nature portfolio

Peer Review File

Genome-wide profiling of highly similar paralogous genes using HiFi sequencing

Corresponding Author: Dr Xiao Chen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this manuscript, the authors apply a previously described method, Paraphase, to a range of duplicated genes across samples from various populations. As part of the validation efforts, the authors show that Paraphase completely matches with orthogonal methods across 21 clinical samples and 8 disease-causing genes; as well as show very high trio concordance across 36 trios. Nevertheless, I think the paper will benefit from a comparison with other methods, including general-purpose CNV-calling tool Parascopy. In general, the paper is well written and shows interesting conclusions, but would benefit from addressing the following concerns:

Major: On line 66 the authors cite Parascopy (Prodanov & Bansal, Nat.Comm. 2022). I am one of the Parascopy developers, and, therefore, most likely biased towards it, but it seems that the overall pipeline (pool reads to one of the genes and call variants) is similar. In addition, Parascopy is able to detect paralog-specific copy number even in the absence of a flanking region or tandem repeats, and can identify paralog-specific variant genotypes (Bioinf. 2023). Could the authors perform the comparison with Parascopy and show that Paraphase produces higher accuracy or finds additional information?

Major: I could not find description, of how the variants are called in this manuscript or in the previous paper. Could the authors provide more details about that? In general, the paper requires a more detailed description of the method If the authors do not consider it important for the main text, they can place it in the Supplementary Methods.

Minor: In the Results, could the authors put more emphasis on which facts were already known before, and which are novel?

Minor: For long reads, which often spans a big portion of the region, lower depth is probably sufficient for a detailed CN prediction. How does Paraphase perform at lower read depth? Could the authors check the accuracy at subsampled datasets?

Minor: The authors should provide a testing dataset, that the users (or reviewers) can use to check the method before running it on their datasets.

Minor: The authors should mention the method runtime and memory usage.

Minor: The authors mention that the reason behind highly similar gene copies is gene conversion (for example line 202). However, this can also be a result of a recent duplication event. Did the authors perform any analysis to differentiate these two potential causes?

Minor: In Fig.1B it does not make sense to abbreviate Illumina as ILMN, just use the full word.

Minor: Table 1 requires a caption. Also, please mention the number of samples there, as well.

Minor: Methods-Validation (line 490) – “pathogenic variants in 8 disease-causing genes that were previously validated by orthogonal methods” – needs reference to how they were validated.

Minor: Line 170: SMN1 is a bad example, because it does not always have 4 copies even across the healthy individuals.

Maybe a better way to phrase this and the next sentence is to say “vast majority” instead of “all samples”, although even then SMN1 would not fit that description. Otherwise, the authors may have wanted to say “many samples have more than 2 copies”?

Minor: Currently, Figure 5 is hard to understand. The authors should consider modifying the caption and the figure itself to make it easier to understand. One can try to zoom in a specific point of interest; to use four shades (two different greens and two different purples) for four haplotypes; mention that C4A/B have short gene versions to explain deletions in the haplotypes, or completely crop this part out. Also, if possible, make text on the genome browser panels bigger.

Minor: Many supplementary figures require better description in the caption (for example Fig.S1).

Advise: not necessary, but the authors may consider improving Fig.2 by removing all or most of the grid lines and reducing whitespace around the gene names.
I usually use the following configuration in ggplot2:
theme(panel.grid = element_blank(),
strip.text = element_text(margin = margin(t = 2, b = 2)))
Also, it makes more sense to have even numbers on the X axis (0, 2, 4… or 0, 4, 8).

(Remarks on code availability)
The README is comprehensive and well structured, and the tool was easy to install. Nevertheless, the authors should provide a mock dataset (see main comments) to facilitate the tool testing.

Reviewer #2

(Remarks to the Author)
Authors Chen et al. present interesting work demonstrating the utility of Paraphase at a genome-wide scale. They specifically interrogate 160 gene families with high sequence identity (within family) and show that it’s possible to better resolve genomic regions and variants using phased long-read sequencing. In principle, this has been shown in a previous paper, but here they expand their analyses genome wide and present some valuable findings, including the copy number comparisons across populations.

Overall, I feel like the paper presents valuable results. My biggest concerns are that there are a lot of minor issues and that I think the paper could be much more thorough and impactful than it currently is. Also, the methods are not even close to sufficient.

To the authors: overall, I think it’s great work and Paraphase is a major improvement over standard approaches for highly similar gene regions. I intend my comments to be constructive with the hope that your work will be even stronger.

Minor (important but easily addressed)
1. Would be good to include a reference for the following statement in the introduction: “These high rates of gene conversion promote 63 sequence exchange between SDs further increasing the errors in read alignment.”
2. Maybe I’m misunderstanding Figure 1A, but I feel like it could be more clear and detailed. e.g., the last phased read is labeled as a “copy number change” but I don’t see any indication in the read that there’s a difference.
3. Would be helpful to clarify the following statement. Specifically, the authors state “76% have a median MAPQ <= 20”. Are the authors are referring to MAPQs for reads across the entire gene family? Similar question for “genomic segments”. Would be great to be more specific. Here’s the sentence of interest: “For short-read data, MAPQs are extremely low (76.4% have a median MAPQ<=20, and 98.8% have genomic segments with MAPQ<=20), indicating the difficulty of mapping short reads to these regions.”
4. Just as a suggestion for Figure 1B, I’d probably use a color brighter than grey for the histograms. They’re easy to miss because the eye is drawn to the red dots and we can’t see the density of the dots at position (60, 0). You might even consider doing the histograms in red and the dots in grey so the histograms become the focus. I would also specifically call attention to the large number of dots at that position in the figure legend.
5. I have several questions regarding the following statement: “There are 25 (15.6%) gene families where the median MAPQ is 60 and there are no bases with MAPQ<=20 in the long-read data. These are either regions where the homology extends less than the HiFi read length of ~15-20 kb or regions included in Paraphase for fusion calling between less similar paralogs (see Methods).”
a. I don’t understand the first sentence. Can you clarify? For example, for the median MAPQ of 60, are we talking about short or long-read data? The last part of the sentence says “in long-read data”, but not sure if that’s only referring to the second statistic given in that sentence. i.e., are we comparing MAPQs between short and long-read data?
b. Authors state there are “no bases” with MAPQ<=20 in the long-read data, but MAPQs are assessed to entire reads, not bases. I don’t know what is meant.
c. The second sentence is also confusing to me. My guess is that the MAPQ of 60 in the first sentence is referring to the long-read data, and this is just saying the MAPQ is so good because the long reads span the entire region. I also don’t understand why the second portion of the second sentence has any effect on the MAPQs. Please clarify this here, instead of only pointing readers to the methods.

6. The authors use “MLPA” without defining it.
7. I wouldn’t use abbreviations in section titles (e.g., “CN variability of gene families”)

Major
1. Overall, I feel like the paper contains a lot of valuable findings, but I also feel like each section could be much stronger in both writing and figures.
2. Most sections of the methods are missing a lot of detail. Additionally, methods need to include sections outlining how samples and data were selected/obtained and analyzed (with clear specifics).
3. Authors need to make all scripts accessible and easy to execute.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
In their manuscript "Genome wide profiling of highly similar paralogous genes using HiFi sequencing" Chen et al describe a genome wide analysis for the characterization of highly similar paralogous genes, at population level by the application of the recently developed method Paraphase. Althoug Paraphase was previously conceptulized and described in https://doi.org/10.1016/j.ajhg.2023.01.001, here the authors include a substantial amount of novel data and insights by expanding the range of application of the method to a collection of 316 genes (only one gene previously) and by performing genome-wide and population-wide analyses for a more systematic characterization of highly similar paralogous genes.
The paper is very well written and easy to read. The study is highly interesting, and provides important methodological insights for the study of some clinically relevant genes -among the other things. In my opinion however several key aspects of the method (optimal range of applicability, limitations) and the rationale behind some analytical choices need to be clarified and/or reconsidered prior to the acceptance for publication in Nature communications.

A list of the most critical points is reported below:

1) In my opinion the authors should refrain to use the term "gene family" throughout the text. As correctly stated in the title what they analyse is highly similar (probably recent) SD and paralogous genes. Gene family is a broader concept in evolutionary biology. Members of the same gene family do not necessarily need to be highly similar at sequence level. Please revise

2) As far as I understood, paraphase does not derive the set of candidate SD regions de novo from the reads and is bound to analyse a set of target, highly similar, genomic sequences for which SD have been previously detected and archetype genes defined. Authors identified these regions based on sequence similarity searches in the hg38 reference assembly. Human protein coding genes were used as a query. I have to admit that I am puzzled by the choice of the hg38 reference genome for this domain of application. The "recent" T2T genome assemblies and/or even better the Human Pangenome Reference do provide a notionally improved representation of repetive and duplicated regions. See https://doi.org/10.1038/s41586-023-05895-y and https://doi.org/10.1126/science.abj6965 (both cited in the introduction). Could the authors please clarify and explain?
2.1) Since haplotype level assemblies are available for at least some of the 259 individuals included in the analyses performed by Chen et al., one would expect that the haplotypes reconstructed by paraphase should be compared with those represented in the assemblies. Are the reconstructed haplotypes coherent? Are all the candidate segmental duplications correctly represented?
2.2) If the haplotypes reconstructed by paraphase are substantially equivalent to those included in the Human Pangenome Reference assemblies, which are then the main advantages of using paraphase? is it more computationally efficient? is it more accurate?
2.3) 2.1 could complement/add to the section "validation of paraphase calls"

3) The range of application of the paraphase should be better described/clarified in the text. The authors focus only on SD of at least 10Kb in size, and with 99% sequence identity. Is there a specific reason for that? Is this the range were paraphase works optimally? Is this due to the size of the reads and/or accuracy of the sequencing technology? Can the authors please elaborate? Is there a minimum level of divergence beyond which phase error become more frequent.
3.1) I am not myself a fan of simulated data, but it is my opinion that the sensitivity and specificity of pharapase in reconstructing haplotypes should somehow be measured. How do these parameters change w.r.t 1) reads size? 2) level of divergence of the genes? Maybe the authors should perform some controlled analyses with simulated data (or in silico trimmed/treated real data) to address these issue and provide some estimates
3.2) In the section "validation of pharapase calls" an error rate of 0.29% is reported when considering the consistency of reconstructed haplotypes through the analysis of family trios. What was the possible cause of these errors? Could the authors please elaborate? How are errors set apart from - for example- the 7 de novo mutationsin the section "Identification of de novo mutations and gene conversion"?

4) In my opinion figure, also by looking to figure 2A the authors use a very lenient definition for the identification of genes with higly variable CN (or alternatively a very stringent definition for low variability). Genes like NSF, HIC2, H3C14 for example do not show a highly variable copy number in my eyes. Probably the stratification should have a higher granularity. Intuitively at least classes should be formed: 1 invariable; 2; moderately variable; 3 highly variable.

5) If I understood correctly, "gene families" with low diversity are defined as those on having a variability below the less variable 10% of a randomly selected sample of 600 genes. A total of 23 (out of 160) variable gene families are identified by this approach. It should be noted that 23 is circa 14% of 160. This does not seem like a big enrichment w.r.t the expected number. Methods applied/used in this analysis are not adequate in my opinion, due to several reasons:
5.1) As far as I understood substitutions rates were computed as flat substitution rate on the complete gene body. Obviously coding exons, introns and UTRs have different evolutionary rates. This should be kept into account. Otherwise you will simply select the gene with the longest introns as the most variable.
5.2) 10% is a very lenient cut-off. Probably a more stringent cut-off should be used.
5.3) Probably duplicated genes and recently duplicated genes have (very) different selective pressure. It should probably better to stratify lowly variable duplicated genes by examining duplicated genes in isolation.

6) Figure 6 B,C,D. Why was a PCA used? What proportion of the variance is explained by the first 2 principal components. Where additional components examined? Figure 2B: pink dots on the top left seem to be well separated from the light blue dots. Since this PCA includes 3 juxtaposed genes, could the authors repeat the analysis by considering one gene at a time, and demonstrate that the above mentioned pink dot do not associate with only one of the genes. (That would be an interesting observation)

(Remarks on code availability)
The code avaialable through a dedicated github repository and is publicly accessible, the documentation is complete and clear. The tool can be easily installed and executed.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the authors for the work they did on revisions. I think the manuscript has improved and I don't have any remarks to it.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors' responses were satisfactory.

(Remarks on code availability)
I only briefly reviewed the code and GitHub. It appears to be well documented.

Reviewer #3

(Remarks to the Author)
Authors have addressed all the major points of criticism raised by this referee in a thorough and comprehensive manner. It is my opinion that the revised version of the manuscript should be accepted as it is for publication

(Remarks on code availability)
Paraphase can be easily installed and executed

We would like to sincerely thank all reviewers for the insightful and positive feedback. We are very grateful to you for the time you spent reviewing this paper and helping to make it better. Please find responses to your comments below. The corresponding changes are highlighted in blue in the manuscript. Here is also a summary of the most substantial changes:

1. We have strengthened the benchmarking section by comparison of Paraphase variant calls against high quality diploid assemblies and demonstrated the consistent performance of Paraphase as a function of depth, read length and paralog divergence. Also, we performed a comparison with Parascopy and shared the results in our response though we chose not to include this comparison in the updated manuscript.
2. We have incorporated the reviewers' suggestions to improve the writing throughout the manuscript and clarify the Methods section and improve the main figures. In addition, we have gone through and carefully re-read the manuscript especially where the reviewers were confused to clarify and improve the writing.

Reviewer #1 (Remarks to the Author):

In this manuscript, the authors apply a previously described method, Paraphase, to a range of duplicated genes across samples from various populations. As part of the validation efforts, the authors show that Paraphase completely matches with orthogonal methods across 21 clinical samples and 8 disease-causing genes; as well as show very high trio concordance across 36 trios. Nevertheless, I think the paper will benefit from a comparison with other methods, including general-purpose CNV-calling tool Parascopy. In general, the paper is well written and shows interesting conclusions, but would benefit from addressing the following concerns:

> Thank you for your comments and suggestions.

Major: On line 66 the authors cite Parascopy (Prodanov & Bansal, Nat.Comm. 2022). I am one of the Parascopy developers, and, therefore, most likely biased towards it, but it seems that the overall pipeline (pool reads to one of the genes and call variants) is similar. In addition, Parascopy is able to detect paralog-specific copy number even in the absence of a flanking region or tandem repeats, and can identify paralog-specific variant genotypes (Bioinf. 2023). Could the authors perform the comparison with Parascopy and show that Paraphase produces higher accuracy or finds additional information?

> Thanks for suggesting this analysis. We initially didn't do this comparison because we anticipated that Paraphase would be significantly better mostly due to the underlying data type that the methods were designed to use (i.e. long vs short reads). Parascopy is designed for short reads while Paraphase is designed for long reads and thus the tools work very differently by necessity of the differences of the starting data. Most importantly, because it is designed to work with long reads, Paraphase reports fully phased variants on haplotypes spanning complete genes. Getting full gene sequences (hapotypes) makes it straightforward to annotate and interpret variant consequences, as well as to assign gene copies into genes or paralogs. Conversely, for short read based variant calls, phasing between variants is mostly impossible

and assigning unphased variants to genes/paralogs based on paralogous sequence variants (PSVs) can be error-prone, as in many paralog groups, repeat copies are highly similar to each other and can exchange sequences through gene conversion. As a result, many of the PSVs are variable in the populations and can be unreliable for accurately assigning reads. In addition, in highly similar regions between genes and their paralogs, there might not be any PSVs in some regions so that reads/variants cannot be assigned to genes/paralogs with short reads.

Taking one gene-paralog pair, *PMS2* and *PMS2CL*, as an example below, we compared Paraphase and Parascopy variant calls against high quality diploid assemblies in 42 HPRC samples. Note that our comparison here did not assess the phasing between variants, which Paraphase can do but Parascopy cannot (the phasing accuracy of Paraphase was assessed when we did our full comparison to assemblies in our response to Review #3 Comment 2.1). When variants were called against the main repeat copy (*PMS2*), Parascopy successfully called a large portion of the variants. However, when variants were called in a paralog-specific way, i.e. after assigning variants to *PMS2* or *PMS2CL*, Parascopy had a low recall, indicating that many of the variants were assigned incorrectly, i.e. *PMS2* variants assigned to *PMS2CL*, or vice versa. This is likely because in this gene-paralog pair, most PSVs are unreliable for assigning variants to *PMS2* or *PMS2CL*, as frequent gene conversions lead to segments of *PMS2* that look like *PMS2CL*, or vice versa (see Figure 5c of the manuscript). Some samples had zero recall for Parascopy as Parascopy incorrectly reported copy number gains/losses of *PMS2* or *PMS2CL* (due to signals from gene conversion), leading to wrong genotype calls (all samples in this analysis have two copies of *PMS2* and two copies of *PMS2CL*). We decided not to include this analysis in the paper because the relative performance is driven by the physical features of the sequence data and not differences between the two software tools.

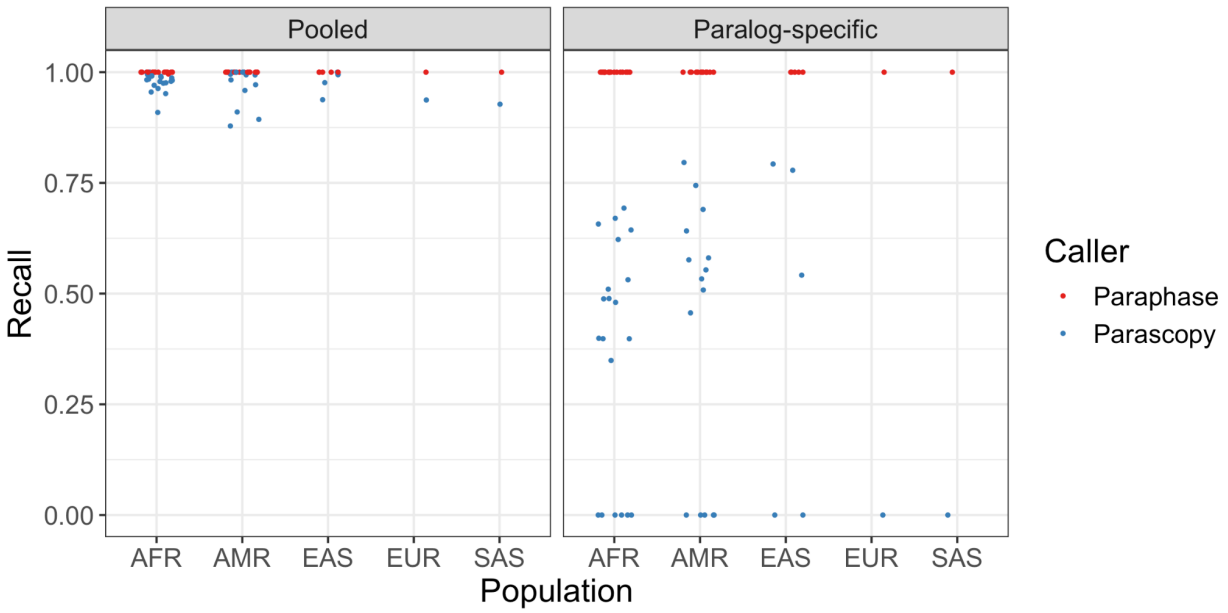


Figure R1. Variant recall of Paraphase and Parascopy against HPRC assemblies in *PMS2*/*PMS2CL*. Each dot represents one sample.

Major: I could not find a description, of how the variants are called in this manuscript or in the previous paper. Could the authors provide more details about that? In general, the paper requires a more detailed description of the method If the authors do not consider it important for the main text, they can place it in the Supplementary Methods.

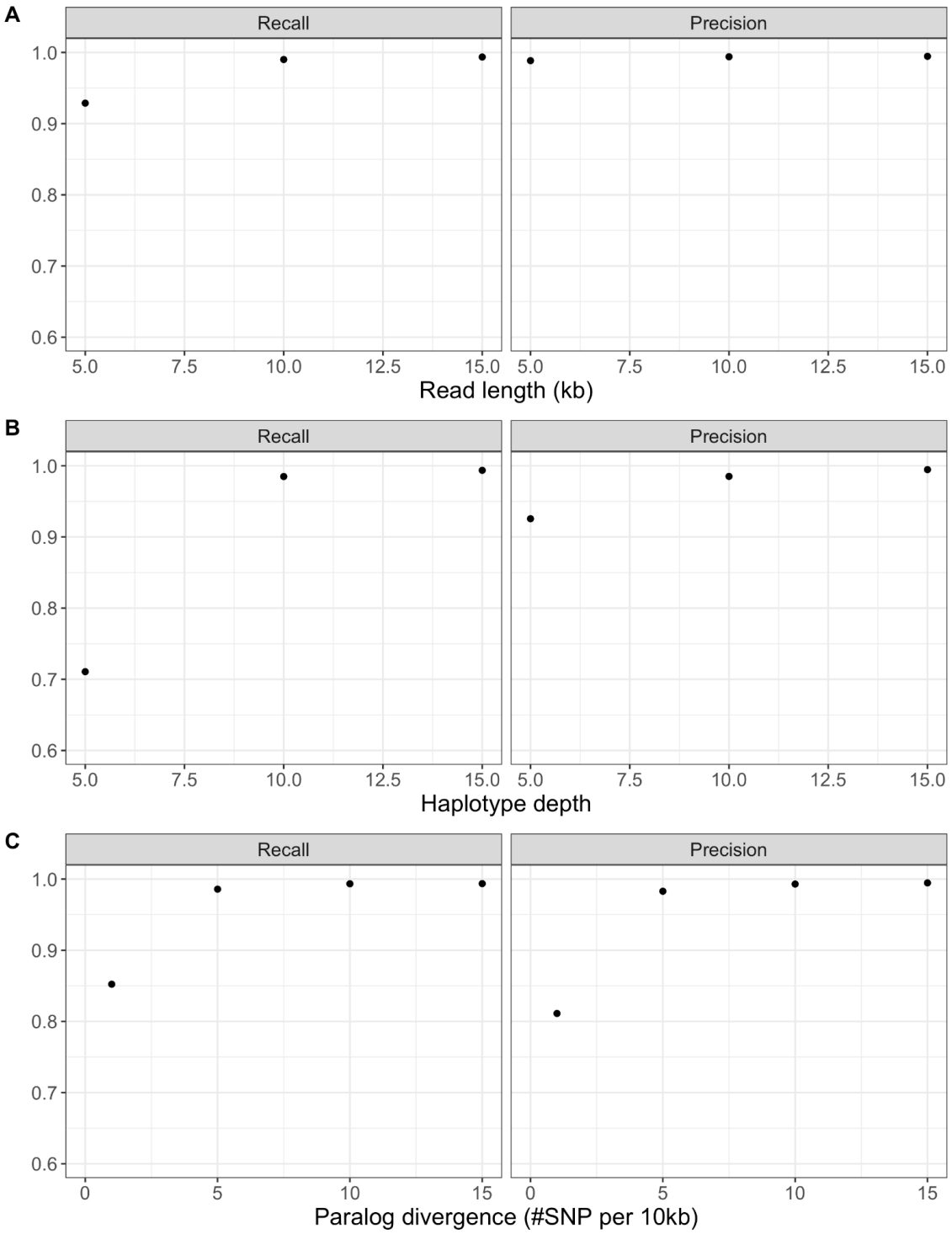
> Thanks for highlighting the need for a better description of how Paraphase works. We have added more details to the Methods section on the variant calling logic in Paraphase (Page 11). Very briefly, reads are re-aligned to the main gene (e.g. *SMN1* in the case of *SMN1*/*SMN2* genotyping), haplotypes are constructed from these alignments first and reads are assigned to the haplotypes that they originate from. For each haplotype, Paraphase takes the consensus of all reads corresponding to the haplotype at each position and variants are called as base differences between the consensus and the reference. This is now described more fully in the updated Methods section.

Minor: In the Results, could the authors put more emphasis on which facts were already known before, and which are novel?

> We have revised the Results section to distinguish known facts and novel findings.

Minor: For long reads, which often spans a big portion of the region, lower depth is probably sufficient for a detailed CN prediction. How does Paraphase perform at lower read depth? Could the authors check the accuracy at subsampled datasets?

> Paraphase detects copy numbers of a paralog group primarily by phasing all gene copies, i.e. haplotypes, of the paralog group. The success of phasing is dependent on multiple factors, such as read length, read depth and the divergence level between haplotypes. We performed a simulation experiment for all of these factors following the suggestions of Reviewer 1 and Reviewer 3 (Supplementary Fig. 4 below) to examine their impact on haplotyping accuracy. Panel B indicates that a haplotype depth of 10X and 15X results in high phasing accuracy, and when depth is lowered to 5X, it's possible to have lower recall and precision. Also see our response to Review #3 Comment 3.1. We have added this analysis into Supplementary Notes.



Supplementary Fig. 4. Paraphase performance on simulation data of varied read length (baseline 15 kb), haplotype depth (baseline 15X) and paralog divergence (baseline 0.15%). Paraphase calls were examined for variant recall and precision at varied read lengths of 5 kb, 10 kb and 15 kb (A), varied haplotype depths of 5X, 10X and 15X (B) and varied paralog divergence of 1, 5, 10 and 15 SNPs per 10 kb (0.01%, 0.05%, 0.1% and 0.15%, respectively) (C), while the other two variables are kept at baseline for each panel.

Minor: The authors should provide a testing dataset, that the users (or reviewers) can use to check the method before running it on their datasets.

> We have added the test dataset to the github page and updated the README accordingly (https://github.com/PacificBiosciences/paraphase/blob/main/docs/demo.md).

Minor: The authors should mention the method runtime and memory usage.

> We have added this information to the manuscript (Line 405, Page 11):
>
> "For a single WGS sample, across 160 paralog groups, Paraphase requires 4Gb memory and typically completes analysis in 90 minutes (1 thread) or 15 minutes (16 threads)."

Minor: The authors mention that the reason behind highly similar gene copies is gene conversion (for example line 202). However, this can also be a result of a recent duplication event. Did the authors perform any analysis to differentiate these two potential causes?

> We list two different types of highly similar paralog groups in Table 2. The highly similar paralog groups that we found on sex chromosomes are all conserved in other primates, suggesting that these are not recent duplication events and gene conversion could be a main force maintaining the high similarity between paralogs. On the other hand, most of the highly similar paralog groups that we found on autosomes are human-specific duplications and thus could be recent. For these genes, it could be a mixed effect of recent duplication, positive selection and gene conversion. We have now altered the text to emphasize that gene conversion is just one of several factors that could be driving this observation and have highlighted that in the Discussion (Line 344, Page 10):
>
> "It is possible that recent duplication and positive selection and/or gene conversion could play a role in the evolution of these genes, preventing sequence divergence and maintaining an elevated gene dosage in humans."

Minor: In Fig.1B it does not make sense to abbreviate Illumina as ILMN, just use the full word.

> We have revised the figure accordingly.

Minor: Table 1 requires a caption. Also, please mention the number of samples there, as well.

> We have revised the table and its caption accordingly (Page 22).

Minor: Methods-Validation (line 490) – "pathogenic variants in 8 disease-causing genes that were previously validated by orthogonal methods" – needs reference to how they were validated.

> We have added the reference (Höps *et al.* 2024, Ref #52).

Minor: Line 170: SMN1 is a bad example, because it does not always have 4 copies even across the healthy individuals. Maybe a better way to phrase this and the next sentence is to say “vast majority” instead of “all samples”, although even then SMN1 would not fit that description. Otherwise, the authors may have wanted to say “many samples have more than 2 copies”?

> Thanks for this suggestion and we have removed “like *SMN1/SMN2*” from the sentence (Line 176, Page 6). The paragraph is about using population copy numbers to confirm the accuracy of the reference genome, i.e. to identify false duplication regions in the reference genome and we agree that *SMN1* is too variable to be a good example here.

Minor: Currently, Figure 5 is hard to understand. The authors should consider modifying the caption and the figure itself to make it easier to understand. One can try to zoom in a specific point of interest; to use four shades (two different greens and two different purples) for four haplotypes; mention that C4A/B have short gene versions to explain deletions in the haplotypes, or completely crop this part out. Also, if possible, make text on the genome browser panels bigger.

> We have simplified the figure and improved the caption (Page 21).

Minor: Many supplementary figures require better description in the caption (for example Fig.S1).

> We have reviewed and improved the captions of supplementary figures.

Advise: not necessary, but the authors may consider improving Fig.2 by removing all or most of the grid lines and reducing whitespace around the gene names.
I usually use the following configuration in ggplot2:
theme(panel.grid = element_blank(),
strip.text = element_text(margin = margin(t = 2, b = 2)))
Also, it makes more sense to have even numbers on the X axis (0, 2, 4… or 0, 4, 8).

> We agree and have revised the figure accordingly.

Reviewer #1 (Remarks on code availability):

The README is comprehensive and well structured, and the tool was easy to install. Nevertheless, the authors should provide a mock dataset (see main comments) to facilitate the tool testing.

> We have added the test dataset to the GitHub page and updated the README.

Reviewer #2 (Remarks to the Author):

Authors Chen et al. present interesting work demonstrating the utility of Paraphase at a genome-wide scale. They specifically interrogate 160 gene families with high sequence identity (within family) and show that it’s possible to better resolve genomic regions and variants using phased long-read

sequencing. In principle, this has been shown in a previous paper, but here they expand their analyses genome wide and present some valuable findings, including the copy number comparisons across populations.

Overall, I feel like the paper presents valuable results. My biggest concerns are that there are a lot of minor issues and that I think the paper could be much more thorough and impactful than it currently is. Also, the methods are not even close to sufficient.

To the authors: overall, I think it's great work and Paraphase is a major improvement over standard approaches for highly similar gene regions. I intend my comments to be constructive with the hope that your work will be even stronger.

> Thank you for your comments and suggestions.

Minor (important but easily addressed)
1. Would be good to include a reference for the following statement in the introduction: "These high rates of gene conversion promote 63 sequence exchange between SDs further increasing the errors in read alignment."

> We have now added references (Dumont *et al.* 2015 and Vollger *et al.* 2023, Ref #9 and #10).

2. Maybe I'm misunderstanding Figure 1A, but I feel like it could be more clear and detailed. e.g., the last phased read is labeled as a "copy number change" but I don't see any indication in the read that there's a difference.

> Thanks for highlighting the confusing nature of this figure. We have updated the figure legend to make it more clear (Line 678, Page 20). In this example, the third copy of the paralog gene indicates that there is a copy number gain (two copies are expected for a diploid genome).

3. Would be helpful to clarify the following statement. Specifically, the authors state "76% have a median MAPQ <= 20". Are the authors referring to MAPQs for reads across the entire gene family? Similar question for "genomic segments". Would be great to be more specific. Here's the sentence of interest: "For short-read data, MAPQs are extremely low (76.4% have a median MAPQ<=20, and 98.8% have genomic segments with MAPQ<=20), indicating the difficulty of mapping short reads to these regions."

> We agree that this section was a little confusing and have revised the Methods section to better describe how this calculation was performed (Line 429, Page 12). We have also revised the term to be "summary MAPQ" to slightly differentiate from traditional read-specific MAPQ. The summary MAPQ values were calculated in the following way:
>
> 1. We calculate the median MAPQ for all of the reads overlapping a particular base position across all 20 samples (e.g. if each genome is sequence to 30x we would expect this to be the median of 20x30=600 total reads). This metric is termed base MAPQ.

2. We then calculate the median of this value (base MAPQ) over all of the bases for the relevant paralog group and call this the summary MAPQ. ”

In summary, each base has a base MAPQ, and each paralog group has a summary MAPQ value, which is the median of all the base MAPQ values. The percentages in the sentence of interest refer to the percentages out of all paralog groups. So 76.4% of the paralog groups have a summary MAPQ<20. We have updated the sentence to make it more clear. We have also revised “genomic segments” to simply “bases”.

The sentence of interest now reads (Line 118, Page 4):
“For short-read data, MAPQs are extremely low (76.4% of the paralog groups have a summary MAPQ<=20, and 98.8% of the paralog groups have some bases with a base MAPQ<=20), indicating the difficulty of mapping short reads to these regions.”

4. Just as a suggestion for Figure 1B, I’d probably use a color brighter than grey for the histograms. They’re easy to miss because the eye is drawn to the red dots and we can’t see the density of the dots at position (60, 0). You might even consider doing the histograms in red and the dots in grey so the histograms become the focus. I would also specifically call attention to the large number of dots at that position in the figure legend.

Great suggestions. We have changed the figure based on the reviewer’s suggestions.

5. I have several questions regarding the following statement: “There are 25 (15.6%) gene families where the median MAPQ is 60 and there are no bases with MAPQ<=20 in the long-read data. These are either regions where the homology extends less than the HiFi read length of ~15-20 kb or regions included in Paraphase for fusion calling between less similar paralogs (see Methods).”
a. I don’t understand the first sentence. Can you clarify? For example, for the median MAPQ of 60, are we talking about short or long-read data? The last part of the sentence says “in long-read data”, but not sure if that’s only referring to the second statistic given in that sentence. i.e., are we comparing MAPQs between short and long-read data?

This sentence only refers to long-read data and we have revised the sentence to make it more clear as below (Line 122, Page 4):

“For long-read data, there are 25 (15.6%) paralog groups where the summary MAPQ is 60 and there are no bases with a base MAPQ<=20.”

b. Authors state there are “no bases” with MAPQ<=20 in the long-read data, but MAPQs are assessed to entire reads, not bases. I don’t know what is meant.

This refers to our definition of how we are assessing the MAPQ values for these regions (Please see our response to your comment #3).

c. The second sentence is also confusing to me. My guess is that the MAPQ of 60 in the first sentence is referring to the long-read data, and this is just saying the MAPQ is so good because the long reads span the entire region. I also don't understand why the second portion of the second sentence has any effect on the MAPQs. Please clarify this here, instead of only pointing readers to the methods.

> This sentence and the sentence that follows try to explain why these regions that seem to have no alignment problem are included in Paraphase. We have updated the sentences to make it more clear (Line 124, Page 4):
>
> "These are either regions where the sequence similarity is high but the homology extends less than the HiFi read length of ~15-20 kb, or regions which have lower sequence similarity but are included in Paraphase for fusion calling (see Methods). Paraphase analysis can still improve the performance in these high MAPQ regions because: 1) even reads with high MAPQ can be misaligned due to reference genome artifacts, common CNVs and high rates of gene conversion, 2) gene fusions are hard to detect because split alignments are unlikely to happen in regions of homology and 3) lower MAPQs will be expected in data with shorter read length, such as in HiFi hybrid capture data."

6. The authors use "MLPA" without defining it.

> We have included the full name of MLPA (Line 77, Page 3). Thank you for pointing it out.

7. I wouldn't use abbreviations in section titles (e.g., "CN variability of gene families")

> We have removed the abbreviations (Line 162, Page 5).

Major

1. Overall, I feel like the paper contains a lot of valuable findings, but I also feel like each section could be much stronger in both writing and figures.

> We have gone through the entire manuscript and revised the text to clarify both the suggested areas highlighted by the reviewers and also other places where appropriate. In addition, we have improved the figures throughout the paper.

2. Most sections of the methods are missing a lot of detail. Additionally, methods need to include sections outlining how samples and data were selected/obtained and analyzed (with clear specifics).

> We have added more details to the Methods section, particularly around sample collection and data processing.

3. Authors need to make all scripts accessible and easy to execute.

> We have made the scripts accessible on zenodo (https://doi.org/10.5281/zenodo.14003613).

Reviewer #3 (Remarks to the Author):

In their manuscript "Genome wide profiling of highly similar paralogous genes using HiFi sequencing" Chen et al describe a genome wide analysis for the characterization of highly similar paralogous genes, at population level by the application of the recently developed method Paraphase. Althoug Paraphase was previously conceptulized and described in https://doi.org/10.1016/j.ajhg.2023.01.001, here the authors include a substantial amount of novel data and insights by expanding the range of application of the method to a collection of 316 genes (only one gene previously) and by performing genome-wide and population-wide analyses for a more systematic characterization of highly similar paralogous genes.
The paper is very well written and easy to read. The study is highly interesting, and provides important methodological insights for the study of some clinically relevant genes -among the other things. In my opinion however several key aspects of the method (optimal range of applicability, limitations) and the rationale behind some analytical choices need to be clarified and/or reconsidered prior to the acceptance for publication in Nature communications.

> Thank you for your comments and suggestions.

A list of the most critical points is reported below:

1) In my opinion the authors should refrain to use the term "gene family" throughout the text. As correctly stated in the title what they analyse is highly similar (probably recent) SD and paralogous genes. Gene family is a broader concept in evolutionary biology. Members of the same gene family do not necessarily need to be highly similar at sequence level. Please revise

> Thanks for this suggestion and we have revised the term to “paralog group”.

2) As far as I understood, paraphase does not derive the set of candidate SD regions de novo from the reads and is bound to analyse a set of target, highly similar, genomic sequences for which SD have been previously detected and archetype genes defined. Authors identified these regions based on sequence similarity searches in the hg38 reference assembly. Human protein coding genes were used as a query. I have to admit that I am puzzled by the choice of the hg38 reference genome for this domain of application. The "recent" T2T genome assemblies and/or even better the Human Pangenome Reference do provide a notionally improved representation of repetive and duplicated regions. See https://doi.org/10.1038/s41586-023-05895-y and https://doi.org/10.1126/science.abj6965 (both cited in the introduction). Could the authors please clarify and explain?

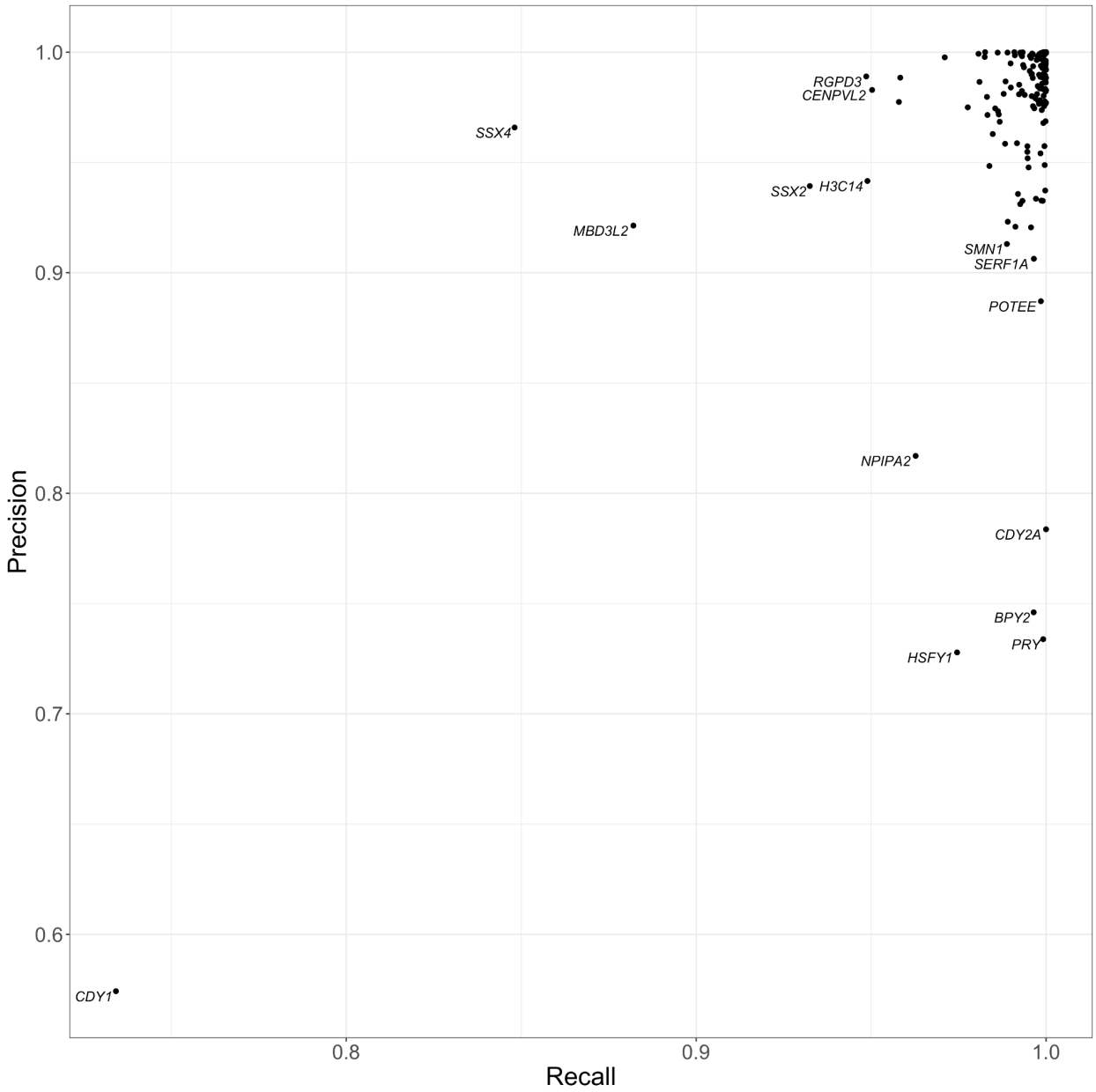
> We primarily chose to work on hg38 because this is currently the most widely used reference genome, especially in the clinical space. Paraphase is designed to call variants in segmental duplications, especially clinically relevant ones, in hg38. So we focus on analyzing segmental duplication regions in hg38 to capture all paralog groups that could result in alignment errors in hg38. The same set of analyses can be applied to incorporate segmental duplications identified in the CHM13 T2T reference genome (and future assembled genomes) to enable a population-level analysis of segmental duplications in the future.

2.1) Since haplotype level assemblies are available for at least some of the 259 individuals included in the analyses performed by Chen et al., one would expect that the haplotypes reconstructed by paraphase should be compared with those represented in the assemblies. Are the reconstructed haplotypes coherent? Are all the candidate segmental duplications correctly represented?

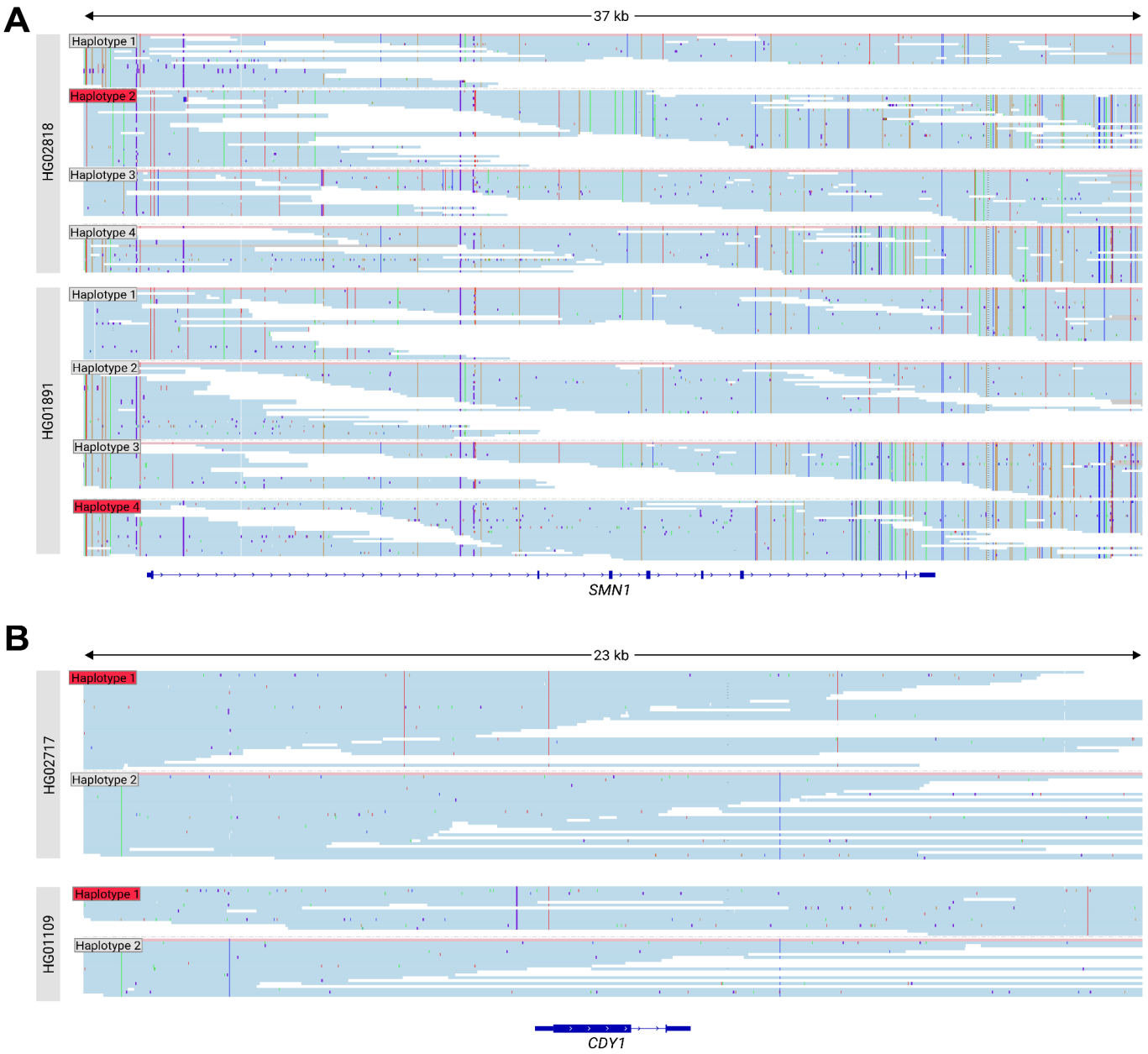
Thank you for raising this important issue. We compared Paraphase calls with high quality, phased diploid assemblies in 47 HPRC samples and summarized recall and precision (taking assembly variant calls as the truth) in Supplementary Fig. 1 below. Paraphase small variant calls were consistent with the assembly in the majority of paralog groups (defining the assembly as the ground truth, 82.4% of paralog groups have >95% recall and >95% precision). The low precision in some paralog groups appears to be driven by missing gene copies in the assemblies. We highlight this in Supplementary Fig 2, with a couple examples where the numbers of copies of the paralog group differ between Paraphase and the assemblies. In each case, we see reads that align directly to the Paraphase-only haplotype (highlighted in red) showing clearly that the Paraphase-only copy is present in the sequence data. Since these are highly similar paralog groups, some gene copies may have been collapsed in the assembly, leading to missing genes and incorrect copy number.

Additionally, the low recall in some paralog groups such as *CDY1, SSX4, MBD3L2 and SSX2* is related to the observation that sometimes a segment of a single HiFi read is directly incorporated into the assembly such that the assembly inherits all (or most) of the errors from that single read. Supplementary Fig. 3 highlights a few examples where the assembly appears to be consistent with a single read but inconsistent with all the other reads overlapping a position. As a result, the assembly calls many more indels than Paraphase, which does not have this problem because it takes the consensus sequence of all the reads that originate from a haplotype. This could be an area for improvement for future assemblers.

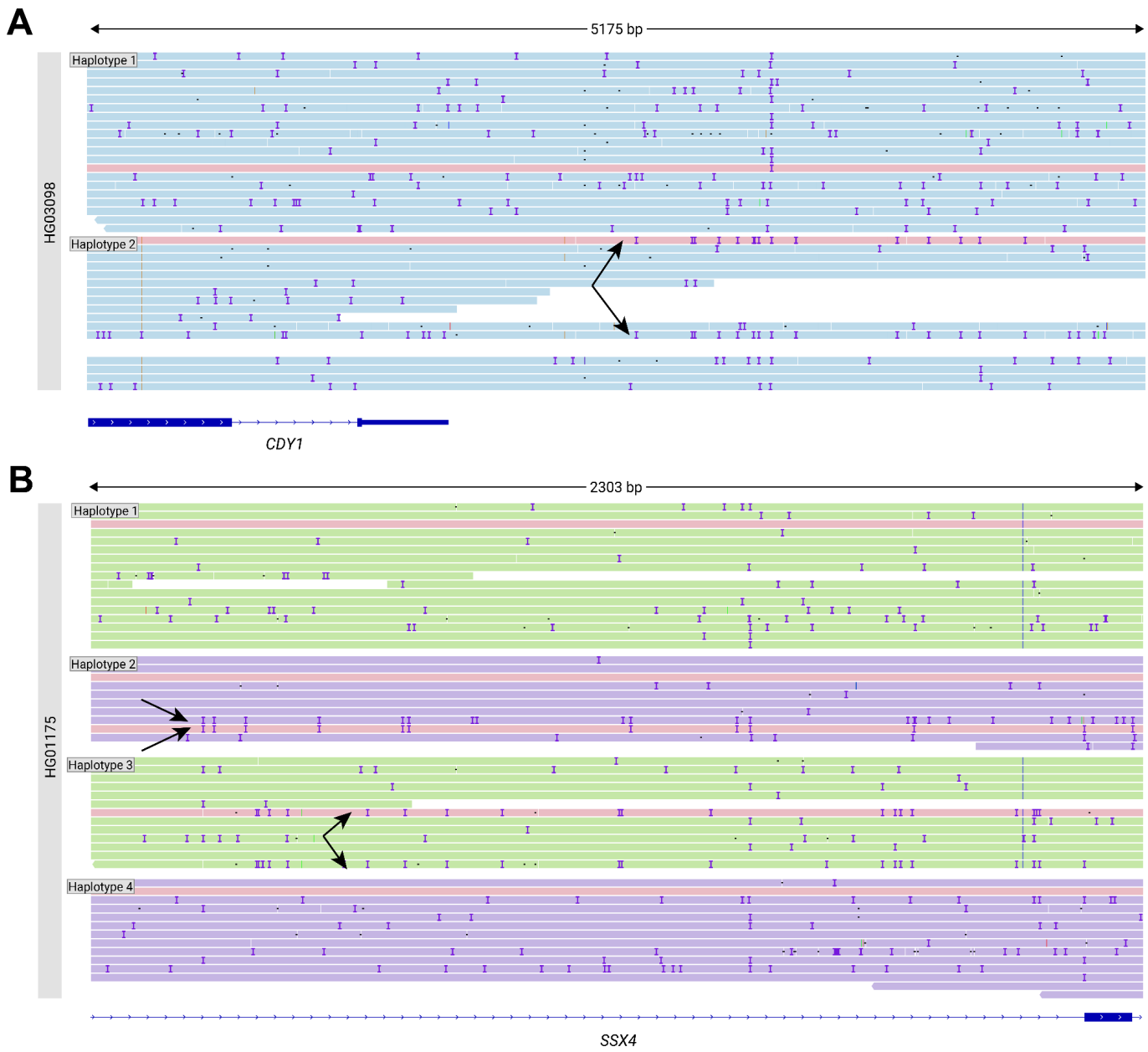
This comparison shows that the assembly can be inaccurate especially in regions where different copies of the segmental duplication are highly similar to each other.



Supplementary Fig. 1. Recall and precision of Paraphase variant calls against high quality diploid assemblies in 47 HPRC samples. Each dot denotes the values of a paralog group.



Supplementary Fig. 2. Examples of missing gene copies in assemblies, leading to reduced precision of Paraphase in some paralog groups. Each sample has one Paraphase-called haplotype, labeled in red-filled rectangles, that does not have any matching segment in the assemblies. Two samples are shown for the *SMN1* paralog group (A) and two samples are shown for the *CDY1* paralog group (B). HiFi reads are grouped by the haplotypes called by Paraphase and the assemblies are shown (colored in pink) along with their corresponding Paraphase haplotype.



Supplementary Fig. 3. Examples of base errors in assemblies, leading to reduced recall of Paraphase in some paralog groups. One sample is shown for the *CDY1* paralog group (A) and one sample is shown for the *SSX4* paralog group (B). HiFi reads are grouped by the haplotypes called by Paraphase (haplotypes from two alleles are colored in green and purple for *SSX4*). Segments from contigs of the assemblies are shown together with the haplotypes they correspond to and are colored in pink. For each haplotype, Paraphase takes the consensus sequence of all reads that originate from it and it is expected that the consensus matches the assembly. Black arrows highlight assembly contigs (pink) that have incorporated sequencing errors, mostly indels, from a single HiFi read (non-pink read). This leads to errors in the assembly and leads to reduced recall of Paraphase when considering the assembly as the ground truth.

2.2) If the haplotypes reconstructed by paraphase are substantially equivalent to those included in the Human Pangenome Reference assemblies, which are then the main advantages of using paraphase? is it more computationally efficient? is it more accurate?

There are four primary reasons why a Paraphase approach is preferred over assembly.

1. Related to your previous comment, we have observed cases where Paraphase is more accurate than assemblies for genotyping some paralog groups, especially those that are highly similar.
2. In addition to standard variants, Paraphase's haplotyping approach can also identify mosaicism or reveal methylation status of haplotypes, while assemblies may lose this information.
3. High quality assemblies require more than one sequencing technology and are thus not always available for any sample, while Paraphase provides a HiFi-only approach for resolving those complex paralogous genes.
4. Paraphase is also more tolerant to lower depth and shorter read length than assemblers. As we showed in our response to your comment 3.1 below, Paraphase can work down to less than 20x depth. High quality assemblies require 60x combining both long and ultra long reads and often trio information.

2.3) 2.1 could complement/add to the section "validation of paraphase calls"

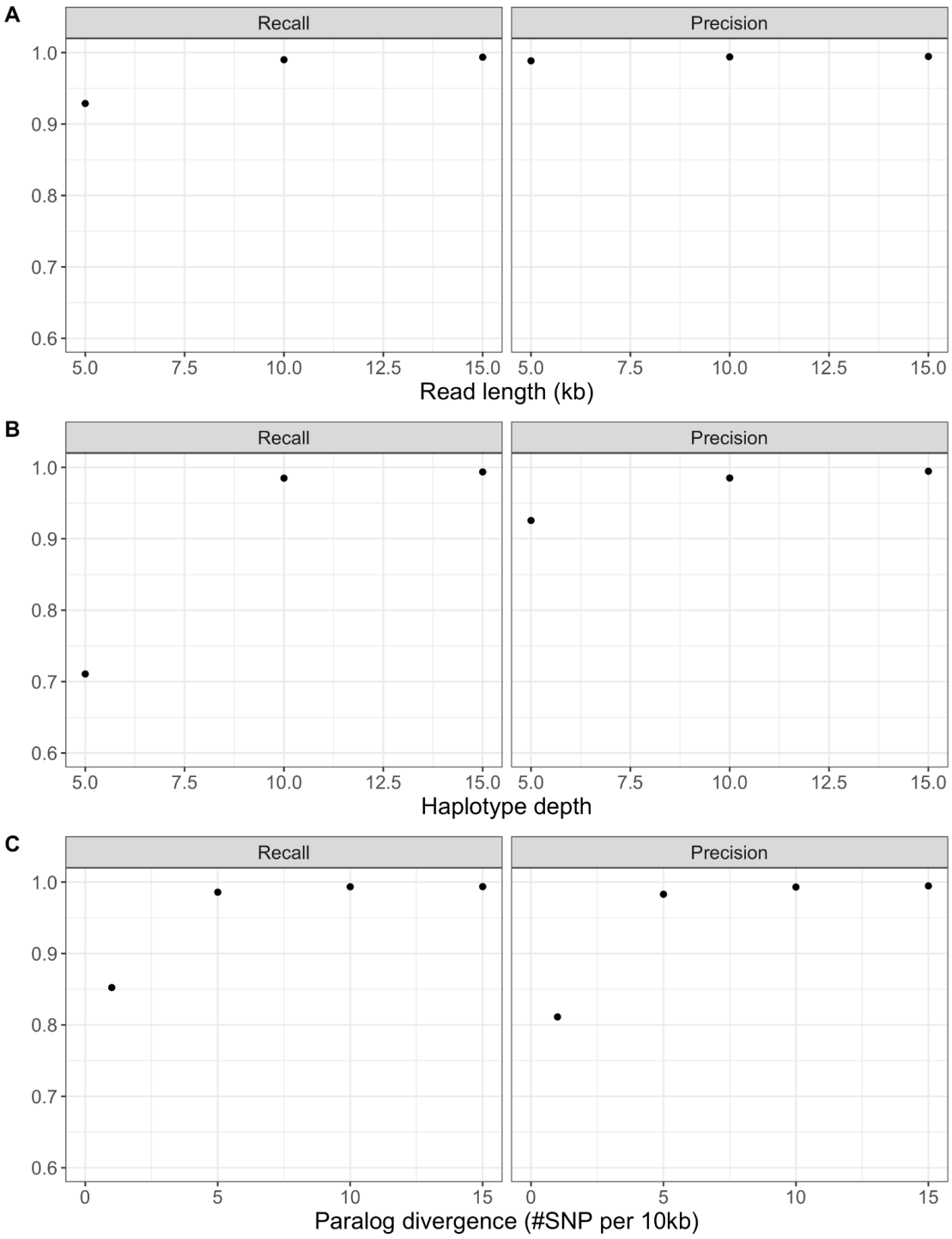
We have added the comparison against assemblies as part of the validation section (Line 147, Page 5) and more details to Supplementary Notes.

3) The range of application of the paraphase should be better described/clarified in the text. The authors focus only on SD of at least 10Kb in size, and with 99% sequence identity. Is there a specific reason for that? Is this the range where paraphase works optimally? Is this due to the size of the reads and/or accuracy of the sequencing technology? Can the authors please elaborate? Is there a minimum level of divergence beyond which phase error become more frequent.

We chose to focus Paraphase on SDs of at least 10 kb in size and >99% sequence identity because these are the regions where aligners could align reads incorrectly (i.e. the homology region is comparable with the length of a HiFi read). For SDs that do not meet the criteria, genes and paralogs are likely to be different enough from each other so that there won't be alignment problems and variant calling using a standard HiFi workflow is sufficient, and thus we do not think that Paraphase is needed in these regions.

3.1) I am not myself a fan of simulated data, but it is my opinion that the sensitivity and specificity of pharapase in reconstructing haplotypes should somehow be measured. How do these parameters change w.r.t 1) reads size? 2) level of divergence of the genes? Maybe the authors should perform some controlled analyses with simulated data (or in silico trimmed/treated real data) to address these issue and provide some estimates

> We performed a simulation experiment following the suggestion (Supplementary Fig. 4 below) to examine the effects of read length, read depth and the divergence level on haplotyping accuracy. The simulation result indicates that at standard WGS parameters (10-15 kb read length and 10-15X haplotype depth) and when paralogs are reasonably diverged from each other (>0.05% divergence, i.e. 1 SNP every 2 kb), Paraphase can achieve high accuracy. Phasing accuracy could be impacted at short read length (5 kb), lower haplotype depth (5X) or low paralog divergence level (0.01%, i.e. 1 SNP every 10 kb). Real world data is of course more complicated, particularly by the fact that paralog divergence can be variable from region to region. There could exist local regions that are very scarce in base differences between haplotypes and are thus more vulnerable to lower read depth and shorter read length. We have added this simulation result into Supplementary Notes.



Supplementary Fig. 4. Paraphase performance on simulation data of varied read length (baseline 15 kb), haplotype depth (baseline 15X) and paralog divergence (baseline 0.15%). Paraphase calls were examined for variant recall and precision at varied read lengths of 5 kb, 10 kb and 15 kb (A), varied haplotype depths of 5X, 10X and 15X (B) and varied paralog divergence of 1, 5, 10 and 15 SNPs per 10 kb (0.01%, 0.05%, 0.1% and 0.15%, respectively) (C), while the other two variables are kept at baseline for each panel.

3.2) In the section "validation of paraphase calls" an error rate of 0.29% is reported when considering the consistency of reconstructed haplotypes through the analysis of family trios. What was the possible cause of these errors? Could the authors please elaborate? How are errors set apart from - for example- the 7 de novo mutations in the section "Identification of de novo mutations and gene conversion"?

> These errors were identified by manual examination of the Paraphase-called haplotypes and their supporting reads. We defined an error as a haplotype that is not fully supported by the reads. The possible causes of the errors are switch errors, i.e. segments from two haplotypes are incorrectly joined into one haplotype or cases where Paraphase didn't identify all of the haplotypes. Conversely, the de novo mutations were supported by the reads but were not consistent between the parents and child. We have added more details to this part of the section in the main text (Line 141, Page 5):
>
> "Upon examining the 55 inconsistent cases, 43 (0.29%) are not fully supported by reads and thus determined as Paraphase errors (e.g. switch errors or missed haplotypes in the parent). The remaining 12 (0.081%) inconsistent haplotypes are fully supported by reads, and thus represent true recombination or *de novo* events (See "Identification of *de novo* mutations and gene conversion" section)."

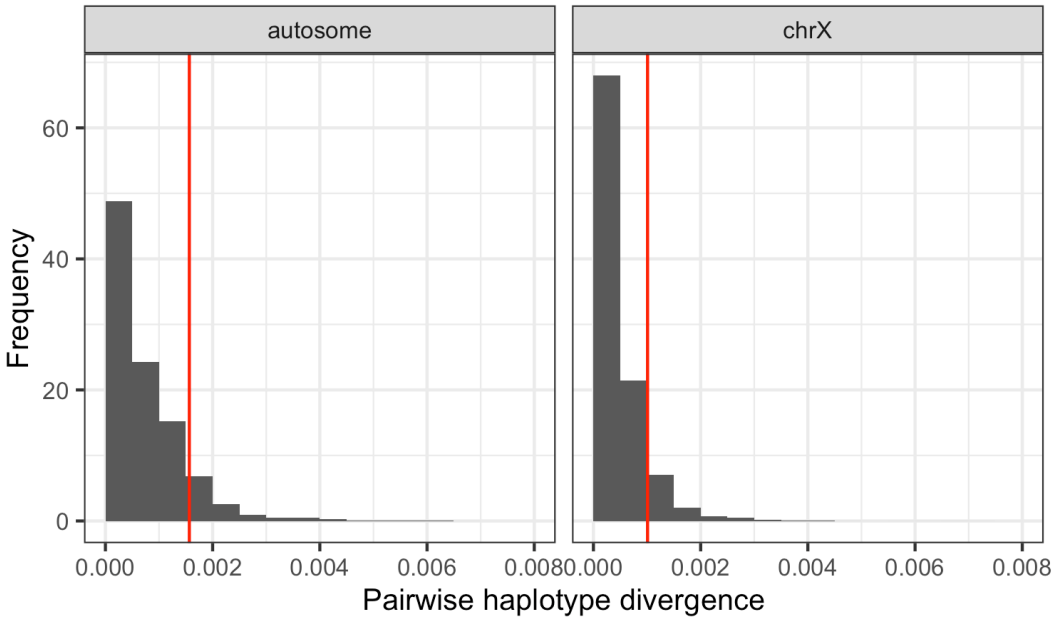
4) In my opinion figure, also by looking to figure 2A the authors use a very lenient definition for the identification of genes with higly variable CN (or alternatively a very stringent definition for low variability). Genes like NSF, HIC2, H3C14 for example do not show a highly variable copy number in my eyes. Probably the stratification should have a higher granularity. Intuitively at least classes should be formed: 1 invariable; 2; moderately variable; 3 highly variable.

> We have updated Figure 2a to three classes as suggested by the reviewer: low variability (>90% of individuals have the mode CN), medium variability (80-90% of individuals have the mode CN) and high variability (<80% of individuals have the mode CN).

5) If I understood correctly, "gene families" with low diversity are defined as those on having a variability below the less variable 10% of a randomly selected sample of 600 genes. A total of 23 (out of 160) variable gene families are identified by this approach. It should be noted that 23 is circa 14% of 160. This does not seem like a big enrichment w.r.t the expected number. Methods applied/used in this analysis are not adequate in my opinion, due to several reasons:

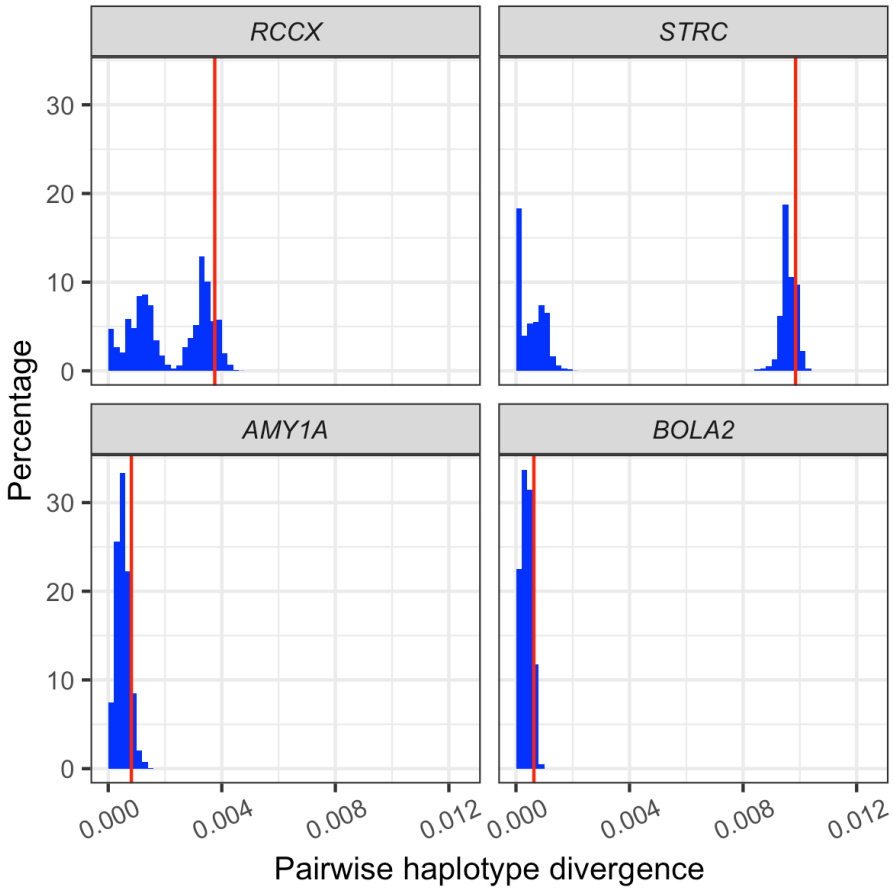
> It's clear that the analysis we performed needs to be clarified. We have improved the Methods section to make it more clear (Line 484, Page 14). The goal of this analysis is to find the expected divergence level for alleles of genes that do not have paralogs, and use this cutoff to identify paralog groups where the divergence between genes (between a gene and its paralog) is similar to the divergence within genes (between an allele and another allele of the same gene).

To get the background distribution, we look at genes without paralogs. For each gene we calculated the pairwise divergence between any two haplotypes of the same gene in the population. We merged these divergence values across 400 randomly selected autosomal genes (or 200 chrX genes) into one distribution. To get our cutoff value, we then identified the 90% percentiles of these distributions (shown by the red vertical lines in Supplementary Fig. 17, copied below). This gives us the expectations for the divergence we expect to see in single genes. Then we turn to paralog groups targeted by Paraphase and want to find paralog groups where, on average, the haplotypes observed in a gene and its paralog are as similar as two haplotypes from the same gene, by comparing the 90% percentiles of the pairwise divergence values to the cutoff values.



Supplementary Fig. 17. Distributions of allelic sequence divergence for genes without paralogs.

For each paralog group, we calculated the distribution of the divergence values between any two haplotypes of the same paralog group analyzed by Paraphase. For example, in the *SMN1*/*SMN2* paralog group, we calculated pairwise divergence between any two haplotypes, which could be *SMN1* or *SMN2*. This gives us a distribution for each paralog group (examples can be found in Supplementary Fig. 9, copied below). When the gene and its paralog diverge significantly (e.g. RCCX and *STRC* shown below, upper panel), we see two peaks corresponding to gene-gene (paralog-paralog) divergence and gene-paralog divergence, and the 90% percentile (shown by the red vertical lines) is naturally much higher than the background expectation shown in Supplementary Fig. 17. When the gene and its paralog are very similar (e.g. *AMY1A* and *BOLA2* shown below, lower panel), the 90% percentile of the pairwise divergence is similar to the background levels. We define paralog groups with low within-group diversity as those whose 90% percentile is lower than the 90% percentile of the background distributions. It should be noted that this is a conservative approach and is not expected to identify all paralog groups with low within-group diversity.



Supplementary Fig. 9. Histograms of pairwise sequence divergence for paralog groups with differentiated genes and paralogs (upper panel) vs. paralog groups with low within-group diversity (lower panel).

5.1) As far as I understood substitutions rates were computed as flat substitution rate on the complete gene body. Obviously coding exons, introns and UTRs have different evolutionary rates. This should be kept into account. Otherwise you will simply select the gene with the longest introns as the most variable.

The pairwise divergence was indeed calculated on the complete gene body. Our goal is to identify paralog groups where genes and paralogs are highly similar, indicating that the high sequence similarity is being maintained by some forces like gene conversion. In this initial analysis, the paralog groups that we identified are highly similar throughout the gene body. In the Discussion (Line 347, Page 10, copied below) we have pointed out that future analyses could focus on identifying more local, smaller regions of a gene impacted by gene conversion and one can study the differences between exons, introns and UTRs as suggested by the reviewer. To fully quantify the differences between introns and exons, we expect that much larger sample numbers will be needed.

“Beyond paralog groups with low within-group diversity throughout the entire gene body, one future direction is to extend this analysis to identify local low-diversity regions resulting from

gene conversions, such as the gene conversion found in Exons 1-6 of *SMN1/SMN2* (Supplementary Fig. 8), and Exon 15 of *PMS2* (Figure 5c, Supplementary Fig. 13)."

5.2) 10% is a very lenient cut-off. Probably a more stringent cut-off should be used.

Please see our response to your comment #5 above.

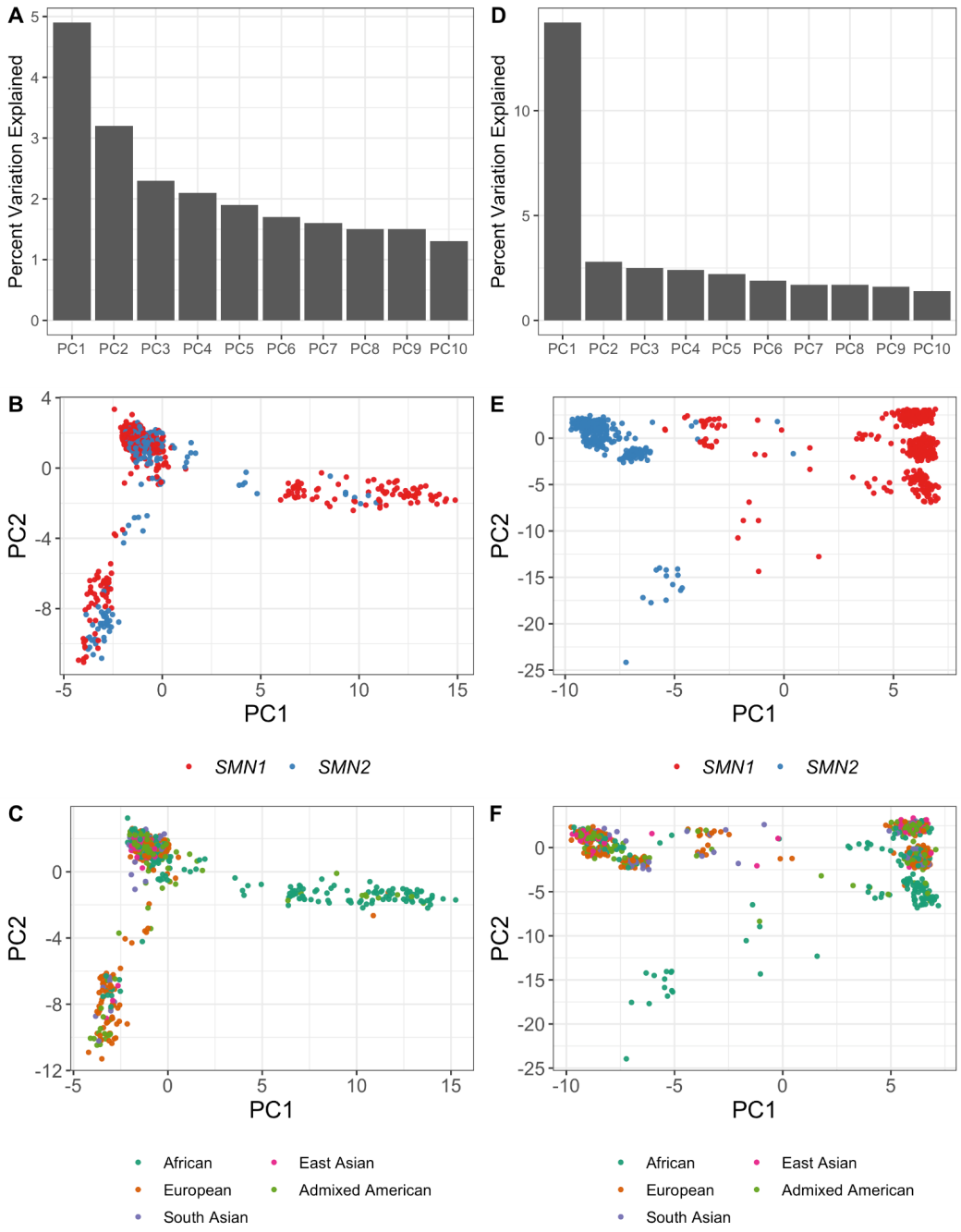
5.3) Probably duplicated genes and recently duplicated genes have (very) different selective pressure. It should probably better to stratify lowly variable duplicated genes by examining duplicated genes in isolation.

We agree that the age of the duplication is an important factor to consider. We list two different types of highly similar paralog groups in Table 2. The highly similar paralog groups that we found on sex chromosomes are all conserved in other primates, suggesting that these are not recent duplication events and gene conversion could be a main force maintaining the high similarity between paralogs. On the other hand, most of the highly similar paralog groups that we found on autosomes are human-specific duplications and thus could be recent. For these genes, it could be a mixed effect of recent duplication, positive selection and gene conversion. Dissecting between these potential mechanisms could be a direction for future studies.

6) Figure 6 B,C,D. Why was a PCA used? What proportion of the variance is explained by the first 2 principal components. Where additional components examined? Figure 2B: pink dots on the top left seem to be well separated from the light blue dots. Since this PCA includes 3 juxtaposed genes, could the authors repeat the analysis by considering one gene at a time, and demonstrate that the above mentioned pink dot do not associate with only one of the genes. (That would be an interesting observation)

We chose to use PCA because it provides a nice visual representation of whether haplotypes from a paralog group fall into distinct clusters (corresponding to different genes) based on their sequences. Each dot is the sequence of a haplotype from a paralog group. For a normal pair of gene and paralog, there should exist distinct paralogous sequence variants (PSVs) between them so they can be easily separated into distinct clusters on a PCA plot (see Supplementary Fig. 8E, copied below, for Exons 7-8 of *SMN1/SMN2* or Supplementary Fig. 13B for Exon 12 of *PMS2/PMS2CL*).

In Figure 3b we are plotting the haplotypes in PCA space for each of the genes (*AMY1A*, *AMY1B*, and *AMY1C*) and the dots are colored based on the gene (*AMY1A*=green; *AMY1B*=orange; *AMY1C*=purple). Thus the "pink dots" (orange) all come from the same gene (*AMY1B* in this case). Haplotypes of the three genes do not form distinct clusters, but overlap with each other, indicating that the three genes are not distinguishable in sequence. There undoubtedly exist some slight differences in the haplotypes (due to population history) for these genes, but this difference is not striking. The PCs and their proportion of variance explained can be found in Supplementary Fig. 10. The proportion of variance explained is low for each PC, indicating that all the sequences examined here are highly similar to each other.



Supplementary Fig. 8. Principal component analysis (PCA) of Exons 1-6 and Exons 7-8 sequences in *SMN1/SMN2*. (B) *SMN1* and *SMN2* are indistinguishable in sequence in Exons 1-6. (E) *SMN1* and *SMN2* are differentiated in sequence in Exons 7-8. (Legend for other panels are omitted here.)

Reviewer #3 (Remarks on code availability):

The code avaialable through a dedicated github repository and is publicly accessible, the documentation is complete and clear. The tool can be easily installed and executed.