HiFi long-read genomes for difficult-to-detect, clinically relevant variants

Authors

Wolfram Höps, Marjan M. Weiss, Ronny Derks, ..., Alexander Hoischen, Christian Gilissen, Lisenka E.L.M. Vissers

Correspondence

christian.gilissen@radboudumc.nl (C.G.), lisenka.vissers@radboudumc.nl (L.E.L.M.V.)

Detecting pathogenic germline variants in the clinic remains technically challenging. We analyzed 145 previously identified, hard-to-detect variants in 100 samples using HiFi long-read sequencing (LRS). 93% of these variants could be recovered in the data, highlighting LRS as a promising single technology for the diagnosis of rare diseases.

Höps et al., 2025, The American Journal of Human Genetics *112*, 450–456
February 6, 2025
https://doi.org/10.1016/j.ajhg.2024.12.013

CellPress

REPORT

HiFi long-read genomes for difficult-to-detect, clinically relevant variants

Wolfram Höps,[1,2,6] Marjan M. Weiss,[1,2,6] Ronny Derks,[1,2] Jordi Corominas Galbany,[1] Amber den Ouden,[1,2] Simone van den Heuvel,[1] Raoul Timmermans,[1] Jos Smits,[1] Tom Mokveld,[3] Egor Dolzhenko,[3] Xiao Chen,[3] Arthur van den Wijngaard,[4] Michael A. Eberle,[3] Helger G. Yntema,[1] Alexander Hoischen,[1,2,5,6] Christian Gilissen,[1,2,6,*] and Lisenka E.L.M. Vissers[1,2,6,*]

Summary

Clinical short-read exome and genome sequencing approaches have positively impacted diagnostic testing for rare diseases. Yet, technical limitations associated with short reads challenge their use for the detection of disease-associated variation in complex regions of the genome. Long-read sequencing (LRS) technologies may overcome these challenges, potentially qualifying as a first-tier test for all rare diseases. To test this hypothesis, we performed LRS (30× high-fidelity [HiFi] genomes) for 100 samples with 145 known clinically relevant germline variants that are challenging to detect using short-read sequencing and necessitate a broad range of complementary test modalities in diagnostic laboratories. We show that relevant variant callers readily re-identified the majority of variants (120/145, 83%), including ~90% of structural variants, SNVs/insertions or deletions (indels) in homologous sequences, and expansions of short tandem repeats. Another 10% ($n = 14$) was visually apparent in the data but not automatically detected. Our analyses also identified systematic challenges for the remaining 7% ($n = 11$) of variants, such as the detection of AG-rich repeat expansions. Titration analysis showed that 90% of all automatically called variants could also be identified using 15-fold coverage. Long-read genomes thus identified 93% of challenging pathogenic variants from our dataset. Even with reduced coverage, the vast majority of variants remained detectable, possibly enhancing cost-effective diagnostic implementation. Most importantly, we show the potential to use a single technology to accurately identify all types of clinically relevant variants.

The >7,000 rare diseases (RDs) known to date collectively present a common healthcare issue. More than 70% of RDs are genetic in origin, and their molecular genetic diagnosis is important for patients and families.[1] Comprehensive diagnostics of rare genetic diseases requires a complex mix of diverse testing modalities. Many diagnostic laboratories still apply traditional approaches, such as karyotyping, fluorescence *in situ* hybridization (FISH), genomic microarrays, Southern blotting, multiplex ligation probe amplification (MLPA), and Sanger sequencing, leading to complex, long-lasting, and often expensive testing cascades. Over the last decade, next-generation sequencing (NGS) techniques, particularly exome and genome sequencing (ES and GS, respectively), have emerged as more generic clinical tests.[2,3] While GS represents the most successful first-tier test to date, a recent systematic study showed that short-read sequencing (SRS)-based GS would not be suited to replace all other diagnostic approaches for up to 25% of all patients referred for testing.[4] For this group of patients, technical limitations associated with short-read technologies prevent the robust identification of certain clinical variant types. These include, but are not limited to, short tandem repeat (STR) expansions, complex structural rearrangements (such as translocations, complex structural variations [SVs], and mobile element insertions [MEIs]), variants in segmental duplications, and genes with homologies and pseudogenes. Other factors that make the detection of specific variants with short-read whole-genome sequencing (SR-WGS) challenging include extreme GC content and homopolymer stretches. Moreover, certain readouts such as methylation are entirely missing in SR-WGS and require separate test modalities. Long-read sequencing (LRS) may, however, overcome many of these challenges. Emerging LRS technologies include PacBio high-fidelity (HiFi) LRS, Oxford Nanopore Technologies (ONT) nanopore-based sequencing, and synthetic long-read technologies such as iCLR.[5] These technologies have matured to enable population-level sequencing[6–8] and have led to discoveries in RD research.[9] A systematic assessment of their application in clinical diagnosis has, however, yet to be performed.

As a first step toward estimating the clinical utility of whole-genome HiFi LRS, we performed an analysis for 145 known clinically relevant variants from 100 samples, specifically enriched for variants that are challenging or impossible to identify by SR-WGS (Figures 1A, 1B, and S1; Table S1), to identify technological benefits and limitations of relevance for clinical use. Of note, 42 of the 100

[1]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; [2]Radboudumc Research Institute for Medical Innovation, Radboud University Medical Center, Nijmegen, the Netherlands; [3]Pacific Biosciences, Menlo Park, CA, USA; [4]Department of Clinical Genetics, Maastricht University, Maastricht, the Netherlands; [5]Department of Internal Medicine, and Radboud Expertise Center for Immunodeficiency and Autoinflammation and Radboud Center for Infectious Disease (RCI), Radboud University Medical Center, Nijmegen, the Netherlands
[6]These authors contributed equally
*Correspondence: christian.gilissen@radboudumc.nl (C.G.), lisenka.vissers@radboudumc.nl (L.E.L.M.V.)
https://doi.org/10.1016/j.ajhg.2024.12.013.

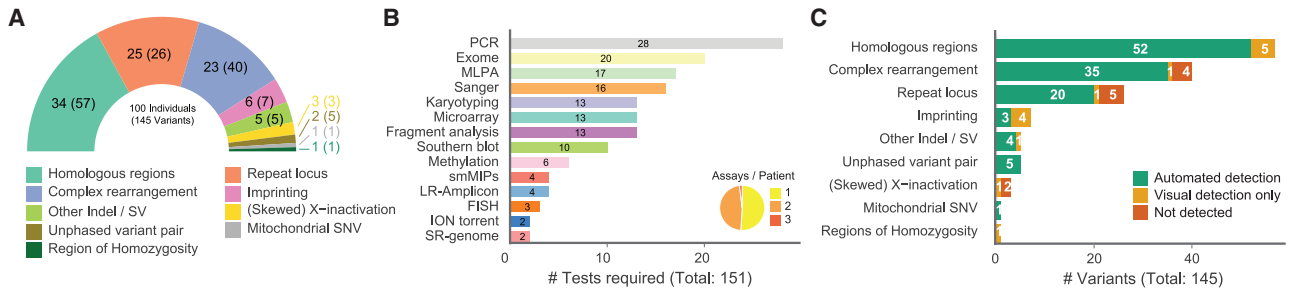

Figure 1. Samples, variants, and LRS-based recovery

(A) Pie chart depicting the cohort composition by variant type for all 145 variants. The number of samples is indicated within parentheses.

(B) Different test modalities (y axis) that were used in a diagnostic laboratory to identify all 145 clinically relevant variants in the 100 selected samples (x axis). The number of assays required per patient is shown in an inlay pie chart.

(C) Sensitivity of LRS by automated variant detection and visual inspection for all 145 variants from 100 analyzed samples (x axis), stratified by *a priori* known disease-associated variant types (y axis). LRS-based detection rates are indicated in green (detected by a variant caller, 83% [$n = 120$]), orange (detection by visual read inspection, 10% [$n = 14$]), and red (undetected variant, 7% [$n = 11$]).

samples, containing 70 variants, were previously included in a clinical utility study using 30× short-read genomes.[4] This earlier study demonstrated the challenges of SR-WGS variant detection, as only 29 of the 70 variants could be identified using automated variant calling algorithms. Automated detection was challenging because of the type of variant (e.g., STRs and SVs) or their location within the genome (e.g., homologous regions). The remaining 58 samples were chosen to contain variants expected to be equally challenging to call, e.g., due to being located in hard-to-map regions of the genome. Hence, the total series included 25 samples with STR expansions, 34 samples with variants characterized by homology and/or pseudogenes, and 23 samples with (complex) structural events. The remaining 18 samples included a variety of other clinically relevant variants, such as imprinted loci and mtDNA variation.

Here, we used a single SMRT cell on a PacBio Revio system for each sample, expected to generate ±30× coverage. All samples were processed in the same fashion, according to the manufacturer's instructions (PacBio, Menlo Park, CA, USA). In brief, 7 μg high-molecular-weight gDNA was sheared on Megaruptor 3 (Diagenode, Liège, Belgium) to a target size of 15–18 kb, libraries were prepared with SMRTbell prep kit 3.0 (PacBio), size-selected >10 kb on the BluePippin (Sage Science, Beverly, MA, USA), and sequenced for 24 h on the Revio system (ICS 12.0.4). Samples were then analyzed using a bioinformatics pipeline that incorporates a variety of software tools for long-read sequencing (LRS) data analysis. Alignment (pbmm2 v.1.10.0) of HiFi reads was performed against the GRCh38 reference genome while generating haplotigs by performing *de novo* assembly with Hifiasm (v.0.15.3). Structural variants (PBSV v.2.9.0) and small variants (DeepVariant v.1.5.0) were called with phasing information (HiPhase v.0.10.1) and annotated using publicly available databases. An additional analysis for copy-number variants (CNVs) based on read depth was performed using HiFiCNV (v.0.1.6). STRs were called (TRGT v.0.4.0), visualized (TRVZ v.0.4.0), and annotated using an in-house pipeline. Specific variant calls for paralogs and pseudogenes were called using Paraphase (v.2.2.3). Methylation calls were generated using pb-cpg-tools (v.2.3.1). When the expected variant was not detected by the respective software, the sequencing data were visually inspected using the Integrated Genomics Viewer (IGV; v.2.16.2) for region(s) containing the variant(s), taking into account a suitable genomic window. The mode of visual confirmation depended on the variant type, and variants were considered visually confirmed when the sequencing reads showed a clear deviant pattern in read mapping, methylation profiles, or variant allele frequencies. All procedures followed were in accordance with the ethical standards of the Medical Ethics Review Committee Arnhem-Nijmegen, and proper informed consent was obtained.

Across all 100 samples, we obtained a median output of 94.0 Gb of data, achieving a median genome-wide coverage of 29.7× with an average read length of 15.35 kb (Table S2). Of the 145 variants analyzed, 120 (83%) were detected fully automatically by the relevant variant calling software (Figure S2). This included, among others, 61 SVs, 20 STR expansions, and 40 single-nucleotide variants (SNV) and indels (Figures 2A, 2B, 2D, and 2E), the majority of which affected loci with homologous sequences (Table S1; Figure 1C). The 20 repeat expansions ranged from 16 to >150 additional repeat units in 12 different genes. In some cases, HiFi genomes also provided additional molecular insights that were not obtained from the traditional clinical tests. For instance, targeted *de novo* assembly by HiFiasm (v.0.15.3) and visualization by NAHRwhals v.1.4[10] allowed us to resolve the structure of a complex genomic rearrangement encompassing *CEP85L* (MIM: 618865) and *MCM9* (MIM: 610098) (Figure 2F). In another case (P1-C1), two additional, smaller duplications were identified adjacent to the known pathogenic variant, and in a final case involving the *OPN1LW*/*OPN1MW* (MIM: 300822/300821) homologous gene cluster (P11-F11), the sequence context of *OPN1MW2*

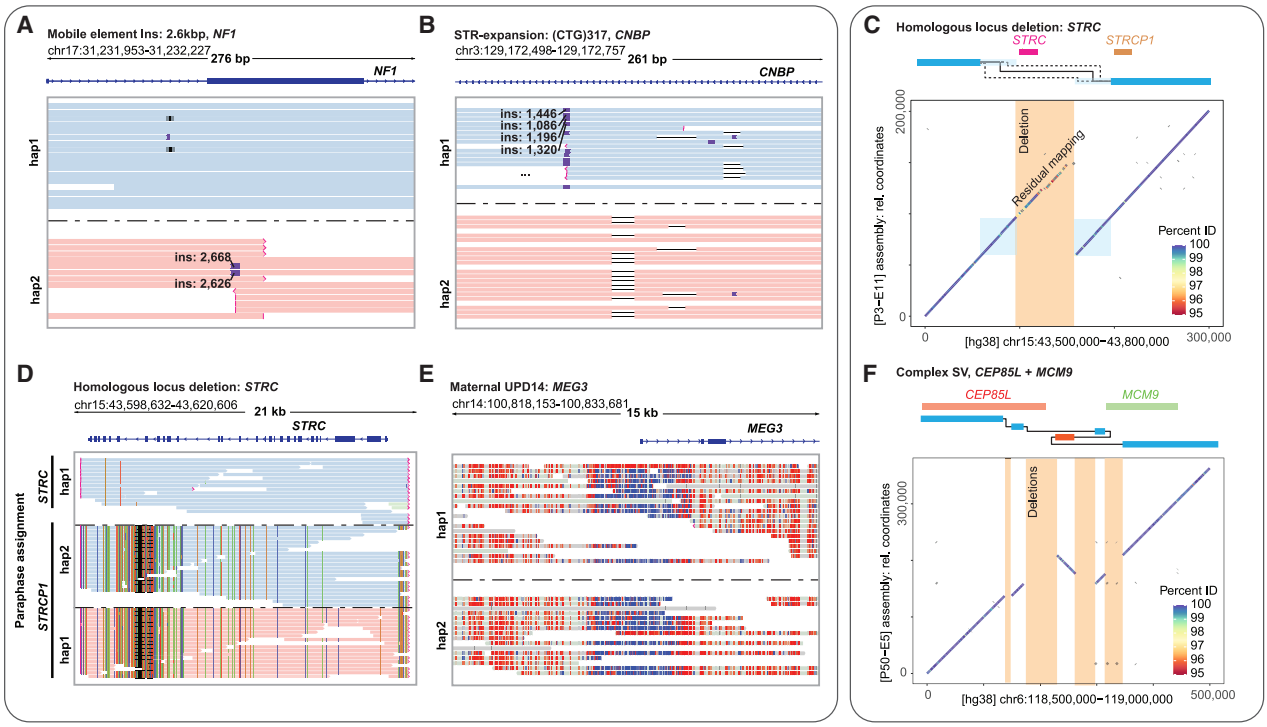

Figure 2. Examples of variants identified in an automated fashion or by visual inspection
(A and B) IGV screenshots of long-read sequencing data for specific variants from samples P50-A1 and P4-H11, respectively. Reads are colored by phase.
(C) Visualization of the *de novo* assembly of a deletion of *STRC*, with the pseudogene *STRCP1* intact for sample P3-E11. Using the mapping quality metric, the breakpoint can be narrowed down to a ~30 kbp window (light blue squares). A schematic view of the genes and assembly mapping is indicated on top, and raw dot-plot mappings of GRCh38 (x axis) vs. the assembled region (y axis) are displayed on the bottom.
(D) The same variant from (C) visualized with Paraphase. Reads are grouped by inferred (pseudo)gene identity. Only one haplotype of *STRC* is observed, thus indicating a deletion of the other allele.
(E) An imprinting defect on the maternal chromosome 14 due to a uniparental heterodisomy for sample P50-G3. Reads are colored by methylation status, with blue indicating unmethylated CpGs and red methylated CpGs.
(F) *De novo* assembly of a locus containing a ~200 kbp complex genomic rearrangement for sample P50-E5.

was revisited as a result of the entire copy having been resolved at the base-pair level. While similar benefits of SR-WGS have been reported from direct comparisons of short-read genomes to exomes,[11] the ability to analyze even these complex loci of the human genome using LRS provides great promise for clinical care.

For an additional 10% ($n = 14$) of variants (Figure S2; Table S1), automated detection failed, requiring either manual inspection of aligned sequencing reads ($n = 8$) or visualization of processed data files in the absence of dedicated callers ($n = 6$; Figures S3–S5). Manual inspection successfully identified variants in the *OPN1LW/OPN1MW* ($n = 3$) and *CFH* (MIM: 235400) ($n = 2$) gene clusters, as well as a complex structural rearrangement involving complex/repetitive regions, one of the STRs (in *CNBP* [MIM: 116955]; Figure S3), and an MEI inserted in *NF1* (MIM: 613113). Once again, molecular insights were further enhanced by *de novo* assemblies, for instance, discriminating between the deletion of *STRC* (MIM: 606440) and its *STRCP1* pseudogene copy in one of the samples (P3-E11; Figure 2C; supplemental material and methods). We expect that further improvements in structural variant detection based on *de novo* assembly should allow for the automatic detection of these variants in the future. It may, however, already be beneficial for long-read technologies to use the most accurate human reference genome for mapping purposes, such as the one presented by the telomere-to-telomere (T2T) consortium.[12] Indeed, in one of the samples (P50-E2), an unbalanced translocation involving chromosomes 13 and Y, the variants could only be fully resolved when mapping to the T2T reference instead of GRCh38. The fact that this case of a translocation could only be fully resolved using the T2T reference genome adds to the notion of substantial benefits of a complete reference genome.[12] While not all tools and annotations are available yet for this reference, we believe that its integration as a "supplementary" analysis or for follow-up of particular cases should already be considered in clinical care, particularly when long reads are available.

For six variants, an evaluation of detection could only be performed after additional computational analyses of the output files generated by the standardized variant callers included in our pipeline (supplemental material and methods). Three of these variants concerned regions of

homozygosity (ROHs) for which no appropriate calling algorithm was available (Figure S4). For these, we visualized homozygosity by plotting variant allele frequencies of SNPs within the region (Figure S4; supplemental material and methods), allowing the detection of each of them. We anticipate that methods for automatically detecting ROHs and uniparental disomies (UPDs[13]) from SR-WGS can be readily adapted for long-read data, enabling the automated detection of these variants. The other three variants concerned various percentages of skewed X inactivation (P5-H5, 80%/20%; P50-A6, 90%/10%; and P50-E8, 100%/0%). To detect these variants, we devised a method to calculate the ratio of methylation per allele in all female samples (supplemental material and methods). Only one of three showed a clear deviation in the methylation pattern confirming skewing (P50-E8; Figure S5), whereas the other two could not be discriminated from the other 49 tested female samples (Figure S6). This suggests that LRS either lacks the sensitivity to detect skewing below 80% (at 30× coverage) or that the standard-of-care method to detect skewed X-inactivation involving the methylation status of a single gene (*AR*) may not be predictive for the methylation status of the entire X chromosome. However, targeted interpretation of the *AR* locus in the LRS data of these two samples also did not suggest skewing. Notably, two other samples (P2-E4, P1-C1) identified by our analysis as potentially having skewed X inactivation had pathogenic structural variants on the X chromosome. It is to be expected that new methods for the automated calling of methylation defects from LRS data will further enhance the utility of LRS in a clinical setting.

Of all 145 variants, 11 (7%) could not be detected in HiFi genomes by automated callers, visual inspection, or additional computational analyses. We investigated whether a common factor could explain these detection failures. Two of the variants were related to the unverifiable skewed X inactivation described above. Five variants (3%) involved GA-based repeat expansions, including $(GAA)_n$ in *FXN* and $(AAGGG)_n$, $(AAAGG)_n$, or $(ACAGG)_n$ in *RFC1*, suggesting a systematic issue (Table S1). For these variants, HiFi sequencing suffered from reduced quality, resulting in fewer high-quality reads being available for HiFi read generation, leading to insufficient sequencing coverage in these regions for variant calling. This reduced quality for GA-repeat regions may result from the formation of non-B DNA conformations, hampering the DNA polymerase and reducing read length and quality.[14] Based on this assumption, we hypothesized that if non-HiFi quality reads could be "rescued," then additional coverage could be added for these regions, potentially allowing for better calling of repeat lengths. Indeed, by manually adding "rejected," low-quality reads to the HiFi data, we increased the coverage for *FXN*, enabling the automatic detection of the pathogenic repeat (Figure S7; Table S3). However, this approach did not work for *RFC1* repeats where coverage remained too low to detect the expansion in all four cases. Notably, sufficient coverage is observed at these repeat regions for wild-type alleles, with reduced coverage only apparent in cases of repeat expansions. This observation supports our hypothesis of allelic dropout due to technical difficulties as a consequence of the repeat expansion. Hence, detecting reduced coverage at these repeat loci could serve as an indicator of a potential repeat expansion, warranting further follow-up. This proxy detection method of repeat expansions might be useful when HiFi genomes are used as a first-tier test. Alternatively, given the clinically recognizable phenotypes associated with most GA-based repeat expansions (e.g., rare neurological movement disorders), a targeted approach, such as the PureTarget technology, may also be considered, as the increase in read depth for those loci would potentially compensate for the loss of high-quality reads, allowing automated calling.

For the remaining four variants that could not be detected, detection was hampered by the fact that each variant was associated with breakpoints in highly repetitive regions (P4-C4 and P3-F6) or segmental duplications >50 kb in size (P1-C1 and P9-B1). For P4-C4, a translocation between the acrocentric p arm of chromosome 22 and the repeat-rich regions of the Y chromosome, and P3-F6, involving a Robertsonian translocation affecting the acrocentric p arms of chromosomes 13 and 14, we attempted re-alignment using the T2T reference genome, but this did not recover the variants. P1-C1 and P9-B1 concerned unbalanced translocation events whose breakpoints fell inside segmental duplications, leading to their incorrect classification as simple CNVs. Detecting these four variants may require further algorithmic development or the use of long-range DNA information that can span regions >50 kb, as recently demonstrated by others.[15]

We next focused our attention on a technical evaluation of the role of sequencing coverage. Several studies have assessed minimum coverage thresholds for LRS.[16] However, these studies relied on gold standards of unselected variants that were also detectable with SR-WGS, were based on non-human models,[16] or alternatively focused on one specific class of variants.[17] Given our unique heterogeneous set of validated variants, we saw an opportunity to estimate the sensitivity for detection in relation to the depth of coverage. To this end, we implemented an *in silico* downsampling experiment (supplemental material and methods). Three samples (P13-G4, P50-G5, P50-H7) with an initial coverage of less than 20-fold were excluded from this analysis. For the remaining samples, BAM files were subsampled to target coverages of 10×, 15×, and 20× using the "*samtools view -s*" command (samtools v.1.11; supplemental material and methods). This process was repeated 10 times for each target coverage, resulting in a total of 30 subsampled readsets per sample, or 10 per target coverage. Downstream processing and variant calling were performed analogously to the original datasets, resulting in independent variant callsets for each permutation. In the original dataset, these 97 samples contained 117 biological variants that were detectable by automated

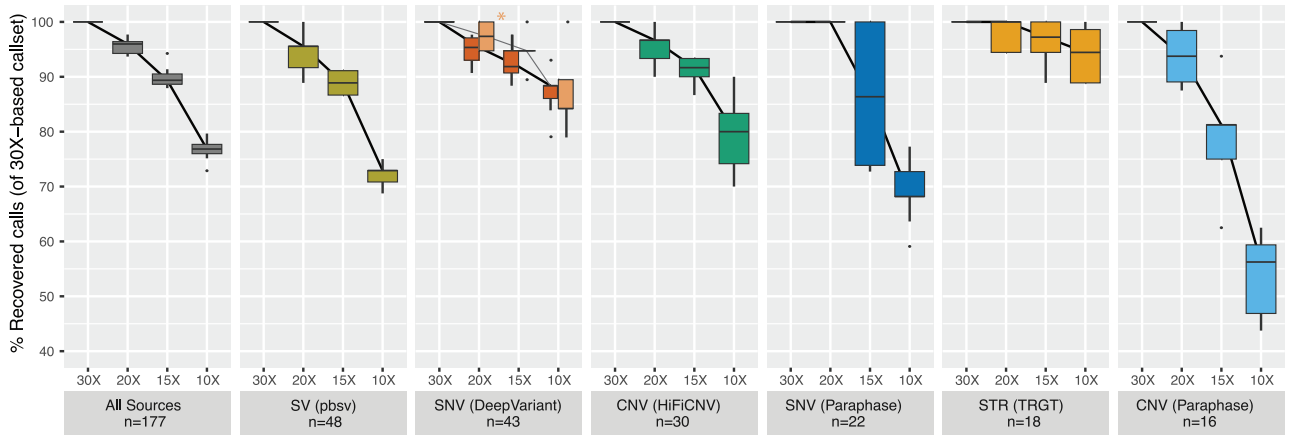

Figure 3. Variant recall in titration experiments
Results of automated variant detection per variant and calling tool for different genome-wide coverage levels (10×, 15×, 20×, and 30×) based on the initial 177 calls (Table S4). Boxplots are based on 10 random selections of different reads from the original 30× coverage sample. For SNVs, we distinguished between all SNVs (in red; $n = 43$) and SNVs not overlapping a homologous region (in light orange with asterisk; $n = 18$).

callers. Due to cross-calling between different tools (e.g., HiFiCNV and pbsv both calling certain CNVs; Table S1), those 117 variants were represented by 177 individual variant calls, which were tested for recall separately (Table S4). To account for potential small-variant incongruencies, e.g., caused by (micro)homology near breakpoints, we tested and implemented custom similarity thresholds for all classes of variants (Figures S8–S10; supplemental material and methods).

We found a reduced discovery rate from the initial 177 variants identified at ∼30×, with median detection rates dropping to 170 (96.0%), 158.5 (89.5%), and 136 (76.8%) for 20×, 15×, and 10× coverages, respectively (Figure 3; Table S4). Further stratification by variant type and caller revealed the sharpest decline in repeat-associated CNVs and SNVs (recalled by Paraphase) (−18.8% and −13.6% in 15× vs. 30×), as well as SVs (−10.4%), CNVs (−8.3%), and SNVs/indels in homologous regions (−10.4%). STRs (−2.8%) and SNVs outside of homologous regions (−5.3%) were less affected (Table S4). We also noted that the reduced sensitivity for SNVs at 20× coverage was primarily driven by those within homologous regions, whereas sensitivity for SNVs outside these regions diminished from 15× coverage and lower. Our results suggest that 30× or higher coverage may be needed to capture difficult-to-detect variants and harness the full potential of LRS. However, we also observe that at 15×, the primary impact is on the sensitivity for detecting variants in homologous regions. In contrast, the sensitivity of SNVs, CNVs, SVs, and STRs in other genomic regions is reduced only by about 10%. Depending on the circumstances, the ability to sequence more samples at the same cost may outweigh this reduced sensitivity.

Overall, our data show that currently available HiFi variant-calling methods could potentially lead to an automated detection rate of 83% based on the variants we tested, with an additional 10% visually apparent in the data that could likely be recovered using specialized or improved calling algorithms. Nonetheless, it should be noted that these numbers do not readily reflect a performance metric of LRS for clinical use. The clinically relevant variants were known *a priori*, allowing us to focus on these loci and sidestep variant prioritization. Moreover, we enriched for difficult-to-detect, clinically relevant variants to test the possibilities and limitations of the technology, suggesting that the actual sensitivity for unselected samples in a clinical setting will likely be higher. Although sensitivity and specificity parameters are essential for the implementation of LRS in routine diagnostic care, determining these will require a different study design, such as a prospective evaluation of LRS in parallel to the standard of (genetic) care. This would also allow us to perform a cost-effectiveness analysis of LRS for routine clinical use. Irrespectively, our outcomes are already of particular relevance in the context of genetic diagnostic laboratories that are considering short- and long-read approaches for generic first-tier germline testing for RDs. As mentioned, a subset of 42 samples tested here (70 variants) were previously also used to technically benchmark short-read genome sequencing. This unique setup allows for a direct comparison between SRS and LRS for these difficult-to-detect regions of the genome. In short-read genomes, 29/70 variants (41%) were automatically detected compared to 62/70 variants (89%) in HiFi long-read genomes (Tables S5 and S6). Whereas 21 variants in the short-read data could be recovered by visual inspection of the aligned reads, this approach is only feasible for small genomic loci or a limited set of genes, requiring a strong clinical diagnosis to guide interpretation.[18] We note, however, that these 21 variants may potentially have been picked up in short reads if other calling tools had been used. The remaining 20 variants were completely undetected in

short-read genomes. We re-evaluated all these loci in short-read and long-read reference data, demonstrating that at least 90% of them are indeed unlikely to be recoverable in short-read genomes, most commonly due to overlap with segmental duplications (supplemental material and methods; Figures S11 and S12).

Along with previous studies that have demonstrated the ability to identify molecular diagnoses in a significant percentage of previously undiagnosed cases,[17,19,20] our results suggest that HiFi genomes may be a more attractive first-tier, generic assay for germline testing in RDs. Nevertheless, there are still several limiting factors to a straightforward adoption in clinical settings. Most importantly, LRS is still several times more expensive than SRS, although our results suggest that less coverage may be required for LRS than for SRS to obtain similar sensitivity. The current lack of scalability of LRS capacity and automation additionally poses a major challenge for clinical implementation at the moment. Ongoing developments in LRS technology suggest, however, that these limitations are likely to be temporary.

In summary, our study provides detailed insights into the abilities and limitations of LRS to identify the full spectrum of clinically relevant genome variations. Although prospective studies are still needed, our results show that LRS has the potential to be implemented as a first-tier diagnostic workflow for germline testing, potentially replacing all current tests for diagnosing individuals with RDs.

Data and code availability

Our study is performed on sensitive patient data, which are subject to institutional review board (IRB) restrictions. We are thus unable to submit primary genotype data to a public database. All data available for sharing, particularly information about pathogenic variants, are included as Tables S1, S2, S3, S4, S5, and S6. We are open to legitimate data requests, which will be processed according to national and institutional guidelines. The code generated to create raw versions of all main figures is available at https://github.com/WHops/revio1figures (https://doi.org/10.5281/zenodo.14036018). Code generated for the comparison of datasets in the downsampling experiment (pre-processing for Figure 3) can also be obtained through GitHub at https://github.com/WHops/lrs100_downsample/ (https://doi.org/10.5281/zenodo.14035709).

Acknowledgments

We thank Dr. R. Blok, T. Hofste, Dr. L. Haer-Wigman, Dr. E. Kamsteeg, Dr. N. de Leeuw, Dr. D. Lugtenberg, Dr. T. Rinne, Dr. A. Simons, Dr. C. Harteveld, and R. Smeets for useful discussions. We thank the Klinisch Genetisch Centrum Nijmegen, the Radboud Genome Technology Center, and the Netherlands X-omics Initiative NWO (project 184.034.019) for technical and financial support. The aims of this study contribute to the Solve-RD project (to A.H., C.G., and L.E.L.M.V.), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 779257. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101150006.

Declaration of interests

T.M., E.D., X.C., and M.A.E. are employees and shareholders of Pacific Biosciences, a company commercializing DNA sequencing technologies. Pacific Biosciences also kindly provided part of the reagents required for this study.

Web resources

OMIM, http://www.omim.org

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.12.013.

Received: August 31, 2024
Accepted: December 12, 2024
Published: January 13, 2025

References

1. Nguengang Wakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur. J. Hum. Genet. *28*, 165–173. https://doi.org/10.1038/s41431-019-0508-0.
2. Wojcik, M.H., Lemire, G., Berger, E., Zaki, M.S., Wissmann, M., Win, W., White, S.M., Weisburd, B., Wieczorek, D., Waddell, L.B., et al. (2024). Genome Sequencing for Diagnosing Rare Diseases. N. Engl. J. Med. *390*, 1985–1997. https://doi.org/10.1056/NEJMoa2314761.
3. Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., et al. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. Nature *583*, 96–102. https://doi.org/10.1038/s41586-020-2434-2.
4. Schobers, G., Derks, R., den Ouden, A., Swinkels, H., van Reeuwijk, J., Bosgoed, E., Lugtenberg, D., Sun, S.M., Corominas Galbany, J., Weiss, M., et al. (2024). Genome sequencing as a generic diagnostic strategy for rare disease. Genome Med. *16*, 32. https://doi.org/10.1186/s13073-024-01301-y.
5. Gorzynski, J.E., Marwaha, S., Reuter, C.M., Jensen, T., Ferrasse, A., Raja, A., Fernandez, L., Kravets, E., Carter, J., Bonner, D., et al. (2024). Clinical application of Complete Long Read genome sequencing identifies a 16kb intragenic duplication in EHMT1 in a patient with suspected Kleefstra syndrome. Preprint at medRxiv. https://doi.org/10.1101/2024.03.28.24304304.
6. Schloissnig, S., Pani, S., Rodriguez-Martin, B., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., et al. (2024). Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. Preprint at bioRxiv. https://doi.org/10.1101/2024.04.18.590093.
7. Gustafson, J.A., Gibson, S.B., Damaraju, N., Zalusky, M.P., Hoekzema, K., Twesigomwe, D., Yang, L., Snead, A.A., Richmond, P.A., De Coster, W., et al. (2024). Nanopore sequencing of 1000 Genomes Project samples to build a comprehensive catalog of human genetic variation. Preprint at medRxiv. https://doi.org/10.1101/2024.03.05.24303792.

8. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet. *53*, 779–786. https://doi.org/10.1038/s41588-021-00865-4.
9. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front. Genet. *10*, 426. https://doi.org/10.3389/fgene.2019.00426.
10. Höps, W., Rausch, T., Jendrusch, M., Human Genome Structural Variation Consortium HGSVC, Korbel, J.O., and Sedlazeck, F.J. (2024). Impact and characterization of serial structural variations across humans and great apes. Nat. Commun. *15*, 8007. https://doi.org/10.1038/s41467-024-52027-9.
11. van der Sanden, B.P.G.H., Schobers, G., Corominas Galbany, J., Koolen, D.A., Sinnema, M., van Reeuwijk, J., Stumpel, C.T.R.M., Kleefstra, T., de Vries, B.B.A., Ruiterkamp-Versteeg, M., et al. (2023). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur. J. Hum. Genet. *31*, 81–88. https://doi.org/10.1038/s41431-022-01185-9.
12. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2022). A complete reference genome improves analysis of human genetic variation. Science *376*, eabl3533. https://doi.org/10.1126/science.abl3533.
13. Yauy, K., de Leeuw, N., Yntema, H.G., Pfundt, R., and Gilissen, C. (2020). Accurate detection of clinically relevant uniparental disomy from exome sequencing data. Genet. Med. *22*, 803–808. https://doi.org/10.1038/s41436-019-0704-x.
14. Mellor, C., Perez, C., and Sale, J.E. (2022). Creation and resolution of non-B-DNA structural impediments during replication. Crit. Rev. Biochem. Mol. Biol. *57*, 412–442. https://doi.org/10.1080/10409238.2022.2121803.
15. Guarracino, A., Buonaiuto, S., de Lima, L.G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Gerton, J.L., et al.; Human Pangenome Reference Consortium (2023). Recombination between heterologous human acrocentric chromosomes. Nature *617*, 335–343. https://doi.org/10.1038/s41586-023-05976-y.
16. Lee, H., Kim, J., and Lee, J. (2023). Benchmarking datasets for assembly-based variant calling using high-fidelity long reads. BMC Genom. *24*, 148. https://doi.org/10.1186/s12864-023-09255-y.
17. Noyes, M.D., Harvey, W.T., Porubsky, D., Sulovari, A., Li, R., Rose, N.R., Audano, P.A., Munson, K.M., Lewis, A.P., Hoekzema, K., et al. (2022). Familial long-read sequencing increases yield of de novo mutations. Am. J. Hum. Genet. *109*, 631–646. https://doi.org/10.1016/j.ajhg.2022.02.014.
18. Corominas, J., Smeekens, S.P., Nelen, M.R., Yntema, H.G., Kamsteeg, E.-J., Pfundt, R., and Gilissen, C. (2022). Clinical exome sequencing-Mistakes and caveats. Hum. Mutat. *43*, 1041–1055. https://doi.org/10.1002/humu.24360.
19. Steyaert, W., Sagath, L., Demidov, G., Yépez, V.A., Esteve-Codina, A., Gagneur, J., Ellwanger, K., Derks, R., Weiss, M., den Ouden, A., et al. (2024). Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. Preprint at medRxiv. https://doi.org/10.1101/2024.05.03.24305331.
20. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics *31*, 2601–2606. https://doi.org/10.1093/bioinformatics/btv201.