Dear Dr. Galtier,

Please find enclosed a revised version of the manuscript entitled "*Efficient k-mer based curation of raw sequence data: application in Drosophila suzukii*". I sincerely appreciate the efforts you and the two reviewers (Marie Cariou and Denis Baurain) have made in evaluating this paper and I wish to thank you for the positive feedback and very helpful and constructive comments. I read and considered all comments with great attention and tried to clarify as much as possible all the issues raised by the reviewers, modifying some parts of the manuscript where necessary.

Below, I provide a detailed point-by-point responses (each concern is in green italics and the corresponding response follows in black with text from the manuscript highlighted in orange). I also provide a marked-up copy of the manuscript, with all changes since the original submission highlighted in blue (and all deleted portions highlighted in red).

I hope you and the reviewers are satisfied with this revised version of the manuscript and look forward to your final decision.

Yours, sincerely

Mathieu Gautier

Recommander Comments to Author:

EC: *I have an additional comment, also briefly mentioned by one reviewer. Besides contamination, there might be biological reasons why a given sample contains sequence reads assigned to a different species, namely hybridization and gene flow. How is the newly introduced method expected to behave when reproductive isolation between the analyzed species is incomplete? In particular, is there a risk that the method partially erases the signal of gene flow, if actually present? I think these questions could deserve a specific discussion as gene flow is quite common in nature and the focus of many population genomic studies.*

I completely agree with your comment and think this is an important aspect. In a sense, this amounts to interpreting "contamination" as gene flow between closely related species (if any). Note that gene flow between populations belonging to the same species may have virtually no effect on the results, since all "migrant" sequences of a sample can actually be defined here with respect to the reference genomes used to build the species-discriminating kmers (i.e., such a migrant sequence can only be left unassigned if it contains a variable position with all dictionary kmers that map to the same genomic position). Conversely, the analysis of *D. suzukii* and *D. subpulchrella* individuals in the present study allows to illustrate, to some extent, both the possibilities of the approaches to assess the amount of interspecific gene flow (on a genome-wide level) and their limitations (mostly related to incomplete representation of reference assemblies and incomplete lineage sorting). I have also tried to improve the discussion of this issue (L780-L826, reproduced below) in relation to the one comment from reviewer 2 (R2C6):

"Although two different D. suzukii genome assemblies were used to build the species-discriminating k–mer dictionary, all (pure) D. suzukii Ind-Seq and Pool-Seq samples showed a small but non-negligible fraction of their sequences (from 1.14% to 2.78%) assigned to D. subpulchrella by the most stringent criterion. Because i) the D. suzukii reference genome assemblies were derived from isofemale lines established from individuals sampled in the North American (5) and European (23) invaded areas; and ii) D. subpulchrella has not been yet described (to our knowledge) outside the Asian native range of D. suzukii; it is highly unlikely that this pattern is the result of pervasive gene flow between the two species,but rather can be explained by the close phylogenetic relationship between the two species. Indeed, some D. subpulchrella-discriminating k–mers may actually map to orthologous regions not represented in

the D. suzukii reference assemblies and/or capture shared genetic variation between the two species due to incomplete lineage sorting (ILS). Including more reference assemblies (e.g., from different strains) for each target species may be considered as a valuable strategy to improve both the sensitivity (by 'positive filtering' of the discriminating k–mers that capture intraspecific genetic variation) and specificity (by 'negative filtering' of the incompletely sorted k–mers). The optimal number of representative assemblies is thus likely to both depend on the relatedness of the selected target species and for each target species on their genetic diversity. Alternatively, the misassigned short read sequences found in the analyzed samples can be included in the construction of the k–mer dictionary, assuming that the considered samples are not contaminated and are 'pure' representatives of the corresponding target species. Such refined target dictionaries may even further allow providing (rough) estimates of the genome-wide level of interspecific gene flow, pr at least the identification of highly admixed individuals. Hence, in the sample of identified D. subpulchrella individuals, if about 2% of the short-read sequences were assigned to D. suzukii (in a similar but reversed pattern as observed for D. suzukii individuals), one (presumably) D. subpulchrella individual had nearly 10% of its sequences assigned to D. suzukii. The status of this sample may be of special interest for fur ther study as it could represent a previously unreported case supporting some recent (i.e., only a few generations back) admixture events between D. suzukii and D. subpulchrella. As discussed by Lalyer et al. (17), if no such recent events have been reported to date, several studies suggest that hybridization has occurred between these two sister species (7)"

Reviewer 1 (Marie Cariou) comments:

This article describes a procedure used to control publicly available sequence data of Drosophila Suzukii for mislabeling and contaminations. The procedure relies on the construction of discriminatory k-mers dictionaries to compare with k-mers present in each dataset. It was performed using the software CLARK, which was created for the taxonomic classification of metagenomic sequences.The procedure efficiently identified 16 mislabeled samples among the 236 individual D. suzukii sequence data and 2 contaminated samples among 22 pool-seq sequence data. I found this approach really interesting and well presented in the manuscript. The author 1) advocates for the routine inclusion of such k-mer based quality check in data quality assessment practices. 2) presents a curated dataset of D. suzukii public sequences, useful for further population genomics studies.

I would like to express my sincere thanks for the positive feedback and the constructive comments and suggestions.

R1C1: *I may have a question regarding the idea that such check should be included in standard quality assessment. In this analysis, the author relied on extensive and curated assemblies genomic data (« high quality assemblies for several dozen of drosophilid genomes »). Here, these numerous genomes also allow to evaluate the "global" efficiency of the approach, but I wonder to what extend such approach could be easily generalized for any species. What would be the author guidelines to perform such check for any genomic dataset ? To say it differently, what would be the minimal external data (in terms of both quality of assembly and taxonomic coverage) required to construct a meaningful dictionary ?*

This is an important question, but it is difficult to provide a general answer. The purpose of this study was to propose and evaluate an approach to assess the level of contamination in a sequencing data set in one species (i.e. D. suzukii), but also applicable to other species for which i) contamination is an issue (in particular due to difficult morphological identification); ii) putative confounding (or contaminating) species were known in advance; and iii) reference genome assemblies are available for them (at least for closely related species). From a practical point of view, the resources provided here (e.g., kmer dictionaries and scripts) are directly applicable to other highly studied

drosophilid species, especially in the field of population genomics (e.g., D. melanagaster or D. simulans), and the above requirements may also be met in other organisms of interest, given the growing number of assembled genomes. For such organisms, the proposed approach can be easily implemented and evaluated in a similar manner as described here. In particular, inspection of the assignment results for a sample of sequence data may provide insight into the performance of a newly built dictionary (e.g., percentage of unassigned and misassigned sequences).

R1C2: *L47-51 the repetition of « the resulting combined datasets » might be avoided.*

The sentence was rephrased (L45-L50):

"However, this increased availability of data comes at the cost of increased heterogeneity in the resulting combined dataset. For example, data sets may combine different sequencing library preparation protocols or technologies that are rapidly evolving with variable sequence quality or coverage."

R1C3: *L237. I think « 305 » should be « 301 », to match the sum listed in the paragraph (43+236 +22), which is also coherent with the number of lines in table S2 and S3 and to the value L331.*

This is correct (thank you very much for reporting this error). The sentence was modified accordingly (L244).

R1C4: *Fig 2B . Are the colors corresponding to target and other (light and dark blue) reversed? I expected the more dispersed and almost bimodal distribution (dark blue), with higher percentage of sequences with no match to correspond to the « other species ».*

This is correct (thank you for reporting this error). The panel in Figure 2B (and the legend) has been changed accordingly.

R1C5: *L314-316 Does this option -s 2 have a strong impact on computation time and fraction of sequences with no matching k-mers?*

For Clarkl, the -s option is actually ignored (i.e., all the kmer dictionary is loaded). Note that the -s 2 is useless when running *clarkl* in the script provided in the Data INRAE repository (*"run_fastp_clarkl_clark_and_summarize_results.sh", L31*):

For Clark, I actually followed the manual's recommendation, which states that:

“The higher this factor [i.e., s option] is, the lower the RAM usage is. The higher this factor is, the higher the classification speed/precision is. However, our experiments show that the sensitivity can be quickly degraded, especially for values higher than 3. In the default mode, this factor is set to 2 because it represents a good trade-off between speed, accuracy and RAM usage.”

I tried to clarify this by modifying the sentence (L318-L323), which now reads:

“In practice, CLARK was run with option -s 2 to load only half of the species-discriminating kmers in the target dictionary, following the manual recommendation indicating that this value `represents a good trade-off between speed, accuracy and RAM usage'. Both CLARK and CLARKL were run with the options -n 1 (i.e., on a single thread) and -m 0 (to compute the confidence score).”

R1C6: *l410 « may thus [be] display »*

This was corrected (L419).

R1C7: *I was able to retrieve the databases, cleaned assemblies and scripts from the Data INRAE repository but I did not attempted to run clark myself. However, they look well formatted and organized. In “run_fastp_clarkl_clark_and_summarize_results.sh”: l20: “cleanning seqeunce” → “cleaning sequences”*

This is correct (thank you for reporting these typos). However, since this is a typo in a comment (i.e., with no effect on the implementation), I decided not to change it in the INRAE data repository. In fact, I would have to upload a new updated archive (which should include the >15Go k-mer dictionaries), while the current one would still be stored due to the repository policy.

Reviewer 2 (Denis Baurain) comments:

In this empirical study on Drosophila whole genome samples, Gautier evaluates the use of the metagenomic classifier CLARK to analyse the contamination structure of short-read datasets by closely related species of the advertised organism and its microbial commensals. The author shows that this approach is both accurate and computationally efficient and, as a byproduct, releases a curated set of >60 population samples of D. suzuki that should be useful in future population genetic studies.

Generally speaking, I enjoyed reviewing this manuscript. The study is well-designed, the text is clear and pleasant to read and the figures are easy to understand. Moreover, the work is extremely well-documented, with most of the study details provided in Supplementary Tables, while data and scripts are made available in a public repository (please note that I did not download the latter to check the actual content). Consequently, my comments are minor and aimed at further clarifying the text when needed. However, I noticed a number of small errors in the reporting of the results. As some of them are quite confusing, I insist that they should be addressed in the revision of the manuscript.

I would like to thank you very much for the positive feedback, the constructive suggestions and the very careful reading of the manuscript.

Scientific questions

R2C1: *lines 173-175: I don't understand if the 101 assemblies of the paper (which are taxonomically diverse) are part of the 129 assemblies on the NCBI portal and, if not, why the former were not preferred to the latter? Was there some global quality assessment of all available assemblies (in the NCBI and elsewhere) prior to taking these decisions?*

I am not sure I fully understand this comment. As detailed in Table 1, the target dictionaries were constructed from 32 assemblies representing 29 drosphilid species and 13 assemblies representing 12 common drosophilid commensal species. In total, 45 assemblies (not 101) were used and downloaded from the NCBI (n=43), ENA (n=1) and Dryad (n=1) repositories. The selection criteria are summarized at the beginning of the M&M section (L153-L172), which has also been clarified and modified per R2C12 comment below:

"Of the 136 reference genome assemblies available for species belonging to the genus Drosophila in the NCBI repository (https://www.ncbi.nlm.nih.gov/datasets/genomes/ accessed in February 2022), 29

were retained based on assembly quality criteria such as contiguity (evaluated with contig N50) and completeness (using BUSCO scores, 19); but also and mostly based on phylogenetic criteria (Figure 1). Our goal was to obtain a good representation of species closely related to D. suzukii, focusing on those belonging to the two subgenera Sophophora and Drosophila that are not unambiguously resolved (see Discussion). For subgroups or groups represented by multiple species (among those with good quality assemblies available), only one target species was selected, favoring the most cosmopolitan or temperate species (12), except for the species most closely related to and likely to be confounded with D. suzukii (e.g., D. subpulchrella and D. biarmipes). To further improve the representation of D. suzukii in the k–mer dictionary, the draft assembly of Ometto et al. (23) was also downloaded from the ENA repository (https://www.ebi.ac.uk/ena/browser/home). "

***R2C2:** lines 197-198 ("widespread lateral gene transfer from Wolbachia"): this raises the issue of whether such transfers should be considered as contamination in this species... and in other species! On a side note, had the species datasets completely devoid of Wolbachia sequences been aggressively curated before public release?*

I agree with the reviewer that this issue is of particular interest, and I was actually hesitant to filter out Wolbachia sequences, especially for D. ananassae. However, I find it more informative and robust (for species assignment of the sample) to annotate Wolbachia sequences separately in the target repositories. This indeed allows i) *"for the rapid identification of Wolbachia-infected samples, which may be of interest for a first rapid screening of drosophilids samples since the set of Wolbachia-discriminating k–mers was built by combining D. simulans and D. melanogaster Wolbachia assemblies"* (as mentioned in the main text L835-L840 and shown in Figure 4); ii) to avoid (mis)assigning sequences from Wolbachia-infected samples (e.g., from D. suzukii) to the wrong species (e.g., D. ananassae or another assembly if not filtered for Wolbachia) which would then overestimate the contamination level. This is indeed what happened in preliminary studies (not shown here) against unfiltered assemblies when screening *D. melanagaster* individuals. Those infected with Wolbachia had a substantial proportion of sequences assigned to *D. ananassae*.

Regarding the side note, I would tend to think that for most (if not all) samples, the submitters have not made any special effort to curate their data for Wolbachia sequences. Samples without Wolbachia sequences may not be infested (e.g., laboratory strains are often treated with antibiotics to remove Wolbachia).

R2C3: *lines 209 ("after filtering out contaminating sequences"): if I understand correctly, Kraken2 was used on whole contigs, not pseudo-reads spliced out of contigs. Then does "filtering out" mean removing these whole contigs (i.e., up to 1.4 Mb in one case)? Was it not possible to preserve more information by only masking the foreign regions of large contigs (assuming they might be chimeric)?*

In fact, here I preferred to remove all contigs from the assembly if they were judged to be contaminated based on the k-mer-based Kraken2 assignment, as this seems to me to be the safer option. Most of the assemblies were of very good quality (see response to R2C1 comment above), making it unlikely that high levels of mosaicism exist in any single contig, even the larger one. In addition, as mentioned in the main text (L196-L204; and detailed in Table S1), the *"contaminating sequences were mostly short, ranging from 110 bp to 1,478,327 bp (median size of 1,522 bp), totaling only 102.7 Mb (i.e. 1.72% of all sequences). It should be noted that Wolbachia-related sequences represented only 6,173,139 bp of the contaminating sequences (6.01%), with the major contributor being the D. ananassae assembly (6,078,940 bp), which may be explained by the widespread lateral gene transfer from Wolbachia described in this species "*. Conversely, most contigs of the assemblies remained correctly assigned to drosophilid species (generally with >95% of assigned k-mers).

Masking sequences would have posed several problems, as the actual contaminant species could generally not be inferred with certainty, although Kraken-2 may provide some insight into a closely related species (e.g., microbial contaminants). As a result, an assembly for the corresponding contaminant species may not be available for masking.

Overall, it should be emphasized that in general only a small fraction of the assemblies were removed (Table 1 and Table S1) without any negative impact on their completeness as assessed by the BUSCO score (Table 1).

R2C4: *lines 213-216: it is mentioned briefly in the Discussion (lines 766-769 and 793-795), but I wonder if "pangenomes" (rather than single strains) would have provided more sensitivity for pathogen and commensal screening. This is an important issue from a practical point of view.*

For the purposes of this study, I was mainly interested in drosophilid assignment, and the inclusion of common pathogens and commensal genomes was mainly illustrative to provide an indirect insight into the original

sample quality (e.g., degraded samples can be expected to contain a substantial amount of sequences assigned to *S. cerevisiae*). In this respect, Wolbachia can be considered an exception, as the identification of infested samples may be of particular interest.
More generally, I expect that a substantial proportion of unassigned sequences may originate from unrepresented microbial species or, I agree, from strains more distantly related to those chosen to build the dictionaries. Since the contamination rates here were estimated as a proportion of assigned reads, the use of pangenomes for microbial species may yield only marginal gains.
If one wants to characterize the sample microbiota, I would rather recommend focusing on unassigned sequences and analyzing them with a dedicated target dictionary or preferably by querying large databases using tools like *Kraken2* that take phylogenetic information into account (see also the answer to comment R2C22 below).

R2C5: *lines 243-244 ("including data on 12 of the 29 target species"): is it on purpose that 17 of the target species are not tested by the samples?*
This is correct. For a more reliable assessment (based on 41 drosophilid target and non-target species), I deliberately chose to focus on WGS data from laboratory strains because they are less likely to be contaminated (and more likely to be representative of the species). The 17 target species not represented are those for which I could not find public WGS data from laboratory strains.

R2C6: *in Table 1: I know that it is suggested in CLARK paper, but I wonder if the representation of some species by multiple assemblies is really harmless in terms of assignment statistics. Similarly, are we sure that the results are not biased in some way when some species are more distant and thus would provide a lot more specific k-mers than groups of highly related species? I did not find a discussion of this issue in CLARK paper, but for the present purposes, knowing the answer would be important. If so, it might be introduced at lines 298-300. Also, a related bit of discussion appears at lines 665-677.*
From my empirical experience, the inclusion of multiple assemblies to represent a given species was beneficial, particularly for the evaluation of the *D. suzukii* samples analyzed here (and *D. melanogaster* pools, not shown in the present manuscript), when compared to preliminary analyses

performed using a target dictionary built with only one assembly for the *D. suzukii*, *D. subpulchrella*, *D. simulans*, and *Wolbachia*. For example, as discussed in L761-L770 of the first version of the manuscript: "*...some D. subpulchrella-discriminating kmers may actually map to orthologous regions not represented in the D. suzukii reference assemblies and/or capture shared genetic variation between the two species (due to incomplete lineage sorting). In both cases, refining the kmer dictionary by including additional reference assemblies for each species, or alternatively the misassigned short read sequences found in the analyzed samples (then assumed to be pure), may help improve sensitivity.*".

So I would tend to think that the higher the number of assemblies per target species, the better. This may indeed (ideally) allow one to build a *kmer* dictionary that captures within-species variation while filtering out *kmers* that are non-specific (due to incomplete lineage sorting) but may appear to be discriminative if too few assemblies are used to represent each species. The example of *D. suzukii* and *D. subpulchrella* is illustrative in the sense that I would expect even more resolution (e.g. in Figure 4) if more than two assemblies (assuming they are not contaminated) had been used to build the kmer dictionary. I have changed the corresponding part in the Discussion to emphasize/clarify this (L793-L811, with some highlighted in bold below; see also response to recommender comment EC):

"*Indeed, some D. subpulchrella-discriminating k–mers may actually map to orthologous regions not represented in the D. suzukii reference assemblies and/or capture shared genetic variation between the two species due to incomplete lineage sorting (ILS).* ***Including more reference assemblies (e.g., from different strains) for each target species may be considered as a valuable strategy to improve both the sensitivity (by 'positive filtering' of the discriminating k–mers that capture intraspecific genetic variation) and specificity (by 'negative filtering' of the incompletely sorted k–mers). The optimal number of representative assemblies is thus likely to both depend on the relatedness of the selected target species and for each target species on their genetic diversity.*** *Alternatively, the misassigned short read sequences found in the analyzed samples can be included in the construction of the k–mer dictionary, assuming that the considered samples are not contaminated and are 'pure' representatives of the corresponding target species*."

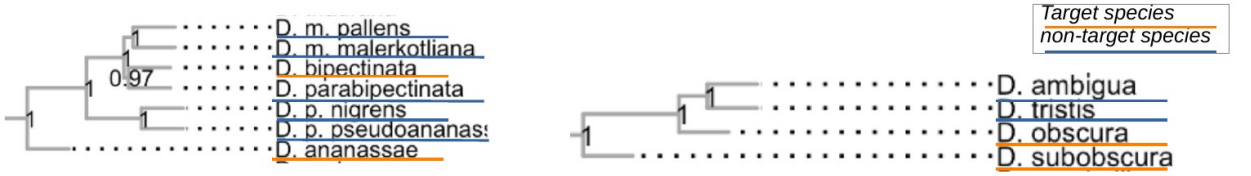
Regarding the second part of the comment, it is indeed important to note (as specifically mentioned in the Discussion) that the target dictionary was constructed with the scanning of a *D. suzukii* sample in mind (and to some extent also *D. melanogaster*), and I agree that assignment of samples from less well represented species must be done with caution, as illustrated by the analysis of the 30 (reference) samples from species not represented in the target dictionary (see e.g. Figures 2 and S4). In practice, the overall percentage of assigned sequences (at the chosen criteria) may be a good indicator of the relevance of the *kmer* dictionary for species assignment of a sample, motivating a re-analysis with a more appropriate custom *kmer* dictionary (or further analyses with other tools). This was clarified in the discussion (L669-L685):

"*However, as illustrated by the assignment of sequences from species closely related to one of the represented groups or subgroups (e.g., ananassae or obscura) but not included in the construction of the k-mer dictionary, species-level assignment provided consistent results about their origin.* ***Yet, assignment of samples to species belonging to groups or subgroups less well represented by the target species should be interpreted with caution, especially when the observed proportion of non-matching k-mers is high (Figure S4). In such cases, analysis with a newly built k-mer dictionary including more closely related species may be valuable. Indeed, our main focus was on the evaluation of D. suzukii samples.*** *We therefore chose to deliberately overrepresent the suzukii subgroup in the k-mer dictionary construction by including the high quality genome assemblies available for D. suzukii, D. subpulchrella, and D. biarmipes.*"

R2C7: *lines 478-494: for the 16 species not represented in the target dictionaries but still assigned to a single target species, 5 are assigned to D. bipectinata (and none to D. ananassae) and 2 to D. obscura (and none to D. subobscura). Is there a phylogenetic reason for this?*

As mentioned above in response to R2C5 comment, the selection of public WGS data for the 30 non-target species (including the 16 with >95% of assigned sequences assigned to a single target species) was biased towards laboratory strains (from Kim et al., 2021). Part of the clustering pattern could thus be explained by such an ascertainment bias.

However, a closer inspection of the phylogeny inferred by Kim et al. (2021) [their Figure 5 based on 250 randomly selected BUSCO genes and using RaxML] is consistent with the pattern observed for the analyzed species belonging to the *ananassae* (n=5 non-target species with *D. bipectina* and *D. ananassae* as target species) and *obscura* (n=2 non target-species with *D. obscura* and *D. subobscura* as target species), as shown in the subtrees below:



adapted from Figure 5 Kim et al. (2021, https://doi.org/10.7554/eLife.66405)

*** Clarity issues**

R2C8: *lines 16-31 in the Abstract are a copy-paste from the end of the Introduction (lines 137-152); maybe rephrase some sentences?*

I agree (thanks for pointing this out). The end of the introduction has been rewritten (and simplified) in order to present the plan of the manuscript in a more focused way (L136-L150):

"To assess contamination in publicly available D. suzukii raw sequencing data, we developed and evaluated a fast and efficient approach based on k–mer-based methods implemented in the software CLARK (24). We first build dictionaries of species-discriminating k–mers from the curated assemblies of 29 target drosophila species and 12 common drosophila pathogens and commensals. WGS data for individual samples representative of both the target and other drosophilid species were then analyzed to evaluate the performance of the proposed approaches, both in terms of run time and accuracy of sequence assignment. Finally, we analyzed publicly available WGS data for the aforementioned 236 Ind-Seq (17) and 32 Pool-Seq (21) samples of the invasive species Drosophila suzukii, allowing us to identify unambiguously contaminated samples"

R2C9: *lines 57-58 ("the characteristics mentioned above have mostly remained"): I don't understand the idea here; please rephrase.*

I agree that the sentence was not clear. I have rewritten it as (L56-L58):

“Nevertheless, such technical characteristics [referring to Pool-Seq or Ind-Seq/ or coverage variation] *can be taken into account in downstream analyses if an appropriate statistical framework is used.”*

R2C10: *lines 87-88 ("but they are not well suited for the analysis of large amounts of samples"): please add a hint about why it is so.*

I have added the following precision L85-L86 (quoting Cornet and Baurain, 2022): *“but they are not well suited for the analysis of large amounts of samples as they require a case-by-case inspection of the results (8)”*.

I must admit that I am not familiar with these methods, which may not even be appropriate for distinguishing closely related species. It seems to me that there are more designed to identify microbial contamination in genome assemblies (but I may be wrong).

R2C11: *line 91 ("the genomes of the putative contaminant species"): this is a bit restrictive (only negative filtering), especially considering that the current study use both positive and negative filtering; please add a bit of nuance. BTW, positive filtering is discussed at lines 700-704.*

This was indeed clarified (L86-L93): “*Reference-based methods consist of aligning sequences to a set of tagged sequences representative of all or part (e.g., genes) of the genomes of candidate species. In practice, this allows for example negative or positive filtering (i.e., removal of contaminating sequences or identification of sequences from some species of interest) of sequencing data (8)*”.

R2C12: *lines 159-160: please explain the logic behind the phylogenetic breadth of the reference sampling to help others (e.g., why also the subgenus Drosophila).*

This part has been reworded and clarified (L157-L170):

“...*29 were retained based on assembly quality criteria such as contiguity (evaluated with contig N50) and completeness (using BUSCO scores, 19); but also and mostly based on phylogenetic criteria (Figure 1). Our goal was to obtain a good representation of species closely related to D. suzukii, focusing on those belonging to the two subgenera Sophophora and Drosophila that are not unambiguously resolved (see Discussion). For subgroups or groups represented by multiple species (among those with good quality assemblies available), only one target species was selected, favoring the most cosmopolitan or temperate species (12), except for the species most closely related to and likely to be confounded with D. suzukii (e.g., D. subpulchrella and D. biarmipes).* ”

R2C13: *lines 161-163 ("for subgroups or groups represented by multiple assemblies, only one species was selected"): ambiguous phrasing: multiple assemblies of the same species or multiple assemblies of different species? In my view, one assembly does not always equate one species.*

This was clarified (see response to R2C12 comment above).

R2C14: *line 186 ("including Wolbachia endosymbionts"): ambiguous wording; is it an exception or a precision?*

This has been rephrased to avoid confusion (L188-L193): "*A contig or scaffold sequence was considered contaminating if it was assigned to a taxonomic identifier unrelated to any drosophilid species. Note that contigs assigned to Wolbachia endosymbionts were also flagged as contaminating, as we chose to consider Wolbachia specifically here (see below)*"

R2C15: *lines 230-232 ("Building the k-mer dictionary took 2h46min"): such timings are quite useless without some idea of the CPU architecture; please specify it.*

I agree and have added the following specification: *"Building the k-mer dictionary (on a single thread of a cluster node equipped with a processor Intel® Xeon® CPU E5-2683 v4 @2.10GHz) took 2h46m..."* (L236-L237) and also in the legend of Table 2.

R2C16: *line 307 (and around): in CLARK paper, the confidence score is only computed based on the two top-matching sequences, not all; please check.*

This is correct. Thank you for pointing out this incorrect definition of Clark's confidence score in the text. I have changed it accordingly in the M&M section (L307-L316; I left here latex formula for clairity):

"*More specifically, for a given sequence, let t_1 and t_2 be the target species with the highest and second highest number $k_q(t_1)$ and $k_q(t_2)\leq k_q(t_1)$ of matching \kmers, respectively. If no species-discriminating \kmer was found in the sequence (i.e., $k_q(t)=0$ for all target species t), the sequence is unassigned. If $k_q(t_1)>0$, the sequence is assigned to species t_1 with a `confidence score' defined as $c_q(t_1)=\frac{k_q(t_1)}{k_q(t_1)+k_q(t_2)}$, noting that*

$c_q(t_1)=1$ if all the matching kmers are assigned to t_1 (i.e., $k_q(t)=0$ for all $t\neq t_1$).". The Figure 2 legend was also modified.

R2C17: *lines 349-350 ("sequence length was representative of typical short read datasets"); please state that datasets here include a variable mixture of merged and unmerged reads (if I understand correctly).*

I agree. I added the following precision in parenthesis (L360):

"*The sequence length was representative of typical short read datasets, with a sample mean length* ***(after merging overlapping reads)*** *ranging from 92.7 bp to 287 bp* "

R2C18: *line 379 ("averaging 24.5%"): why to report a mean here and everywhere else median values? Is there a specific reason?*

I agree. This was corrected and the median value is now given for consistency (L387-L389):

"*The percentage of sequences with no matching kmer (i.e., not assignable) was similar between \clark (ranging from 2.29% to 85.5%* ***with a median value of 20.1%****) and \clarkl (ranging from 4.07% to 86.1%* ***with a median value of 15.7%****)*"

R2C19: *Tables S4/S5 (and lines 414-415): "assignable (and assigned) sequences" should be better defined (see also my comment below for line 325). "% assigned sequences (with at least one matching kmer)" in head of Col E is confusing because either a) it should complement Col D "% seq with no matching kmer" [since a sequence either has zero or at least one matching k-mer (= assignable?)] or b) Col E actually reports the fraction of assignable sequences that are assigned (at >=5/6 and >=0.95 thresholds?). Please clarify.*

This was clarified. Changed the definition of assignable the main text to "*assignable sequences* ***(i.e., containing at least one kmer matching the dictionary of target species discriminating kmers)***" (L425-L426).

Table and column legends (see also response to R2C39 comment below) were modified. More precisely, the two corresponding Supplementary Table legends (for Tables S4 and S5) now read: "*Table S4 CLARK assignment results at nk>4 and c>0.95 filtering criteria (for the 301 samples)*" and "*Table S5 CLARK-l assignment results at nk>4 and c>0.95 filtering criteria (for the 301 samples)*" and their ColE legend now reads: "*% of*

assigned sequences (i.e., with nk>4 and c>0.95) among all assignable ones (i.e., with at least one matching kmer)".

R2C20: *in legend of Figure 2 ("corresponding target dictionary"): why "corresponding" here? There is only one global dictionary per method, correct?*

Yes this is correct. I removed "corresponding" to avoid confusion.

R2C21: *lines 503-505 ("capture less than 30% of the assigned sequences"): the text does not exactly match what is shown in Figure S4 (rather 40% for Doshi, Dprui and Dbock while Dcard is not cited). Why such a discrepancy with Figure 3?*

There was indeed an error in the script that generated Figure S4 (thank you for catching it) that made panels A and B the same, and corresponding to the ClarkL results. The new version of Figure S4 has been corrected (and Figure S4A is now consistent with Figure 3).

R2C22: *lines 527-529: if I count correctly, 5 Ind-Seq samples are not mentioned in this part (236-215-16 = 5). Four of them are cited when discussing Wolbachia contamination, but not the last one: US-Nc2_CF1. Anything to say about it?*

This individual (US-Nc2_CF1) is the one with 9.58% *S. cerevisiae* contamination (mentioned at the end of the section, L622). For clarity, I actually decided to organize this (mostly descriptive) section by first focusing on i) sample mislabeling/contamination by drosophilid species for IndSeq (215 uncontaminated and 16 reassigned among the 236) and PoolSeq (17 uncontaminated and 5 contaminated); and then ii) microbial contamination (the 4 CN-Dan and US-Nc2_CF1 showing >5% contamination with *S. cerevisiae*, i.e. <95% of sequences assigned to D. suzukii).

R2C23: *lines 708-710 (about filtering based on k-mers): I agree with the assertion, but it seems ironic that target contigs were filtered with Kraken2 in the present study. It should be explicitly reminded here to avoid the feeling.*

I agree with the referee and have added some comments on this topic to clarify (L722-L738): "*For sequence filtering purposes, however, such approaches must be used with caution because they rely on species-discriminating k–mers and thus may leave a substantial fraction of sequences unassigned. More advanced (and computationally expensive)*

methods may then be valuable, such as the one implemented in CLARK-S (23), which allows some mismatches in k-mer matching to improve the sensitivity of sequence assignment, ***or even KRAKEN (32, 33), which was used here to identify contaminating contigs in the assemblies of the target species. Indeed, this program can rely on k-mers shared by several species for sequence assignment, and not only species discriminating k-mers, since all the k-mers of the target dictionary (possibly built from very large databases such as the NCBI nt) are mapped to the nodes of a phylogenetic tree (species discriminating k-mers to terminal nodes and shared k-mers to internal nodes***"

R2C24: *lines 732-737 (about contaminated Pool-Seq samples): was this issue known prior to the current study? If not, this would be useful to state it.*

This was indeed not known (unfortunately). I slightly modified the first sentence to suggest it (L758): "*Two of the 22 Pool-Seq samples of \citep{Olazcuaga2020} collected in the Asian native area were also,* ***and unexpectedly****, found to be contaminated with dsubL individuals*".

R2C25: *legend of Figure S2: why "Total assignment time"? I guess it includes sample loading time, but this is not mentioned in the main text. Is it what this means?*

This was clarified in the legend: "*Time spent for assigning all the sequences (i.e., excluding time for loading the k-mer dictionary) with...*". This indeed makes the corresponding sentence from the main text clearer (L377-L379): "*Given the size of the data sets, most of the analysis time was spent on sequence assignment which was almost linearly related to the number of sequences (Figure S2)*".

Mild suggestions

R2C26: *line 142 (and elsewhere): were assigned => were re-assigned [to emphasize the original assignment error?]*

This has been modified accordingly (L20).

R2C27: *lines 184,189-190: choose between "contaminating" and "contaminated"? In the present case, they are used interchangeably and this might be confusing.*

This has been corrected (L185, L192, L196).

R2C28: *Figure 1: why two species names in bold? Besides, for consistency with, e.g., willistoni, I would add the subgroup guinaria and virilis in the figure (especially because "subgroup virilis" is used in the text).*
Figure 1 has been modified. Only the name for D. suzukii is now in bold (this is now mentioned in the figure legend: "*D. suzukii is highlighted in bold*.") as this species is the focal species for the application. According to the NCBI taxonomy used here (see Figure 1 legend), there are no subgroups within the *virilis* and *guinaria* groups. I have changed the text to refer to the *virilis* group (rather than the subgroup).

R2C29: *lines 494-495: these samples => these 16 samples [for clarity and maybe it would be useful to color them differently in Figure S4]*
I agree with the referee and have edited both the sentence (L504) and Figure S4 (which shows the 16 samples in black and the 14 others in gray).

R2C30: *lines 499-500: the most represented species => the most closely related represented species [also check y axis in Figure S4].*
The sentence has been clarified (L508-L512): "*For the other samples from the most distantly related species, both the highest observed assignment rate (to a target species) and the percentage of sequences with no matching kmer clearly suggested that the target repository was not representative*". Figure S4 legend (and y-axis label) has also been modified for clarification.

R2C31: *line 539: 1. 71% => 1.71%*
Corrected (L549).

R2C32: *Figure S1B (y axis): %% overlapping => % overlapping*
Corrected (note: y-axis is now the percentage of non-overlapping reads, see R2C38 comment).

Reporting errors

R2C33: *line 6: 32 => 22*

Corrected (L6).

R2C34: *line 114: n=8 => n=6*

Corrected (L116).

R2C35: *in the Excel file, Tables S2 and S3 are reversed*

Corrected. I also noticed that the sheet names were truncated (due to a conversion problem from ods to xlsx with the OpenOffice software that I used). I changed the format to xls (instead of xlsx) and check conversion.

R2C36: *line 250: n=3 => n=4 [and "missing Illumina HiSeq X Ten (PE150) (n=1)"]*

Corrected (L257).

R2C37: *line 261 ("all sequenced on an Illumina HiSeq4000 in PE150 mode") => except 30 samples sequenced in PE100*

Corrected (“*These were all sequenced on an Illumina HiSeq4000 in PE150 (n=201) or PE100 (n=35) mode*”) (L268-L269).

R2C38: *Tables S2/S3: column headers for timing values are incorrect, which makes the section about run times extremely confusing; please check and fix! Moreover, the head of the column for overlapping values has the word "Non" in it, which (wrongly) suggests that these numbers are "non-overlapping reads" (see also line 281 in main text).*

This has been corrected (thank you for reporting these errors). Regarding overlapping reads, the percentages reported in the tables and main text were actually non-overlapping reads (this has been corrected in the main text). Figure S1B has also been changed to show the % of non-overlapping reads (instead of overlapping reads) for consistency (see also answer to comment R2C32 above).

R2C39: *line 325 (and elsewhere): I am not sure that it is an error, but to me, >1 and >5 mean "at least two" and "at least 6", respectively. Is*

it what is meant here? The issue is important because the section about the proportion of assigned sequences is difficult to understand with this doubt in mind (see comment above).

This was actually nk>=1 or nk>=5 (and c>0.9 and c>0.95) (see the *summary_csv.awk* script from the Zenodo repository). I have clarified this throughout the text (and changed the legend panel in Figure 2C).

R2C40: *Figure 2B: I am pretty sure that there is an error in the order of the first two violin plots. Target sp and Other sp are probably reversed because, as such, they neither match the text (lines 389-398) nor Figure 2A. Color key is right though. Please check and fix!*

This is correct (see also response to R1C4 comment from reviewer 1 above). The figure has been changed accordingly.