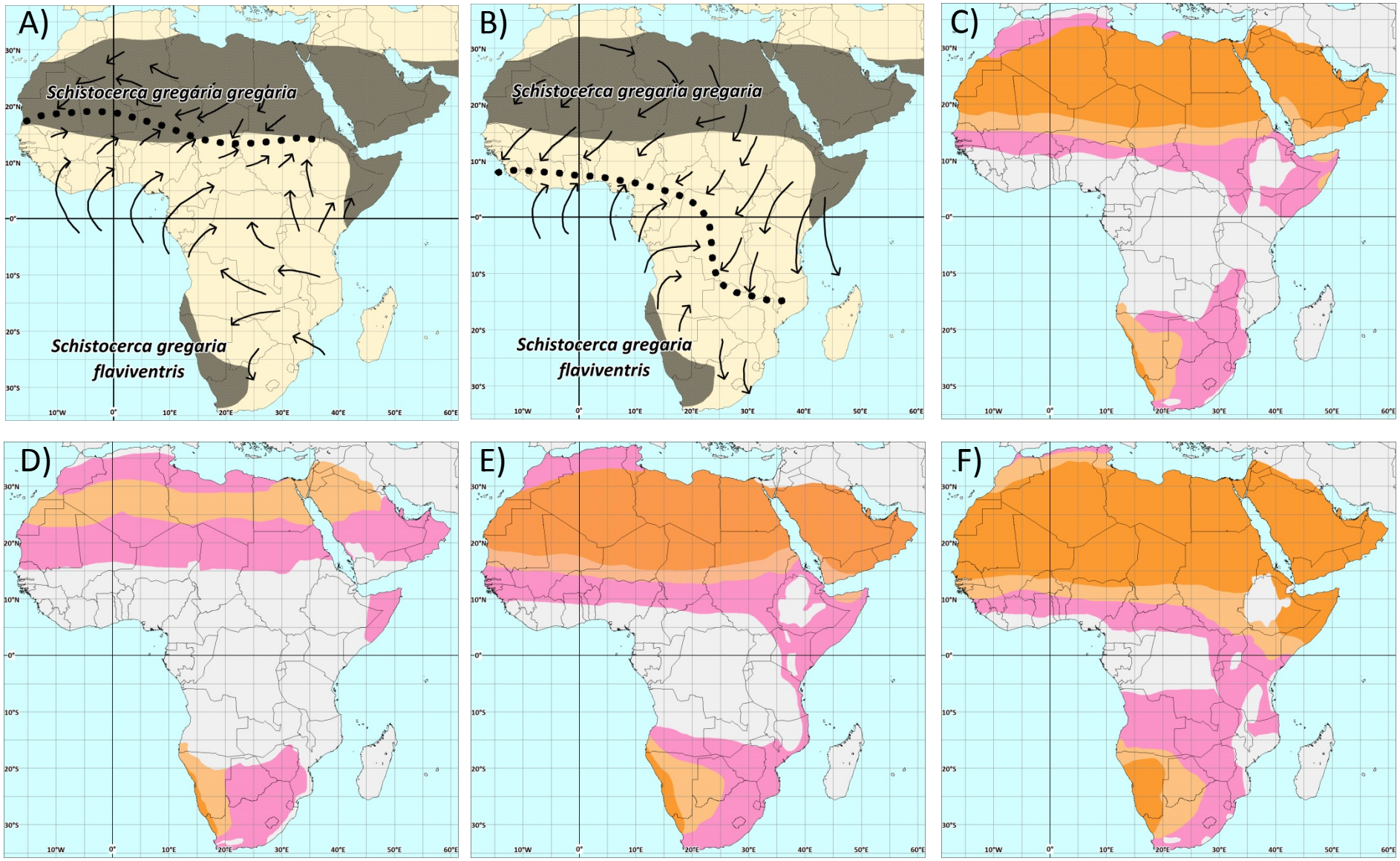
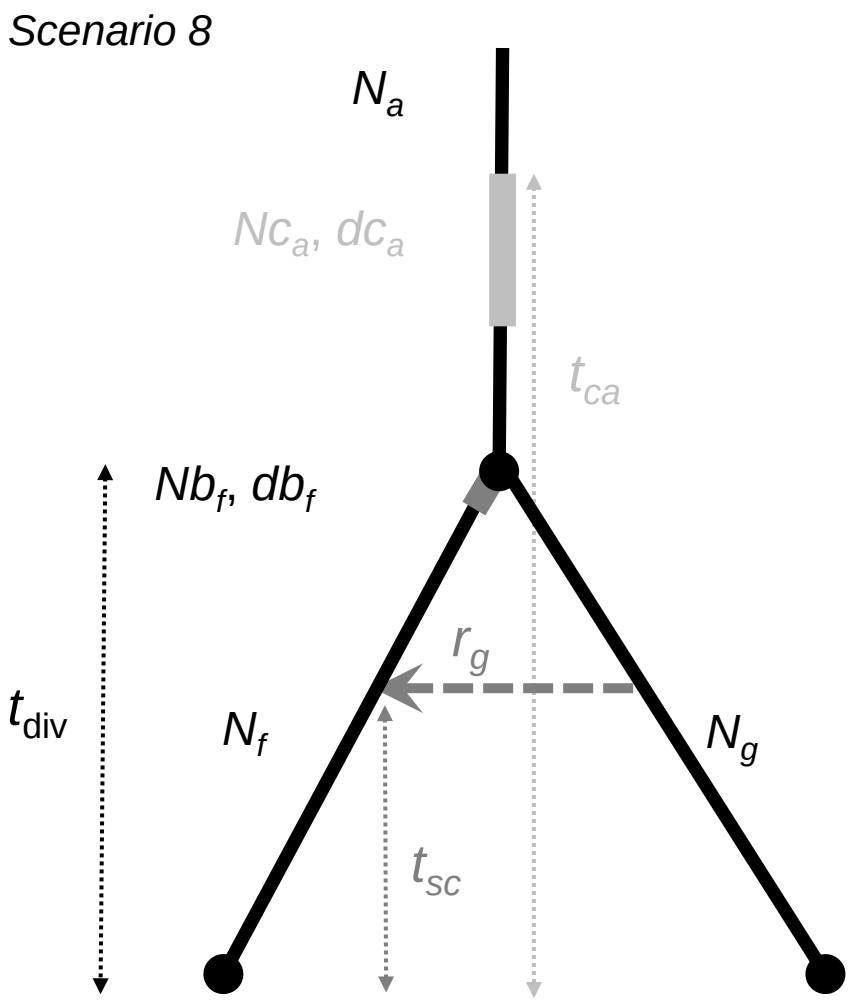

A)
Schistocerca gregaria gregaria
Schistocerca gregaria flaviventris
B)
Schistocerca gregaria gregaria
Schistocerca gregaria flaviventris
C)
D)
E)
F)
30°N
20°N
10°N
0°
10°S
20°S
30°S
10°W
0°
10°E
20°E
30°E
40°E
50°E
60°E

Fig. 2



	Evolutionary event		
Scenario	c_a	*b*	*sc*
1	no	no	no
2	no	yes	no
3	yes	no	no
4	yes	yes	no
5	no	no	yes
6	no	yes	yes
7	yes	no	yes
8	yes	yes	yes

Fig. 3

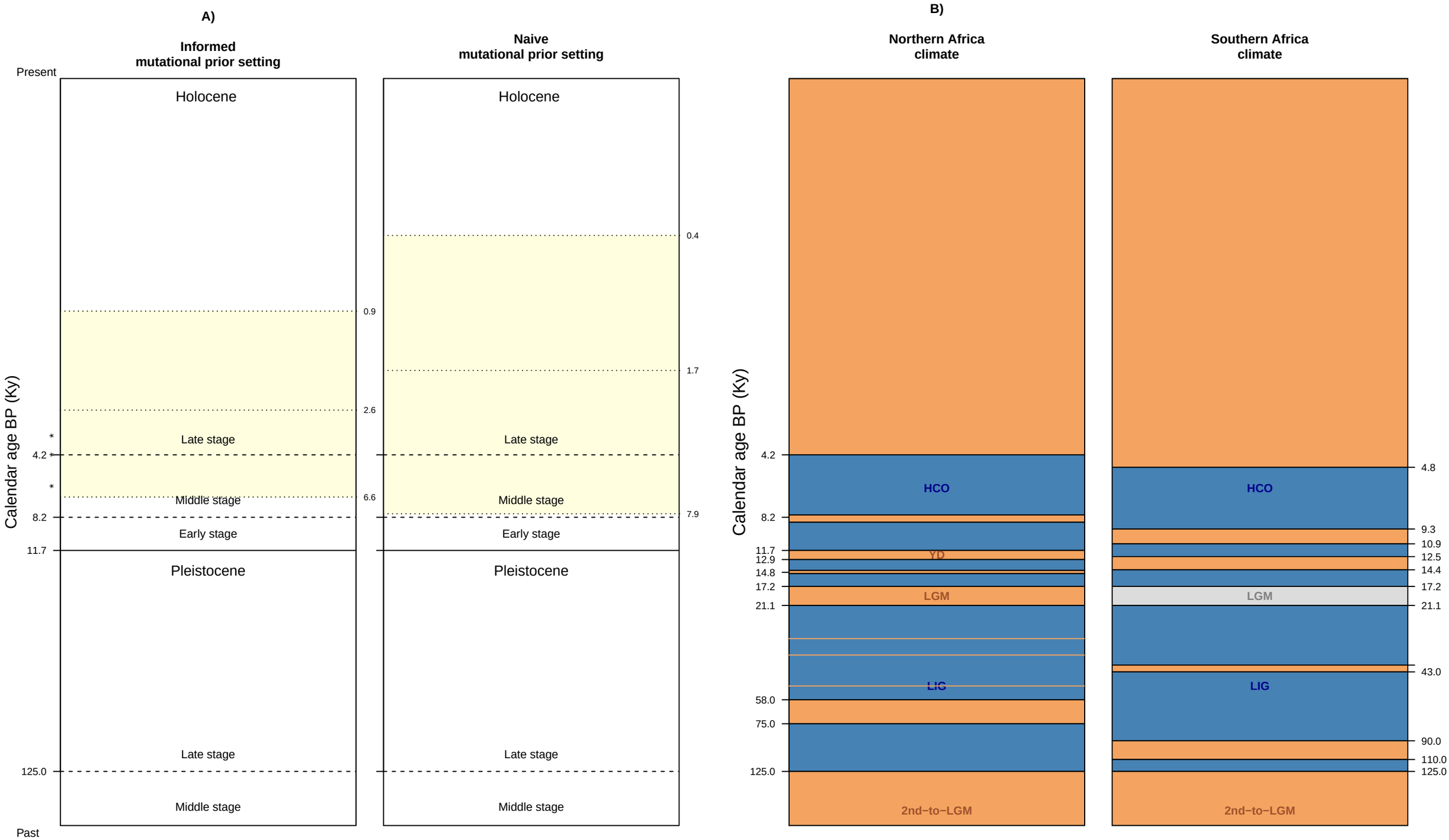
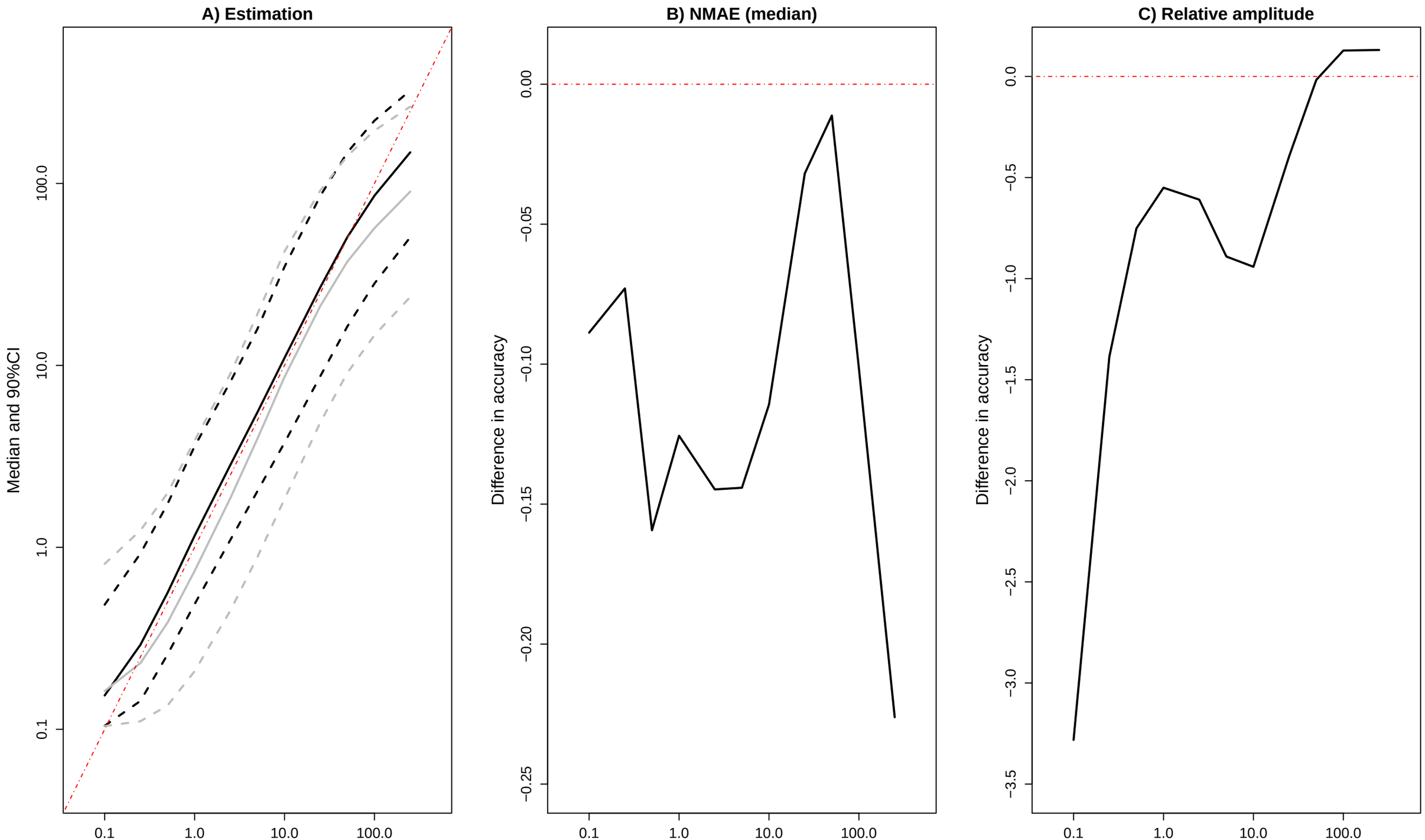

A)
Informed mutational prior setting
Naive mutational prior setting
Present
Past
Calendar age BP (Ky)
Holocene
Pleistocene
Late stage
Middle stage
Early stage
0.9
2.6
6.6
0.4
1.7
7.9
4.2
8.2
11.7
125.0
B)
Northern Africa climate
Southern Africa climate
HCO
YD
LGM
LIG
2nd−to−LGM
4.2
8.2
11.7
12.9
14.8
17.2
21.1
58.0
75.0
125.0
4.8
9.3
10.9
12.5
14.4
43.0
90.0
110.0

Fig. 4


A) Estimation
Median and 90%CI
0.1
1.0
10.0
100.0
B) NMAE (median)
Difference in accuracy
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
C) Relative amplitude
Difference in accuracy
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5

Tables

Table 1. Scenario choice when analyzing groups of scenarios or scenarios separately.

Mutational prior setting	Analyses of groups of scenarios Group 1=no *b* vs. group 2=*b*	Group 1=no c_a vs. group 2= c_a	Group 1=no *sc* vs. group 2= *sc*	Analysis of scenarios separately
Informed				
Prior error rate	10.2%	24.9%	23.5%	47.9%
Posterior probability (selected group or individual scenario)	0.965 (*b*)	0.746 (c_a)	0.742 (no *sc*)	0.584 (scenario 4)
Naive				
Prior error rate	11.1%	26.7%	24.4%	50.2%
Posterior probability (selected group or individual scenario)	0.950 (*b*)	0.704 (c_a)	0.775 (no *sc*)	0.547 (scenario 4)

Scenarios were grouped based on the presence or not of a bottleneck in *S. g. flaviventris* (*b or no b*), a population size contraction in ancestor (c_a *or no* c_a) and a secondary contact with asymmetrical genetic admixture from *S. g. gregaria in*to *S. g. flaviventris* (*sc or no sc*). We reported values for prior error rates and posterior probabilities of the selected group of scenarios or individual scenario, averaged over ten replicate analyses. The local posterior error rate (corresponding to a confidence measure of the selected scenario given the observation.) can be computed as 1 minus the posterior probability (see Supplementary material S1 for details). The number of records for each reference datasets simulated from DIYABC was set to 100,000 and the number of RF- trees was 3,000.

Table 2. Parameter estimation under the best supported scenario (scenario 4).

	Posterior values using an informed mutational prior setting		Posterior values using a naive mutational prior setting		Prior values	
	Median	90% CI	Median	90% CI	Median	90% CI
t_{div}	7,723	2,785 – 19,708	5,235	1,224 – 23,845	1,212	124 – 73,795
t_{ca} / t_{div}	2.75	1.11 – 35.47	4.55	1.24 – 51.66	12.17	1.24 – 762.26
N_f / N_g	5.43	0.52 – 25.56	4.71	0.37 – 31.45	1.00	0.12 – 8.11
db_f / Nb_f	1.06	0.49 – 2.41	1.09	0.45 – 3.11	1.00	0.13 – 7.57

t_{div}: time of divergence between the two desert locust subspecies (in number of generations); t_{ca}: time of ancestral population size contraction; N_f : stable effective population size of *S. g. flaviventris*; N_g: stable effective population size of *S. g. gregaria;* db_f : duration of the bottleneck event; Nb_f : effective population size during the bottleneck event. For each evolutionary parameter, we reported posterior point estimates averaged over ten replicate analyses. CI: credibility interval. The number of records for each reference datasets simulated from DIYABC was set to 100,000 and the number of RF-trees was 2,000.

Table 3. Accuracy in parameter estimation under the best supported scenario (scenario 4).

Mutational prior setting	t_{div}	t_{ca} / t_{div}	N_f / N_g	db_f / Nb_f
Informed				
Prior NMAE	0.359	1.077	1.726	0.299
Posterior NMAE	0.369	0.596	1.332	0.323
Naive				
Prior NMAE	0.542	1.258	1.824	0.340
Posterior NMAE	0.571	0.921	1.382	0.391

Accuracy has been measured with the normalized mean absolute error (NMAE) which corresponds to the mean of the absolute difference between the point estimate of the parameter (here the median) and the (true) simulated value divided by the (true) simulated value. NMAE measures were computed using out-of-bag predictions either on the whole data space defined by the prior distributions (prior NMAE) or conditionally to the observed dataset (posterior NMAE); see Supplementary material S1 for details. t_{div}: time of divergence between the two desert locust subspecies (in number of generations); t_{ca}: time of ancestral population size contraction; N_f : stable effective population size of *S. g. flaviventris*; N_g: stable effective population size of *S. g. gregaria;* db_f : duration of the bottleneck event; Nb_f : effective population size during the bottleneck event. For each evolutionary parameter, reported error estimates were averaged over ten replicate analyses. The number of records for each reference datasets simulated from DIYABC was set to 100,000 and the number of RF-trees was 2,000.

Supplementary material online for the manuscript: « A young age of subspecific divergence in the desert locust »

Marie-Pierre Chapuis[1,2], Louis Raynal[3,4], Christophe Plantamp[5], Laurence Blondin[6], Jean-Michel Marin[3,4] and Arnaud Estoup[4,7]

[1]CIRAD, CBGP, Montpellier, France.

[2]CBGP, CIRAD, INRA, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France.

[3] IMAG, Univ de Montpellier, CNRS, Montpellier, France.

[4]Institut de Biologie Computationnelle (IBC), Montpellier, France.

[5]ANSES, Laboratoire de Lyon, France.

[6]CIRAD, BGPI, Montpellier, France.
[7]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France.

Outline:

Supplementary material S1. Overview of the used ABC Random Forest (ABC-RF) methods.

Supplementary material S2. Details on results from ABC-RF treatments using an informed mutational prior setting.

Supplementary material S3. Details on results from ABC-RF treatments using a naive mutational prior setting.

Supplementary material S4. On the influence of climatic cycles on the potential range variation of the desert locust *Schistocerca gregaria.*

Supplementary material S5. Details on results from ABC-RF treatments when assuming uniform priors for the three time period parameters of the studied scenarios.

Supplementary material S6. Details on the set of summary statistics used for ABC-RF treatments and effect of the number of simulated datasets recorded in the reference table and of the number of trees in the random forest.

Supplementary Material S1: Overview of the used ABC Random Forest (ABC-RF) methods

In this supplementary material, we provide readers with an overview of the Approximate Bayesian Computation Random Forest (hereafter ABC-RF) methods used in the present paper. We invite the reader to consult Pudlo et al. (2016), Estoup et al. (2018), and Raynal et al. (2018) for more in-depth explanations.

ABC framework

Let $\mathbf{y}$ denote the observed data and $\boldsymbol{\theta}$ a vector of parameters associated to a statistical model whose likelihood is $f(. \mid \boldsymbol{\theta})$. Under the Bayesian parametric paradigm the posterior distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

is of prime interest. It characterizes the distribution of $\boldsymbol{\theta}$ given the observation $\mathbf{y}$ and can be interpreted as an update of the prior distribution $\pi(\boldsymbol{\theta})$ by the likelihood of $\mathbf{y}$. The likelihood is hence pivotal, but unfortunately intractable in the evolutionary scenarios (models) we consider in the present study, as well as in many other evolutionary studies. As a matter of fact, the underlying Kingman's coalescent process (Kingman, 1982) does not allow a close expression for the likelihood because all the possible genealogies and mutational process yielding $\mathbf{y}$ should be considered. To solve this issue, some likelihood-free methods have been developed using the fact that, even though the likelihood is not available, generating artificial (i.e. simulated) data for a given value of $\boldsymbol{\theta}$ is much easier if not feasible (e.g. Beaumont (2010). Approximate Bayesian computation (ABC) is one of them (Beaumont et al., 2002).

In a nutshell, ABC consists in generating parameters $\boldsymbol{\theta}'$ and associated pseudo-data $\mathbf{z}$ from the scenario, and accepting $\boldsymbol{\theta}'$ as a realization from an *approximated* posterior if $\mathbf{z}$ is similar to $\mathbf{y}$. In standard ABC treatments, the notion of similarity is defined through the use of a distance ρ to compare $\eta(\mathbf{z})$ and $\eta(\mathbf{y})$, where $\eta(.)$ is a projection of the data in a lower dimensional space of summary statistics. Only pseudo-data providing distance lower than a threshold ϵ are retained. The choice of ρ, $\eta(.)$ and ϵ is a major issue in ABC (Beaumont, 2010).

ABC-RF is a recently derived ABC approach based on the supervised machine learning tool named Random Forest (Breiman, 2001), which has as major advantage to avoid the three above-mentioned difficulties. Initially introduced in Pudlo et al. (2016) for model choice and then extended to parameter inference in Raynal et al. (2018), ABC-RF relies on the use of random forests on a set of simulated pseudo-data according to the generative Bayesian models under consideration. Let consider M Bayesian parametric models. For a given model index $m \in \{1, \ldots, M\}$, a prior probability $\mathbb{P}(\mathcal{M} = m)$ is defined, with $\boldsymbol{\theta}_m$ its associated parameters and $f_m(\mathbf{y} \mid \boldsymbol{\theta}_m)$ its likelihood. The generation process of a reference table made of H elements is described in Algorithm 1.

Algorithm 1: Generation of a reference table with H elements

```
for j ← 1 to H do
    Generate m^(j) from the prior P(M = m)
    Generate θ_{m^(j)} from the prior π_{m^(j)}(.)
    Generate z^(j) from the model f_{m^(j)}(. | θ_{m^(j)})
    Compute η(z^(j)) = (η_1(z^(j)), ..., η_d(z^(j)))
end
```

The output takes the form of a matrix containing simulated model indexes, parameters and summary statistics, as described below

$$\begin{bmatrix} m^{(1)} & \theta_{m^{(1)}} & \eta_1(\mathbf{z}^{(1)}) & \eta_2(\mathbf{z}^{(1)}) & \dots & \eta_d(\mathbf{z}^{(1)}) \\ m^{(2)} & \theta_{m^{(2)}} & \eta_1(\mathbf{z}^{(2)}) & \eta_2(\mathbf{z}^{(2)}) & \dots & \eta_d(\mathbf{z}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m^{(H)} & \theta_{m^{(H)}} & \eta_1(\mathbf{z}^{(H)}) & \eta_2(\mathbf{z}^{(H)}) & \dots & \eta_d(\mathbf{z}^{(H)}) \end{bmatrix}.$$

ABC-RF for model choice

The ABC-RF strategy for model choice is described in Algorithm 2. The output is the affectation of $\mathbf{y}$ to a model (scenario), this decision being made based on the majority class of the RF tree votes.

Algorithm 2: ABC-RF for model choice

Input : a reference table used as learning set, made of H elements, each one composed of a model index $m^{(H)}$ and d summary statistics. A possibly large collection of summary statistics can be used, including some obtained by machine-learning techniques, but also by scientific theory and knowledge

Learning : construct a classification random forest $\hat{m}(\cdot)$ to infer model indexes

Output : apply the random forest classifier to the observed data $\eta(\mathbf{y})$ to obtain $\hat{m}(\eta(\mathbf{y}))$

The selected scenario is the one with the highest number of votes in his favor. In addition to this majority vote, the posterior probability of the selected scenario can be computed as described in Algorithm 3.

Algorithm 3: ABC-RF computation of the posterior probability of the selected scenario

Input : the values of $\mathbb{I}\left\{m^{(h)} \neq \hat{m}(\eta(\mathbf{z}^{(h)}))\right\}$ for the trained random forest and corresponding summary statistics of the reference table, using the out-of-bag classifiers

Learning : construct a regression random forest $\hat{\mathbb{E}}(.)$ to infer $\mathbb{E}\left(\mathbb{I}\left\{m \neq \hat{m}(\eta(\mathbf{y}))\right\} \mid \eta(\mathbf{y})\right)$

Output : an estimate of the posterior probability of the selected model $\hat{m}(\eta(\mathbf{y}))$

$$\hat{\mathbb{P}}\left(m = \hat{m}(\eta(\mathbf{y})) \mid \eta(\mathbf{y})\right) = 1 - \hat{\mathbb{E}}\left(\mathbb{I}\left\{m \neq \hat{m}(\eta(\mathbf{y}))\right\} \mid \eta(\mathbf{y})\right)$$

Such posterior probability provides a confidence measure of the previous prediction at the point of interest $\eta(\mathbf{y})$. It relies on the building of a regression random forest designed to explain the model prediction error. More specifically, and as a first step, posterior probability computation makes use of out-of-bag predictions of the training dataset. Because each tree of the random forest is built on a bootstrap sampling of the H elements of the reference table (i.e. the training dataset), there is about one third of the reference table that remains unused per tree, and this ensemble of left aside datasets corresponds to the "out-of-bag". Thus, for each pseudo-data of the reference

table, one can obtain an out-of-bag prediction by aggregating all the classification trees in which the pseudo-data was out-of-bag. In a second step, the out-of-bag predictions $\hat{m}(\eta(\mathbf{z}^{(h)}))$ are used to compute the indicators $\mathbb{I}\left\{m^{(h)} \neq \hat{m}(\eta(\mathbf{z}^{(h)}))\right\}$. These 0 - 1 values are used as response variables for the regression random forest training, for which the explanatory variables are the summary statistics of the reference table. Predicting the observed data thanks to this forest allows the derivation of the posterior probability of the selected model (Algorithm 3). Note that using the out-of-bag procedure prevents over-fitting issues and is computationally parsimonious as it avoids the generation of a second reference table for the regression random forest training.

Model grouping A recent useful add-on to ABC-RF has been the model-grouping approach developed in Estoup et al. (2018), where pre-defined groups of scenarios are analysed using Algorithm 2 and 3. The model indexes used in the training reference table are modified in a preliminary step to match the corresponding groups, which are then used during learning phase. When appropriate, unused scenarios are discarded from the reference table. This improvement is particularly useful when a high number of individual scenarios are considered and have been formalized through the absence or presence of some key evolutionary events (e.g. admixture, bottleneck, ...). Such key evolutionary events allow defining and further considering groups of scenarios including or not such events. This grouping approach allows to evaluate the power of ABC-RF to make inferences about evolutionary event(s) of interest over the entire prior space and assess (and quantify) whether or not a particular evolutionary event is of prime importance to explain the observed dataset (see Estoup et al. (2018) for details and illustrations).

ABC-RF for parameter estimation

Once the selected (i.e. best) scenario has been identified, the next step is the estimation of its parameters of interest under this scenario. The ABC-RF parameter estimation strategy is described in Algorithm 4 and takes a similar structure to Algorithm 2. The idea is to use a regression random forest for each dimension of the parameter space (i.e. for each parameter). For a given parameter of interest, the output of the algorithm is a vector of weights $\mathbf{w_y}$ that can be used to compute posterior quantities of interest such as expectation, variance and quantiles. $\mathbf{w_y}$ provides an empirical posterior distribution for $\theta_{m,k}$; see Raynal et al. (2018) for more details.

Algorithm 4: ABC-RF for parameter estimation

Input : a vector of $\theta_{m^{(h)},k}$ values (i.e. the k-th component of $\theta_{m^{(h)}}$) and d summary statistics

Learning : construct a regression random forest to infer parameter values

Output : apply the random forest to the observed data $\eta(\mathbf{y})$, to deduce a vector of weights $\mathbf{w_y} = \{w_{\mathbf{y}}^{(1)}, \ldots, w_{\mathbf{y}}^{(H)}\}$, which provides an empirical posterior distribution for $\theta_{j,k}$
$\mathbf{w_y}$ is used to compute the estimators of the mean, the variance and the quantiles of the parameter of interest

$$\hat{\mathbb{E}}(\theta_{m,k} \mid \eta(\mathbf{y})), \;\; \hat{\mathbb{V}}(\theta_{m,k} \mid \eta(\mathbf{y})), \;\; \hat{\mathbb{Q}}_{\alpha}(\theta_{m,k} \mid \eta(\mathbf{y}))$$

Global prior errors

In both contexts, model choice or parameter estimation, a global quality of the predictor can be computed, which does not take the observed dataset (about which one wants to make inferences) into account. Random forests make it possible the computation of errors on the training reference table, using the out-of-bag predictions previously described in the section "ABC-RF for model choice".

For model choice, this type of error is called the prior error rate, which is the mis-classification error rate computed over the entire multidimensional prior space. It can be computed as

$$\frac{1}{H}\sum_{h=1}^{H}\mathbb{I}\left\{m^{(h)} \neq \hat{m}(\eta(\mathbf{z}^{(h)}))\right\}.$$

For parameter estimation, the equivalent is the prior mean squared error (MSE) or the normalised mean absolute error (NMAE), the latter being less sensitive to extreme values. These errors are computed as

$$\begin{aligned}\text{MSE} &= \frac{1}{H}\sum_{h=1}^{H}\left(\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}\right)^2,\\ \text{NMAE} &= \frac{1}{H}\sum_{h=1}^{H}\left|\frac{\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}}{\theta_{m^{(h)},k}}\right|.\end{aligned}$$

They can be perceived as Monte Carlo approximation of expectations with respect to the prior distribution.

Local posterior errors

In the present paper, we propose some posterior versions of errors, which target the quality of prediction with respect to the posterior distribution. As such errors take the observed dataset $\eta(\mathbf{y})$ into account, we mention them as local posterior errors.

For model choice, the posterior probability provided by Algorithm 3 is a confidence measure of the selected scenario given the observation. Therefore

$$1 - \hat{\mathbb{P}}\left(m = \hat{m}(\eta(\mathbf{y})) \mid \eta(\mathbf{y})\right)$$

directly yields the posterior error associated to $\eta(\mathbf{y})$: $\hat{\mathbb{P}}\left(m \neq \hat{m}(\eta(\mathbf{y})) \mid \eta(\mathbf{y})\right)$.

For parameter estimation, when trying to infer on $\theta_{m,k}$, a point-wise analogous measure of a local error can be computed as the posterior expectations

$$\mathbb{E}\left(\left(\theta_{m,k} - \hat{\theta}_{m,k}\right)^2 \mid \eta(\mathbf{y})\right) \qquad \text{and} \qquad \mathbb{E}\left(\left|\frac{\theta_{m,k} - \hat{\theta}_{m,k}}{\theta_{m,k}}\right| \mid \eta(\mathbf{y})\right). \tag{1}$$

We approximate these expectations by

$$\sum_{i=1}^{H} w_{\mathbf{y}}^{(h)}\left(\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}\right)^2 \qquad \text{and} \qquad \sum_{i=1}^{H} w_{\mathbf{y}}^{(h)}\left|\frac{\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}}{\theta_{m^{(h)},k}}\right|.$$

We again uses the out-of-bag information to compute $\hat{\theta}_{m^{(h)},k}$, hence avoiding the (time consuming) production of a second reference table, and assume that the weights $\mathbf{w_y}$ from the regression random forest are good enough to approximate any posterior expectations of functions of $\theta_{m,k}$: $\mathbb{E}(g(\theta_{m,k}) \mid \eta(\mathbf{y}))$.

Another more expensive strategy to evaluate the posterior expectations (1) is to construct new regression random forests using the out-of-bag vector of values

$$\left(\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}\right)^2 \qquad \text{or} \qquad \left|\frac{\theta_{m^{(h)},k} - \hat{\theta}_{m^{(h)},k}}{\theta_{m^{(h)},k}}\right|,$$

depending on the targeted error. The observation $\eta(\mathbf{y})$ is then given to the forests, targeting the expectations (1).

Note that the values $\hat{\theta}_{m^{(h)},k}$ in the previous formulas can be replaced by either the approximated posterior expectations $\hat{\mathbb{E}}(\theta_{m^{(h)},k} \mid \eta(\mathbf{y}))$ or the posterior medians $\hat{\mathbb{Q}}_{50\%}(\theta_{m^{(h)},k} \mid \eta(\mathbf{y}))$, again using the out-of-bag information, to provide the local posterior errors. We found that both in the present paper (see main text, Materials and Methods section) and for various tests that we carried out on different inferential setups and datasets (results not shown), the posterior median provides a better accuracy of parameter estimation than the posterior expectation (aka posterior mean). This trends also holds for global prior errors that can be computed using either the mean or the median as point estimates.

As final comment, it is worth noting that so far a common practice consisted in evaluating the quality of prediction (for model choice or parameter estimation) in the neighborhood of the observed dataset, that is around $\eta(\mathbf{y})$ and not exactly for $\eta(\mathbf{y})$. For model choice, Estoup et al. (2018) use the so called posterior predictive error rate which is an error of this type. In this case, some simulated datasets of the reference table close to the observation are selected thanks to an Euclidean distance, then new pseudo-observed datasets are simulated using similar parameters, on which is computed the error (see also Lippens et al., 2017, for a similar approach in a standard ABC framework). However, the main problem of processing this way is the difficulty to specify the size of the area around the observation, especially when the number of summary statistics is large. We therefore do not recommend the use of such a "neighborhood" error anymore, but rather to compute the local posterior errors detailed above as the latter measured prediction quality exactly at the position of interest $\eta(\mathbf{y})$.

References

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406.

Beaumont, M. A., Zhang, W., and Balding, D. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Estoup, A., Raynal, L., Verdu, P., and Marin, J.-M. (2018). Model choice using Approximate Bayesian Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *Journal de la Société Française de Statistique*, 159(3).

Kingman, J. F. C. (1982). On the Genealogy of Large Populations. *Journal of Applied Probability*, 19(A):27–43.

Lippens, C., Estoup, A., Hima, M. K., Loiseau, A., Tatard, C., Dalecky, A., Bâ, K., Kane, M., Diallo, M., Sow, A., Niang, Y., Piry, S., Berthier, K., Leblois, R., Duplantier, J. M., and Brouat, C. (2017). Genetic structure and invasion history of the house mouse (Mus musculus domesticus) in Senegal, West Africa: a legacy of colonial and contemporary times. *Heredity*, 119(2):64–75.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2018). ABC random forests for Bayesian parameter inference. *Bioinformatics*. to appear.

Supplementary material S2. Details on results from ABC-RF treatments using an informed mutational prior setting.

Table S2.1. Scenario choice for each of the ten replicate analyses using an informed mutational prior setting.

We report values for the proportion of votes, prior error rates and posterior probabilities of the best scenario on ten replicate analyses based on ten different reference tables. Scenarios are depicted in Figure 2. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF-trees was 3,000. The scenario 4 was the best supported for all replicate analyses: it involves a bottleneck event in *S. g. flaviventris* right after divergence, a population size contraction in the ancestral population and not any secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*.

Reference table	Best scenario	Votes (scenario 1)	Votes (scenario 2)	Votes (scenario 3)	Votes (scenario 4)	Votes (scenario 5)	Votes (scenario 6)	Votes (scenario 7)	Votes (scenario 8)	Prior error rate	Posterior probability (best scenario)
1	scenario 4	0.010	0.170	0.032	0.629	0.005	0.021	0.018	0.114	0.480	0.625
2	scenario 4	0.012	0.209	0.018	0.583	0.007	0.026	0.023	0.123	0.480	0.589
3	scenario 4	0.011	0.199	0.037	0.608	0.011	0.028	0.017	0.090	0.479	0.597
4	scenario 4	0.006	0.150	0.021	0.639	0.012	0.018	0.034	0.121	0.475	0.627
5	scenario 4	0.010	0.134	0.017	0.669	0.005	0.044	0.023	0.098	0.478	0.613
6	scenario 4	0.009	0.220	0.026	0.563	0.019	0.026	0.030	0.106	0.477	0.546
7	scenario 4	0.007	0.176	0.033	0.532	0.006	0.023	0.040	0.184	0.479	0.565
8	scenario 4	0.020	0.220	0.035	0.506	0.004	0.031	0.018	0.164	0.480	0.536
9	scenario 4	0.010	0.186	0.039	0.572	0.012	0.041	0.034	0.106	0.480	0.521
10	scenario 4	0.013	0.126	0.035	0.620	0.005	0.019	0.021	0.163	0.478	0.618
All	**scenario 4**	**0.011**	**0.179**	**0.029**	**0.592**	**0.008**	**0.028**	**0.026**	**0.127**	**0.479**	**0.584**

Table S2.2. Estimation of the divergence time between *S. g. gregaria* and *S. g. flaviventris* for the ten replicate analyses processed using an informed mutational prior setting under the best supported scenario (scenario 4).

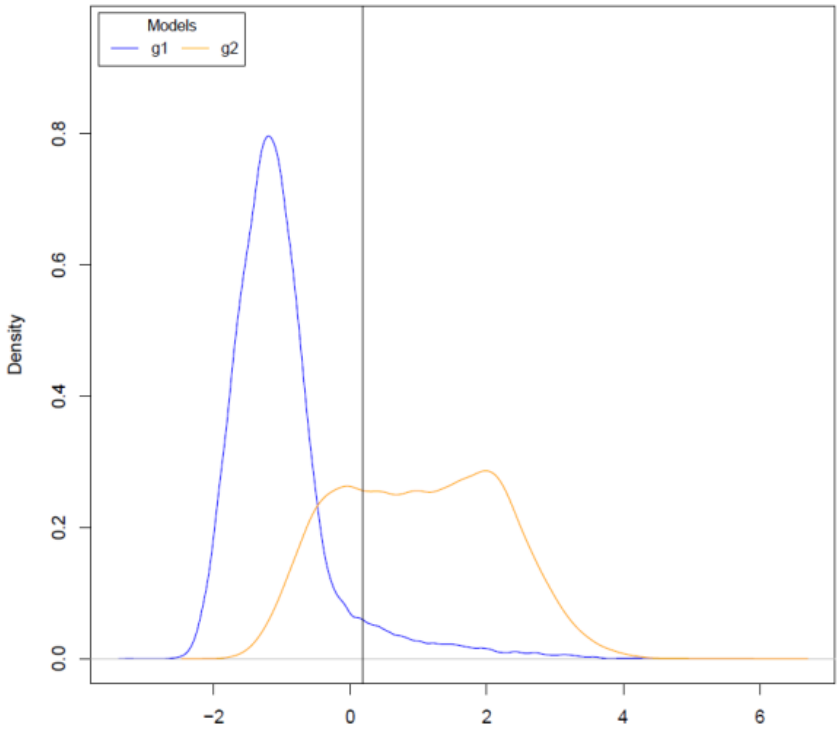
Replicate analyses have been processed on different reference tables. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF-trees was 2,000. Divergence times are given in number of generations (G). SD stands for standard deviations computed from the ten values of median, 5% quantile and 95% quantile estimated from the ten replicate analyses.

t_{div} (*G*)	Median	q5%	q95%
reference table 1	7440.0	2485.0	19380.0
reference table 2	8257.1	2668.0	21086.0
reference table 3	7930.3	2771.0	20310.9
reference table 4	7301.0	2888.0	19639.5
reference table 5	7598.6	2376.0	18260.6
reference table 6	7426.0	2975.7	19704.7
reference table 7	7776.0	3190.0	19290.2
reference table 8	7960.0	2812.0	19664.2
reference table 9	7552.3	2717.0	20685.0
reference table 10	7991.0	2966.9	19060.7
Mean	**7723.2**	**2785.0**	**19708.2**
SD	**307.3**	**240.5**	**817.8**

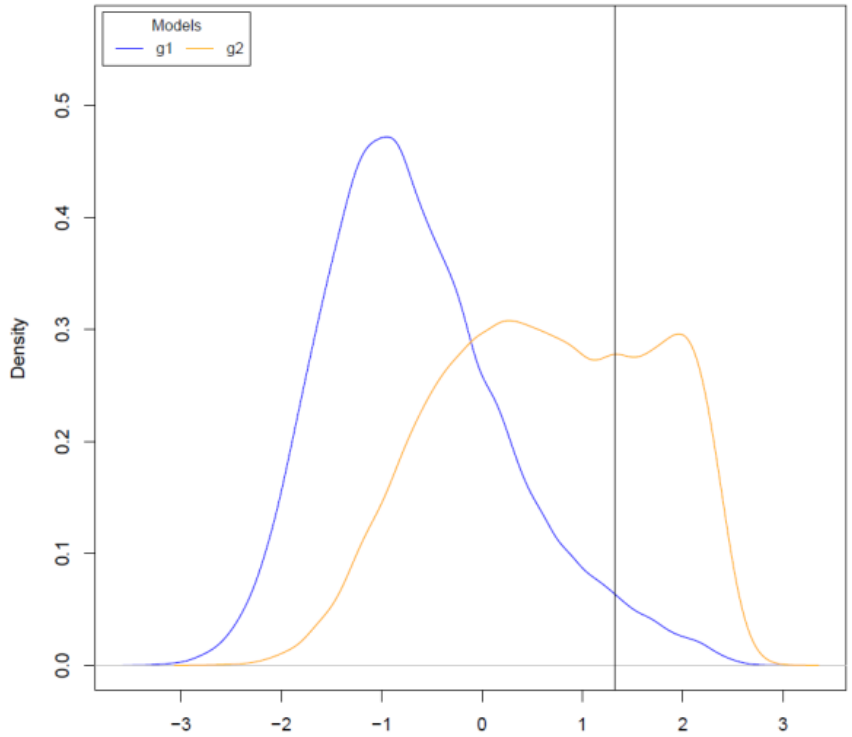
Figure S2.1. Projection on a single (when analyzing pairwise groups of scenarios) or on the first two LDA axes (when analyzing the eight scenarios separately) of the observed dataset and the datasets recorded in the reference table simulated using an informed mutational prior setting.

Colors correspond to group of scenarios or individual scenarios. The location of the desert locust observed dataset is indicated by a vertical black line or a star. Scenarios were grouped based on the presence or not of a bottleneck in *S. g. flaviventris* (*b or no b*), a population size contraction in ancestor (c_a *or no* c_a) and a secondary contact with asymmetrical genetic admixture from *S. g. gregaria in*to *S. g. flaviventris* (*sc or no sc*). When considering the whole set of eight scenarios separately (d), the projected points substantially overlapped for at least some of the scenarios. This suggests an overall low power to discriminate among scenarios considered. Conversely, considering pairwise groups of scenarios, one can observe a weaker overlap of projected points (at least for (a) and (b)) suggesting a stronger power to discriminate among groups of scenarios of interest than when considering all scenarios separately. One can note that the location of the observed dataset (indicated by a vertical line) suggests an association with the scenario group with a bottleneck event in *S. g. flaviventris* and with the scenario group with a population size contraction in the ancestral population.

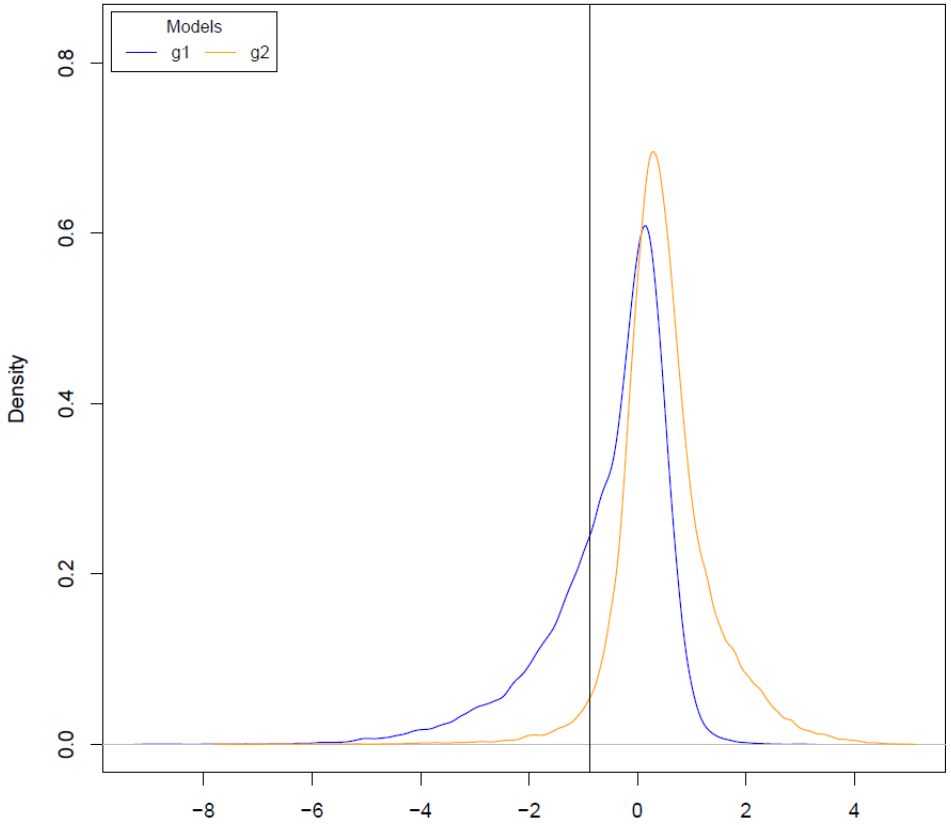
(a) Scenario group g1= *no b* vs. group g2 = *b*



(b) Scenario group g1 = *no* c_a vs. group g2 = c_a



(c) Scenario group g1 = *no sc* vs. group g2 = *sc*



(d) All eight scenarios considered separately

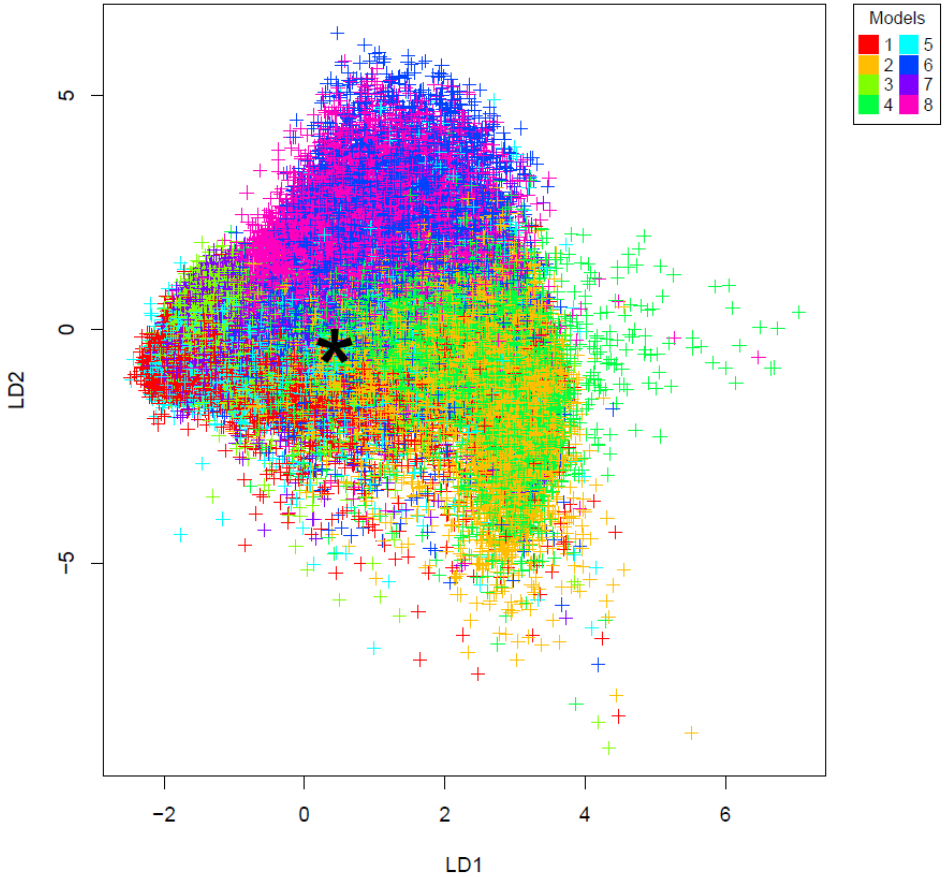
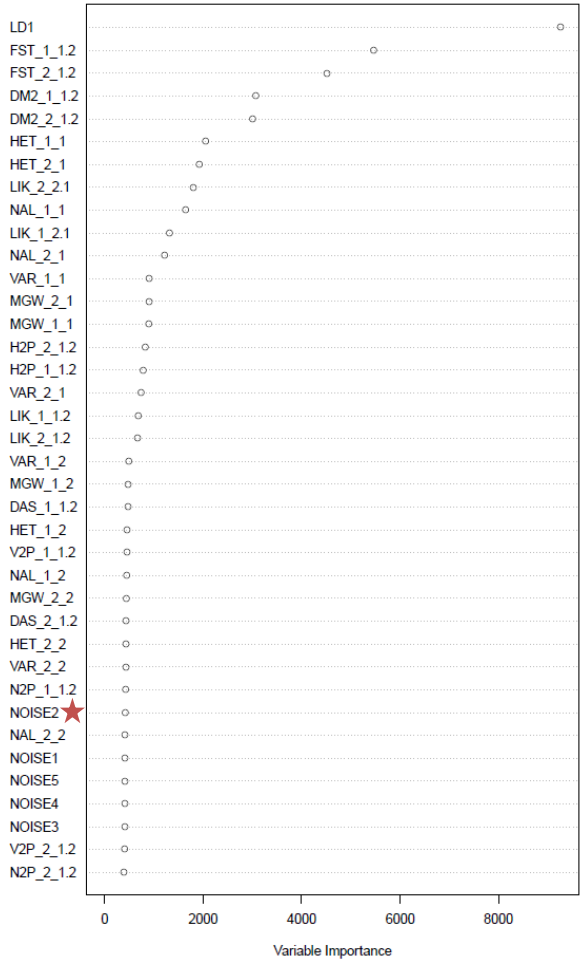


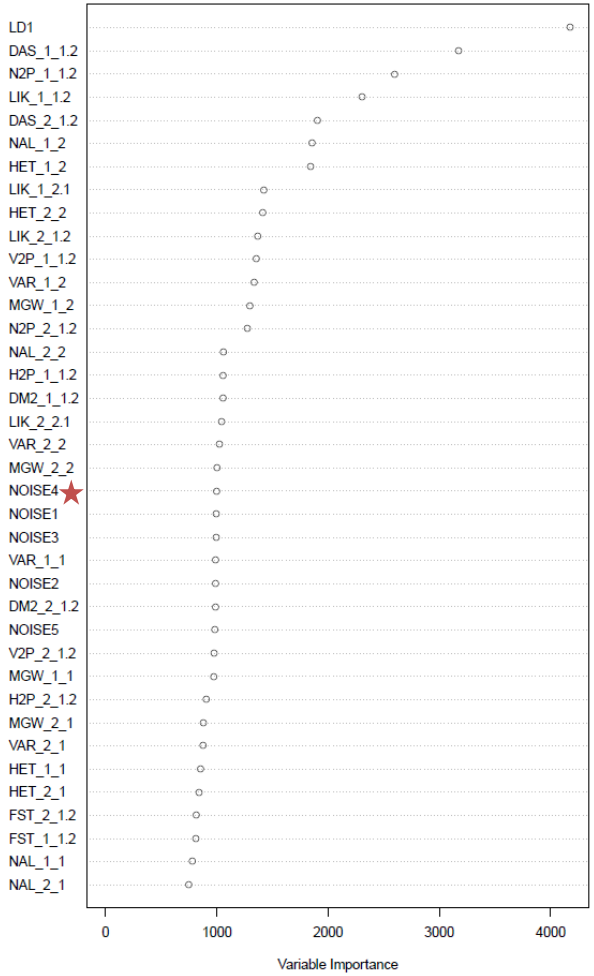
Figure S2.2. Contributions of ABC-RF summary statistics when choosing between groups of scenarios using an informed mutational prior setting.

The contribution of each 32 summary statistics and one LDA axis is evaluated as the total amount of decrease in the Gini criterion (variable importance on the x-axis). The higher the contribution of the statistics, the more informative it is in the inferential process. The microsatellite set and subspecies sample are indicated at the end of each statistics by indices *k_i* for single population statistics and *k_i.j* for two population statistics, with *k*=1 for the set of untranscribed microsatellites or *k*=2 for the set of transcribed microsatellites, and *i*(*j*)=1 for the *S. g. flaviventris* subspecies or and *i*(*j*)=2 for the *S. g. gregaria* subspecies. See Table S6.1 for details on the summary statistics abbreviations. Five noise variables, randomly drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 were added to the set of summary statistics processed by RF, in order to evaluate from which amount of decrease in the Gini criterion the summary statistics computed from our genetic datasets were not informative anymore (indicated by a red star).

(a) Scenario group g1=no *b* vs. group g2 = *b*



(b) Scenario groups g1 = no c_a vs group. g2 = c_a



(c) Scenario groups g1 = no *sc* vs. group g2 = *sc*

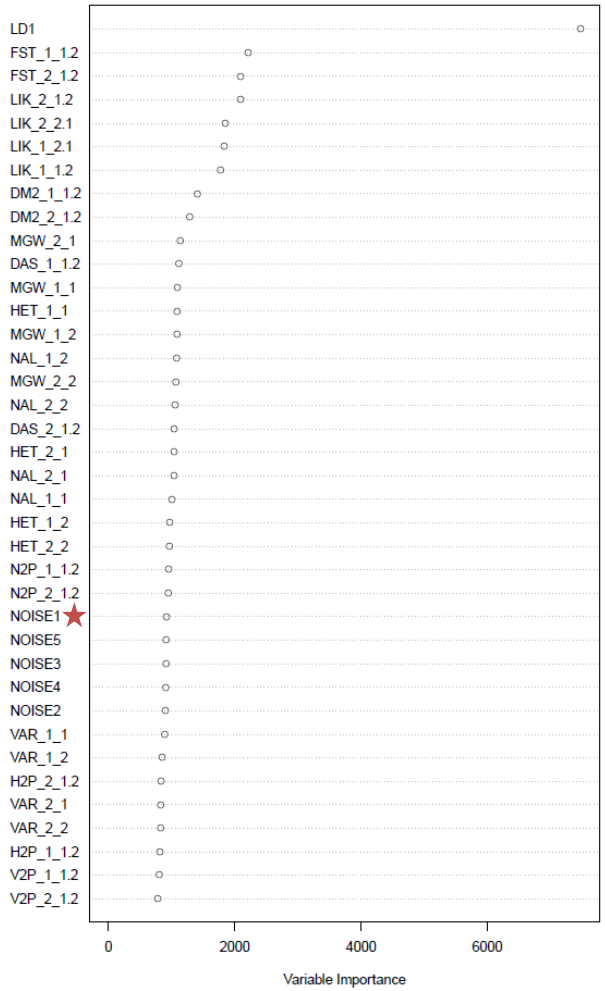


Figure S2.3. Contributions of ABC-RF summary statistics when choosing among the eight individual scenarios using an informed mutational prior setting.

The contribution of each 32 summary statistics and one LDA axis is evaluated as the total amount of decrease in the Gini criterion (variable importance on the x-axis). The higher the contribution of the statistics, the more informative it is in the inferential process. The microsatellite set and subspecies sample are indicated at the end of each statistics by indices *k_i* for single population statistics and *k_i.j* for two population statistics, with *k*=1 for the set of untranscribed microsatellites or *k*=2 for the set of transcribed microsatellites, and *i*(*j*)=1 for the *S. g. flaviventris* subspecies or and *i*(*j*)=2 for the *S. g. gregaria* subspecies. See Table S6.1 for details on the summary statistics abbreviations. Five noise variables, randomly drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 were added to the set of summary statistics processed by RF, in order to evaluate from which amount of decrease in the Gini criterion the summary statistics computed from our genetic datasets were not informative anymore (indicated by a red star).

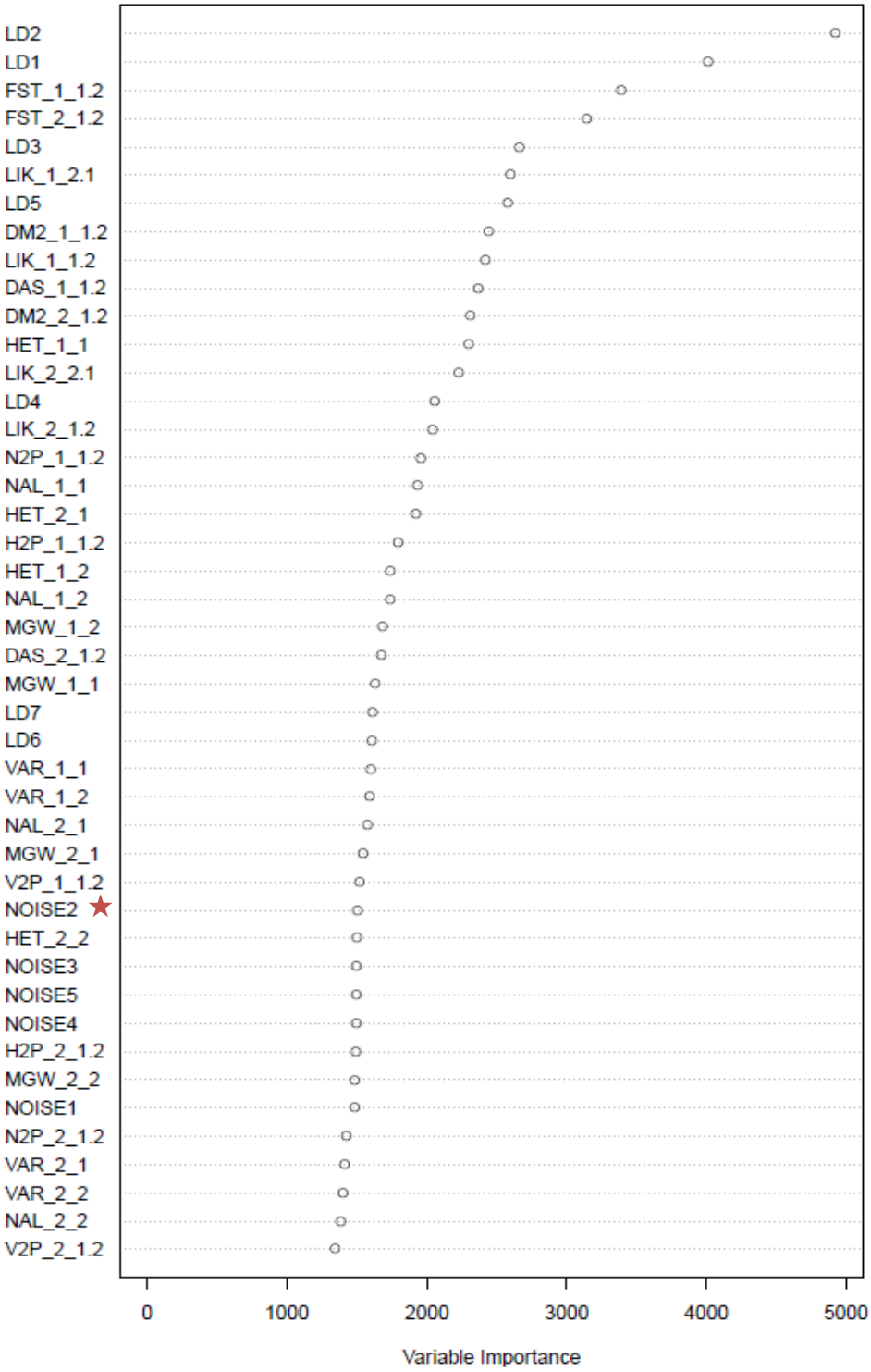
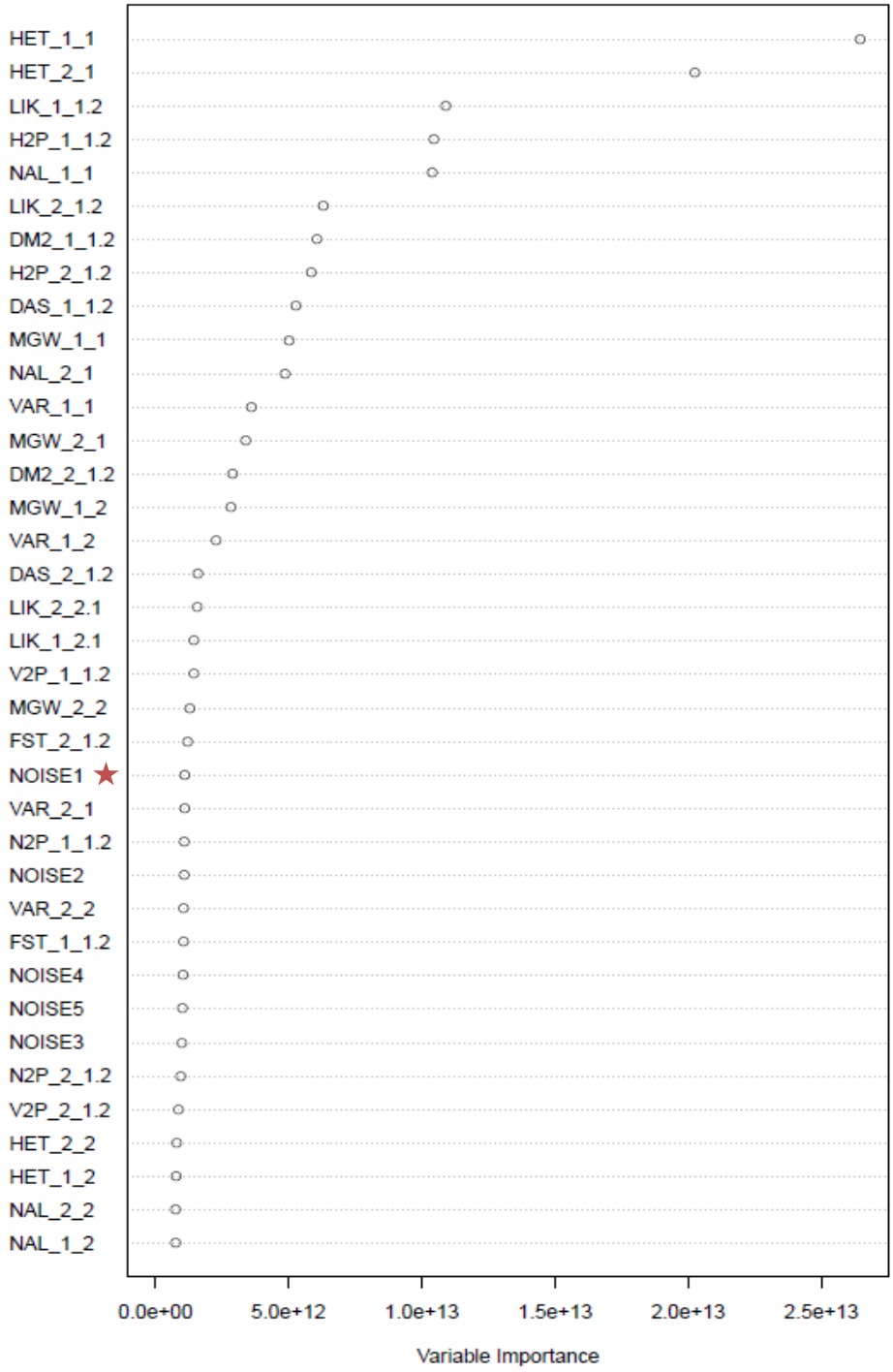


Figure S2.4. Contributions of ABC-RF summary statistics when estimating the divergence time between the two desert locust subspecies using an informed mutational prior setting under the best supported scenario (scenario 4).

The contribution of each 32 summary statistics is evaluated as the total amount of decrease of the residual sum of squares, divided by the number of trees, (variable importance on the x-axis). The higher the contribution of the statistics, the more informative it is in the inferential process. The microsatellite set and subspecies sample are indicated at the end of each statistics by indices *k_i* for single population statistics and *k_i.j* for two population statistics, with *k*=1 for the set of untranscribed microsatellites or *k*=2 for the set of transcribed microsatellites, and *i*(*j*)=1 for the *S. g. flaviventris* subspecies or and *i*(*j*)=2 for the *S. g. gregaria* subspecies. See Table S6.1 for details on the summary statistics abbreviations. Five noise variables, randomly drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 were added to the set of summary statistics processed by RF, in order to evaluate from which amount of decrease in the variable importance criterion the summary statistics computed from our genetic datasets were not informative anymore (indicated by a red star).



Supplementary material S3. Details on results from ABC-RF treatments using a naive mutational prior setting.

Table S3.1. Scenario choice for the ten replicate analyses using a naive mutational prior setting.

We report values for the proportion of votes, prior error rates and posterior probabilities of the best scenario on ten replicate analyses based on ten different reference tables. Scenarios are depicted in Figure 2. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF-trees was 3,000. The scenario 4 was the best supported for all replicate analyses: it involves a bottleneck event in *S. g. flaviventris* right after divergence, a population size contraction in the ancestral population and not any secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris.*

Reference table	Best scenario	Votes (scenario 1)	Votes (scenario 2)	Votes (scenario 3)	Votes (scenario 4)	Votes (scenario 5)	Votes (scenario 6)	Votes (scenario 7)	Votes (scenario 8)	Prior error rate	Posterior probability (best scenario)
1	scenario 4	0.029	0.198	0.029	0.593	0.007	0.022	0.024	0.098	0.501	0.544
2	scenario 4	0.015	0.149	0.025	0.680	0.008	0.022	0.020	0.080	0.503	0.597
3	scenario 4	0.035	0.218	0.037	0.581	0.003	0.028	0.044	0.055	0.503	0.444
4	scenario 4	0.014	0.142	0.022	0.661	0.003	0.018	0.029	0.112	0.504	0.578
5	scenario 4	0.021	0.199	0.028	0.578	0.009	0.045	0.017	0.103	0.505	0.478
6	scenario 4	0.007	0.143	0.027	0.692	0.009	0.020	0.018	0.082	0.500	0.610
7	scenario 4	0.009	0.193	0.033	0.620	0.014	0.037	0.017	0.077	0.503	0.554
8	scenario 4	0.038	0.171	0.036	0.614	0.013	0.025	0.027	0.076	0.504	0.565
9	scenario 4	0.019	0.151	0.101	0.540	0.009	0.052	0.034	0.094	0.500	0.518
10	scenario 4	0.005	0.220	0.031	0.600	0.015	0.029	0.025	0.074	0.500	0.583
All	**scenario 4**	**0.019**	**0.178**	**0.037**	**0.616**	**0.009**	**0.030**	**0.026**	**0.085**	**0.502**	**0.547**

Table S3.2. Estimation of the divergence time between *S. g. gregaria* and *S. g. flaviventris* for ten replicate analyses using a naive mutational prior setting under the best supported scenario (scenario 4).

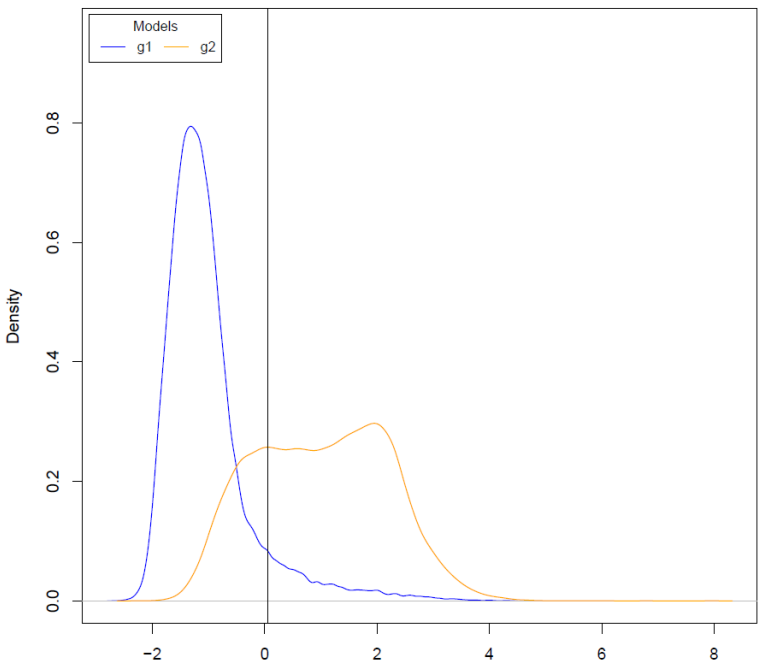
Replicate analyses have been processed on different reference tables. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF-trees was 2,000. Divergence times are given in number of generations (G). SD stands for standard deviations computed from the ten values of median, 5% quantile and 95% quantile estimated from the ten replicate analyses.

t_{div} (*G*)	Median	q5%	q95%
reference table 1	4778.0	1201.0	21090.1
reference table 2	4649.8	1102.0	24906.4
reference table 3	4980.0	1192.0	22502.1
reference table 4	5425.7	1155.0	22348.2
reference table 5	5103.3	1309.0	23837.2
reference table 6	5879.8	1167.0	23822.8
reference table 7	5067.0	1395.0	28037.6
reference table 8	4449.0	1143.0	19775.3
reference table 9	5495.8	1344.0	26723.8
reference table 10	6519.1	1235.0	25410.8
Mean	**5234.8**	**1224.3**	**23845.4**
SD	**619.0**	**95.4**	**2530.7**

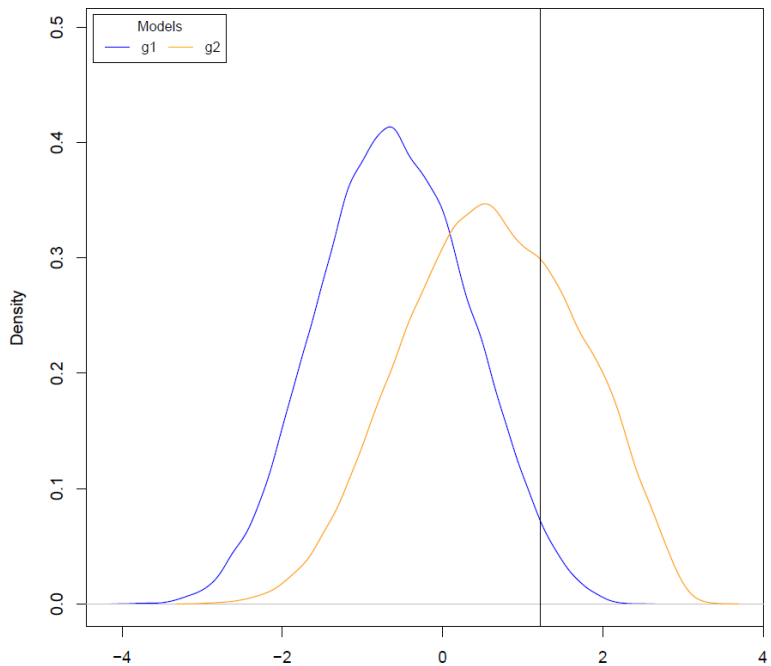
Figure S3.1. Projection on a single (when analyzing pairwise groups of scenarios) or on the first two LDA axes (when analyzing the eight scenarios separately) of the observed dataset and the datasets recorded in the reference table simulated using a naive mutational prior setting.

Colors correspond to group of scenarios or individual scenarios. The location of the desert locust observed dataset is indicated by a vertical black line or a star. Scenarios were grouped based on the presence or not of a bottleneck in *S. g. flaviventris* (*b or no b*), a population size contraction in ancestor (c_a *or no* c_a) and a secondary contact with asymmetrical genetic admixture from *S. g. gregaria in*to *S. g. flaviventris* (*sc or no sc*). When considering the whole set of eight scenarios separately (d), the projected points substantially overlapped for at least some of the scenarios. This suggests an overall low power to discriminate among scenarios considered. Conversely, considering pairwise groups of scenarios, one can observe a weaker overlap of projected points (at least for (a) and (b)) suggesting a stronger power to discriminate among groups of scenarios of interest than when considering all scenarios separately. One can note that the location of the observed dataset (indicated by a vertical line) suggests an association with the scenario group with a bottleneck event in *S. g. flaviventris* and with the scenario group with a population size contraction in the ancestral population.

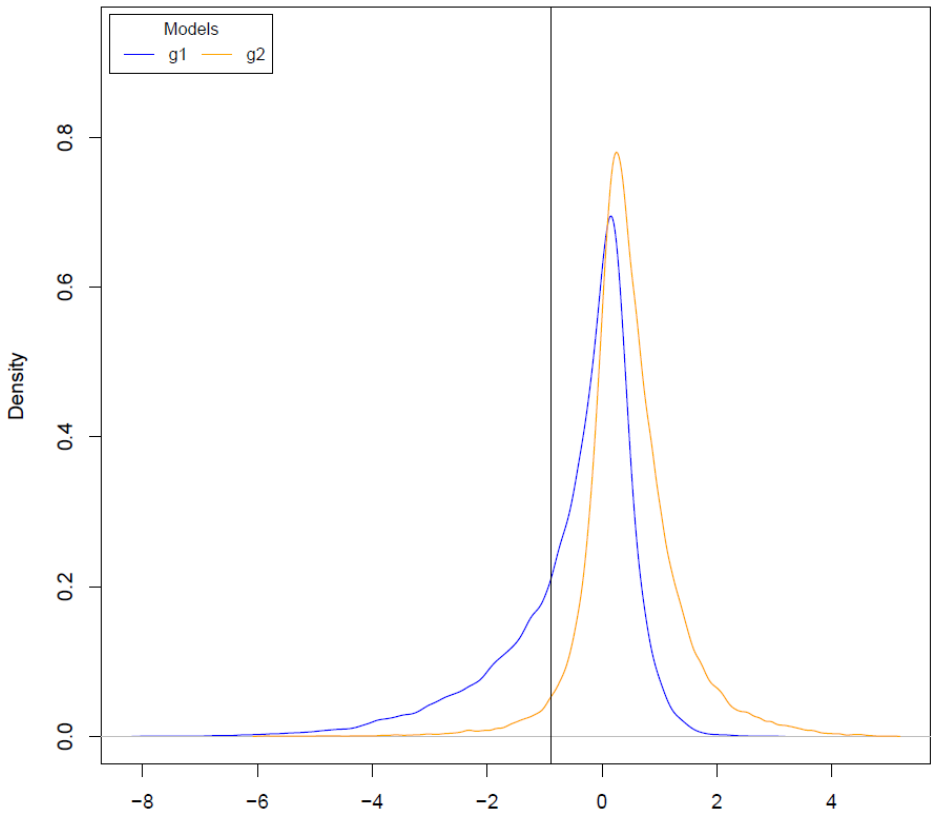
(a) Scenario group g1 = no *b* vs. group g2 = *b*



(b) Scenario group g1 = no c_a vs. group g2 = c_a



(c) Scenario group g1 = no *sc* vs. group g2 = *sc*



(d) All eight scenarios considered separately

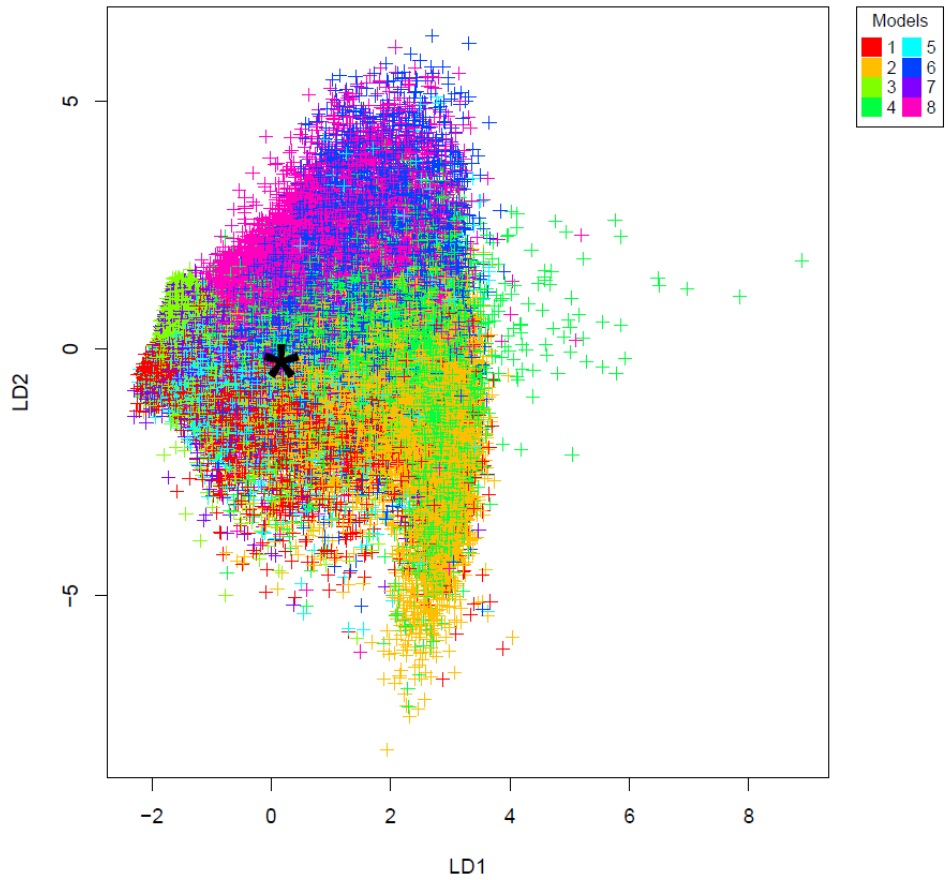


Figure S3.2. Contributions of ABC-RF summary statistics when choosing among the eight individual scenarios using a naive mutational prior setting.

The contribution of each 32 summary statistics and one LDA axis is evaluated as the total amount of decrease in the Gini criterion (variable importance on the x-axis). The higher the contribution of the statistics, the more informative it is in the inferential process. The microsatellite set and subspecies sample are indicated at the end of each statistics by indices *k_i* for single population statistics and *k_i.j* for two population statistics, with *k*=1 for the set of untranscribed microsatellites or *k*=2 for the set of transcribed microsatellites, and *i*(*j*)=1 for the *S. g. flaviventris* subspecies or and *i*(*j*)=2 for the *S. g. gregaria* subspecies. See Table S6.1 for details on the summary statistics abbreviations. Five noise variables, randomly drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 were added to the set of summary statistics processed by RF, in order to evaluate from which amount of decrease in the Gini criterion the summary statistics computed from our genetic datasets were not informative anymore (indicated by a red star).

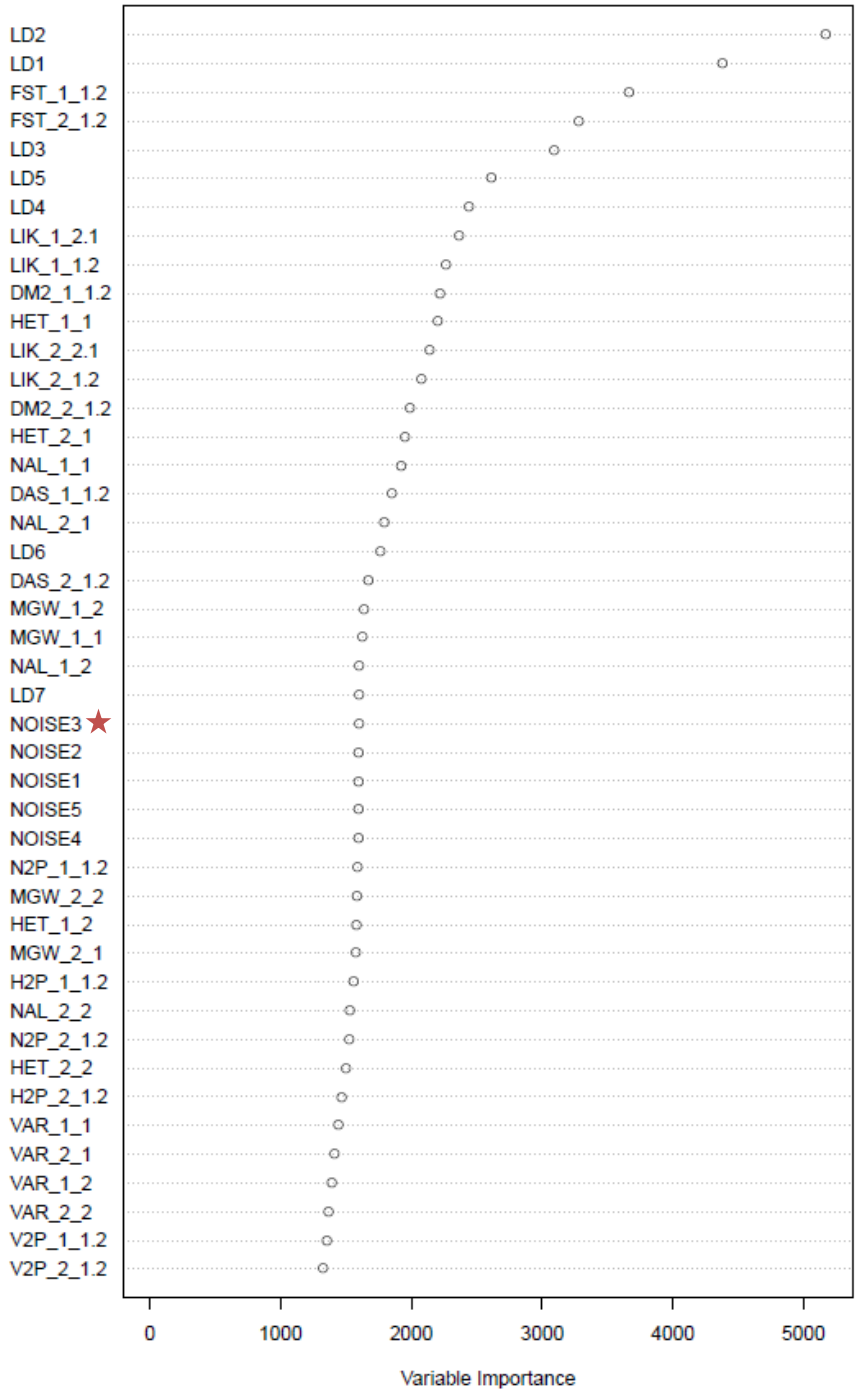
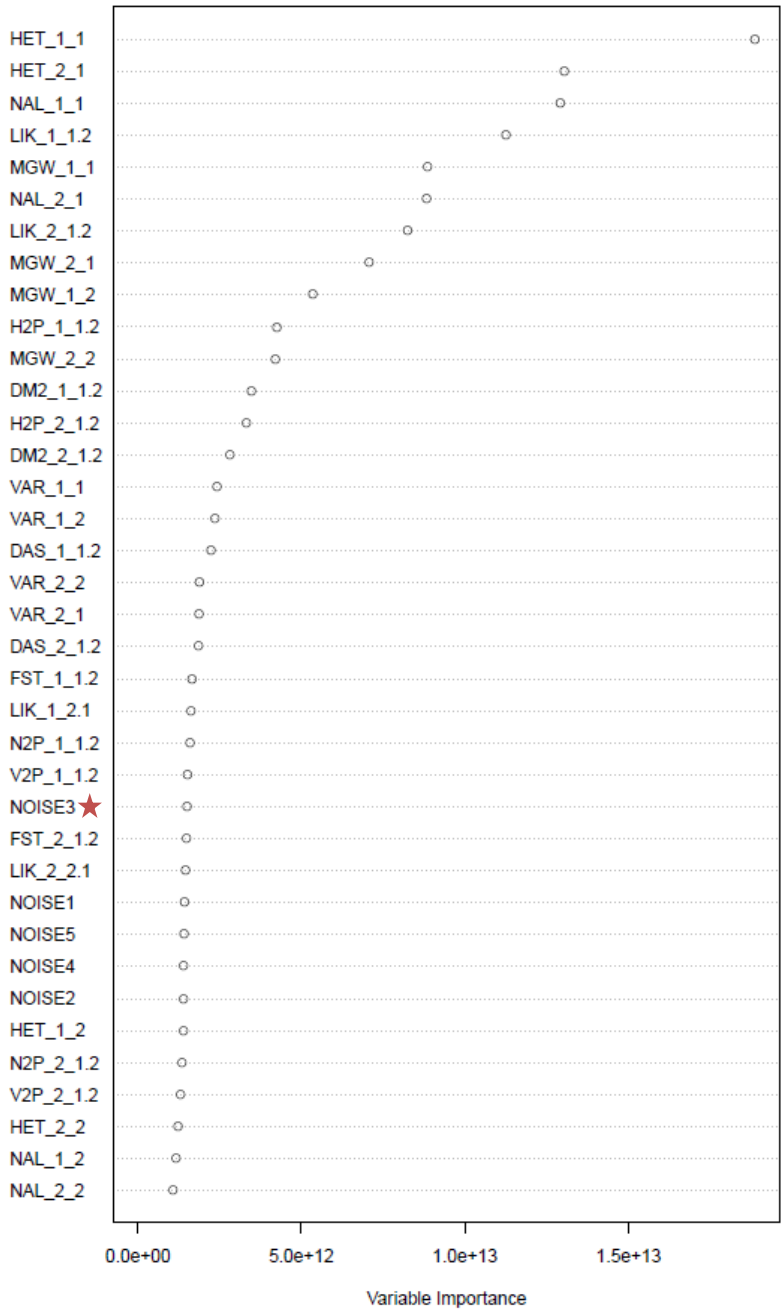


Figure S3.3. Contributions of ABC-RF summary statistics when estimating the divergence time between the two desert locust subspecies using a naïve mutational prior setting under the best supported scenario (scenario 4).

The contribution of each 32 summary statistics is evaluated as the total amount of decrease of the residual sum of squares, divided by the number of trees, (variable importance on the x-axis). The higher the contribution of the statistics, the more informative it is in the inferential process. The microsatellite set and subspecies sample are indicated at the end of each statistics by indices *k_i* for single population statistics and *k_i.j* for two population statistics, with *k*=1 for the set of untranscribed microsatellites or *k*=2 for the set of transcribed microsatellites, and *i*(*j*)=1 for the *S. g. flaviventris* subspecies or and *i*(*j*)=2 for the *S. g. gregaria* subspecies. See Table S6.1 for details on the summary statistics abbreviations. Five noise variables, randomly drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 were added to the set of summary statistics processed by RF, in order to evaluate from which amount of decrease in the variable importance criterion the summary statistics computed from our genetic datasets were not informative anymore (indicated by a red star)



Supplementary material S4. On the influence of climatic cycles on the potential range variation of the desert locust *Schistocerca gregaria.*

It may appear surprising, at least at first sight, that the southern colonization of the desert locust did not occur during one of the major glacial episodes of the last Quaternary cycle, since these periods are characterized by a more continuous range of the desert locust (see paleo-vegetation maps in Figures 1E and 1F in the main text). In particular, during the last glacial maximum (LGM, -14.8 Ky to -26 Ky), the Sahara desert extended hundreds of km further South than at present and annual precipitation were lower (i.e. ~200–1,000 mm/year). Several hypotheses explain why our evolutionary scenario choice procedure provided low support to the possibility of a birth of the locust subspecies *S. g. flaviventris* at older periods. First, we cannot exclude that our microsatellite genetic data allow making inferences about the last colonization event only. The probabilities of choosing scenarios including a genetic admixture event after the split were the lowest, with a posterior predictive error of 16.1% (see Table 1 in the main text). The recent North-to-South colonization event selected by our ABC-RF treatment may hence have blurred traces of older colonization events.

Second, while there is large evidence that much of Africa was drier during the last glacial phase, this remains debated for southwestern Africa (see the gray coloration in Figure 3B in the main text). Some climate models show that at least some parts of this region, such as the Kalahari Desert, may have experienced higher rainfall than at present (Cockcroft 1987; Ganopolski *et al.* 1998; Chase and Meadows 2007). Such regional responses to glacial cycles may have prolonged until the middle Holocene. In particular, the northern Younger Dryas (i.e. -12.9 to -11.7 Ky) can be correlated only partly with an arid period in the southern hemisphere (i.e. -14.4 to -12.5 Ky). Such older climate episodes in antiphase between

hemispheres (see the sandy brown coloration in Figure 3B in the main text) may have prevented from either a successful North-to-South migration event or a successful establishment and spread in the new southern range.

Third, although semi-desert and desert biomes were more expanded than at present during the LGM, extreme aridity and lowered temperatures may have actually been unfavorable to the species. For example, mean temperatures lowered by 5 to 6°C in both southern-western Africa (Stute and Talma 1997) and Central Sahara (Edmunds *et al.* 1999).The maintenance of desert locust populations depends on the proximity of areas with rainfalls at different seasons or with the capacity to capture and release water. For instance, in the African northern range, breeding success of locust populations relies on seasonal movements between the Sahel-Saharan zones of inter-tropical convergence, where the incidence of rain is high in summer, and the Mediterranean-Saharan transition zone, with a winter rainfall regime (Rainey and Waloff 1951). In addition, adult migration and nymphal growth of the desert locust are dependent upon high temperature (Roffey and Magor 2003). It is hence possible then that the conjunction of hyper-aridity with intense cold could not easily support populations of the desert locust, despite the high extent of their migrations.

While ABC-RF analyses did not support that the Quaternary climatic history explained the subspecific divergence in the desert locust, they provided evidence for the occurrence of a large contraction of the size of the ancestral population preceding the divergence. Using the median as a point estimate, we estimated that the population size contraction in the ancestor could have occurred at a time about three fold older than the divergence time between the subspecies. This corresponds to the African humid period in the early and middle stages of the Holocene, though the large credibility interval also included the last interglacial period of the Pleistocene (Figure 3B in the main text). Such population size contraction was likely induced by the severe(s) contraction(s) of deserts that prevailed prior the estimated divergence between the two subspecies. Interestingly, these humid periods were

more intense and prolonged in northern Africa, which corresponded to the presumed center of origin of the most recent common ancestor (Scott 1993; Partridge 1997; Shi *et al.* 1998).

References cited

Chase BM, Meadows ME (2007) Late Quaternary dynamics of southern Africa's winter rainfall zone. *Earth Sci Rev*, **84**, 103-138.

Cockcroft MJ, Wilkinson MJ, Tyson PD (1987) The application of a present-day climatic model to the Late Quaternary in southern Africa. *Climate Change*, **10**, 161-181.

Edmunds WM, Fellman E, Goni IB (1999) Lakes, groundwater and palaeohydrology in the Sahel of NE Nigeria: evidence from hydrogeochemistry. *J Geol Soc Lond*, **156**, 345–355.

Ganopolski A, Rahmstorf S, Petoukohov V, Claussen M (1998) Simulation of modern and glacial climates with a coupled global model of intermediate complexity. *Nature*, **391**, 351-356.

Partridge TC (1997) Cainozoic environmental change in southern Africa, with special emphasis on the last 200 000 years. *Progr Phys Geog*, **21**, 3-22.

Rainey RC, Waloff Z (1951) Flying locusts and convection currents. *Anti-Locust Bull*, **9**, 51-70.

Roffey J, Magor JI (2003) Desert Locust population parameters. Desert Locust Field Research Stations, Technical Series, 30, 29 p. FAO, Rome, Italy.

Scott L (1993) Palynological evidence for late Quaternary warming episodes in Southern Africa. *Palaeogeogr Palaeocl*, **101**, 229-235.

Shi N, Dupont LM, Beug H-J, Schneider R (1998) Vegetation and climate changes during the last 21 000 years in S.W. Africa based on a marine pollen record. *Veg Hist Archaeobot*, **7**, 127-140.

Stute M, Talma AS (1997) Isotope techniques in the study of past and current environmental changes in the hydrosphere and the atmosphere. IAEA Vienna Symposium 1997, Isotopic techniques in the study of environmental change. International Atomic Energy Agency, Vienna, pp. 307–318.

Supplementary material S5. Details on results from ABC-RF treatments when assuming uniform priors for the three time period parameters of the studied scenarios.

Table S5.1. Scenario choice for each of the ten replicate analyses using an informed mutational prior setting and uniform priors for the three time period parameters of the studied scenarios.

We empirically evaluated the influence of shape of prior distributions for the time periods on our inferences by conducting all ABC-RF analyses assuming a set of uniform priors bounded between 100 and 500,000 generations. In the main document, prior values for time periods were drawn from log-uniform distributions bounded between 100 and 500,000 generations. We report values for the proportion of votes, prior error rates and posterior probabilities of the best scenario on ten replicate analyses based on ten different reference tables. Scenarios are depicted in Figure 2. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF-trees was 3,000. The scenario 4 was the best supported for all replicate analyses: it involves a bottleneck event in *S. g. flaviventris* right after divergence, a population size contraction in the ancestral population and not any secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*.

Reference table	Best scenario	Votes (scenario 1)	Votes (scenario 2)	Votes (scenario 3)	Votes (scenario 4)	Votes (scenario 5)	Votes (scenario 6)	Votes (scenario 7)	Votes (scenario 8)	Prior error rate	Probability (best scenario)
1	scenario 4	0.012	0.150	0.041	0.590	0.010	0.058	0.036	0.103	0.697	0.635
2	scenario 4	0.016	0.207	0.071	0.491	0.017	0.028	0.055	0.115	0.695	0.604
3	scenario 4	0.020	0.320	0.029	0.452	0.011	0.027	0.041	0.101	0.696	0.483
4	scenario 4	0.031	0.205	0.061	0.494	0.012	0.045	0.045	0.107	0.698	0.499
5	scenario 4	0.044	0.230	0.053	0.422	0.018	0.056	0.021	0.157	0.697	0.458
6	scenario 4	0.012	0.253	0.034	0.505	0.012	0.049	0.031	0.105	0.697	0.597
7	scenario 4	0.040	0.128	0.114	0.465	0.022	0.057	0.070	0.104	0.697	0.602
8	scenario 4	0.016	0.250	0.055	0.439	0.014	0.028	0.040	0.159	0.696	0.520
9	scenario 4	0.016	0.220	0.061	0.528	0.009	0.030	0.027	0.109	0.697	0.610
10	scenario 4	0.026	0.163	0.031	0.538	0.005	0.049	0.038	0.150	0.693	0.621
Mean	**scenario 4**	**0.023**	**0.212**	**0.055**	**0.492**	**0.013**	**0.043**	**0.040**	**0.121**	**0.696**	**0.563**

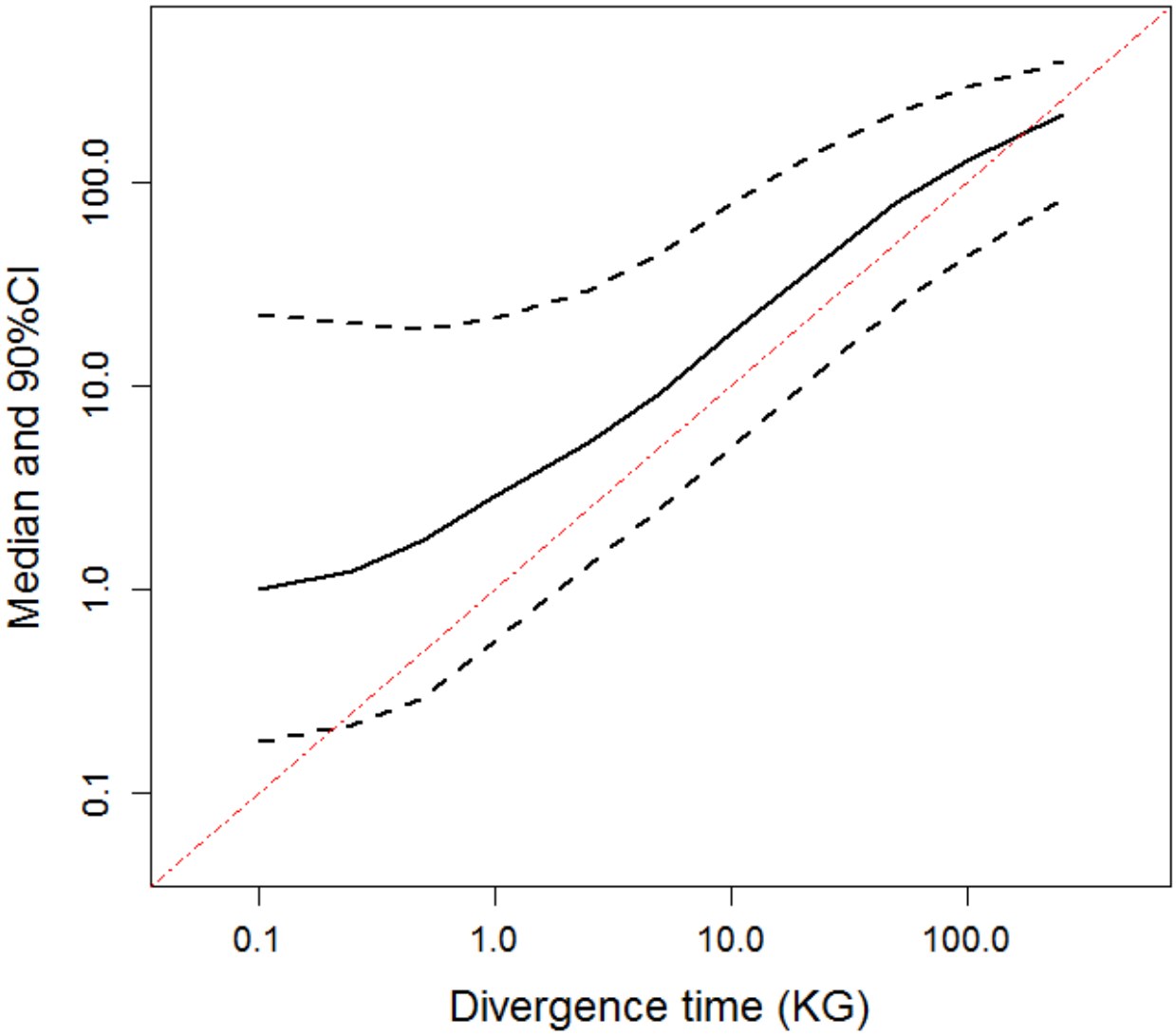
Table S5.2. Estimation of the divergence time between *S. g. gregaria* and *S. g. flaviventris* for ten replicate analyses using an informed mutational prior setting and uniform prior distributions for the three time period parameters under the best supported scenario (scenario 4).

We empirically evaluated the influence of shape of prior distributions for the time periods on our inferences by conducting all ABC-RF analyses assuming a set of uniform priors bounded between 100 and 500,000 generations. Median value and 90% CI for priors are 146,936 and 13,195 – 498,867, respectively. Replicate analyses have been processed on different reference tables. For each reference table, the number of datasets simulated using DIYABC was set to 100,000 and the number of RF- trees was 2,000. Divergence times are given in number of generations. SD stands for standard deviations computed from the ten values of median, 5% quantile and 95% quantile estimated from the ten replicate analyses.

t_{div} (*G*)	Median	q5%	q95%
reference table 1	10236.6	3678.5	25457.0
reference table 2	7603.0	2311.2	23020.3
reference table 3	8500.0	2788.8	19401.0
reference table 4	10598.5	3564.4	24444.5
reference table 5	9226.0	3147.9	21286.6
reference table 6	9665.0	3381.5	24445.2
reference table 7	8675.5	2572.3	26375.8
reference table 8	10281.6	2919.8	27349.0
reference table 9	8845.0	3271.9	24094.7
reference table 10	7909.7	2040.1	21333.4
Mean	**9154.1**	**2967.6**	**23720.8**
SD	**1027.5**	**541.6**	**2476.4**

Figure S5.1. Estimation of the time since divergence between the two desert locust subspecies as a function of time scales using an informed mutational prior setting and uniform prior distributions for the three time period parameters under the best supported scenario (scenario 4).

Simulated datasets (5,000 par divergence time) were generated for fixed divergence time values of 100 ; 250 ; 500; 1,000 ; 2,500; 5,000 ; 10,000 ; 25,000 ; 50,000; 100,000; and 250,000 generations. The median (plain lines) and 90% credibility interval (dashed lines), averaged over the 5,000 datasets, are represented. Divergence time values are in number of generations.



Supplementary material S6. Details on the set of summary statistics used for ABC-RF treatments and effect of the number of simulated datasets recorded in the reference table and of the number of trees in the random forest.

Table S6.1. Summary statistics provided by DIYABC and values computed from the observed microsatellite dataset.

	Summary statistics	Observed values at untranscribed markers	Observed values at transcribed markers
S. g. gregaria **one-sample statistics**	*NAL*	28.8	15.8
	HET	0.92	0.79
	VAR	36.1	13.3
	MGW	0.92	0.86
S. g. flaviventris **one-sample statistics**	*NAL*	23.4	14.4
	HET	0.86	0.69
	VAR	33.4	16.7
	MGW	0.96	0.95
Two-samples statistics	*FST*	0.04	0.12
	DAS	0.07	0.16
	$LIK_{Sgg \rightarrow Sgf}$	3.61	2.82
	$LIK_{Sgf \rightarrow Sgg}$	3.20	2.55
	DM2	22.7	12.4
	N2P	35.0	21.1
	H2P	0.91	0.79
	V2P	40.2	18.2

NAL: mean number of alleles; *HET*: mean expected heterozygosity; *VAR*: variance of allele sizes in base pairs; *MGW*: M index of Garza and Williamson (2001); *FST*: pairwise differentiation estimator of Weir and Cockerham (1984); *DAS*: shared allele distance (Chakraborty and Jin 1993); *LIK*: the mean index of classification (Rannala and Moutain, 1997; Pascual et al. 2007); *DM2*: distance of Golstein et al. (1995). *N2P*, *H2P* and *V2P*: *NAL*, *HET* and *VAR* statistics computed after pooling the two population samples. Note that five “noise variables”, randomly

drawn into uniform distributions bounded between 0 and 1, and denoted NOISE1 to NOISE5 in the concerned illustrations, were added to the set of summary statistics processed by RF, in order to evaluate which summary statistics of our genetic datasets were informative in our different inferential ABC-RF settings, when conducting scenario choice or parameter estimation. Such noise variables do not alter ABC-RF inferences (see Marin et al. 2018; Raynal et al. 2019).

Table S6.2. Effect of the number of simulated datasets in the reference table on scenario choice using an informed mutational prior setting.

.

n_{ref}	50,000		80,000		90,000		100,000	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Votes (scenario 1)	0.011	0.006	0.010	0.006	0.009	0.004	0.011	0.004
Votes (scenario 2)	0.197	0.035	0.192	0.034	0.182	0.029	0.179	0.034
Votes (scenario 3)	0.033	0.009	0.029	0.009	0.031	0.008	0.029	0.008
Votes (scenario 4)	0.572	0.032	0.580	0.051	0.592	0.046	0.592	0.050
Votes (scenario 5)	0.009	0.006	0.008	0.005	0.008	0.005	0.008	0.005
Votes (scenario 6)	0.029	0.012	0.030	0.008	0.029	0.009	0.028	0.009
Votes (scenario 7)	0.021	0.006	0.024	0.008	0.025	0.006	0.026	0.008
Votes (scenario 8)	0.127	0.028	0.126	0.031	0.123	0.027	0.127	0.032
Prior error rate	**0.486**	**0.001**	**0.480**	**0.002**	**0.479**	**0.002**	**0.479**	**0.001**
Posterior probability of the best model	**0.573**	**0.032**	**0.576**	**0.035**	**0.581**	**0.035**	**0.584**	**0.039**

Scenarios are depicted in Figure 2. The number of records in the reference datasets (n_{ref}) simulated from DIYABC varied from 50,000 to 100,000. We report mean and standard deviation values for the proportion of votes for each scenario, and for prior error rates and posterior probabilities of the best scenario for ten replicate analyses. Replicate analyses have been processed on different reference tables. The number of RF-trees was 3,000. The scenario 4 was the best supported for all replicate analyses: it involves a bottleneck event in *S. g. flaviventris* right after divergence, a population size contraction in the ancestral population and not any secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris.*

Table S6.3. Effect of the number of simulated datasets in the reference table on posterior point estimation values (A) and estimation accuracy (B) of the divergence time between *S. g. gregaria* and *S. g. f laviventris* under the best supported scenario (scenario 4) and using an informed mutational prior setting.

(A)

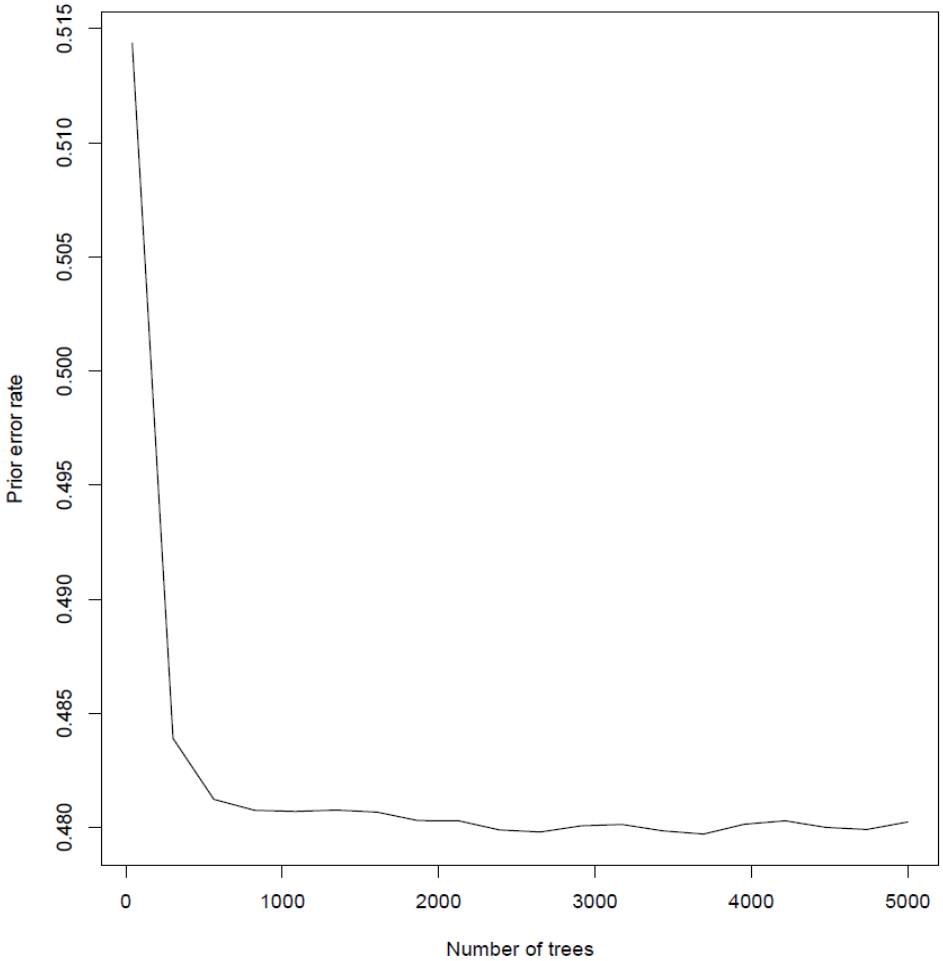
n_{ref}		50,000		80,000		90,000		100,000	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Posterior estimations	Median	7731.8	448.9	7691.8	357.5	7724.8	318.0	7723.2	307.3
	q5%	2697.0	129.7	2706.8	235.2	2764.4	172.9	2785.0	240.5
	q95%	20295.5	1763.6	19711.6	1451.3	19508.7	1264.1	19708.2	817.8

(B)

n_{ref}		50,000	80,000	90,000	100,000
Accuracy measures	Prior NMAE	0.378	0.365	0.362	0.359
	Posterior NMAE	0.375	0.370	0.365	0.369

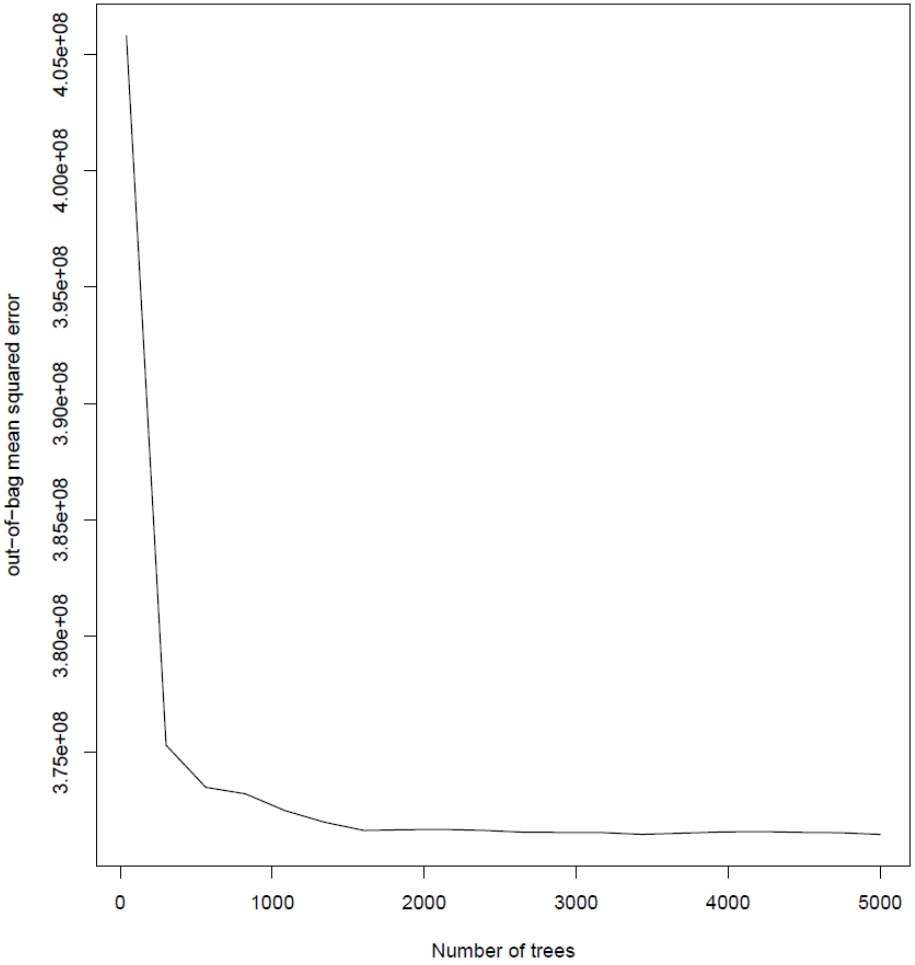
The number of records in the reference datasets (n_{ref}) simulated from DIYABC varied from 50,000 to 100,000. The number of RF-trees was set to 2,000. (a) Replicate analyses have been processed on ten reference tables. (b). The normalized mean absolute error (NMAE) is the absolute difference between the point estimate (here the median) and the (true) simulated value divided by the (true) simulated value.

Figure S6.1. Effect of the number of RF-trees for scenario choice.



We here illustrate the effect of the number of trees in the forest on the prior error rate using an informed mutational prior setting and considering the eight compared scenarios separately. The number of datasets in the reference table simulated using DIYABC was 100,000. The shape of the curve shows that the prior error rate stabilizes for a number of RF-trees > 2,000.

Figure S6.2. Effect of the number of RF-trees for parameter estimation.



We here illustrate the effect of the number of trees in the forest on the RF mean square error of the divergence time between *S. g. gregaria* and *S. g. flaviventris* under the selected scenario 4 and using an informed mutational prior setting. The number of datasets in the reference table simulated using DIYABC was 100,000. The shape of the curve shows that the prior error rate stabilizes for a number of RF-trees > 1,500