Montpellier, the 24th of November 2019

Dear recommenders of PCI Evolutionary Biology,

Please find a revised version of the manuscript entitled "A young age of subspecific divergence in the desert locust, *Schistocerca gregaria*". We are most grateful to the reviewers for their very helpful and constructive suggestions. We have neatly included those suggestions in a revised (and hopefully considerably improved) version of the manuscript. Please find (below) and as attached file, a document which details the comments of the reviewers and outlines our revisions.

Sincerely,

Marie-Pierre Chapuis, on behalf of all authors

Editor

Dear Dr. Chapuis,

Three reviewers have assessed your manuscript. We both agree with reviewers that the study addresses several important issues (eg., the evolutionary history of agriculturally important species in the relatively understudied geographic region) by using carefully formulated analysis. From a technical point of view, it provides a useful statistical framework to discriminate possible demographic scenarios by using microsatellite markers.

At the same time, all three reviewers had numerous but constructive comments about the analyses and interpretation of the data. They pointed out that there are two or three different sections, such as the biological question of dating the divergence between the two subspecies and the methodological approach, but these are not optimally integrated. One possibility is to reduce the methodological details (by moving them to Supplementary Materials) and to focus more on the more biological themes in the main text. Other possible solutions are, of course, welcome.

A couple of minor comments:
1) Fig 1. It seems to me that the dark orange and light orange legend might be exchanged. Does the dark orange represent deserts ("extreme deserts" sensu Adams and Faure 1997) and the light orange represent xeric shrublands ("semi-deserts" sensu Adams and Faure 1997)?

=> We thank the recommenders to report this typo, which is now corrected.

2) Why distinguishing between untranscribed and transcribed microsatellites, if they have been previously shown to be independent and under neutrality (lines 544)? Have I missed the explanation?

=> In Chapuis *et al.* (2015) - *Mol. Ecol.* 24: 6107-6119, we showed in the desert locust that the mean rate of mutation of transcribed dinucleotide microsatellites was half the rate found at untranscribed dinucleotide microsatellites. This was, at least partly, explained by a shorter mean allele length, which is predicted to limit the rate of mutation, and was already reported in natural populations of animal and plant species. Large repeat expansions are often known to be detrimental inside or near genes, causing for instance genetic diseases. In addition, the mutational model deviated from that usually considered for untranscribed dinucleotide microsatellites, with most mutations involving multistep changes that avoid disrupting the reading frame. In other words, evolutionary patterns and rates of mutations differ in genic regions and inter-genic regions, because

of strong negative selection to maintain the long-term stability of genic regions fixed in populations. This background selection is independent of other short-term and fine-tuned selective processes that allow adaptation to environment in natural populations, to which we referred at line 554 when we indicated 'selective neutrality of these loci' (i.e. FST-based tests of divergent and balancing selection using DETSEL). Accordingly, we considered here different prior values of the mean rate of mutation (μ) and the mean parameter of the geometric distribution of the length in number of repeats of mutation events (P) for the two types of microsatellite loci.

We look forward to the revised manuscript.

Sincerely,
Dr. Concetta Burgarella
Dr. Takeshi Kawakami

Reviewer 1

This preprint by Chapuis et al. estimates the divergence time of two subspecies of an African locust species, Schistocerca gregaria. The authors compared present day species distribution to the projected past distribution of associated habitats to formulate competing demographic models. They then used fast-evolving miscrosatellite markers and ABC-random forest inference for model selection and parameter estimation. The authors estimate a young subspecific divergence time with highest support for the demographic scenario with a bottleneck in the southern and ancestral population. I believe the authors did a thorough job in the analyses to infer the best demographic scenario and estimate parameters. The authors provide a good empirical application of the abcrf method for demographic modeling which is an important contribution to make these types of inference more time-efficient and robust to correlated summary statistics.

My main concern with the manuscript at this state is that the main objective gets lost in the details of the ABC-RF methods. The manuscript initially reads as though the priority is to accurately date the divergence time of the two locus subspecies but then shifts priority to the utility of the ABC-RF method. As it is written, these two focuses are not connected cohesively in one story. If the focus is to lean towards the young divergence time of the subspecies (as the title suggests), there needs to be stronger background as to why estimating this parameter is particularly relevant. Why these two subspecies in particular? Why now? What new questions or avenues of research would open up? The divergence time of two subspecies would already be assumed to be quite young so the estimate should be made to be more broadly relevant outside of this species or at least expand on how this is particularly important. If the focus is to lean towards the application of ABCRF, then it would be important to actually include this in the title and have more of a focus in the introduction. It would then also be important to discuss what might be a novel contribution of applying the methods to this particular study system and question. Although it may already be discussed briefly, it would important to further emphasize why this method provides a better or more time efficient alternative to traditional methods and how this particular study supports that.

=> We thank the reviewer for his organization/formatting suggestions. Reviewer 2 also expressed, in his first comment, the same main concern and suggested to decide on a single focus for the MS in order to ease its reading. Thus, we replied to both comments below (in reviewer 2' section, first comment).

Furthermore, the direct benefit of the paleo-veg inference is not immediately obvious. It might be helpful to include examples of demographic scenarios that was ruled out by taking this first step. It is also not clear from the methods if the inference of the past distribution was done qualitatively or quantitatively. From the methods section, it seems that the present day distribution map was qualitatively matched with particular habitat and this habitat was used as a proxy to infer past distribution. If this were to be included, it would be important to run this in a more quantitative manner by conducting niche modeling (ex. MAXENT) to infer the present day distribution from occurrence points to then find the relevant bioclimatic variables and the project these to the distribution of the variables in the past. Before conducting such analyses, it would first be important to pin down the main focus and then add why it is relevant to include paleo-veg information.

=> As suggested by the reviewer, we complemented botanical models with species distribution modeling along with climatic reconstructions of past temporal windows. Climate scenarios are only available for the mid-Holocene (HCO) and Last Glacial Maximum (LGM) periods, and the Younger Dryas is thus not represented for these analyses. In addition, as it was mentioned in the introductive section of the former version of the MS, uncertainties related to current and past extrapolations of climate over large areas of Africa are often unknown (e.g. Rowell et al. 2016). Indeed, 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 long time periods involving several thousand years into the past (Chase and Meadows 2007; Dupont 2011). Therefore, we chose to keep relying on distribution projections based on paleo-vegetation to formalize our evolutionary scenarios.

Results from the climatic modeling effort do not fully match vegetation reconstructions presented in the main text. In brief, the predicted distribution during the HCO and the LGM are very similar to the current distribution. In particular, paleo-vegetation maps during the LGM show large expansion of semi-desert and desert biomes, which may have been favorable to the desert locust. However, paleo-climate modeling suggests for this period an extreme aridity in northern Africa and lowered temperatures in the entire continent, which may have actually been unfavorable to the species. Interestingly, this was already discussed in the sub-section 'On the influence of climatic cycles' of the Discussion based on literature. Thus, we now refer to these new results based on climate models in this sub-section (see lines 364-369 in the revised version) and included them as a supplemental material S2.

Lastly, the relevance of the 'evolution of phase polyphenism' in the discussion section escapes me. Although it is an interesting point, it seems out of place without being mentioned at all in the introduction. I

think this should either be removed or brought it earlier as a way to emphasize the relevance of this system which goes beyond the estimate of the divergence time.

=> We removed this section of the MS.

In the end, these concerns are mainly with the broad organization and relevance of the paper. This preprint has good potential contributions if the story is pinned down.

Figure 1. The colors need a legend and it would be beneficial to include a small title for each panel (A-F) in the figure itself.

=> This is now done.

Figure 2. The evolutionary events (c, b, sc) should be written out and the variable be in the parentheses so this would be more informative.

=> The figure 2 was completely re-designed following the recommendations of reviewers 1 and 3. We believe that this figure is now much more explicit and easy to understand even for scientists unfamiliar with this type of analysis.

Reviewer 2

This is an impressive manuscript that focuses on several important subjects in evolutionary biology, and it certainly should be / will be publishable in a major journal(s) following some revision. The authors address three overlapping subjects, each of which represents a major theme: 1) biology of Schistocerca gregaria, a species of considerable economic importance during its swarm phase ; 2) the question of population divergence in animal species that disperse very effectively over long distances ; 3) using molecular and quantitative methods to estimate the ancestry of populations that have diverged recently and cannot be studied with standard / classical phylogenetic approaches. As currently written, the manuscript integrates all three themes and is quite long, even without the supplementary material sections at the end. Thus, the authors should carefully weigh the positive and negative points of a single article of 'monograph' format versus 2 or 3 separate articles. And if they opt for a single monograph, the journal to which it would be submitted needs much consideration. Themes 2 and 3 are far too important to be 'buried' in a monograph devoted to Schistocerca biology, the worldwide pest status of this species notwithstanding.

=> We thank the reviewer for his organization/formatting suggestions. We reply to his comment and the first comment of reviewer 1 below in the same response. Indeed, both reviewers express the same main concern of deciding on a single focus in order to shorten the MS and ease the reading. This recommendation concerns mainly the Introduction section. First, we decided to de-emphasize the methodological aspects of the MS (application of ABC-RF) by removing the former 'New methods' section next to the Introduction. Note that this change also addresses the comment 5 of the reviewer 3. We also removed all data and text devoted to the quantitative evaluation of the gain in incorporating independent information in the mutational prior

setting. This shortens significantly the MS and makes the results section, tables and figures, much easier to follow. We rewrote, in the 'Discussion' section only, a section on the importance of our study as a first example of application of ABC-RF for estimating the parameters of interest under the best scenario and focused only on first-time novelties (i.e., computation of a posterior local error, assessment of the divergence time threshold above which posterior estimates are biased, etc) (see the sub-section 'Statistical advances by means of ABC Random Forest' at lines 382-408). Second, we removed the 'Discussion' sub-section on the implication of our results on the evolution of phase polyphenism in *Schistocerca gregaria*, by this way also addressing the comment 3 of the reviewer 1. Third, we re-phrased the end of the Introduction to connect cohesively challenges in estimating accurately the ancestry of populations that have diverged recently and the utility of some molecular (fast-evolving markers informed for their rates of mutations) and quantitative (ABC-RF) methods to address this challenge. Note that this introductive text was also leaned because the former lines 62-70 were removed: indeed, the quantitative modeling of the past climatic distributions of the desert locust is now addressed in the Discussion section (see response to the comment 2 of the reviewer 1).

Specific points:

Treatment of theme 2 would be improved by comparison with other species exhibiting similar ecologies and evolutionary histories. First, the origin of the New World Schistocerca species should be discussed, as it is argued (see papers by R.F. Chapman et al) that they are all descended from the Old World (African) Schistocerca gregaria : A single trans-Atlantic founder event to NE Brazil, followed by (adaptive) radiation. The difference between this case and that treated by the authors of the submitted manuscript is that the founder event in Brazil is a bit older (Pleistocene). Interestingly, none of the New World Schistocerca species exhibit a change from solitary to swarm phase. Second, the Monarch butterfly in North America exhibits a population structure that is roughly similar to Schistocerca gregaria in Africa: The major population is found in eastern North America, and a small population is found in on the West Coast. Both populations undergo an annual north-south migration each year, and admixture is believed to be minimal.

=> The reviewer states as a fundamental question that of population divergence in animal species that disperse very effectively over long distances. We agree with this suggestion and reckon it was not underlined enough. This theme became important in light of the unexpected result of a recent divergence time, which cannot be explained by climatically-induced shifts in vegetation/distribution range (i.e. the main hypothesis under disjoint species distributions so far). The role of dispersal in the disjoint distribution and genetic divergence of the desert locust subspecies was also supported by the result of a strong support for a bottleneck event in the nascent subspecies. This is the reason why this theme/question was not the anchor of the 'Introduction' section and the focus of our MS, and we still think it is not its place. Instead, we addressed further this theme in the 'Discussion' section, adding a dedicated sub-section '*On the role of dispersal on subspecific divergence*'. We included the reviewer recommendations to refer to the trans-Atlantic flights of the desert locust in the past history, and their implications in

terms of divergence and speciation (see lines 305-308 and lines 324-326 in the revised version). In addition, a quick literature review on the Monarch butterfly in North America revealed that there is an absence of genetic divergence between the two disjoint distributions of the Monarch subspecies (see Pfeiler et al. 2017 J Heredity; Brower and Jeansonne 2014 Ann Entomol Soc Am; Lyons et al. 2012 Mol Ecol). Consequently, we did not include this example.

Some of the points discussed in the supplementary material deserve integration in the main body of the manuscript. For example, the question raised in S4 on the possibility of Pleistocene colonization of southern Africa is too critical for relegation to an addendum, which is unlikely to be read. And regarding this possibility, one explanation is that such colonization had occurred but went extinct. This type of scenario has been proposed for the West Coast population of the Monarch butterly in North America.

=> This is now integrated in the main text.

Evaluation of the 8 different evolutionary scenarios for recent (late Holocene) divergence of Schistocerca gregaria is extremely difficult to follow in Figure 4 and the Tables. I recommend a much simpler presentation of the ABC-RF information in the Figure and Tables, with details placed in the supplementary materials.

=> Figure 4 and tables were simplified thanks to the removal from the manuscript of the side question of the methodological gain of including independent information on mutational prior setting. We do think that both figures and tables from the main document are now easy to follow, with most of methodological details provided in the Supplement document.

The writing is generally clear, particularly in the beginning of the manuscript, but there are placed where clarity could be / should be improved. I attach an annotated pdf with some suggestions.

=> We thank the reviewer for his editorial corrections that we incorporated within the new version of the MS. We appreciate his gesture and time.

Michael Greenfield

Reviewer 3

The manuscript « a young age of sub-specific divergence in the desert locust Schistocerca gregaria » by Chapuis et al. is a well performed RF-ABC analysis aiming at inferring the most likely divergence history of the species and estimating the associated demographic parameters.
The paper is well written and the author were careful in their analysis, providing justification for the choice of most of their demographic scenario, carefully assessing the robustness of model choice and parameter estimation, which the RF based analysis made easy. The used of the vegetation map at different time periods helps a lot the reader who is unfamiliar with the system to draw expectation regarding the possible scenario of divergence. Generally, more study of these king are needed in nonmodel species prior to make adaptive hypotheses.

For now I only have few remarks related to the choice of scenario that may perhaps be improved:

i) The author did not allow for bidirectional secondary contact, but only an asymmetric secondary contact (which is modeled like a single discrete admixture pulse in the present study). Although the author provided some verbal argument in the discussion and methods, I think that a formal test of models with bidirectional secondary contact vs unidirectional asymmetric contact might be relevant. Once the best models among the two will be chosen, the author could then compare it to models without contact. I think it would make a more rigorous example of how to test for this process without relying too much on priors. Moreover, the conclusions drawn here should not be affected given the nearly absent lack of support for admixture.

=> We followed the reviewer's suggestion and included in our model choice analyses scenarios including a bidirectional admixture event, in addition to the former scenarios including unidirectional admixture events. Thus, we now analyze a set of 12 scenarios instead of 8 scenarios, which are graphically depicted in Figure 2. Results are presented in Table 2 and details in the Supplemental document. As predicted by the reviewer, these 4 new scenarios including a bidirectional admixture event were not supported by the ABC-RF model choice analysis.

ii) Related to this, I find the use of the term secondary contact a little confusing: it seems to me that the author simulated a single discrete admixture pulse at a particular point in time and not really a secondary contact that would last for several generation of ongoing gene-flow (see. e.g. Roux et al. 2013 and 2016 for appropriate definitions). It would have been nice to test for a model with true secondary contacts rather than single discrete admixture pulse and I would be curious to see if the results remains the same or not. (I would expect the model choice not to be affected, but it would be more rigorous to test explicitly for it).

=> First of all, we have reconsidered our wording by removing any mention to a "secondary contact" and referring to a "discrete genetic admixture event". The reviewer suggests including scenarios with gene-flow (scenarios with secondary contact that would last for several generations of ongoing gene-flow – as mentioned in this comment - and scenarios with a split with initial ongoing gene-flow - as mentioned below). First, it is worth noting however that a secondary contact with several generations of ongoing gene-flow can be advantageously modeled by a punctual admixture event (if the number of generations with ongoing gene-flow is not too large) as we did here with DIYABC. Second, we unfortunately do not know a statistical framework other than DIYABC that takes into account mutational features that are realistic for microsatellite loci of grasshoppers. DIYABC indeed allows considering complex and specific mutational models such as the symmetric generalized stepwise mutation model (GSM) (Zhivotovsky et al. 1997), insertions or deletions in the flanking regions of the microsatellite sequence, and allow considering altogether several sets of markers (here, genomic- *versus* transcriptomic microsatellites characterized by distinct mutational features), etc. We believe that these features have more impact on the global genetic diversity within populations and differentiation between populations than modeling past migration that last several generations instead of a single-generation event of admixture.

iii) When looking at the evolution of suitable areas on the maps, I was left with the impression that the species range has been continuous (on longer time scale) and may have then been progressively split into two units over a very recent time period (linked to climate variation, as explained by the authors). Under this scenario, there would not be a split

without gene-flow, rather I think this could be approximated by a model of split with initial ongoing gene-flow (bi-directional and with the possibility for asymmetry; also referred to as ancient migration models). Although it would even be better if the authors could test for a model of progressive decrease in gene flow following split time, until the model converged to a model of strict isolation. I think this could be more realistic than the "single long distance migration event of a small fraction of the ancestral population" which is modeled through a bottleneck. Excluding model of initial gene-flow would enriched the discussion line 365 – 382.

=> Please see our response to the previous comment. In addition, we must stress out that considering "a single long distance migration event of a small fraction of the ancestral population" is quite realistic in this atypical species, characterized by impressive high dispersal ability and some spectacular long-distance migration events during outbreaks. We realized that this was not obvious for readers unfamiliar with locusts. We therefore now detail further this biological peculiarity in the discussion of the revised version of our ms.

Aside from these general remarks, most of my comments are suggestions to improve the manuscripts.

Other comments:

During the reading, I've been wondering several times how many individuals were used, what were the levels of within subspecies genetic diversity, and what were levels of genetic differentiation. I also wondered if their was significant genetic structure within subspecies. This information is only indirectly provided in the methods section through reference to previously published articles. A short paragraph on these topics would be relevant at the beginning of the results.

=> The supplemental table that describes values for summary statistics (including He and Fst) is now in the main document referred as to table 1. Following the reviewer recommendation, we also added a paragraph at the beginning of the 'Results' section (lines 124-132 of the revised version) describing the main descriptive results: "Table 1 shows the values of the summary statistics obtained from the observed population dataset consisting in two unstructured pooled samples of the subspecies *S. g. gregaria* and *S. g. flaviventris*. A total of 170 individuals (*i.e.*, 80 and 90 individuals for *S. g. gregaria* and *S. g. flaviventris*, respectively) were genotyped at 23 microsatellite markers derived from either genomic DNA (14 loci) or messenger RNA (9 loci) resources (hereafter referred to as untranscribed and transcribed microsatellite markers, respectively). The level of differentiation between the two subspecies (as measured by the parameter F_{ST}) was 0.04 and 0.12 for untranscribed and transcribed microsatellite markers, respectively. The level of genetic diversity was higher within *S. g. gregaria* (+7 and 14% for the mean number of alleles and expected heterozygosity, respectively)."

Also, the discussion was interesting, maybe the author could provide us with some more discussion regarding the advantage and limits of using RF-ABC (this is mostly in the Supp Mat S1). While microsatellite are less and less used in evolutionary biology, being progressively replaced by SNPs, maybe the authors should also highlight the fact that such approach is relevant also for SNP/genome-based inference of evolutionary history?

=> We added in the Discussion section, a sub-section entitled "Statistical advances by means of ABC Random Forest" which discussed main advantages of ABC-RF for our specific question of the estimation of divergence time

between subspecies of the desert locust. This sub-section provides more or less similar information to the former "New methods" section that was removed (cf. response to the comment 1 of the reviewer 2), and now notably mention the relevance of our statistical framework for the treatments of massive simulation data, including for inferences using single nucleotide polymorphisms (i.e., SNPs) obtained from new generation sequencing technologies.

Line 144-145 and 148-155 : One of the first empirical study using RF-ABC was Rougemont et al. (2016). Applying the RF algorithm to the problem of model choice, they find several advantages to the RF algorithm, but with similar difficulties in discriminating complex scenario compared to the neural-network based ABC procedure.

=> We added the reference provided by the reviewer (now lines 391 and 574).

Minor comments:
Line 156 – 165 on model grouping: I am not sure this is entirely new. For instance, Roux et al. (2016) performed grouping of model with gene flow (IM,SC) against models without gene flow (SI, AM). Similarly Leroy et al. (2017) compared groups of IM, SC, AM, and SI models and then statistically compared the best alternative version of the previous "best" model. I guess this procedure is already widely used in the ABC literature.

=> First, we added the two references suggested by the reviewer. In addition, it is worth noting that the new version of the MS does not emphasize anymore model grouping as a new addition or a rare case of application in our study as it was suggested before. Rather, we now refer to it as a recent approach.

Fig 2: Maybe provide separate figures, to ease the understanding for people unfamiliar with these sort of analysis? (or in Supp Mat?).

=> The figure 2 was completely re-designed following the recommendations of reviewers 1 and 3. We believe that this figure is now much more explicit and easy to understand even for scientists unfamiliar with this type of analysis.

Line 218- 221: maybe in the text provide the number of the scenario that are compared against each other, it would also eased the understanding of which scenarios are compared without the need to go reading the table 1 directly.
i) was 1+ 3 +5 +7 vs 2 + 4 +6 +8
ii) was 1+2+5+6 vs 3,4,7,8, and
iii) was 1+2+3+4 vs 5+6+7+8.
It becomes clear only after reading the methods.
Although I would present them in the same order of the columns of the table in Fig2, (this would also follow the order of events in time (contraction, then bottleneck, then gene-flow)).

=> We renamed each scenario by meaningful acronyms made up with initials of each event in their order in time: S for split, C for contraction, B for bottleneck, A for admixture, etc (see Figure 2). In addition, for each group model analysis, we referred to all IDs of scenarios of each group in the text and in the tables.

Line 280 : Reference for « average of three generations per years »? (it is only provided in the Methods).

=> The reference Roffey and Magor 2003 has been provided each time the average of 3 generations per year was mentioned.

Fig 1B is not mentioned (should be with Fig1A or remove?)

=> Fig 1b is mentioned line 297.

Methods:
Prior choice: It is still not clear to me how Tca was chosen. I think diyABC only allows single admixture event, however, how likely is this? It may be more relevant to allow for a broader time period.

=> We assume that the reviewer actually refers to the time of admixture (and not of ancestral contraction). There is indeed no time period since it corresponds to a point time (a single generation). Cf. response to comment 2 of the reviewer.

Also, the lower bound of the prior on admixture parameter seem to be very narrow. Why not letting it varying to lower value (i.e. closer to 0). As in models with gene flow?

=> We used a lower value of 0.05 for the prior of the admixture rate in order to have clear distinction between i) scenarios with an admixture event and ii) scenarios without an admixture event. Indeed, if we had included the value of 0 or very low values close by, then scenarios referred as to with an admixture event would actually include the possibility of an absence of admixture event.

Line 548 – 553 : it would be nice to have an overview of the dataset, perhaps with a few summary statistics related to the pooled samples (e.g. genetic diversity within pool of subspecies, Fst between subspecies,etc).

=> We replied (positively) to this comment in our response to the comment ii) of the reviewer 3.