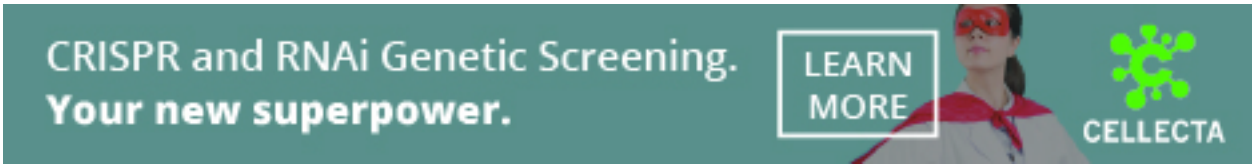
Analytical validation of germline small variant detection using long-read HiFi genome sequencing

Nathan Hammond, Linda Liao, Pun Wai Tong, et al.

Genome Res. published online April 11, 2025
Access the most recent version at doi:10.1101/gr.278836.123



Analytical validation of germline small variant detection using long-read HiFi genome sequencing

Nathan Hammond[1*], Linda Liao[1], Pun Wai Tong[1], Zena Ng[1], Thuy-Mi P. Nguyen[2], Chandler Ho[1], Yao Yang[1,2], Stuart A. Scott[1,2]

[1] Clinical Genomics Laboratory, Stanford Medicine, Palo Alto, CA 94304.
[2] Department of Pathology, Stanford University, Stanford, CA 94305.

Current affiliation:
[*] Influx Bio, San Francisco, CA 94124.

Running Title: Long-read HiFi genome sequencing validation

Manuscript Category: Regular Article
Abstract: 234
Tables: 3
Figures: 3

Key words: Clinical genome sequencing; long-read sequencing; SMRT sequencing; HiFi sequencing; germline genetic testing; diagnostic testing; medical genomics; LRS Special Issue.

Sources of support: S.A.S was supported in part by NIH/NHGRI grant U01HG011762.

Disclosures: N.H. is currently an employee of Influx Bio; all other authors declare no conflicts of interest.

CORRESPONDENCE TO:
Stuart A. Scott, PhD
Professor,
Department of Pathology
Stanford University
Palo Alto, CA 94305
Telephone: 650-724-0973
Email: sascott@stanford.edu

ABSTRACT

Long-read sequencing has the capacity to interrogate difficult genomic regions and phase variants; however, short-read sequencing is more commonly implemented for clinical testing. Given the advances in long-read HiFi sequencing chemistry and variant calling, we analytically validated this technology for small variant detection (single nucleotide variants, insertions/deletions; SNVs/indels; <50bp). HiFi genome sequencing was performed on DNA from reference materials and clinical specimen types, and accuracy results were compared to short-read genome sequencing data. HiFi genome sequencing recall and precision across Genome in a Bottle (GIAB)-defined non-difficult and difficult genomic regions (high confidence) for SNVs were >99.9% and >99.7%, respectively, and for indels were >99.8% and >99.1%, respectively. Moreover, HiFi genome sequencing outperformed short-read genome sequencing on overall SNV/indel F1-score accuracy at all paired sequencing depths, which were further stratified across 100 total GIAB-defined genomic regions for a comprehensive evaluation of performance. Of note, HiFi genome sequencing F1-scores for SNVs and indels surpassed 99% at ~15× and ~25×, respectively. In addition, high confidence small variant concordance across all HiFi genome sequencing reproducibility assessments (two specimens, three independent sequencing datasets) were >99.8% for SNVs and >98.6% for indels, and average high confidence small variant concordance between paired blood, saliva, and swab specimens were all >99.8%. Taken together, these data underscore that long-read HiFi genome sequencing detection of SNVs and indels is very accurate and robust, which supports the implementation of this technology for clinical diagnostic testing.

INTRODUCTION

Genome sequencing has evolved from a widely used research platform to a comprehensive clinical test at selected medical centers and laboratories (Belkadi et al. 2015; Costain et al. 2020), with the capacity to sequentially interrogate regions of the genome based on new evidence and/or clinical indication (Rehm 2017; Costain et al. 2018; Bick et al. 2019; Yang et al. 2024). Short-read sequencing is the most commonly implemented platform for genome sequencing; however, long-read sequencing is rapidly emerging as an alternative platform with notable benefits over short-read sequencing (Logsdon et al. 2020; Cohen et al. 2022; Conlin et al. 2022). For example, long-read genome sequencing has improved interrogation of clinically significant regions, including structural variants, repeat expansions, homologous gene families and the *HLA* region, as well as the inherent benefit of variant phasing (Ardui et al. 2018; Ameur et al. 2019). The two primary long-read sequencing chemistries currently available are single molecule real-time (SMRT; HiFi) and nanopore sequencing, which recently have been employed by the Telomere-to-Telomere (T2T) Consortium to more comprehensively characterize the CHM13 human reference assembly (Nurk et al. 2022).

Long-read SMRT sequencing has been shown to generate highly accurate high-fidelity (HiFi) read lengths of ~10-25 kb using the Sequel II platform (Pacific Biosciences) (Wenger et al. 2019; Hon et al. 2020). More recently, long-read HiFi genome sequencing has been used to expand the small variant benchmarks in the commonly leveraged Genome in a Bottle Consortium (GIAB) reference material samples to include difficult-to-map regions and segmental duplications that are inherently challenging for short reads (Wagner et al. 2022a). In addition, comparisons of sequencing platforms and variant calling strategies have recently been reported by the

PrecisionFDA Truth Challenge V2, which found long-read HiFi genome sequencing to outperform both short-read and long-read nanopore sequencing with genome-wide variant calling accuracy (Olson et al. 2022).

Although long-read HiFi genome sequencing has improved accuracy and haploblock phasing performance compared to short-read sequencing, its adoption into clinical genetic testing laboratories is only now emerging. Of note, resources for validating new clinical sequencing assays are available from the College of American Pathologists (CAP), Association for Molecular Pathology (AMP) (Aziz et al. 2015; Roy et al. 2018), and related professional consortia (Gargis et al. 2012; Matthijs et al. 2016; Santani et al. 2017; Santani et al. 2019); as well as benchmarking reference materials from the GIAB/National Institute of Standards and Technology (NIST) consortium,(Zook et al. 2019) and the Global Alliance for Genomics and Health Benchmarking (GA4GH) Team (Krusche et al. 2019). Therefore, to facilitate the implementation of clinical long-read HiFi genome sequencing, our initial effort was centered on a robust analytical validation of germline small variant (SNV/indel) detection using current best practices and benchmarking resources.

RESULTS

Long-Read HiFi Genome Sequencing Small Variant Accuracy

HiFi genome sequencing accuracy (SNV/indel; <50bp) was evaluated by sequencing seven GIAB/NIST reference material samples (average 31.1×). Benchmarking was performed using tools and practices recommended by the GA4GH (Krusche et al. 2019), and hap.py version v0.3.15 was used to compare observed results with the published truth set version v4.2.1.

Accuracy was measured by recall (i.e., sensitivity) and precision (i.e., positive predictive value), which were stratified by high confidence genomic regions as defined by GIAB/NIST (Zook et al. 2019). SNV/indel detection across the GIAB reference samples was highly accurate, as the average recall and precision for all seven samples were >99.9% for SNVs and >99.8% for indels across the non-difficult genomic regions (**Table 1**). As expected, recall and precision were slightly lower when interrogating genomic regions with known sequencing challenges (low complexity, low mappability, segmental duplications); however, average recall and precision across all difficult genomic regions were still >98.9% and >98.7% for SNVs and indels, respectively. Of note, in addition to the low complexity, low mappability, and segmental duplication regions highlighted in **Table 1**, the GIAB 'Difficult regions' are defined broadly to include low or high GC content (<25% or >65%), bad promoter regions, false duplications, and other difficult genomic regions (Krusche et al. 2019). The average recall and precision across all genomic regions were >99.7% and >99.1% for SNVs for indels, respectively (**Table 1**). Among the discordant HiFi small variants with the GIAB truth set, homopolymers were the most common source of error, with ~75% of indel errors located in or adjacent to a homopolymer run.

Long-Read HiFi and Short-Read Genome Sequencing Small Variant Accuracy

In addition to accuracy benchmarking across the high confidence GIAB/NIST regions (i.e., low complexity, low mappability, segmental duplications, all difficult regions, not in any difficult region, all), HiFi genome sequencing performance was further evaluated across 100 GIAB-defined subregions of the human genome (Zook et al. 2019). F1-scores were generated and stratified by variant type (i.e., SNVs/indels), and results compared to paired analyses with publicly available short-read genome sequencing data (average 40.3×). As illustrated in **Figure**

1A-F, HiFi genome sequencing small variant F1-scores were superior to short-read genome sequencing small variant F1-scores across 22 informative genomic subregions within the categories of low mappability, homopolymers, tandem repeats and GC content, which were most notable for indels of increasing size. In addition, HiFi genome sequencing variant calling accuracy across the recently reported Challenging Medically Relevant Gene (CMRG) truth set (Wagner et al. 2022b) was also interrogated and compared with short-read genome sequencing accuracy, which identified HiFi's advantage over short-read in detection of >15bp deletions (87.76% vs. 67.81% recall), >15bp insertions (84.44% vs. 74.81% recall) but otherwise showed roughly similar performance in the CMRG regions (**Supplemental Table S1**).

Genome Sequencing Depth Stratification and Accuracy

To evaluate HiFi genome sequencing small variant accuracy at different sequencing depths, small variant recall and precision were assessed across a series of sequencing depths downsampled from NA24385 (HG002). As expected, HiFi genome-wide small variant accuracy was reduced at low sequencing depths (**Supplemental Table S2**); however, HiFi genome sequencing F1-scores for SNVs and indels surpassed 99% at ~15× and ~25×, respectfully. In comparison, short-read genome sequencing F1-scores for SNVs and indels surpassed 99% at ~20× and ~35×, respectfully (**Figure 2**). At sequencing depths of ~30×, HiFi genome sequencing F1-scores for SNVs and indels were 99.8% and 99.1%, respectfully; and short-read sequencing F1-scores for SNVs and indels were 99.8% and 98.7%, respectfully (**Figure 2A-C**).

Long-Read HiFi Genome Sequencing Concordance and Reproducibility

HiFi genome sequencing library preparation included both manual and automated workflows, and minor updates were introduced to the laboratory procedure to optimize the automated workflow. Workflow and procedure updates were validated by measuring small variant accuracy and/or concordance with reference material and specimen samples as appropriate. As detailed in **Supplemental Tables S3 and S4**, the accuracy and concordance of the workflow updates (e.g., fragment depletion using the SRE XS and SRE kits) were consistent with paired manual library preparation results, which supported the implementation of these workflow improvements. In addition, a novel validation strategy of two Miro Canvas instruments was accomplished by low-depth NA12878 benchmarking comparisons using a single SMRTcell of data (~9× each) (**Figure 3A-B**), which was supported by consistent quality metrics between manual and automated workflows, and concordant benchmarking results from a subsequent full depth (27.9×) Miro Canvas library preparation (**Supplemental Tables S3 and S4**).

HiFi genome sequencing reproducibility was evaluated by comparing reference sample results to two independent publicly available datasets (NA12878/HG001, NA24385/HG002; see **Materials and Methods**) and measuring F1-score concordance across all three datasets (average 32.9×). Genome-wide SNV/indel non-reference genotype concordance was stratified by genomic context, which ranged from ~98-99.9% (**Table 2**); however, HiFi genome sequencing reproducibility was reduced when assessed across regions not considered high confidence by GIAB (**Supplemental Table S5**). Non-reference genotype concordance for the high confidence RefSeq CDS regions across all reproducibility and repeatability assessments were >99.8% and >99.3% for SNVs and indels, respectively (**Table 2**), indicating that HiFi genome sequencing small variant detection is robust and precise.

Long-Read HiFi Genome Sequencing Specimen Validation

Germline specimens were validated by subjecting paired blood, saliva, and swab samples to HiFi genome sequencing and evaluating SNV/indel concordance. As expected, concordance between specimens was reduced for SNVs/indels when evaluating difficult genomic regions as saliva-based specimens are known to harbor bacterial DNA that interferes with sequencing (Trost et al. 2019; Yao et al. 2020). However, the average paired SNV/indel concordance between all three specimen types were >99% across all high confidence genomic regions (**Table 3**). Although specimen concordance was reduced when assessed across regions not considered high confidence by GIAB (**Supplemental Table S6**), these validation results indicate that HiFi genome sequencing of saliva and swab specimens are consistent with blood for germline SNV/indel detection.

DISCUSSION

To facilitate the implementation of diagnostic long-read HiFi sequencing, we executed an analytical validation plan that was centered on comprehensively evaluating HiFi genome sequencing for germline SNV/indel detection and specimen types that are used for clinical testing in medical genetics. Results were stratified by variant type and GIAB-defined genomic regions to better inform overall performance, which ultimately determined that HiFi genome sequencing is accurate and robust. The accuracy of germline small variant detection in non-difficult genomic regions across reference materials was >99.9% for both SNVs and indels, and small variant detection accuracy in GIAB-defined difficult regions was >99.5% and >98.8% for SNVs and indels, respectively. These analytical validation analyses underscore the accuracy of

long-read HiFi genome sequencing for detecting germline SNV/indels (<50 bp), which supports the implementation of this technology for clinical genetic testing. In addition, quality control (QC) thresholds for clinical long-read HiFi genome sequencing based on CAP requirement MOL.36151 are suggested in **Supplemental Table S7**; however, these should be considered preliminary recommendations, as clinical laboratories should leverage their own experience and data to define internal QC metrics.

Analytical validation is a critical assessment of any new clinical laboratory test, which is defined by CAP Checklists and other state, federal, and/or professional requirements/recommendations. Test performance specifications include reportable range, accuracy, reproducibility/repeatability, sensitivity/specificity, and other relevant performance characteristics. For long-read HiFi genome sequencing we adopted the definitions for 'Reportable Range' and 'Reference Range (Reference Interval)' based on clinical high-throughput sequencing guidelines (Gargis et al. 2012; Santani et al. 2017). However, for a more comprehensive assessment of sequencing performance, reportable range was measured genome-wide but strategically stratified by distinct genomic regions as defined by GIAB/GA4GH. The genomic regions implemented in this validation included high level strata (low complexity, low mappability, segmental duplications, all difficult regions, not in any difficult regions, all high confidence regions, RefSeq CDS regions), as well as the more specific genomic subregions defined by GIAB/NIST and GA4GH (Krusche et al. 2019). These regions were intersected with our SNV/indel performance results, as well as genome sequencing specimen validation data (CAP requirement MOL.31015), as deemed appropriate based on analysis context and intended use.

Sequencing accuracy is a rapidly evolving area that is driven by continual improvements in available chemistries and informatic algorithms developed for calling germline variants. As an integral component of validating clinical sequencing-based platforms (Roy et al. 2018), benchmarking small variant accuracy (i.e., recall, precision, F1) is supported through the GIAB/NIST/GA4GH resources (Majidian et al. 2023; Olson et al. 2023), which recently has been catalyzed by PrecisionFDA challenges that more comprehensive evaluations of sequencing-based variant calling (Zook et al. 2019; Olson et al. 2022). Our analytical validation of HiFi genome sequencing is consistent with the most recent PrecisionFDA V2 challenge, which concluded that long-read HiFi sequencing coupled with machine learning-based variant calling tools (Pei et al. 2021; Olson et al. 2022) was superior to short-read genome sequencing using graph-based variant calling.

It is important to note that our reported accuracy results reflect not only the sequencing platforms evaluated but also the variant calling methods used. Given that it was beyond the scope of our study to perform a full comparison of bioinformatics techniques, we selected best-practice tools with high accuracy in challenging genomic regions for each sequencing platform, as demonstrated by PrecisionFDA V2. Our validation also included a detailed evaluation of performance across GIAB-defined genomic stratifications, which highlighted long-read HiFi sequencing accuracy across challenging regions and particularly among indel variants. Of note, our targeted sequencing depth was ≥30×, consistent with recommendations from the Medical Genome Initiative (Marshall et al. 2020), and as expected small variant accuracy was reduced at lower depths. However, it is notable that 99% accuracy was surpassed at ~15-25× for long-read

HiFi genome sequencing compared to ~20-35× for short-read genome sequencing.

In addition to accuracy, HiFi genome sequencing small variant reproducibility was also interrogated by measuring non-reference genotype concordance between datasets. Concordance across all replicates in high confidence GIAB-defined regions ranged from 99.84-99.91% for SNVs and 97.66-99.30% for indels, indicating that small variant calling is very robust. Given that exome reproducibility/repeatability is typically higher than that observed with genome sequencing due to the more narrow region interrogated (Linderman et al. 2014), we also stratified our genome results by RefSeq CDS regions, which resulted in highly concordant small variant calling across all replicates in the high confidence regions (all: 99.85%/99.37%; non-difficult: 99.91%/99.73%, for SNVs/indels) and non-high confidence regions (all: 98.22%/93.13%; non-difficult: 99.66%/99.19%, for SNVs/indels).

Of note, the GIAB high confidence regions encompass 81.6% of the autosomal GRCh38 human genome, which translates to ~2.52 Gb across the seven GIAB reference samples. The remaining 18.4% of non-high confidence autosomal bases (~567.2 Mb) represent subregions of the genome (and Chromosomes X and Y) that are difficult to benchmark given the uncertainty in underlying truth set (Zook et al. 2016; Zook et al. 2019; Zook et al. 2020). The concordance results across reference materials and specimen types in our validation study were reduced in the genome-wide analyses (i.e., including non-high confidence regions) compared to the concordance results limited to the high confidence genomic regions, most notably in the GIAB-defined difficult regions (low complexity, low mappability, segmental duplications, etc.). These metrics were considered acceptable, as 40% of the variants in these regions had genotype quality scores of

<Q20 compared to less than 1% of variants in the non-difficult regions before filtering, and the increase in variant numbers was much greater in the difficult regions than the non-difficult regions in the non-high confidence regions (2.51× vs 1.08×). As such, these thorough reproducibility analyses together indicate that long-read HiFi genome sequencing is highly robust across the high confidence regions of the human genome; however, variants identified in the GIAB non-high confidence difficult regions in a clinical setting would likely require independent confirmation if reportable.

Another critical CAP requirement for test implementation is validating the specific specimen types used for clinical processing (MOL.31015), which for germline genetic testing typically includes peripheral blood and/or saliva specimens. To satisfy this requirement, paired specimens were subjected to HiFi genome sequencing and concordance was measured across the genome. Despite known challenges with using oral saliva samples for sequencing due to the presence of competing bacterial DNA (Krusche et al. 2019; Trost et al. 2019), concordance between paired blood, saliva, and assisted saliva (swab) specimens ranged from 99.82-99.92% for all SNVs/indels across high confidence genomic regions. However, it is notable that ~99% of reads aligned to the reference genome for blood and cell line specimens, compared to ~93% for assisted saliva specimens and 86% for saliva, resulting in lower average depth (blood, cell lines: 33×; assisted saliva: 28×; saliva: 25×). As such, additional sequencing of oral samples to compensate for unmapped reads (as defined by QC thresholds) may be warranted in clinical production. These specimen validation study results indicate that our HiFi genome sequencing procedure and pipeline generates highly comparable results between peripheral blood and DNA

297 isolated from saliva and assisted saliva, which supports their use as acceptable clinical specimens
298 for this test.
299
300 Of note, copy number variant (CNV) detection by sequencing is routinely implemented among
301 clinical laboratories (Kadalayil et al. 2015; Rajagopalan et al. 2020), and GIAB/NIST has
302 developed consensus germline structural variant (SV) calls from HG002 (NA24385) (Zook et al.
303 2020). However, this dataset is an integration of 68 callsets from multiple algorithms and four
304 different sequencing technologies, each with their own strengths and weaknesses, and as a result
305 it does not currently include robust duplication calls or SVs >100 kb (Whitford et al. 2019).
306 Although long-read HiFi genome sequencing has been shown to be highly effective at CNV/SV
307 detection (Chaisson et al. 2019; Mahmoud et al. 2019; Aganezov et al. 2020), these variants were
308 considered out of scope for this initial analytical validation; however, they are currently being
309 evaluated for a subsequent analytical work product.
310
311 In conclusion, long-read HiFi genome sequencing (≥30×) was analytically validated for germline
312 SNV/indel detection, which supports the implementation of this platform as a robust technology
313 for clinical genetic testing. Of note, practical factors for sequencing platform selection were
314 intentionally excluded from this analytical validation, including cost, labor and sequencing time,
315 as these variables were not applicable to analytical performance testing. This validation also did
316 not explicitly include ‘clinical performance characteristics’ as defined by CAP (MOL.31590), as
317 these analyses were reserved for subsequent validation of clinically significant germline variants.
318 As such, these analytical validation data provide the infrastructure for long-read HiFi genome

sequencing-based detection of germline variation, which supports the use of this innovative technology for clinical diagnostic testing.

MATERIALS and METHODS

Analytical Validation Specimens

High molecular weight (HMW) reference material DNA samples were acquired from the Coriell Institute for Medical Research (Camden, NJ), which included seven benchmarking samples from the GIAB/NIST consortium. Peripheral blood was collected in EDTA vacutainer tubes using standard practices and DNA isolated using the Maxwell RSC Buffy Coat DNA Kit (Promega Corporation) according to manufacturer instructions. Saliva samples were collected using the Oragene Dx OGD-500 kit (DNA Genotek, Ottawa, ON, Canada) or the assisted saliva (swab) Oragene Dx OGD-575 kit (DNA Genotek). DNA was isolated from saliva specimens using Maxwell RSC Stabilized Saliva DNA Kit (Promega Corporation) according to manufacturer instructions. All validation samples and sequencing metrics are summarized in **Supplemental Tables S7-S10**.

Long-Read HiFi Genome Sequencing

Library Preparation and Long-Read HiFi Sequencing

Genomic DNA was analyzed with the Femtopulse Genomic DNA 165 kb kit (Agilent, Santa Clara, CA) to confirm adequate quantity of HMW DNA. Approximately 3-10 ug of DNA was mechanically sheared to 10-20 kb using the Megaruptor Shearing kit (Diagenode, Denville, NJ), with the DNAFluid+ kit (Diagenode) employed for viscous samples. Library preparation was performed through either a manual or an automated workflow with the SMRTbell Prep Kit

according to manufacturer instructions, including end repair, A-tailing, adapter ligation, purification with SMRTbell cleanup beads (Pacific Biosciences, Menlo Park, CA), and nuclease treatment.

Manual library preparation included purification of sheared gDNA with SMRTbell cleanup beads, SMRTbell library generation, followed by size selection on Blue Pippin (Sage Science, Beverly, MA) to remove fragments <10 kb and purification with Ampure PB beads (Pacific Biosciences). Automated library preparation included small fragment depletion using the Short Read Eliminator (SRE) XS kit (<10 kb) or the SRE kit (<25 kb) (Pacific Biosciences) as needed prior to shearing, followed by purification with SMRTbell cleanup beads and library preparation using the Miro Canvas (Miroculus, San Francisco, CA). Manual and automated workflow SMRTbell libraries were both quantified by the Qubit dsDNA assay kit (Invitrogen, Waltham, MA) and bound to sequencing polymerase using the Sequel II Binding Kit (Pacific Biosciences). Long-read HiFi genome sequencing was performed on the Sequel IIe system (Pacific Biosciences) with a 30-hour movie collection time, and each sample was sequenced on three SMRTcells except where otherwise noted.

Publicly Available Data

To evaluate internal long-read HiFi genome sequencing reproducibility, selected publicly available long-read HiFi genome sequencing data for the NA12878/HG001 and NA24385/HG002 reference materials were acquired from the National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nlm.nih.gov/giab): PacBio_SequelII_CCS_11kb

(NA12878), HudsonAlpha_PacBio_CCS (NA12878) (Zook et al. 2016), PacBio_CCS_15kb_20kb_chemistry2 (NA24385), PacBio_SequelII_CCS_11kb (NA24385).

Long-Read HiFi Sequencing Bioinformatics Pipeline and Variant Calling

HiFi reads were generated using SMRTLink 10.2 software and the Circular Consensus Sequencing mode. For analyses requiring downsampled data, subsampling was executed with SAMtools v1.18 (Danecek et al. 2021). Alignment and variant calling were performed using a modified version of the PacBio HiFi-human-WGS-WDL pipeline (https://github.com/PacificBiosciences/HiFi-human-WGS-WDL), with the following steps: alignment of HiFi sequencing reads to Genome Reference Consortium Human Build 38 (GRCh38) using pbmm2 v1.7.0 (Hon et al. 2020); small variant calling using DeepVariant v1.4.0 and the PacBio machine learning model included with the software (Poplin et al. 2018); and calculating aligned read depth with mosdepth v0.2.9 (Pedersen and Quinlan 2018). BCFtools v1.20 was used to filter variants, removing all SNV calls with QUAL < 20 (Danecek et al. 2021). Finally, Picard Tools v2.27.4 was used to calculate quality yield metrics (https://broadinstitute.github.io/picard), alignment summary metrics, and variant calling metrics. For analysis of CMRG genes, a modified version of the GRCh38 build was used, wherein false duplications were masked and decoy contigs were added for falsely collapsed duplications (Behera et al. 2023).

Short-Read Genome Sequencing

Publicly Available Data

To compare long-read HiFi and short-read genome sequencing accuracy, publicly available short-read genome sequencing data for the seven GIAB benchmarking reference materials were acquired from the National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nlm.nih.gov/giab): NIST_NA12878_HG001_HiSeq_300x, NIST_Illumina_2x250bps (NA24385, NA24143, and NA24149), HG005_NA24631_son_HiSeq_300x, NA24694_Father_HiSeq100x, and NA24695_Mother_HiSeq100x.

Short-Read Sequencing Bioinformatics Pipeline and Variant Calling

For analyses requiring downsampled data, subsampling was executed with SAMtools v1.18 (Danecek et al. 2021). Alignment, germline small variant calling, and calculation of quality control metrics were performed with the Illumina DRAGEN Germline Pipeline v4.2.4 in BaseSpace, using the HG38 alt-masked multi-genome graph reference.

Analytical Validation Strategy

The HiFi genome sequencing small variant analytical validation plan followed Laboratory Developed Test (LDT) guidelines as defined by the CAP and AMP (Jennings et al. 2009; Aziz et al. 2015; Roy et al. 2018), the American College of Medical Genetics and Genomics (ACMG) (Rehm et al. 2013), high-throughput sequencing recommendations from professional consortia (Gargis et al. 2012; Matthijs et al. 2016; Santani et al. 2017; Santani et al. 2019), and the Clinical Laboratory Evaluation Program at the Wadsworth Center, New York State Department of Health. The plan was centered on determining the analytical performance characteristics of HiFi genome sequencing for use as a diagnostic technology, as well as defining standard operating

409 procedures (SOPs), quality control/quality assurance procedures, and validating small variant
410 detection and specimen types.
411

412 ## DATA ACCESS

413 All GIAB reference material and blood/saliva sample HiFi genome sequencing aligned bam
414 datasets that were generated in this study have been submitted to the NCBI BioProject database
415 (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1143955.
416

417 ## COMPETING INTEREST STATEMENT

418 N.H. is currently an employee of Influx Bio; all other authors declare no conflicts of interest.
419

420 ## ACKNOWLEDGEMENTS

421 The authors would like to thank Stanford Health Care, Stanford Children's Health, and Pacific
422 Biosciences (Menlo Park, CA) for their programmatic support. S.A.S was supported in part by
423 NIH/NHGRI grant U01HG011762. Project conceptualization: N.H., Y.Y., S.A.S.; Data
424 acquisition: N.H, L.L., P.W.T., Z.N.; Data analysis, interpretation, and management: N.H,
425 P.W.T., Z.N., C.H., T.P.N., Y.Y., S.A.S.; Drafting and revision: N.H., Y.Y., S.A.S.; Final
426 approval: N.H, L.L., P.W.T., Z.N., C.H., T.P.N., Y.Y., S.A.S.
427

428 ## REFERENCES
429

430 Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, Lee I, Kirsche M, Wappel
431 R, Kramer M et al. 2020. Comprehensive analysis of structural variants in breast cancer
432 genomes using single-molecule sequencing. *Genome Res* **30**: 1258-1273.

433 Ameur A, Kloosterman WP, Hestand MS. 2019. Single-Molecule Sequencing: Towards Clinical
434 Applications. *Trends Biotechnol* **37**: 72-85.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* **46**: 2159-2168.

Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DG et al. 2015. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* **139**: 481-493.

Behera S, LeFaive J, Orchard P, Mahmoud M, Paulin LF, Farek J, Soto DC, Parker SCJ, Smith AV, Dennis MY et al. 2023. FixItFelix: improving genomic analysis by fixing reference errors. *Genome Biol* **24**: 31.

Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**: 5473-5478.

Bick D, Jones M, Taylor SL, Taft RJ, Belmont J. 2019. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet* **56**: 783-791.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B et al. 2022. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* doi:10.1016/j.gim.2022.02.007.

Conlin LK, Aref-Eshghi E, McEldrew DA, Luo M, Rajagopalan R. 2022. Long-read sequencing for molecular diagnostics in constitutional genetic disorders. *Hum Mutat* **43**: 1531-1544.

Costain G, Jobling R, Walker S, Reuter MS, Snell M, Bowdin S, Cohn RD, Dupuis L, Hewson S, Mercimek-Andrews S et al. 2018. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur J Hum Genet* **26**: 740-744.

Costain G, Walker S, Marano M, Veenma D, Snell M, Curtis M, Luca S, Buera J, Arje D, Reuter MS et al. 2020. Genome Sequencing as a Diagnostic Test in Children With Unexplained Medical Complexity. *JAMA Netw Open* **3**: e2018109.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**.

Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA et al. 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* **30**: 1033-1036.

Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**: 399.

Jennings L, Van Deerlin VM, Gulley ML, College of American Pathologists Molecular Pathology Resource C. 2009. Recommended principles and practices for validating clinical molecular pathology tests. *Arch Pathol Lab Med* **133**: 743-755.

Kadalayil L, Rafiq S, Rose-Zerilli MJ, Pengelly RJ, Parker H, Oscier D, Strefford JC, Tapper WJ, Gibson J, Ennis S et al. 2015. Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* **16**: 380-392.

Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S et al. 2019. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**: 555-560.

Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, Schadt EE. 2014. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics* **7**: 20.

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**: 597-614.

Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 246.

Majidian S, Agustinho DP, Chin CS, Sedlazeck FJ, Mahmoud M. 2023. Genomic variant benchmark: if you cannot measure it, you cannot improve it. *Genome Biol* **24**: 221.

Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, Rowsey R, Klee EW, Liu P, Worthey EA et al. 2020. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *NPJ Genom Med* **5**: 47.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M et al. 2016. Guidelines for diagnostic next-generation sequencing. *Eur J Hum Genet* **24**: 1515.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al. 2022. The complete sequence of a human genome. *Science* **376**: 44-53.

NYSDOH. Laboratory Standards: Clinical Laboratory Evaluation Program. In *https://wwwwadsworthorg/regulatory/clep/clinical-labs/laboratory-standards*, doi:https://www.wadsworth.org/regulatory/clep/clinical-labs/laboratory-standards.

Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet* **24**: 464-483.

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O et al. 2022. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genom* **2**.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867-868.

Pei S, Liu T, Ren X, Li W, Chen C, Xie Z. 2021. Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief Bioinform* **22**.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983-987.

Rajagopalan R, Murrell JR, Luo M, Conlin LK. 2020. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med* **12**: 14.

Rehm HL. 2017. Evolving health care through personal genomics. *Nat Rev Genet* **18**: 259-267.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E et al. 2013. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* **15**: 733-747.

Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV et al. 2018. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* **20**: 4-27.

Santani A, Murrell J, Funke B, Yu Z, Hegde M, Mao R, Ferreira-Gonzalez A, Voelkerding KV, Weck KE. 2017. Development and Validation of Targeted Next-Generation Sequencing Panels for Detection of Germline Variants in Inherited Diseases. *Arch Pathol Lab Med* **141**: 787-797.

Santani A, Simen BB, Briggs M, Lebo M, Merker JD, Nikiforova M, Vasalos P, Voelkerding K, Pfeifer J, Funke B. 2019. Designing and Implementing NGS Tests for Inherited Disorders: A Practical Framework with Step-by-Step Guidance for Clinical Laboratories. *J Mol Diagn* **21**: 369-374.

Trost B, Walker S, Haider SA, Sung WWL, Pereira S, Phillips CL, Higginbotham EJ, Strug LJ, Nguyen C, Raajkumar A et al. 2019. Impact of DNA source on genetic variant detection from human whole-genome sequencing data. *J Med Genet* **56**: 809-817.

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N et al. 2022a. Benchmarking challenging small variants with linked and long reads. *Cell Genom* **2**.

Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang YC, Gupta R, Wenger AM, Rowell WJ et al. 2022b. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**: 672-680.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155-1162.

Whitford W, Lehnert K, Snell RG, Jacobsen JC. 2019. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. *J Biomed Inform* **94**: 103174.

Yang Y, del Gaudio D, Santani A, Scott SA. 2024. Applications of genome sequencing as a single platform for clinical constitutional genetic testing. *GIM Open* doi:10.1016/j.gimo.2024.101840.

Yao RA, Akinrinade O, Chaix M, Mital S. 2020. Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients. *BMC Med Genomics* **13**: 11.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**: 160025.

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38**: 1347-1355.

Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY et al. 2019. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**: 561-566.

Table 1. Long-read HiFi genome sequencing SNV/indel (<50 bp) accuracy.

		Genome-Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
SNVs	Count	160865	190416	120916	584743	2718604	3303346	14059	20593
	Recall	99.05%	97.84%	96.63%	98.93%	99.91%	99.74%	99.86%	99.56%
	Precision	99.54%	99.81%	99.67%	99.80%	99.99%	99.95%	99.98%	99.96%
Insertions (1-5bp)	Count	128237	4241	4596	140934	69118	210639	46	109
	Recall	98.57%	98.50%	98.62%	98.64%	99.90%	99.05%	99.71%	99.44%
	Precision	98.42%	99.14%	98.98%	98.53%	99.94%	98.99%	100.00%	99.35%
Insertions (6-15bp)	Count	14435	342	486	15460	6496	22004	5	42
	Recall	98.67%	98.39%	98.77%	98.70%	99.73%	99.00%	100.00%	99.58%
	Precision	98.98%	98.52%	98.87%	99.02%	99.94%	99.29%	100.00%	99.25%
Insertions (≥16bp)	Count	2415	122	136	2747	2016	4776	2	13
	Recall	98.76%	96.91%	97.86%	98.80%	99.80%	99.22%	85.71%	97.52%
	Precision	99.16%	97.23%	98.15%	99.17%	99.86%	99.46%	100.00%	96.63%
Deletions (1-5bp)	Count	136024	4600	4507	148953	69677	218129	59	142
	Recall	98.78%	98.45%	98.26%	98.82%	99.87%	99.15%	99.48%	99.00%
	Precision	98.89%	99.18%	98.91%	98.95%	99.93%	99.25%	99.72%	99.49%
Deletions (6-15bp)	Count	16664	466	544	17758	6647	24190	10	50
	Recall	98.68%	98.11%	98.29%	98.68%	99.57%	98.93%	100.00%	99.74%
	Precision	98.88%	98.64%	98.40%	98.90%	99.86%	99.15%	100.00%	99.74%
Deletions (≥16bp)	Count	3424	152	116	3681	1621	5170	2	12
	Recall	99.33%	98.53%	98.15%	99.32%	99.63%	99.42%	85.71%	100.00%
	Precision	99.49%	98.93%	98.15%	99.49%	99.88%	99.61%	100.00%	100.00%
All indels	Count	285386	9840	10200	313715	155551	469067	125	362
	Recall	98.78%	98.43%	98.42%	98.82%	99.86%	99.16%	99.65%	99.26%
	Precision	98.70%	99.09%	98.89%	98.78%	99.93%	99.14%	99.88%	99.38%
SNVs and indels	Count	462064	200340	131301	914275	2874179	3788255	14184	20961
	Recall	98.81%	97.87%	96.78%	98.85%	99.91%	99.65%	99.86%	99.55%
	Precision	98.99%	99.77%	99.61%	99.42%	99.99%	99.85%	99.98%	99.95%

indels: insertions/deletions; RefSeq CDS: NCBI Reference Sequence gene coding sequence; SNVs: single nucleotide variants.

Table 2. Long-read HiFi genome sequencing SNV/indel (<50 bp) reproducibility (GIAB high confidence).

		Genome-Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
SNVs	**Count**	170332	181786	113753	584487	2721331	3305818	14023	20467
	Concordance	99.10%	99.51%	99.30%	99.54%	99.95%	99.88%	99.91%	99.85%
Insertions (1-5bp)	**Count**	139069	4085	4495	151456	69083	221146	48	115
	Concordance	97.41%	99.38%	99.26%	97.61%	99.94%	98.35%	99.66%	99.17%
Insertions (6-15bp)	**Count**	15335	335	490	16341	6487	22882	5	43
	Concordance	98.13%	98.99%	99.31%	98.23%	99.96%	98.74%	100.00%	100.00%
Insertions (≥16bp)	**Count**	2527	118	119	2805	1847	4664	2	12
	Concordance	98.22%	97.70%	98.01%	98.31%	99.84%	98.93%	100.00%	95.22%
Deletions (1-5bp)	**Count**	150504	4459	4313	163229	69693	232407	58	141
	Concordance	98.13%	99.35%	99.21%	98.25%	99.97%	98.77%	99.72%	99.41%
Deletions (6-15bp)	**Count**	17426	457	552	18499	6652	24940	10	47
	Concordance	98.38%	99.17%	98.96%	98.46%	99.96%	98.86%	100.00%	100.00%
Deletions (≥16bp)	**Count**	3525	160	101	3771	1583	5220	3	16
	Concordance	99.21%	98.92%	99.01%	99.26%	99.96%	99.50%	100.00%	100.00%
All indels	**Count**	311652	9518	9877	339360	155318	494492	125	370
	Concordance	97.90%	99.32%	99.21%	98.05%	99.95%	98.64%	99.73%	99.37%
SNVs and indels	**Count**	498720	191398	123824	940587	2876675	3817077	14148	20841
	Concordance	98.27%	99.50%	99.30%	98.96%	99.95%	99.71%	99.91%	99.84%

indels: insertions/deletions; RefSeq CDS: NCBI Reference Sequence gene coding sequence; SNVs: single nucleotide variants.

Table 3. Long-read HiFi genome sequencing SNV/indel (<50 bp) paired specimen concordance (GIAB high confidence).

		Genome-Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
Blood vs Saliva	**Count**	246228.5	158890	94436.5	639126	2753708.5	3392621.5	13785	19515.5
	Concordance	99.40%	99.54%	99.34%	99.60%	99.94%	99.88%	99.94%	99.84%
Blood vs Swab	**Count**	246296.5	159100	94574	639544.5	2754294.5	3393623	13781	19527.5
	Concordance	99.52%	99.57%	99.37%	99.66%	99.96%	99.90%	99.94%	99.85%
Swab vs Saliva	**Count**	246296.5	159100	94574	639544.5	2754294.5	3393623	13781	19527.5
	Concordance	99.30%	99.47%	99.34%	99.54%	99.94%	99.87%	99.93%	99.84%

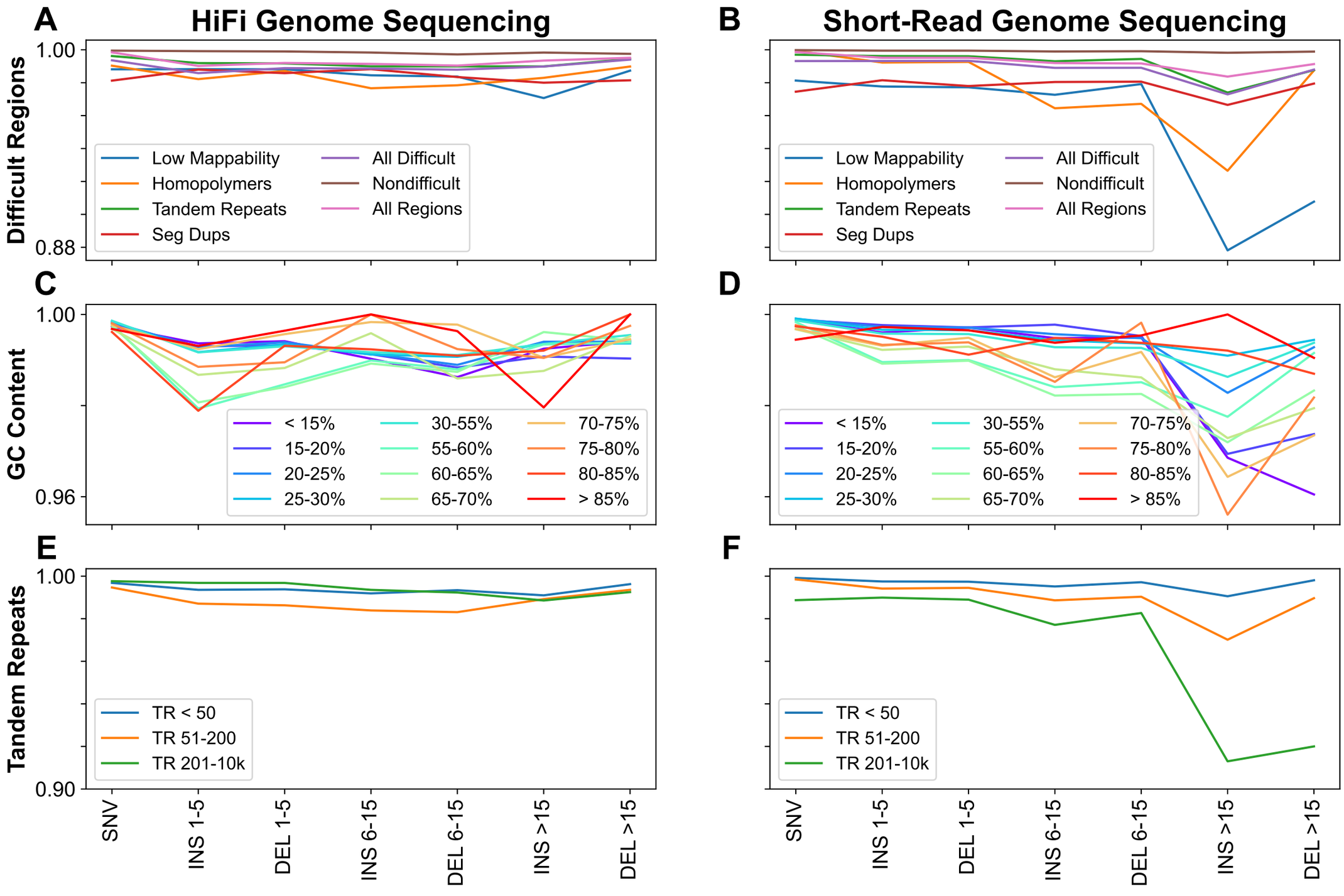
RefSeq CDS: NCBI Reference Sequence gene coding sequence.

FIGURE LEGENDS

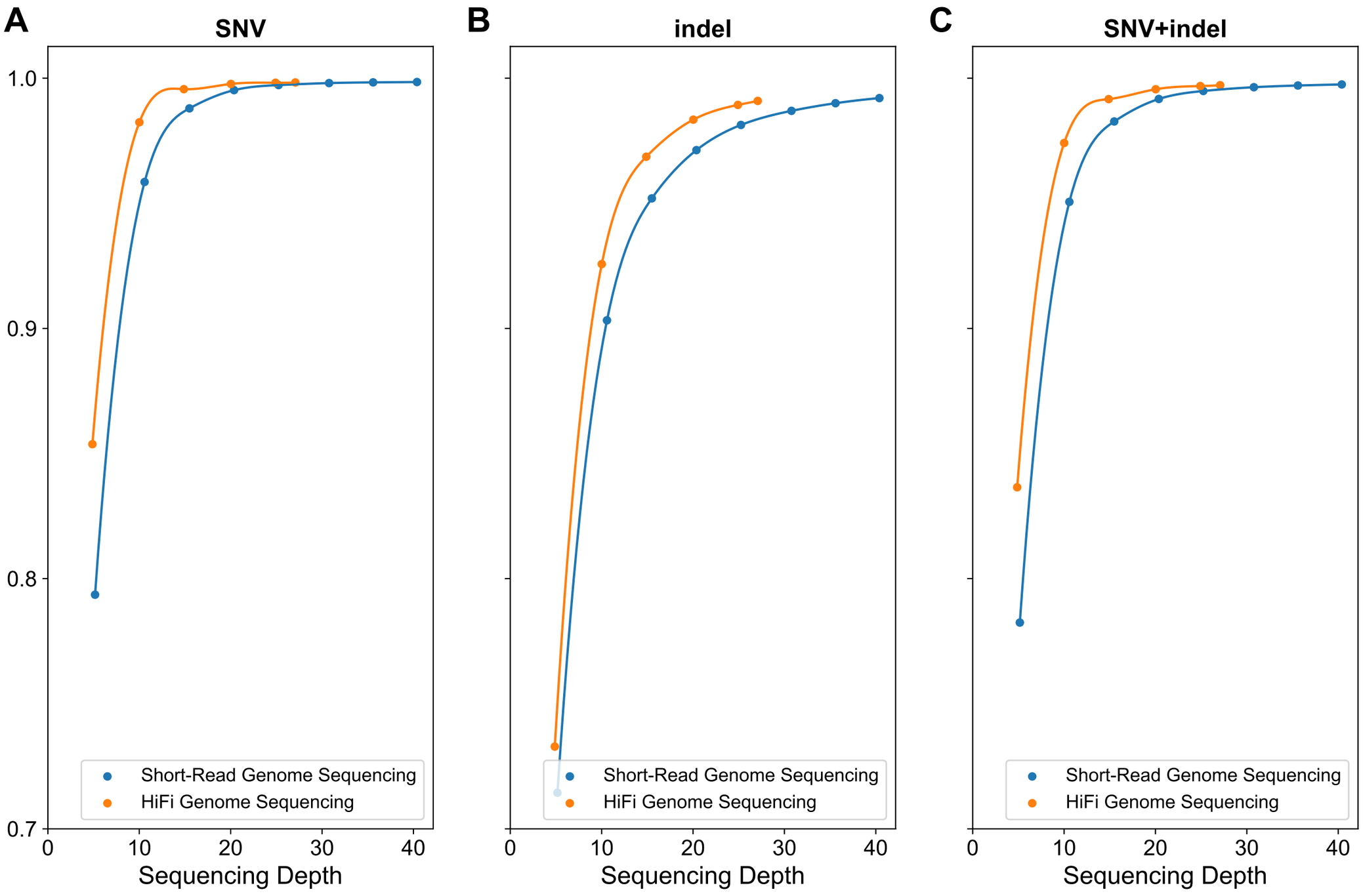
Figure 1. SNV/indel accuracy across GIAB-defined genomic regions. Small variant F1-scores for HiFi genome sequencing (**A, C, E**) and short-read genome sequencing (**B, D, F**) across 22 of the 100 interrogated GIAB-defined genomic regions. Results summarized by difficulty (mappability, homopolymers, tandem repeats, segmental duplications, all difficult, not in any difficult, all regions), GC content (<15% to >85%) and tandem repeats (<50 bp, 51-200 bp, 201-10,000 bp), and stratified by variant type (SNV, indels 1-5 bp, 6-15 bp, and ≥16 bp).

Figure 2. SNV/indel accuracy and depth stratification. Plots of F1-scores across sequencing depths and comparing HiFi genome sequencing and short-read genome sequencing: (**B**) single nucleotide variants (SNVs); (**B**) insertions/deletion variants (indels); and (**C**) SNVs and indels combined.

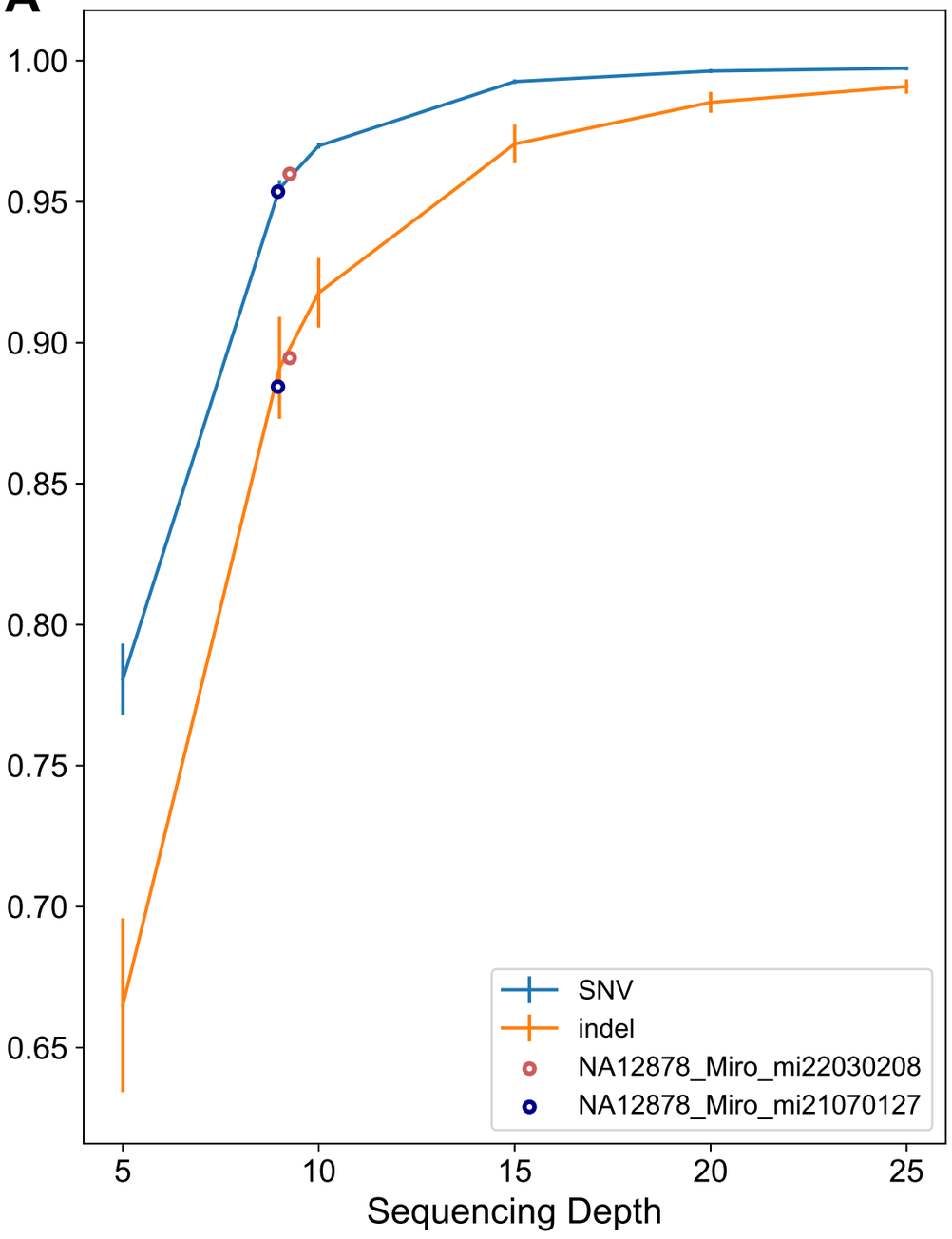
Figure 3. Automated library preparation validation with low depth HiFi genome sequencing. Two Miro Canvas instruments were validated using a single SMRTcell of data (~9× each) with small variant (SNV/indel) benchmarking of NA12878 and fitting results to reference curves defined by manual library preparation of three GIAB reference samples (NA12878, NA24385, NA24631). Error bars represent standard deviations. Automated library preparation results were considered acceptable if Miro Canvas recall (**A**) and precision (**B**) values were equivalent or greater than the average manual preparation reference material accuracy results at comparable depths.


A
HiFi Genome Sequencing
B
Short-Read Genome Sequencing
Difficult Regions
1.00
0.88
Low Mappability
Homopolymers
Tandem Repeats
Seg Dups
All Difficult
Nondifficult
All Regions
C
D
GC Content
1.00
0.96
< 15%
15-20%
20-25%
25-30%
30-55%
55-60%
60-65%
65-70%
70-75%
75-80%
80-85%
> 85%
E
F
Tandem Repeats
1.00
0.90
TR < 50
TR 51-200
TR 201-10k
SNV
INS 1-5
DEL 1-5
INS 6-15
DEL 6-15
INS >15
DEL >15

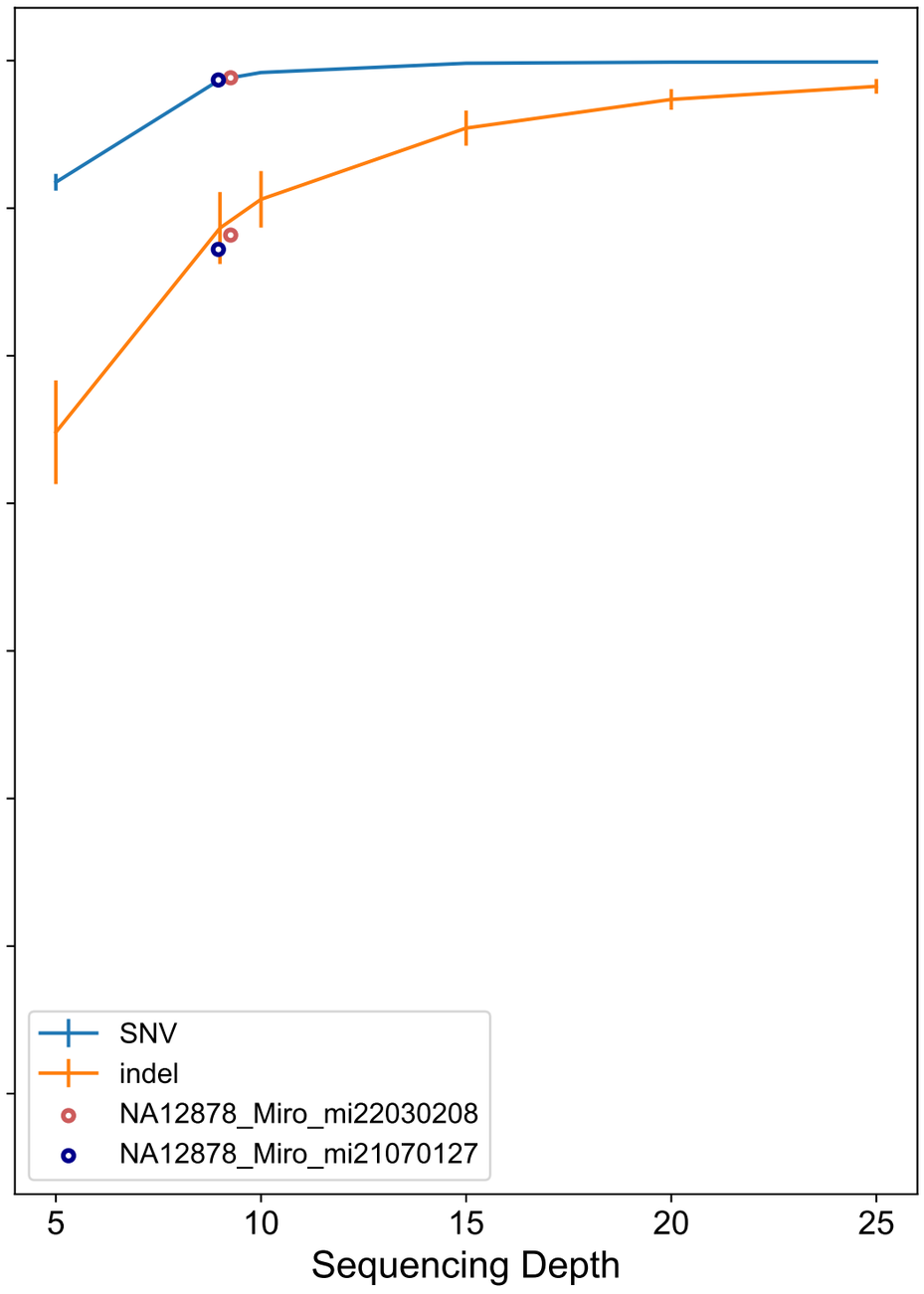
F1 score vs Depth



A


Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
5
10
15
20
25
Sequencing Depth
SNV
indel
NA12878_Miro_mi22030208
NA12878_Miro_mi21070127

B


Precision
5
10
15
20
25
Sequencing Depth
SNV
indel
NA12878_Miro_mi22030208
NA12878_Miro_mi21070127