Cell Genomics Transparent Peer Review Record

Title

Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq

Vahid Akbari[1,2+], Vincent C.T. Hanlon[3+], Kieran O'Neill[1], Louis Lefebvre[2], Kasmintan A. Schrader[2,4], Peter M. Lansdorp[2,3*] & Steven J.M. Jones[1,2*]

Summary

Initial submission:	Received : 5/20/2022
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 3 Revision invited : 7/8/2022 Revision received : 9/8/2022
Second round of review:	Number of reviewers: 2 Accepted : 11/29/2022
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1:

General Comments:
This study by Akbari et al presents a valuable tool in the ability to haplotype phase nanopore long reads, and then assign them to maternal and paternal haplotypes in the absence of parental trio data, for human genomes. To do this, the authors first assign long reads to haplotype 1 and 2 bins using short reads from Strand-Seq data, a technique that sequences Watson and Crick strands separately; second, they assigned those reads to specific chromosomes by mapping them back to the GRCh38 human reference; and third, they use imprinted differential parental methylation information imprinted captured in the long read data to assign which haplotype belongs to which parent. For the parental assignment, this depends on known differentially methylated regions of each chromosomes from each the mother and father that is mostly conserved differences between the sexes in the human population. Since these parental origin methylation differences do not cover not the entire chromosome, assembly and phasing needs to be complete before parental assignment.

Overall, I think this is an important advance that needs to be told. But, I do think the authors present the story in a manner that implies a bigger advance than what was done. The figures and some of the text implies that the nanopore reads were assembled de novo. This was not done, although this could be possible. Instead GRCh38 was used as a reference to bring reads together to the same chromosomes. So this is close to a reference based assembly, although no assembly was conducted, as far as I can tell. This needs to be made clearer.

Also the paper needs to make clearer that methylation data alone in the absence of pre-phasing won't work, due to a sparser parental differential methylation along chromosomes. In the discussion, one possible approach that they should mention is using an approach like hifiasm Hi-C (citation 3) to create a denovo fully phased assembly, and then use nanopore or pacbio methylation data to assign chromosome parental origin. Another potential approach tested in the Human Pangenome Reference Consortium (HRPC) preprint by Jarvis et al 2021, is using Strand-Seq to help with a haplotype phased denovo long read assembly, and then as in this study to use the methylation data to assign parental origin. I think it will be important to mention this HPRC paper and Wang et al Nature 2022 for the source of the HG002 and HRPC data used in the current study, as well as the some of the predictions about StrandSeq demonstrated in the current study.

Specific Comments:
Need line numbers to make reviewing easier.

Differential parental methylation won't be lost in cell culture? Should state this specifically. Should also briefly explain how differential parental methylation is achieved in an organism.

In the introduction, need to explain the trio approach for haplotype phasing and parental assignment. The authors assume that the readers will know about this, when this is not the case.

Are figure 1b and c cartoon examples of what the final phasing and parental assignments should look like, or are they real genome result? This needs to be specified. If real, then is there redundancy with Figure 2?

In figure 2, need to explain what do the color differences mean for the bars - red, black, and gray.

The authors should compare the SNV and indel variant call accuracy achieved here with Strand-Seq phasing and the HG002 assembly in the HPRC bioRxiv study.

The 61.3% of the indels that match the ground truth dataset is low, compared to over 98% in the HG002 assembly from the HRPC. This is because of the nanopore bias of making indel errors. Doesn't this error rate impact that mapping of the reads to GRCh38? All the more reason that the authors need to mention some alternatives in the discussion as to how to get a functionally useful parental assigned genome assembly without such high error error rates. This can be done with polishing the Nanopore data with Illumina (hinted at) or Pacbio HiFi reads, or using HiFi reads in place of nanopore reads of methylation and assembly.

In the discussion, should mention what would be needed to apply the approach of this study to non-human species.

Methods, should cite Jarvis et al bioRxiv 2022 and Wang et al Nature 2022 for the HPRC papers and source of some of the data. Citation 42 is no an HRPC study.

Reviewer #2: In this article, Akbari et al. have developed a chromosome-scale haplotyping method that can consistently discern paternal and maternal haplotypes by integrating known imprinted differentially methylated regions (iDMRs) with Nanopore DNA sequencing data, Nanopore-derived DNA methylation information, and Strand-seq data. The authors have identified a way of assigning parent-of-origin to haplotypes. However the bulk of the work described in this manuscript is not innovative as it relies heavily on several key publications and datasets from the last 2 years. Garg et al. (2021) and Porubsky et al. (2021) both established chromosome-level phasing using PacBio long-read sequencing and they successfully resolved the H1 and H2 haplotypes. In this manuscript, Akbari et al. take this one step further by assigning maternal and paternal identities to H1 and H2, utilizing some of the very same datasets Garg and Porubsky generated. Akbari et al. take Nanopore DNA sequencing data, Nanopore-derived DNA methylation information, and Strand-seq data and integrated it with a well-validated set of known iDMRs to come up with their new software for assigning parent-of-origin to chromosome-level hapolotyes. It is important to note that this pipeline relies on a methylation phasing tool Akbari et al. developed in 2021 called NanoMethPhase which phases reads and CpG methylation info from Nanopore sequencing data, with help from trio datasets.

References:

Garg et al. Nature Biotechnology volume 39, pages 309-312 (2021)
Chromosome-scale, haplotype-resolved assembly of human genomes

Porubsky et al. Nature Biotechnology volume 39, pages 302-308 (2021)
Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads

Akbari, V., et al. Genome Biol 22, 68 (2021).
Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase.
https://doi.org/10.1186/s13059-021-02283-5

Minor comments

Akbari et al.'s methodology uses a lot of arbitrary parameters. It would be helpful for the authors to try to justify their parameter criteria for at least some of the critical analytical steps in determining iDMR and parent of origin.

Reviewer #3: The authors present an elegant method to obtain chromosome-length haplotypes, assigned as paternal or maternal, from nanopore long-read sequencing combined with Strand-seq in a single individual (obviating the need for trio sequencing). Accuracy is demonstrated by the correct haplotype assignment of all autosomes for 5 individuals. Supported by a detailed tutorial on GitHub, this powerful method is likely to be of broad interest.

There is a risk of circular reasoning, because the same cell lines used to define the iDMRs (Abkari et al. bioRxiv 2021) are used to validate the mat/pat haplotyping method. So it is possible that "spurious" DMRs specific to those samples are not generalisable to new samples (and cell types). The authors do take care in validating the DMRs by alternative means, but it is not bulletproof (partial methylation in WGBS/array can have various explanations besides imprinting). This limitation should at least be acknowledged and discussed. Ideally the authors could validate the technique on samples that didn't form part of the "training set".

Minor comments:
- The results section is terse. Some of the results that stand out in the figures (chr9 centromeres in a few samples, the inversion on chr8 in HG002) are not mentioned at all in this section, and only briefly alluded to in the discussion. They would benefit from additional explanation. Can some of the differences in SNV+indel phasing accuracy between samples be explained by nanopore read lengths, or Strand-seq depth?
- Plots of iDMRs locations and contributions (Fig2 and supps), although aesthetically pleasing by their symmetry, are not the easiest to read. In addition to the "1"s pointing the wrong way for Haplotype 1, I was repeatedly mixing up the start and end of the chromosomes, trying to match Haplotype 1 and 2 DMRs. It would be easier to keep the normal orientation for Haplotype 1.
- As noted, some imprinted DMRs vary between tissues and/or individuals. Can anything be said about the cell types used in this study, and how different cell sources might influence the efficacy of the approach?
- The applicability to other species would be limited by the distribution of iDMRs. In mice, not all autosomes harbour iDMRs.
- Could the method be extended to other haplotype markers (e.g. Hi-C) and long reads (e.g. Pacbio, since it's now better at 5mC detection)?
- in the Discussion, it is chromosome 3 that has only 2 DMRs, rather than chromosome 2.
- the inversion for HG002 chromosome 8 should refer to Supp Fig 13 rather than 12.

Authors' response to the first round of review

Reviewer #1
General Comments:

This study by Akbari et al presents a valuable tool in the ability to haplotype phase nanopore long reads, and then assign them to maternal and paternal haplotypes in the absence of parental trio data, for human genomes. To do this, the authors first assign long reads to haplotype 1 and 2 bins using short reads from Strand-Seq data, a technique that sequences Watson and Crick strands separately; second, they assigned those reads to specific chromosomes by mapping them back to the GRCh38 human reference; and third, they use imprinted differential parental methylation information imprinted captured in the long read data to assign which haplotype belongs to which parent. For the parental assignment, this depends on known differentially methylated regions of each chromosomes from each the mother and father that is mostly conserved differences between the sexes in the human population. Since these parental origin methylation differences do not cover not the entire chromosome, assembly and phasing needs to be complete before parental assignment.

Overall, I think this is an important advance that needs to be told. But, I do think the authors present the story in a manner that implies a bigger advance than what was done. The figures and some of the text implies that the nanopore reads were assembled de novo. This was not done, although this could be possible. Instead GRCh38 was used as a reference to bring reads together to the same chromosomes. So this is close to a reference

based assembly, although no assembly was conducted, as far as I can tell. This needs to be made clearer.

We thank Reviewer #1 for their detailed review of this manuscript. As they imply, trio-free chromosome-length phasing has been described in several studies in recent years (typically using Hi-C or Strand-seq): our contribution is a method that determines, for each chromosome, which haplotypes correspond to the maternally- and paternally-inherited homologs (as shown in Figure 1). We have tried to emphasize this to address the concern that we have exaggerated the novelty of the work. In particular, we partially reworded the first paragraph of the introduction to better explain current phasing practices. We have also supplied some missing citations in the caption for Figure 1, which should help convey that chromosome-length phasing itself is not novel.

We did not do de novo genome assembly at all for this manuscript, and we now state this explicitly in the Discussion to make it clearer to the reader. One reason for this is that genome assembly takes longer computationally, which means longer turnaround times that might impact the clinical applications of this method. Instead, we called variants by aligning reads to a reference genome, and then we phased reads and variants using Strand-seq reads (also aligned to a reference genome).

Also the paper needs to make clearer that methylation data alone in the absence of pre-phasing won't work, due to a sparser parental differential methylation along chromosomes. In the discussion, one possible approach that they should mention is using an approach like hifiasm Hi- C (citation 3) to create a denovo fully phased assembly, and then use nanopore or pacbio methylation data to assign chromosome parental origin. Another potential approach tested in the Human Pangenome Reference Consortium (HRPC) preprint by Jarvis et al 2021, is using Strand- Seq to help with a haplotype phased denovo long read assembly, and then as in this study to use the methylation data to assign parental origin. I think it will be important to mention this HPRC paper and Wang et al Nature 2022 for the source of the HG002 and HRPC data used in the current study, as well as the some of the predictions about StrandSeq demonstrated in the current study.

The comment about the necessity of pre-phasing, since iDMRs are scarce, is a very helpful one. We have included text to this effect in the Results:

"Before iDMRs can be used to assign PofO to homologs, chromosome-length haplotypes must be constructed. This is because iDMRs cover only a small fraction of the autosomal genome (estimated in this study to be 0.014%) and are not necessarily phased relative to variants"

Trio-free diploid assemblies, produced with long reads combined with Hi-C or Strand-seq, could certainly be a substitute for our chromosome-length variant haplotypes, and they could be assigned PofO using methylation information at iDMRs in the same way. We now describe this possibility in the Discussion (see below).

The Jarvis et al. 2022 bioRxiv preprint1 demonstrates an updated diploid genome assembly method, which necessarily creates chromosome-length haplotypes that could in theory be used as a starting point for trio-free parent-of-origin assignment (although the Jarvis et al. preprint did not do so, nor discuss any detection of methylation or its use). In the Discussion, we now mention the possibility of adapting the PofO phasing method described in our paper to diploid genome assembly. We also added the citations to Jarvis et al. and Wang et al. 20222 as recommended, to provide other methods of establishing chromosome level haplotypes along with our own method, described by Falconer et al.3.

Specific Comments:
Need line numbers to make reviewing easier.

We added these.

Differential parental methylation won't be lost in cell culture? Should state this specifically. Should also briefly explain how differential parental methylation is achieved in an organism.

In this study, we use cell-line-derived-nanopore data and we do identify differential methylation at the iDMRs established in both cell-line and non-cell-line studies—this is what allowed us to call PofO correctly for all autosomes (verified by trio data). This demonstrates that cell culture does not remove iDMRs. To clarify, we did not perform cell culture for the samples in which methylation was detected by nanopore (although presumably Coriell did). The cell culture done for Strand-seq is for haplotyping purposes only, which has no bearing on methylation detection.

In future, we plan to use fresh samples for nanopore and PofO phasing (e.g., blood as per the Introduction). As per the reviewer's suggestion, we now state in the Introduction that differential methylation is still detectable in cell lines.

We have also added some text to the introduction to briefly explained how iDMRs come to be (new text underlined):

"A striking exception to this paradigm is the parental information provided by consistent differences in DNA methylation between maternally- and paternally-inherited alleles at imprinted differentially methylated regions (iDMRs). This differential methylation is either established in gametes and escapes the epigenetic reprogramming that follows fertilization or it is established after fertilization5,6, and it persists through adulthood. iDMRs reliably suppress the expression of either maternal or paternal alleles at nearby genes or gene clusters and, crucially, can be detected in cell lines or fresh samples using the unique ion current signature of 5-methyl-cytosine by nanopore sequencing (Oxford Nanopore Technologies)7–10."

In the introduction, need to explain the trio approach for haplotype phasing and parental assignment. The authors assume that the readers will know about this, when this is not the case.

This is a good point. We have added a sentence to the Introduction explaining how trio phasing works:

"These haplotypes can only be assigned PofO if trio information of some kind is available: for example, by comparing previously-discovered alleles for the mother with the child's alleles at heterozygous loci"

Are figure 1b and c cartoon examples of what the final phasing and parental assignments should look like, or are they real genome result? This needs to be specified. If real, then is there redundancy with Figure 2?

Figure 1b is a cartoon, and we now specify this in the caption.

In figure 2, need to explain what do the color differences mean for the bars - red, black, and gray.

We have added this to the figure legend. The black, grey, and white regions on the chromosomes are from a schematic of Giemsa banding (G-banding), which targets AT-rich DNA. The red regions represent centromeres.

The authors should compare the SNV and indel variant call accuracy achieved here with Strand- Seq phasing and the HG002 assembly in the HPRC bioRxiv study.

Jarvis et al's bioRxiv preprint demonstrates some very exciting and much-needed advances towards creating haplotype-aware, full-length reference genomes. However, the goal of their study, while superficially similar to ours, is actually quite different. Whereas they are attempting to create a set of pangenome human reference assemblies and variant benchmarks by combining a broad array of state-of-the-art technologies, we are attempting to demonstrate a very specific potential clinical application, that of tracing variants to a parent of origin solely from a sample taken from the proband. While we are sure that their downstream results will be a useful source of comparison in the form of updated GIAB genome benchmarks, it is difficult to see the utility or appropriateness of a comparison between the studies themselves. For now, F1 scores for the GIAB samples against the current GIAB benchmark are shown in Supplementary Figure 3 of this manuscript. The SNV F1 score for HG002 is 99.6%, comparable to the 99.7% that Jarvis et al. report. The indel F1 score is 78.2%, less than the 98.6% in Jarvis et al., presumably because indel calling is less successful using nanopore data.

The 61.3% of the indels that match the ground truth dataset is low, compared to over 98% in the HG002 assembly from the HRPC. This is because of the nanopore bias of making indel errors. Doesn't this error rate impact that mapping of the reads to GRCh38? All the more reason that the authors need to mention some alternatives in the discussion as to how to get a functionally useful parental assigned genome assembly without such high error error rates. This can be done with polishing the Nanopore data with Illumina (hinted at) or Pacbio HiFi reads, or using HiFi reads in place of nanopore reads of methylation and assembly.

Nanopore does indeed have trouble calling indels (although notably, SNV calling accuracy is now comparable to Illumina data4), but this is unlikely to result in poor mapping of reads to GRCh38. The length of the reads gives ample information for the aligner to map them correctly despite small errors. In the case of indels, they are by definition smaller than 50bp, with the actual error in Nanopore reads usually only a few bases inside homopolymer runs, whereas the reads themselves are three orders of magnitude longer. The nanopore data for the 5 individuals from this study had mapping rates above 98.78% (mean 99.55%; mean 71.7% with MAPQ at least 20). In general, high mapping rates for nanopore reads is reflected in their much greater structural variant calling accuracy compared with short-read technologies5.

We have also added a paragraph to the discussion describing how alternative technologies could be used for PofO phasing. Again, we did not do any genome assembly for this work, but we agree that this could be a viable approach:

"Other sequencing technologies could perhaps provide the DNA methylation, DNA sequence, and long-range phase information required for PofO phasing, or different methods could be used to combine them. For instance, although we did not perform genome assembly for this study, PofO could be assigned to de novo trio-free diploid assemblies15,43 rather than chromosome length haplotypes of small variants. SMRT sequencing (Pacific Biosciences), which now provides accurate DNA methylation as well as DNA sequence44, might be a substitute for nanopore data that provides better indel detection; and long-range phasing

with Hi-C43 could perhaps be used instead of Strand-seq, although if phase switches occur at centromeres16 then chromosome arms that lack iDMRs (16q, 17q, 18p, and 20p) may not be assigned to a PofO."

In the discussion, should mention what would be needed to apply the approach of this study to non-human species.

This is an excellent suggestion, and we have done so (new text underlined):

"Moreover, our approach can potentially be expanded to other mammals. DNA methylation-based (canonical) imprinting has been described in all placental mammals, and genomic maps of iDMRs have been established for a number of species, notably mice and primates6–9. This would require adjusting cell culture and flow cytometry conditions for Strand-seq library preparation to suit non-human cells10, and PofO assignment would be limited to chromosomes with known iDMRs (e.g., not chromosomes 4, 5, 13, 14, 16, and 19 in mice)9."

Methods, should cite Jarvis et al bioRxiv 2022 and Wang et al Nature 2022 for the HPRC papers and source of some of the data. Citation 42 is no an HRPC study.

We now cite Jarvis et al. 2022 (bioRxiv), and we replaced the (former) citation 42 with Wang et al. 2022 (Nature) for the HPRC data.

Reviewer #2
In this article, Akbari et al. have developed a chromosome-scale haplotyping method that can consistently discern paternal and maternal haplotypes by integrating known imprinted differentially methylated regions (iDMRs) with Nanopore DNA sequencing data, Nanopore-derived DNA methylation information, and Strand-seq data. The authors have identified a way of assigning parent-of-origin to haplotypes. However the bulk of the work described in this manuscript is not innovative as it relies heavily on several key publications and datasets from the last 2 years. Garg et al. (2021) and Porubsky et al. (2021) both established chromosome-level phasing using PacBio long-read sequencing and they successfully resolved the H1 and H2 haplotypes. In this manuscript, Akbari et al. take this one step further by assigning maternal and paternal identities to H1 and H2, utilizing some of the very same datasets Garg and Porubsky generated. Akbari et al. take Nanopore DNA sequencing data, Nanopore-derived DNA methylation information, and Strand-seq data and integrated it with a well-validated set of known iDMRs to come up with their new software for assigning parent-of-origin to chromosome-level hapolotyes. It is important to note that this pipeline relies on a methylation phasing tool Akbari et al. developed in 2021 called NanoMethPhase which phases reads and CpG methylation info from Nanopore sequencing data, with help from trio datasets.

We thank Reviewer #2 for their review of our manuscript. We do indeed build on our previous work for this study: as outlined in Figure 1, and as the reviewer correctly remarked, the novelty of the method lies in determining which chromosome-length haplotype originated from which parent, without using trio information. We have added more citations to previous chromosome-length haplotyping methods to emphasize more strongly what is novel and what is not (notably in the caption for Figure 1). For the same reason, we reworded the first paragraph of the introduction to better explain the current state of phasing methods.

References:

Garg et al. Nature Biotechnology volume 39, pages 309-312 (2021) Chromosome-scale, haplotype-resolved assembly of human genomes

Porubsky et al. Nature Biotechnology volume 39, pages 302-308 (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads

Akbari, V., et al. Genome Biol 22, 68 (2021). Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. https://doi.org/10.1186/s13059-021-02283-5

Minor comments
Akbari et al.'s methodology uses a lot of arbitrary parameters. It would be helpful for the authors to try to justify their parameter criteria for at least some of the critical analytical steps in determining iDMR and parent of origin.

This is a good suggestion. At several points in the STAR Methods, we have included more justification for these parameters or thresholds. For instance, we replaced this:

"Only iDMRs with more than 10 detected CpGs and with |α(HP1) - α(HP2)| comprising at least 10% of all detected CpGs were considered for PofO assignment"

With this:

"The length of an iDMR can vary between individuals, and different studies report different start and end positions for the same iDMR. We wished to capture PofO information even when just a small part of an iDMR is imprinted in an individual, while avoiding inferences based on very few CpGs. Therefore, we only used iDMRs with |$\alpha HP1 - \alpha HP2$| comprising at least 10% of all detected CpGs and with more than 10 detected CpGs in total (i.e., $\frac{|\alpha HP1 - \alpha HP2|}{n} > 0.1$ and $n > 10$)."

Reviewer #3
The authors present an elegant method to obtain chromosome-length haplotypes, assigned as paternal or maternal, from nanopore long-read sequencing combined with Strand-seq in a single individual (obviating the need for trio sequencing). Accuracy is demonstrated by the correct haplotype assignment of all autosomes for 5 individuals. Supported by a detailed tutorial on GitHub, this powerful method is likely to be of broad interest.

We thank Reviewer #3 for their thoughtful review of this manuscript.

There is a risk of circular reasoning, because the same cell lines used to define the iDMRs (Abkari et al. bioRxiv 2021) are used to validate the mat/pat haplotyping method. So it is possible that "spurious" DMRs specific to those samples are not generalisable to new samples (and cell types). The authors do take care in validating the DMRs by alternative means, but it is not bulletproof (partial methylation in WGBS/array can have various explanations besides imprinting). This limitation should at least be acknowledged and discussed. Ideally the authors could validate the technique on samples that didn't form part of the "training set".

This is a very important point. Roughly 20% of the iDMRs used for PofO phasing were observed only in previous work by Akbari et al. (now published in eLife 11) using the same cell lines. Although we attempted to validate these iDMRs by checking for partial methylation in WGBS data for other samples, we had overlooked the circularity that

Reviewer #3 correctly identified: namely, that the primary evidence that these loci are imprinted comes from a comparison with the parents' variant callsets, which means that they are likely to support the correct PofO in these five samples even if they turn out to be spurious iDMRs.

We now discuss this possibility at length in the text. Additionally, we evaluated the extent to which PofO phasing relies on iDMRs reported by just one study (which we consider to be less reliable), including the iDMRs reported only by Akbari et al.'s previous work. To do this, we reran PofO assignment using only iDMRs reported by at least two studies. No homologs were misassigned, and only 5.5% of homologs could not be assigned PofO.

Our discussion of this problem in the text is as follows:

"Roughly half of the iDMRs [used for PofO assignment] were reported in at least two studies, while the rest were reported in just one study and confirmed by partial methylation observed among 179 WGBS datasets from 119 blood and 60 tissue samples (see Methods). One potential weakness of the single-study iDMRs is that 38 of them (19.8%) come from a previous study of 12 trios that included the same five trios examined here11, and it is possible that some of these iDMRs might provide misleading or insufficient PofO information when examined in new individuals (i.e., if they are not truly imprinted). We tested the dependence of PofO phasing on the single-study iDMRs by re-running PofO assignment using only the 93 iDMRs found in at least two studies: 208 of 220 autosomal homologs were correctly assigned PofO (94.5%), while chromosome 5 was not assigned PofO for NA19240 and chromosome 12 was not assigned PofO for any individual because it did not have an iDMR. This suggests that PofO phasing is not reliant on poorly characterized iDMRs, likely because all autosomes have at least three iDMRs, with the exception of chromosome 17 which has one and chromosome 3 which has two. This redundancy helps maintain robust PofO assignment even when some putative iDMRs provide weak or conflicting parental information. At a few iDMRs in some samples, for example at TRPC3 on chromosome 4 in NA12878, we detected hypermethylation on the parental allele that is reported to be unmethylated: this might be due to inaccuracies in methylation calling or phasing of nanopore reads, or it could reflect random allelic methylation rather than imprinting. Nevertheless, additional iDMRs on the same chromosome enabled correct PofO assignment, albeit with the lowest confidence score (58.6%).

[...] ultimately the method must be tested on additional trios from diverse genetic populations to determine which chromosomes are troublesome for PofO phasing. Advances in characterizing human DNA methylation may further improve PofO phasing by identifying additional iDMRs on iDMR-poor chromosomes (e.g., chromosome 17), by removing spurious iDMRs, and perhaps even by enabling PofO assignment for the X chromosome in females12."

Minor comments:
- The results section is terse. Some of the results that stand out in the figures (chr9 centromeres in a few samples, the inversion on chr8 in HG002) are not mentioned at all in this section, and only briefly alluded to in the discussion. They would benefit from additional explanation. Can some of the differences in SNV+indel phasing accuracy between samples be explained by nanopore read lengths, or Strand-seq depth?

We have added some text describing these local phasing errors (visible in the Mendelian error figures) to the Results:

"Local phasing errors are indicated by elevated Mendelian error rates at a large common

inversion on chromosome 8 for HG002 (mismatch rate 99.86% for SNVs and 97.05% for indels inside the inversion at chr8: 8120810-12362538), which is the individual with the most phasing errors overall (SNV mismatch rate 0.54%; Table 1), as well as at the centromere for chromosome 9. The latter is in fact a single bin of 1000 SNVs stretched across the centromere, which exaggerates its importance in Supplementary Figure 13."

In general because there are only 5 individuals in this study, it is hard to attribute variation among samples. Moreover, differences in phasing error rates are dominated by the chromosome 8 inversion in HG002, which is not related to Strand-seq depth or nanopore read length. But we notice, not unexpectedly, that there is a clear trend for the number of variants recovered for a sample: samples with better nanopore coverage recovered more indels and SNVs relative to the ground truth datasets. This is especially significant for indel discovery, which as another reviewer noted has a high false negative rate with nanopore reads. We have added some text to this effect to the Results:

"The SNV callsets for each individual included nearly all SNVs in the five corresponding ground truth callsets (M=97.98%, SD=1.67%, range=95.51%-99.64%; "M" mean, "SD" standard deviation; Table 1), while fewer indels were recovered (M=64.01%, SD=8.43%, range=52.69%- 78.18%). For both SNVs and indels, we recovered the greatest proportion of ground truth variants in the individual with the greatest nanopore coverage, while we recovered the smallest proportion in the individual with the least coverage. This suggests that including more nanopore data would be one way to address the high false negative rate for indels."

- Plots of iDMRs locations and contributions (Fig2 and supps), although aesthetically pleasing by their symmetry, are not the easiest to read. In addition to the "1"s pointing the wrong way for Haplotype 1, I was repeatedly mixing up the start and end of the chromosomes, trying to match Haplotype 1 and 2 DMRs. It would be easier to keep the normal orientation for Haplotype 1.

We have re-oriented Fig. 2 and the related supplementary figures so that the p arms are always on the left and the "1"s point the right way.

- As noted, some imprinted DMRs vary between tissues and/or individuals. Can anything be said about the cell types used in this study, and how different cell sources might influence the efficacy of the approach?

Gametic iDMRs are established in the mature gametes and are maintained by DNMT1 in all somatic lineages. We expect these to be very conserved across tissues, cell types, and individuals. Different cell types may have different somatic iDMRs, which are established postfertilization. However, insofar as these are typically controlled by a nearby gametic iDMR, PofO inferences should still be possible. Tissues with limited somatic imprinted differential methylation might be more difficult to assign PofO, and we have added a sentence to this effect in the text (see below; new text underlined). Alternatively, some cell types may even turn out to be better-suited for PofO inference, if they harbour more somatic iDMRs than these cell lines.

"Similarly, true iDMRs may display biological variability that could prevent PofO assignment for some chromosomes in new individuals, or in other tissues or cell types that have fewer or different somatic iDMRs than the cell lines used in this study."

- The applicability to other species would be limited by the distribution of iDMRs. In mice, not all autosomes harbour iDMRs.

This is an excellent point, which we addressed above in our responses to Reviewer #2. We now mention this in the Discussion.

- Could the method be extended to other haplotype markers (e.g. Hi-C) and long reads (e.g. Pacbio, since it's now better at 5mC detection)?

We do think this method could be extended to Hi-C plus PacBio—Reviewer 1 also noted this. We have included a new paragraph in the Discussion outlining some possibilities:

"Other sequencing technologies could perhaps provide the DNA methylation, DNA sequence, and long-range phase information required for PofO phasing, or different methods could be used to combine them. For instance, although we did not perform genome assembly for this study, PofO could be assigned to de novo trio-free diploid assemblies15,43 rather than chromosome-length haplotypes of small variants. SMRT sequencing (Pacific Biosciences), which now provides accurate DNA methylation as well as DNA sequence44, might be a substitute for nanopore data that provides better indel detection; and long-range phasing with Hi-C43 could perhaps be used instead of Strand-seq, although if phase switches occur at centromeres16 then chromosome arms that lack iDMRs (16q, 17q, 18p, and 20p) may not be assigned to a PofO."

- in the Discussion, it is chromosome 3 that has only 2 DMRs, rather than chromosome 2.

We appreciate the correction. We have changed this.

- the inversion for HG002 chromosome 8 should refer to Supp Fig 13 rather than 12.

We appreciate the correction. We have changed this.

Reference list

1. Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., … Miga, K. H. (2022). Automated assembly of high-quality diploid human reference genomes. bioRxiv doi:10.1101/2022.03.06.483034.

2. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., … Consortium, the H. P. R. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature 604, 437–446.

3. Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M. & Lansdorp, P. M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods 9, 1107–1112.

4. Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., Johanson, E., Boja, E., … Zook, J. M. (2022). PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics 2, 100129.

5. Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A. & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods doi:10.1038/s41592-018-0001-7.

6. Renfree, M. B., Hore, T. A., Shaw, G., Marshall Graves, J. A. & Pask, A. J. (2009). Evolution of Genomic Imprinting: Insights from Marsupials and Monotremes. Annu. Rev. Genomics Hum. Genet. 10, 241–262.

7. Cheong, C. Y., Chng, K., Ng, S., Chew, S. B., Chan, L. & Ferguson-Smith, A. C. (2015). Germline and somatic imprinting in the nonhuman primate highlights species differences in oocyte methylation. Genome Res. 25, 611–623.

8. Xie, W., Barr, C. L., Kim, A., Yue, F., Lee, A. Y., Eubanks, J., Dempster, E. L. & Ren, B. (2012). Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. Cell 148, 816–831.

9. Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., … Ritchie, M. E. (2019). Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res. 47,.

10. Hills, M., Falconer, E., O'Neill, K., Sanders, A. D., Howe, K., Guryev, V. & Lansdorp, P. M. (2021). Construction of Whole Genomes from Scaffolds Using Single Cell Strand-Seq Data. International Journal of Molecular Sciences vol. 22.

11. Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M. A., Hirst, M. & Jones, S. J. M. (2022). Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. Elife 11, e77898.

12. Jima, D. D., Skaar, D. A., Planchart, A., Motsinger-Reif, A., Cevik, S. E., Park, S. S., Cowley, M., Wright, F., … Hoyo, C. (2022). Genomic map of candidate human imprint control regions: the imprintome. Epigenetics 1–24 doi:10.1080/15592294.2022.2091815.

Referees' report, second round of review

Reviewer #1: Akbari et al made important improvements to the paper, including additional analyses and explanations. The most important novel discovery is the identification of maternal and paternal haplotypes of all human chromosomes in assembled scaffolds without using parental data, but intrinsic imprinted differential methylation data. Just a few items need fixing, mostly in style and organization.

The text of the results imply that one of the purposes was to generate a de-novo assembly. So the authors should not put clarification that this is not the purpose, in the Discussion. It needs to go in the Results.

In the response about SNV and indel accuracy, I meant for the authors to use the latest HG002 GIAB variant benchmark that was generated in collaboration with the panhuman genome group. But their revisions have effectively done this, and it works fine for the paper.

Line 22. Should mention both mother and father alleles are used in trios. Also cite Koren et al 2018 Nature BioTech for the trio method.

Lines 185-200. This new paragraph should moved from the Discussion to the Results

Line 359. Should cite the Jarvis et al bioRxiv HG002 study for the HG002 data used.

Reviewer #3: I would like to thank the authors for thoughtfully addressing all of my comments. I recommend accepting this excellent work for publication.

Authors' response to the second round of review

Reviewer #1

Akbari et al made important improvements to the paper, including additional analyses and explanations. The most important novel discovery is the identification of maternal and paternal haplotypes of all human chromosomes in assembled scaffolds without using parental data, but intrinsic imprinted differential methylation data. Just a few items need fixing, mostly in style and organization.

We thank the reviewer for their comments.

The text of the results imply that one of the purposes was to generate a de-novo assembly. So the authors should not put clarification that this is not the purpose, in the Discussion. It needs to go in the Results.

We have moved this to the results:
"Our phasing approach used the GRCh38 reference genome, and we did not perform de novo genome assembly of any kind."

In the response about SNV and indel accuracy, I meant for the authors to use the latest HG002 GIAB variant benchmark that was generated in collaboration with the panhuman genome group.
But their revisions have effectively done this, and it works fine for the paper.
Line 22. Should mention both mother and father alleles are used in trios. Also cite Koren et al 2018 Nature BioTech for the trio method.

This is a good suggestion, and we now mention the father's alleles in the sentence in question.
However, trio phasing is a well-established method that pre-dates Koren et al.'s updated approach (e.g., any time two or more variants from a child are assessed in one of their parents), so we do not feel it is appropriate to cite that paper as the source of trio phasing.

Lines 185-200. This new paragraph should moved from the Discussion to the Results

This is a good suggestion. We have done so, and several subsequent paragraphs in the Discussion have been moved to the new Limitations section.

Line 359. Should cite the Jarvis et al bioRxiv HG002 study for the HG002 data used.

We generated the nanopore data for HG002 in house for this study. The Strand-seq libraries were created previously by the Lansdorp Lab and submitted to Genome in a Bottle (GIAB) consortium independently (no relation to the Jarvis et al. preprint). For the ground truth variants, we used the GIAB v4.2.1 VCFs. For these data, GIAB asks users to cite Wagner et al. 2022 (Cell Genomics), which is what we cited here. No other HG002 data were used for this study, so it does not seem appropriate to cite the Jarvis et al. preprint.

Reviewer #3

I would like to thank the authors for thoughtfully addressing all of my comments. I recommend accepting this excellent work for publication.

We thank the reviewer for their recommendation.