nature portfolio

Peer Review File

The Platinum Pedigree: A long-read benchmark for genetic variants

Corresponding Author: Dr Michael Eberle

Version 0:

Decision Letter:

11th Dec 2024

Dear Dr. Eberle,

Your Article, "The Platinum Pedigree: A long-read benchmark for genetic variants", has now been seen by 3 reviewers. As you will see from their comments below, although the reviewers find your work of considerable potential interest, they have raised a number of concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to address these concerns. Regarding Reviewer #1's suggestion to expand the analysis to the T2T-CHM13 genome reference: this expansion is optional, but additional analyses using the T2T-CHM13 reference would be appreciated if feasible. Please also keep us informed about the progress of your submission at Nature.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within 10 weeks. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

IMAGE INTEGRITY
When submitting the revised version of your manuscript, please pay close attention to our Digital Image Integrity Guidelines and to the following points below:

-- that unprocessed scans are clearly labelled and match the gels and western blots presented in figures.
-- that control panels for gels and western blots are appropriately described as loading on sample processing controls
-- all images in the paper are checked for duplication of panels and for splicing of gel lanes.

Finally, please ensure that you retain unprocessed data and metadata files after publication, ideally archiving data in perpetuity, as these may be requested during the peer review and production process or after publication if any issues arise.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see:

https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

SUPPLEMENTARY PROTOCOL
To help facilitate reproducibility and uptake of your method, we ask you to prepare a step-by-step Supplementary Protocol for the method described in this paper. We encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice and report the protocol DOI in the reference list. Nature Portfolio's protocols.io is a free-to-use and open resource for protocols; protocols deposited onto protocols.io are citable and can be linked from the published article. More details can found at protocols.io.

ORCID
Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

The manuscript entitled "The Platinum Pedigree: A Long-Read Benchmark for Genetic Variants" by Kronenberg et al. introduces a benchmark study focusing on variant calling performance, emphasizing completeness and accuracy in complex genomic regions. The authors employed Mendelian inheritance principles within the large CEPH-1463 pedigree (G2 and G3) and various sequencing technologies to create a comprehensive variant truth set. While the topic is of considerable interest and the resources and datasets generated are highly valuable, I have several concerns about the manuscript that need to be addressed in order to justify its publication in Nature Methods and its dissemination to a broad audience of laboratory scientists. Specifically, my concerns pertain to the novelty of the research, the depth of analysis, and the clarity of data presentation. Below, I outline several reservations regarding various aspects of this work.

Major:
The authors mention in the Reporting Summary that this is a companion paper to Porubsky et al., 2024, a preprint under review in Nature. This related study also investigates the four-generation CEPH-1463 pedigree using multiple sequencing technologies but focuses on analyzing de novo mutation rates for SNVs, indels, TRs, and SVs. My primary concern is that the current study does not address a significantly outstanding or interesting question that distinguishes it from Porubsky et al., 2024. Whether considering the new sequencing resources of the CEPH-1463 pedigree data or the germline callsets of genetic variants, the scope of Porubsky et al., 2024 appears to encompass the vast majority of the analyses presented in this manuscript. This redundancy diminishes the perceived originality and significance of the current work.

The analysis for genetic variants is predominantly limited in GRCh38 in this project. Given the increasing interest in the CHM13 T2T reference genome within the field, it would be valuable for the study to include more analysis using this reference genome. Expanding the scope to encompass the CHM13 T2T genome could significantly enhance the comprehensiveness and relevance of the truth set for germline genetic variants in G2 and G3.

The analysis of genetic variants lacks depth for each category of variants. There should be a greater emphasis on complicated genetic variants such as indels, STRs/TRs, inversions, complex SVs, and transposable elements. Employing Mendelian inheritance principles to scrutinize these complicated variants in complex genomic regions would be especially intriguing. Detailed discussions on genotyping error rates and technology discrepancies for these complicated genetic variants would also provide valuable insights.

More comments:
1. Line 82-83: It is not clear whether these numbers refer to all G2 and G3 individuals or just one individual (e.g., NA12878) in G2. Clarification is needed.

2. Although this study is described as a companion to Porubsky et al., 2024, the manuscript should include more information, such as alignment methods and sequencing data matrices. Figures 1 and 2a appear to replicate figures from Porubsky et al., 2024, with only slight modifications. A discussion on whether the hs1 reference genome provides additional information in the inheritance vectors would be beneficial.

3. Is there a reason for not using assembly-based callers like PAV for SNV/indel calling? Also, Table S4 is confusing. A clearer explanation of how the two tables intersect and what numbers are shared between them is necessary.

4. Lines 168-169: What leads to genotyping errors in the inconsistent variants that had the same allele identified in multiple family members? In Table at Line 604, are these numbers for all genetic variants from each technology, or do they exclude tech-specific calls?
5. I like the Table S5 with the 2 reviewers' verdict for the manual inspection of Platinum Pedigree unique small variants.

6. Lines 195-207: Do different sequencing technologies exhibit biases in mismatch/error rates for SNVs or indels, such as ONT having more issues in homopolymer regions? What differences are introduced by the aligners, and is there any bias in SNV and indel calls?

7. Lines 210-220 extensively overlap with results in the "A Multigenerational Variant Callset" section of Porubsky et al., 2024. The manuscript should clearly delineate what is novel in this work.

8. In Figure 3, how many SVs are in the centromere and are deemed confident? Detailed analysis of inversions and transposable elements is needed. Are there any confident germline transposable elements that can be refined in paralogous regions using inheritance vectors?

9. Lines 253-255: The phrase "on average 2.6 different methods" is awkward and can be omitted for clarity.

10. Lines 298-305: Obviously since Ebert et al., 2021 included NA12878 with PacBio HiFi sequencing and assembly, more detailed comparisons with this study should be provided.

11. Line 735: Will the criteria ‘the refine process for the A length difference threshold of 30 bp’ too stringent for defining similarities in SVs?

12. In Figure 4ab, why is there a dip in TR fraction and purity? Is there a biological explanation for this? What is the mutation rate for different lengths of VNTR and STR? Did any reference STR/VNTRs observed in the children exhibit inconsistency in copy numbers (mutations) in the parents?

13. Utility for methods development: The version (v1.5.0) of deepvariant used in calling SNV is not the version used for this part (v1.6.0).

14. Lines 345-346: On the GIAB benchmark set (v4.2.1), a model trained with the Platinum Pedigree NA12878 labels had 8% fewer errors (Table S10). This statement is unclear and hard to understand from Table S10 alone. Provide more detailed explanations of the results in Tables S10 and S11. Also, what is ‘Palladium-trained’ in S10 and S11 table?

Reviewer #2 (Remarks to the Author):

Kronenberg et al. describe an deep analysis of the CEPH-1463 family to create a comprehensive set of truth variants. They use Mendelian inheritance across 3 generations to ensure variants are called accurately. These types of studies create important benchmarks that can be used by the DNA sequencing community to benchmark new wet lab and bioinformatic methods. Overall, I find this study to be well performed and well written. While it is not entirely novel, the corresponding author published a similar paper in 2017 (Eberle et al. 2017). This is a significant update from the 2017 work that includes many new sequencing methods (PacBio, ONT, etc.) and expands the earlier work to include structural variants.

I have only one small issue:
On line 76 you say you used the latest technologies, but I could not find details on how the sequencing was performed, for example was the Illumina data pair end 150 base reads and generated on a Novaseq X? Was the PacBio data generated on a Revio? What version of chip was used for ONT? This sort of data is important to understand how recent the sequencing technology used was.

And one suggestion:
Given the high quality of data and the extensive analyses performed, I’m a little disappointed that nothing was done to try to better characterize/catalog the somatic variants in the cell lines of these samples. I understand that this can never be done perfectly and there will be some variation between different passages, but still, this is a potential cause of many false positives in this type of data and many people do not have access to the original blood samples and so are forced to use the cell lines as there source of DNA. It seems like with phase information generated by PacBio, ONT, and Strand-seq you would be able to find variants that are haplotype specific, but not found in the original blood sample or the other individuals. These would be excellent candidates for somatic variants.

Reviewer #3 (Remarks to the Author):

Recommendation

I would recommend to accept but would suggest to revise for clarity around some of the suggestions mentioned below. Making a more explicit case for applicability outside bioinformatics work (e.g. by showing examples of improvement in phenotypic / medical variant calling) would likely also increase impact / widen the audience of the publication.

Summary of the key results

The authors propose a comprehensive reference set of many different types of genetic variation that is validated through Mendelian inheritance on a large pedigree of individuals and multiple sequencing technologies. They argue that this set can be used to improve accuracy of variant call filtering as well as accurate assessment of variant calls that cover a larger part of the human genome compared to existing reference datasets.

Originality and significance

In my view, the main strength & originality of the publication is in the data resource that was collected and aggregated:

- Data from multiple sequencing technologies is not only a great resource for understanding how their performance differs - specifically in regions of the genome that are difficult to analyse.
- Comprehensive analysis of variant calls across a wide spectrum of variant types, including structural variation and repeats.

I believe that this resource along with the analysis provided is useful beyond specialised audiences in bioinformatics as different DNA sequencing technology becomes part of mainstream diagnostic work, as well as with different variant types being studied as medically relevant. However, that case could be made more explicitly by the paper by showing phenotypic / medically relevant examples that would not be covered by previous truth data.

The analysis itself and methods applied appear largely not novel / have been published separately, although comprehensively applying them to this pedigree dataset on newly collected data as presented here has not been done before to my knowledge.

Discussion of results

I have no concerns about the analysis of small variants and confident regions - this appears mostly in-line with the older Platinum Genomes publication, using better input data from newer technologies. The authors note that analysis of indels in homopolymers & close to SVs is still an issue - it would be good to incorporate this also in a systematic way in the way that confident regions are built & analysed (e.g. would it make sense to exclude homopolymer sections entirely if they contain problematic indels?) - although this is a corner case, it is probably worthwhile as part of an effort that targets the more difficult regions of the genome.

The analysis of the SVs is thorough but does not show a quantitative comparison to other SV sets produced for samples contained in the pedigree. Arguably this is difficult as the set presented is more comprehensive than previously published work - perhaps a simple useful analysis would be a comparative summary of variant counts by regions of the genome with that other datasets (I would assume that there should be some overlap in the variants found if the samples are the same?). Another thing to note is that the variant counts in supplementary table 8 have the same count for NA12878 as for the total (35,662) - I wonder if this is a mistake?

On the TR analysis (Figure 4) I am not sure how to interpret the TR length axis. E.g. a TR length of 0 in the figure — would this be the measured TR length in the sample (so the entire TR would be deleted)? Some clarification would be helpful.

Finally, on the argument of utility for the new truthset by retraining DeepVariant, I wonder whether the variants that are filtered correctly only after re-training are in the regions not covered in the previous training / truth sets - are they truly in more difficult regions, or are these perhaps variants where there was more uncertainty with the smaller training sets beforehand?

Presentation of results

- the paper uses quite a few abbreviations that should be introduced to a wider audience. For example, SD regions in the SV methods section (segmental duplications presumably?).
- Hyperlinks to the references are Paperpile links that require access to the Google doc used for writing the manuscript.

Reviewer #3 (Remarks on code availability):

- License of supplied code is proprietary and not a standard open source license which may limit usability.
- Code compiles and has the required information needed for reproducing the results, however, doing so is not a small effort as it also requires substantial compute & downloading of the data.
- Documentation could be improved to be easier to follow in the right sequence.
- Code documentation and quality is a little variable with some hard-coded data locations remaining in the code - so

reproducing would not work out of the box. However, how to make the changes to get it working is pretty obvious most of the time.
- Conda environment files are provided for environment reproducibility in Python; however, these are machine-specific dumps that might need adjustment. Dockerfiles would probably be more easy to use here.

Version 1:

Decision Letter:

Our ref: NMETH-A58104A

31st Mar 2025

Dear Dr. Eberle,

Thank you for submitting your revised manuscript "The Platinum Pedigree: A long-read benchmark for genetic variants" (NMETH-A58104A). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines. It would also be great if you could provide an update on the manuscript submitted to Nature (Porubsky et al. 2024).

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks or so. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Thank you again for your interest in Nature Methods. Please do not hesitate to contact me if you have any questions. We will be in touch again soon.

Sincerely,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods

Reviewer #1 (Remarks to the Author):

Thank you to the authors for addressing most of my comments and concerns. I only have two remaining points:

The statement, "There were 2,112 Alu, 679 SVA, 324, and 383 LINE-1 element insertions in the family," seems to present aggregated numbers for TEs, rather than at a per-person level. However, the reported SVA number seems questionable.

The new Figure 2B is missing a color legend for SNV.

Reviewer #2 (Remarks to the Author):

Accept, I have no further concerns.

Reviewer #3 (Remarks to the Author):

All my comments from the first round of reviews were addressed. I recommend to accept.

Version 2:

Decision Letter:

12th Jun 2025

Dear Dr Eberle,

I am pleased to inform you that your Article, "The Platinum Pedigree: A long-read benchmark for genetic variants", has now been accepted for publication in Nature Methods. The received and accepted dates will be 2nd Oct 2024 and 12th Jun 2025. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Methods style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email) of someone who will be able to check the proofs and deal with any last-minute problems.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

If you are active on Twitter/X or Bluesky, please e-mail me your and your coauthors' handles so that we may tag you when the paper is published.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Portfolio's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points.

Best regards,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

The manuscript entitled "The Platinum Pedigree: A Long-Read Benchmark for Genetic Variants" by Kronenberg et al. introduces a benchmark study focusing on variant calling performance, emphasizing completeness and accuracy in complex genomic regions. The authors employed Mendelian inheritance principles within the large CEPH-1463 pedigree (G2 and G3) and various sequencing technologies to create a comprehensive variant truth set. While the topic is of considerable interest and the resources and datasets generated are highly valuable, I have several concerns about the manuscript that need to be addressed in order to justify its publication in Nature Methods and its dissemination to a broad audience of laboratory scientists. Specifically, my concerns pertain to the novelty of the research, the depth of analysis, and the clarity of data presentation. Below, I outline several reservations regarding various aspects of this work.

Major:
The authors mention in the Reporting Summary that this is a companion paper to Porubsky et al., 2024, a preprint under review in Nature. This related study also investigates the four-generation CEPH-1463 pedigree using multiple sequencing technologies but focuses on analyzing de novo mutation rates for SNVs, indels, TRs, and SVs. My primary concern is that the current study does not address a significantly outstanding or interesting question that distinguishes it from Porubsky et al., 2024. Whether considering the new sequencing resources of the CEPH-1463 pedigree data or the germline callsets of genetic variants, the scope of Porubsky et al., 2024 appears to encompass the vast majority of the analyses presented in this manuscript. This redundancy diminishes the perceived originality and significance of the current work.

> The Porubsky paper (Nature, in press) is primarily focused on *de novo* mutation and how those rates vary depending on region and class of genetic variation. In particular, that paper focuses on previously inaccessible regions (centromeres, segmental duplications, Y chromosome) and uses T2T assembly-based as well as read-based approaches to access and validate *de novo* germline and somatic mutations in some regions for the first time. Conversely, this paper only explores **germline inherited** variation and develops the methods and analyses behind the production of an extensive variation truth set. Germline, inherited variation and *de novo* variation are, for the most part, exclusive of each other - while *de novo*

variants occurring between G1 and G2 could be included in our truthset, all *de novo* variants arising in G3 would, by definition, fail our inheritance check. Thus, although starting from the same resource, the motivation, the emphasis, and the target audience is very different between these two studies–a robust inherited variant truth set of genetic variants confirmed multigenerationally will be invaluable, for example, to researchers developing new sequencing technologies or clinical labs looking to validate their sequencing pipelines. The materials uniquely presented in this manuscript provide significant benefits to researchers in the following ways (as are also described in the Discussion):

1. **A Resource for Method Development.** The curated genetic map and accompanying open-source software for scoring Mendelian inheritance offer method developers a robust framework to quickly evaluate genotyping accuracy. For instance, the Sawfish structural variant caller was improved using this framework (Saunders et al., 2024).
2. **Comprehensive Benchmarking for Challenging Genomic Regions.** Our more comprehensive small variant truth set enables benchmarks in regions beyond those covered by Genome in a Bottle (GIAB). A direct outcome was the improved accuracy of DeepVariant, a widely used tool, by training it on this expanded truth set.
3. **Insights into Clinically Relevant Genes.** Using long-read datasets, we analyzed clinically significant genes such as *LPA*, *CYP2D6*, *SMN*, and *HLA*. To strengthen this clinical perspective, we have added additional text and figures expanding on these findings.
4. **Validation of clinical pipelines**. Validation of variants beyond SNPs and indels has often been a difficult process, but this truthset includes SNPs, indels, TRs and SVs so that clinical labs can easily validate their WGS pipelines for “all” variant types.

The analysis for genetic variants is predominantly limited to GRCh38 in this project. Given the increasing interest in the CHM13 T2T reference genome within the field, it would be valuable for the study to include more analysis using this reference genome. Expanding the scope to encompass the CHM13 T2T genome could significantly enhance the comprehensiveness and relevance of the truth set for germline genetic variants in G2 and G3.

Thanks for this suggestion. We have constructed two additional truth sets focused on CHM13 T2T: one for small variants and one for structural variants.

These datasets have been uploaded to the Amazon Open Data S3 bucket. In the main text, we briefly compare the recombination and the overall variant counts between the two references noting that the recombination map is, essentially, identical and that we identify fewer variants using CHM13 (**Table S15)**. We followed up on the lower number of variants in CHM13 and confirmed that this is not an indication of significant limitations of either reference. For example, we identified nearly the same number of heterozygous SNPs independent of the reference but more homozygous SNPs when using GRCh38. That difference for homozygous SNPs is expected because our samples are European and CHM13 is primarily a "European" genome while GRCh38 is a mosaic of different ancestry.

We've added this text to the Results:

We also compared how the recombination map is affected by reference genomes, analyzing the data against CHM13 T2T reference. There were two additional short haplotype blocks (chr11:70292011-70389373 and chr8:7823639-8263981) found when using CHM13. The Pearson correlation between the recombination breakpoints between GRCh38 and CHM13 is >0.998.

We've added this text to the methods:

*A truth set for small and structural variants was generated for CHM13 using the same methods applied to GRCh38, allowing for direct comparison (**Table S15**). Despite being a more complete reference genome, CHM13 exhibited fewer small variant calls compared to GRCh38. To investigate this discrepancy, we analyzed heterozygous variant counts across the second and third generations. While heterozygous variant counts were similar, homozygous alternate variant counts were significantly higher on GRCh38 (23%). This suggests that the platinum pedigree family members are genetically closer to CHM13 than to GRCh38, which is a mosaic reference incorporating genomes from multiple individuals, including those of African descent. For structural variants, we observed more deletions but fewer insertions on CHM13, consistent with its status as a more complete reference genome.*

The analysis of genetic variants lacks depth for each category of variants. There should be a greater emphasis on complicated genetic variants such as indels, STRs/TRs, inversions, complex SVs, and transposable elements. Employing Mendelian inheritance principles to scrutinize these complicated variants in complex genomic regions would be especially intriguing. Detailed discussions on genotyping error rates and technology

discrepancies for these complicated genetic variants would also provide valuable insights.

Thank you for the suggestions and we have now performed the following additional analyses:

- We've analysed retroelement insertions by annotating SV insertion sequences with repeat masker, allowing us to identify LINE/SINE/ALU elements.
- To further dive into the variant categories we've stratified the Mendelian concordance of the small variants by region and technology (shown in supplemental **Table S6**). This provides a more thorough view of the performance of different technologies, highlighting that variant calling accuracy varies greatly across the genome.

We've added the following text to main:

*The pedigree concordant rates varied by technology and genomic annotation (**Table S6**), the number of passing variants per genomic annotation is shown in **Fig. 2A**. The stratification of variants by region highlights the need for continued efforts toward improving joint genotyping accuracy in complex regions, e.g. segmental duplications (**Table S6**).*

We annotated insertions in our VCF using repeatmasker, requiring a 90% sequence length match. There were 2112 Alu, 679 SVA, and 383 LINE-1 element insertions in the family.

More comments:
1. Line 82-83: It is not clear whether these numbers refer to all G2 and G3 individuals or just one individual (e.g., NA12878) in G2. Clarification is needed.

This refers to all G2 and G3 individuals and we have added the following sentence for clarification:

Our truth set includes variants in the two parents (NA12878/NA12877), and all eight of their children, represented in a joint VCF file.

2. Although this study is described as a companion to Porubsky et al., 2024, the manuscript should include more information, such as alignment methods and

sequencing data matrices. Figures 1 and 2a appear to replicate figures from Porubsky et al., 2024, with only slight modifications. A discussion on whether the hs1 reference genome provides additional information in the inheritance vectors would be beneficial.

the manuscript should include more information, such as alignment methods and sequencing data matrices

> We have added a table (**Table S1**) that provides detailed information on data generation and mapping, including version numbers. Additionally, for HiFi and Illumina data, we used the HiFi-Human-WGS-WDL and Dragen pipelines, respectively, both of which include comprehensive documentation and version tracking. Additional details about the variant callers are described in the supplemental methods.

A discussion on whether the hs1 reference genome provides additional information in the inheritance vectors would be beneficial.

> The inheritance vectors for the CHM13 T2T (hs1) reference are nearly identical and we've added the following text that describes this:
>
> *We also compared how the recombination map is affected by reference genomes, analyzing the data against CHM13 T2T reference. There were two additional short haplotype blocks (chr11:70292011-70389373 and chr8:7823639-8263981) found when using CHM13. The Pearson correlation between the recombination breakpoints between GRCh38 and CHM13 is >0.998.*

Figures 1 and 2a appear to replicate figures from Porubsky et al., 2024, with only slight modifications

> The reviewer is correct that the pedigree figure (now **Fig. 1A**) is a subset of the figure from the Porubsky paper, but we feel it's required to give context to the experimental design and orient readers. We've made a number of changes to the figures to further distinguish our work from Porubsky and dive deeper into the composition of the new truth set.

- Figure 1 and Figure 2C were combined to make the new **Fig. 1A** and **Fig. 1B**. The new **Fig. 1B** (haplotype map) is completely independent of

Porubsky's paper. We also annotate the number of our recombination breakpoints that were also found with strand-seq data (**Table S3**).

- The original **Fig. 2 A/B** was moved to supplement, as the distribution of recombinants, and their sizes, are still relevant, but as pointed out overlaps Porubsky's paper (although independent).
- We've introduced a new **Fig. 2** that shows the regions where our truth set differs from GIAB, and the composition of the callers across the genome. These figures are more broadly interesting than the prior recombination figure.

3. Is there a reason for not using assembly-based callers like PAV for SNV/indel calling? Also, Table S4 is confusing. A clearer explanation of how the two tables intersect and what numbers are shared between them is necessary.

We initially included PAV in the small variant truthset. However, we discovered that PAV had a low (46%) Mendelian consistency rate. In total PAV generated ~1 million passing SNVs/Indels, compared to the mapping-based callers that generated ~3-5 Million passing variants. We communicated these findings to the tool's author, Peter Audano, who explained that PAV is primarily designed for structural variant (SV) calling, therefore we excluded PAV small variants.

Also, Table S4 is confusing. A clearer explanation of how the two tables intersect and what numbers are shared between them is necessary.

Thank you for highlighting the lack of clarity in Table S4 (now **Table S5**). We've merged the two tables which were mostly identical, we've added brief descriptions in the sheet, and we've added the following table legend to the manuscript:

Table S5. High-Level Summary of Small Variant Calls.
Total variant counts by sequencing technology, including combined Mendelian passing rates for single nucleotide variants (SNVs) and insertions/deletions (INDELs). Reported statistics include the total number of SNVs and INDELs, total variant call format (VCF) records (sites), sites with one or more no-calls, low-quality sites (QV < 20 in VCF), sites with Mendelian inconsistencies, Mendelian-consistent SNVs, Mendelian-consistent INDELs, and sites with a single alternative allele (indicative of errors or de novo variants).

4. Lines 168-169: What leads to genotyping errors in the inconsistent variants that had the same allele identified in multiple family members? In Table at Line 604, are these numbers for all genetic variants from each technology, or do they exclude tech-specific calls?

What leads to genotyping errors in the inconsistent variants that had the same allele identified in multiple family members?

> There are a few possible explanations for a shared allele that is shared but inconsistent. The most likely is that the variant is in a region that is more prone to genotyping errors and thus is correctly genotyped in a few, but not all, of the family members. Another, less common, cause, is that the expected diploid expectation is incorrect at a position because there is a deletion or duplication, and so the genotypes are inconsistent with the inheritance. This was observed in the previous, short-read analysis truth set generated for this pedigree (Eberle et al, 2017)
>
> We have also added the following text to include why variants may be seen multiple times but fail our pedigree filter:
>
> *Indels can be more difficult to genotype both accurately and consistently across the pedigree (Fang et al. 2014), and accordingly our indel truth rate ranged from 72-87% per call set. Similarly, we observe increased failure rates (6-43%, **Table S6**) due to one individual missing a genotype (nocall). The pedigree concordance test requires that all family members are genotyped accurately. Of the variants that did not pass our pedigree inheritance test, without nocalls, 17-23% were identified in a single individual, likely representing false positives or de novo or somatic mutations (Porubsky et al. 2024) (**Table S5**)*

In Table at Line 604, are these numbers for all genetic variants from each technology, or do they exclude tech-specific calls?

> This excludes the tech-specific calls. In building our truth data, we include anything that is pedigree consistent and filter out most errors by removing sites that are inconsistent between technologies. We do not have a similar method to exclude single-technology errors, instead we estimate the single technology errors by comparing the rate at which different technologies call the same variant differently. This table summarizes the three types of observable discrepancies between two or three technologies: (1) genotype conflicts, where the genotypes differ between technologies; (2) differences in alleles at a given site; and (3)

indels that overlap on the same haplotype. Our calculation to estimate the error rate in the single tech calls is described in the subsequent paragraph.

5. I like the Table S5 with the 2 reviewers' verdict for the manual inspection of Platinum Pedigree unique small variants.

It's an important step, thank you for recognizing it.

6. Lines 195-207: Do different sequencing technologies exhibit biases in mismatch/error rates for SNVs or indels, such as ONT having more issues in homopolymer regions? What differences are introduced by the aligners, and is there any bias in SNV and indel calls?

Thank you for the suggestion to analyze mendelian consistency and inconsistency by technology, sequence context, and region. We've added a supplemental **Table S6** to future investigate where 1) mapping issues cause problems, 2) where sequence context e.g. homopolymers cause issues and 3) where structural variants cause issues. We have also revised **Fig. 2**, based on this comment and added the following text:

*Pedigree concordance rates varied by technology and genomic annotation (Table S6), with the number of passing variants for each annotation shown in **Fig. 2A**. The stratification of pedigree concordance by region and technology (**Table S6**) highlights opportunities for improvement in variant calling, particularly in challenging genomic regions including low mapping quality regions or segmental duplications.*

7. Lines 210-220 extensively overlap with results in the "A Multigenerational Variant Callset" section of Porubsky et al., 2024. The manuscript should clearly delineate what is novel in this work.

The flagship paper by Porubsky offers only a rough summary of the total number of variants identified but no deeper analysis. Conversely, this manuscript delves deeply into the specific details of developing this truthset and analysis of the variants including sequence context and technology differences. The improvements we've provided to DeepVariant (via retraining) is one of the notable novel efforts. We've added text to delineate our results and goals from Porubsky et al.

intro:

In this study, we create and analyze a catalog of pedigree concordant genetic variation based on sequence an extended family (two parents and eight children) sequenced with the latest technologies from Illumina, PacBio, and Oxford Nanopore Technologies (ONT) to create a reference catalog across the spectrum of genetic variation.

results:

*For this study, we focus on the haplotypes and variants in G2 that are inherited among the eight children in G3 (**Fig. 1**). In this work, we specifically focus on building the methods and analyses required to generate a highly sensitive truth by tracking haplotypes across multi-generational pedigrees.*

8. In Figure 3, how many SVs are in the centromere and are deemed confident? Detailed analysis of inversions and transposable elements is needed. Are there any confident germline transposable elements that can be refined in paralogous regions using inheritance vectors?

How many SVs are in the centromere and are deemed confident?

In this study, we only have two passing variants in the GRCh38 centromeres (from the PAV caller). This is likely because any variant needs to be genotyped consistently and accurately across all 10 members of the pedigree to pass our strict inheritance requirements. We call this out in the discussion where we describe limitations and future directions where de novo assembly will be required:

Future work will test pedigree assessment of localized de novo assemblies and additional technologies as a way to resolve the larger scale sequence of these regions.

Detailed analysis of inversions and transposable elements is needed. Are there any confident germline transposable elements that can be refined in paralogous regions using inheritance vectors?

We've provided annotations in our VCF for transposable element insertions. To answer your question regarding mobile element insertions in paralogous sequence, we intersected segmental duplication annotations with gene bodies to

come up with a list of paralogous genes, and then intersected the mobile element insertions in these regions. We find 261 full length Alu insertions in paralogous regions, and 10 full length L1 insertions. For example, the gene *ANKRD36* has ~4 regions of homology across the genome with ~90% identity and contains a L1 insertion at chr2:97,158,070-97,158,071.

We've added this text to the main:

We annotated insertions in our VCF using repeatmasker, requiring a 90% sequence length match. There were 2,112 Alu, 679 SVA, and 383 LINE-1 element insertions in the family.

9. Lines 253-255: The phrase "on average 2.6 different methods" is awkward and can be omitted for clarity.

Thank you for the suggestion, we've made the change and this now reads:

The majority (74.9%) of our integrated variants are supported by more than one call set and 25.1% of the calls are unique to only one method.

10. Lines 298-305: Obviously since Ebert et al., 2021 included NA12878 with PacBio HiFi sequencing and assembly, more detailed comparisons with this study should be provided.

Thank you for highlighting this omission. We've done the comparison and added the following main text:

Additionally, we overlapped our SV callset for NA12878 with Ebert et. al. 2021 and found a large amount (76.8%; 18,538) of overlap (Ebert et al. 2021). The majority 84.4% (8898/10537) of SV calls only found in one callset were Tandem Repeats with that differing allele lengths.

11. Line 735: Will the criteria 'the refine process for the A length difference threshold of 30 bp' too stringent for defining similarities in SVs?

We specifically tried different thresholds and adjusting the threshold has virtually no impact on the total number of merged variants, which remains constant. The only observable effect is on genotype consistency, as some supporting variants in tie scenarios may be reassigned to a different primary variant. A length

difference threshold of 30bp was chosen as an optimal balance between maintaining sequence-resolved specificity and avoiding excessive merging.

The reason why this has little impact is that the merging pipeline organizes structural variants (SVs) into groups based on their genomic intervals. Within each group, one variant is designated as the “primary” variant, serving as the anchor, while other overlapping variants are classified as “supporting.” Supporting variants are typically assigned to a single primary variant based on the smallest absolute difference in length. As a result, the ‘length difference threshold’ is rarely applied. However, when a supporting variant is equally close in length to multiple primary variants, the pipeline resolves ties by considering whether a slightly larger length difference could yield a better match—particularly when genotypes are factored in. In such cases, the threshold prevents small length mismatches from overriding genotype concordance.

We have revised the methods section to add clarity to this process.

12. In Figure 4ab, why is there a dip in TR fraction and purity? Is there a biological explanation for this? What is the mutation rate for different lengths of VNTR and STR? Did any reference STR/VNTRs observed in the children exhibit inconsistency in copy numbers (mutations) in the parents?

In Figure 4ab, why is there a dip in TR fraction and purity?

In general, longer tandem repeats tend to be more mutable, which is the most parsimonious reason why we see a drop in purity. Another possibility is long annotations might be broken into multiple repeat units when interrupted by sequence error therefore they would have higher purity. The TR community is working on methods to combine neighboring fragmented pure repeats into a cohesive single repeat (https://www.biorxiv.org/content/10.1101/2024.10.04.615514v1).

What is the mutation rate for different lengths of VNTR and STR?

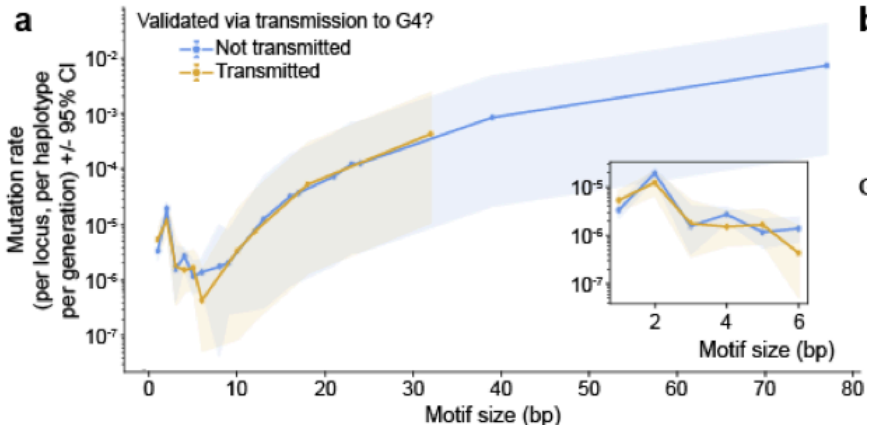
The analysis of de novo rates was analyzed in the flagship paper (here we only looked at inherited variation). Where further study (in more samples) is needed here is the answer to the reviewer’s question from Porubsky (2024 bioRxiv):

“After Element and transmission validation, we found an average of 79.6 TR DNMs (including STRs, VNTRs, and complex loci) per sample and estimated a TR mutation rate across all passing TR loci genome-wide of 5.25×10-6 per locus per haplotype per generation (95% CI = 4.42 - 6.01×10-6)”

“The VNTR (7+ bp motif) mutation rate was 2×10-6 (95% CI= 1.8 – 3.1×10-6)”

What is the mutation rate for different lengths of VNTR and STR?

Correspondingly, this was addressed in Fig. 3A (Porubsky 2024 BioRx):



13. Utility for methods development: The version (v1.5.0) of deepvariant used in calling SNV is not the version used for this part (v1.6.0).

You are correct that this section used a newer version of DeepVariant compared to the one that was used to build the truthset. This project has been ongoing for an extended period, and the transition to version 1.6 occurred mid-project. Importantly, there is almost no change in accuracy between 1.5 and 1.6 when measured against GIAB (DV1.5 case study - SNP F1 0.9997, Indel F1 0.9934. DV1.6 case study SNP F1 0.9997 Indel F1 0.9933), so the gains described in this section are still correct. Importantly, all calling for the truth set was consistent across pedigree (DV 1.5) – ensuring that no technical artifacts were introduced.

14. Lines 345-346: On the GIAB benchmark set (v4.2.1), a model trained with the Platinum Pedigree NA12878 labels had 8% fewer errors (Table S10). This statement is unclear and hard to understand from Table S10 alone. Provide more detailed

explanations of the results in Tables S10 and S11. Also, what is ‘Palladium-trained’ in S10 and S11 table?

Lines 345-346: On the GIAB benchmark set (v4.2.1), a model trained with the Platinum Pedigree NA12878 labels had 8% fewer errors (Table S10). This statement is unclear and hard to understand from Table S10 alone.

We have revised this section to increase clarity.

Also, what is ‘Palladium-trained’ in S10 and S11 table?

The project was initially called Palladium-Genomes, but was later renamed to platinum pedigree by consensus among the co-authors. We have replaced Palladium with Platinum Pedigree everywhere in the text.

Reviewer #2 (Remarks to the Author):

Kronenberg et al. describe an deep analysis of the CEPH-1463 family to create a comprehensive set of truth variants. They use Mendelian inheritance across 3 generations to ensure variants are called accurately. These types of studies create important benchmarks that can be used by the DNA sequencing community to benchmark new wet lab and bioinformatic methods. Overall, I find this study to be well performed and well written. While it is not entirely novel, the corresponding author published a similar paper in 2017 (Eberle et al. 2017). This is a significant update from the 2017 work that includes many new sequencing methods (PacBio, ONT, etc.) and expands the earlier work to include structural variants.

I have only one small issue:
On line 76 you say you used the latest technologies, but I could not find details on how the sequencing was performed, for example was the Illumina data pair end 150 base reads and generated on a Novaseq X? Was the PacBio data generated on a Revio? What version of chip was used for ONT? This sort of data is important to understand how recent the sequencing technology used was.

Thank you for highlighting this important omission. We have now added Table S1, which summarizes the datasets utilized in this study. The table includes the ONT chip version numbers, the PacBio platforms employed, and the accession

details for the Illumina datasets. Additionally, we have expanded the Methods section to provide more comprehensive information on ONT data processing.

And one suggestion:
Given the high quality of data and the extensive analyses performed, I'm a little disappointed that nothing was done to try to better characterize/catalog the somatic variants in the cell lines of these samples. I understand that this can never be done perfectly and there will be some variation between different passages, but still, this is a potential cause of many false positives in this type of data and many people do not have access to the original blood samples and so are forced to use the cell lines as there source of DNA. It seems like with phase information generated by PacBio, ONT, and Strand-seq you would be able to find variants that are haplotype specific, but not found in the original blood sample or the other individuals. These would be excellent candidates for somatic variants.

The analysis of somatic variants was considered in the flagship paper (Porubsky et al, Nature, in press), and cell line artifacts were analyzed in Eberle 2017. The pedigree filtering we used excludes somatic mutations as, by definition, they violate the inheritance vectors. Since our inheritance test focuses on G2/G3 (all blood samples), we avoided the impact of cell line artifacts coming from the first generation. While this data does not allow us to look somatic variants in cell lines, we would like to highlight that somatic variants were explored in the flagship paper:

Porubsky 2024:

"Using flanking SNVs from long-read sequencing data to construct haplotypes as well as allele balance for unphased variants, we categorize de novo variants as either germline DNMs or postzygotic mutations (PZMs) (Fig. 2b, Methods). We classify 17.1% (129/755) of de novo SNVs as PZMs, defined here as somatic mutations occurring very early in development. Of the 311 de novo SNVs in G2 and G3 individuals with offspring, 97.1% of germline events transmit to the next generation, compared with 64.5% of postzygotic
events (Extended Data Fig. 3)."

Reviewer #3 (Remarks to the Author):

Recommendation

I would recommend to accept but would suggest to revise for clarity around some of the suggestions mentioned below. Making a more explicit case for applicability outside bioinformatics work (e.g. by showing examples of improvement in phenotypic / medical variant calling) would likely also increase impact / widen the audience of the publication.

Summary of the key results

The authors propose a comprehensive reference set of many different types of genetic variation that is validated through Mendelian inheritance on a large pedigree of individuals and multiple sequencing technologies. They argue that this set can be used to improve accuracy of variant call filtering as well as accurate assessment of variant calls that cover a larger part of the human genome compared to existing reference datasets.

Originality and significance

In my view, the main strength & originality of the publication is in the data resource that was collected and aggregated:

- Data from multiple sequencing technologies is not only a great resource for understanding how their performance differs - specifically in regions of the genome that are difficult to analyse.
- Comprehensive analysis of variant calls across a wide spectrum of variant types, including structural variation and repeats.

I believe that this resource along with the analysis provided is useful beyond specialised audiences in bioinformatics as different DNA sequencing technology becomes part of mainstream diagnostic work, as well as with different variant types being studied as medically relevant. However, that case could be made more explicitly by the paper by showing phenotypic / medically relevant examples that would not be covered by previous truth data.

> Thank you for the suggestion. We emphasized the value of this for development partly because this is an often overlooked benefit of using pedigrees for methods development. We have added an analysis about using this pedigree to develop truth data for gene-level haplotypes that are clinically important such as PGx star alleles (HLA-A, HLA-B, and CYP2D6), LPA KIV-2 copy number, and segmental duplications (SMN1).

We've added this text to the main:

The platinum pedigree, and the inheritance maps established here, serve as an excellent resource to validate genotypes, even in complex medically-relevant regions. We analysed four loci for demonstrating perfect inheritance within the pedigree (Fig. S12). First we used the diploid genome assemblies (Verkko) to identify the KIV-2 alleles in the pedigree (Fig. S12A). We tested and confirmed accurate segregation of second generation alleles by copy number, ranging from 23 copies (~120 kb) of the 5kb repeat, down to 9 copies (~50 kb). Shorter KIV-2 alleles are associated with higher Lipoprotein(a) concentrations in the blood, which can lead to heart disease. Next we used the Paraphase (Chen et al. 2023) tool to analyse the SMN1 and SMN2 loci; the parental alleles are shown in Fig. S12B. Lastly, we applied StarPhase (Holt et al. 2024) to diplotype critical PGx genes, HLA-A, HLA-B, and CYP2D6 (Fig. S12 C-D), no mendelian violations were detected in the pedigree. These examples reiterate applicability of the pedigree for validating new bioinformatic approaches for complex regions and variant types.

In addition, we've included a sentence in the Discussion where we highlight the value of this resource for validation of clinical pipelines:

Beyond the utility for developers, it can be increasingly difficult to validate the performance of sequencing pipelines to accurately genotype difficult variants, such as SVs and TRs. This benchmark will enable labs to validate the performance of variant calling for SVs and TRs so that they can be implemented in clinical settings.

The analysis itself and methods applied appear largely not novel / have been published separately, although comprehensively applying them to this pedigree dataset on newly collected data as presented here has not been done before to my knowledge.

That is correct, our paper is the first instance of pedigree concordance, beyond trios/quads, using long-read data and covering variants beyond just SNVs and indels. We also like to reemphasize that we've taken care to ensure the pedigree filtering/testing software is built for others to use. We provide the haplotype map in the github repo, and the code takes standard formatted VCF data.

Discussion of results

I have no concerns about the analysis of small variants and confident regions - this appears mostly in-line with the older Platinum Genomes publication, using better input data from newer technologies. The authors note that analysis of indels in homopolymers & close to SVs is still an issue - it would be good to incorporate this also in a systematic way in the way that confident regions are built & analysed (e.g. would it make sense to exclude homopolymer sections entirely if they contain problematic indels?) - although this is a corner case, it is probably worthwhile as part of an effort that targets the more difficult regions of the genome.

> In the discussion, we highlight areas for future work, such as improving homopolymer indel calls and reducing false positives in structural variant detection. To address these challenges, we propose using the assemblies generated in Porubsky et al. 2024 to build an independent truth set by intersecting high-confidence regions from assembly- and read-mapping-based approaches.
>
> To achieve this, we developed a program that traverses the assemblies (relative to GRCh38) in 300 bp windows and performs a pedigree concordance test. This approach identifies regions—not just individual variants—where the assembled haplotypes segregate perfectly within the pedigree.
>
> As part of this study, we now provide three additional resources:
>
> 1. High-confidence assembly regions across the pedigree,
> 2. An assembly-based truth set for NA12878, and
> 3. Intersected high-confidence regions, intended to enhance the specificity of the truth set.
>
> While a full comparison between assembly- and mapping-based approaches is beyond the scope of this manuscript, we plan to investigate this further in future work.

The analysis of the SVs is thorough but does not show a quantitative comparison to other SV sets produced for samples contained in the pedigree. Arguably this is difficult as the set presented is more comprehensive than previously published work - perhaps a simple useful analysis would be a comparative summary of variant counts by regions of the genome with that other datasets (I would assume that there should be some overlap in the variants found if the samples are the same?). Another thing to note is that the variant counts in supplementary table 8 have the same count for NA12878 as for the total (35,662) - I wonder if this is a mistake?

Thank you for the suggestion. We've compared NA12878 SV calls to a high quality callset (Ebert et al 2021), which uses an assembly based approach to detect structural variation.

This text was added to the main:

Additionally, we overlapped our SV callset for NA12878 with Ebert et. al. 2021 and found a large amount (76.8%; 18,538) of overlap (Ebert et al. 2021). The majority 84.4% (8898/10537) of SV calls only found in one callset were Tandem Repeats with that differing allele lengths.

Another thing to note is that the variant counts in supplementary table 8 have the same count for NA12878 as for the total (35,662) - I wonder if this is a mistake?

Thank you for catching this error (failure to remove hom-ref sites for NA12878). The correct number, which is now updated, is 24,171.

On the TR analysis (Figure 4) I am not sure how to interpret the TR length axis. E.g. a TR length of 0 in the figure — would this be the measured TR length in the sample (so the entire TR would be deleted)? Some clarification would be helpful.

Yes, this is the repeat length in the sample and the bins are 50bp each. Thus there may be some positions where the repeat is 0bp in a sample and included in the first bin but this is very rare. We've revised the figure legend to clarify:

Fig. 4. Repeat content characteristics. A) *Proportions of homopolymers, dinucleotide TRs, simple TRs with longer motifs, and complex TRs containing multiple motifs stratified by the mean allele length (binned to nearest 50 bp).* ***B)*** *Allele purities stratified by the mean allele length.* ***C)*** *Example TR alleles of different sizes, purities, and composition.*

Finally, on the argument of utility for the new truthset by retraining DeepVariant, I wonder whether the variants that are filtered correctly only after re-training are in the regions not covered in the previous training / truth sets - are they truly in more difficult regions, or are these perhaps variants where there was more uncertainty with the smaller training sets beforehand?

Thank you for this question and as a first answer, retraining improvement by using the Platinum Pedigree indicates that these are variants that fall in regions not included in other training data. To quantify this, we explored where the gain in accuracy is coming from and found additional calls in homopolymers and tandem

repeats. The following text has been added to the main:

Cataloging the DeepVariant improvements with GIAB genomic stratification reveals that indel error reduction, measured against the two truthsets (GIAB 4.2 and Platinum Pedigree), is most prominent in homopolymers (~60% for both truthsets). This suggests that the original model could not learn to call homopolymers optimally from the original training data, and having more examples allows it to learn better model homopolymer signals. For SNVs, the error reduction differs when considering the different truthsets. We see an 80% improvement in tandem repeats for Platinum Pedigree. In GIAB we find only an 8% reduction in tandem repeats, and most in segmental duplications. This suggests that the Platinum Pedigree tandem repeat training examples may be a relatively more novel set when compared to GIAB.

Presentation of results

- the paper uses quite a few abbreviations that should be introduced to a wider audience. For example, SD regions in the SV methods section (segmental duplications presumably?).

We've gone through and added the full phrases/words in each section, to remind readers of the abbreviations, including Segmental Duplications (SDs).

- Hyperlinks to the references are Paperpile links that require access to the Google doc used for writing the manuscript.

Sorry about this. This was an anomaly of converting from googledocs to word/pdf. In the latest version we are removing the hyperlinks.

Reviewer #3 (Remarks on code availability):

- License of supplied code is proprietary and not a standard open source license which may limit usability.

We have changed the license to MIT: https://opensource.org/license/mit

- Code compiles and has the required information needed for reproducing the results, however, doing so is not a small effort as it also requires substantial compute & downloading of the data.

We agree that downloading and reproducing all of this analysis is a significant undertaking but most of the analyses can be repeated by starting from the VCF files, which greatly reduces the compute burden. Additionally, interested parties can limit themselves to a single sample to assess the truth calls that are provided for each sample.

- Documentation could be improved to be easier to follow in the right sequence.

We have improved the documentation on GitHub so that it is now easier to follow.

- Code documentation and quality is a little variable with some hard-coded data locations remaining in the code - so reproducing would not work out of the box. However, how to make the changes to get it working is pretty obvious most of the time.

We have removed the hard coded paths that can be removed.

- Conda environment files are provided for environment reproducibility in Python; however, these are machine-specific dumps that might need adjustment. Dockerfiles would probably be more easy to use here.

Generally the rust code is fairly portable, and snakemake is a light requirement, however, packaging them into a Docker container would likely be another good option. Currently, we prefer to wait for feedback to ensure that there is enough demand to justify the extra work.

Reviewer #1:
Remarks to the Author:
Thank you to the authors for addressing most of my comments and concerns. I only have two remaining points:

The statement, "There were 2,112 Alu, 679 SVA, 324, and 383 LINE-1 element insertions in the family," seems to present aggregated numbers for TEs, rather than at a per-person level. However, the reported SVA number seems questionable.

> *We have increased the stringency criteria for full length transposable element reporting. We have changed the main text to reflect our updates:*
>
> *There were 1,732 Alu, 81 SVA, and 373 LINE-1 element insertions in the family.*
>
> We have also documented our changes in the methods:
>
> To further filter SVA insertions, we set a minimum size filter of 600 bp. Candidate SVA insertions that met the size restrictions were analyzed using Tandem Repeats Finder (TRF) (version 4.09) to verify that the sequences were not composed entirely of tandem repeats. We used the highest scoring tandem repeat identified by TRF to test the composition of the candidate SVAs. Those that consisted only of tandem repeat sequences were removed from the analysis. To filter *Alu* element insertions that were too small to be full insertions, we set a minimum size filter of 240 bp.

The new Figure 2B is missing a color legend for SNV.

> Thank you for catching that error we've fixed it.

Reviewer #2:
Remarks to the Author:
Accept, I have no further concerns.

> No response required

Reviewer #3:
Remarks to the Author:
All my comments from the first round of reviews were addressed. I recommend to accept.

> No response required