Germline variant calling using Sequencing by Expansion with DeepVariant

1. Introduction

Germline variant calling is a critical process in genomics through which inherited genetic variations present in an individual's DNA are identified. Accurate germline variant calling is essential for understanding the genetic basis of hereditary diseases. Moreover, it plays a foundational role in human genetics, population genetics, evolutionary biology, and the development of genetic screening tools. As such, reliable germline variant detection is a cornerstone of biomedical research.

Sequencing by Expansion (SBX) offers a novel approach that, when coupled with advanced bioinformatics tools like a pangenome aligner and DeepVariant, significantly improves the accuracy of germline variant identification from whole-genome sequencing (WGS) data. A typical SBX-based germline variant calling workflow involves demultiplexing, alignment (potentially using pangenome references), duplicate marking (with specialized tools for single-end data), and variant calling using tools like DeepVariant. Leveraging pangenome aligners enhances mapping accuracy, especially in complex genomic regions, while DeepVariant, a deep learning-based variant caller, improves the precision of variant identification. Performance benchmarking using challenging datasets, like those generated by the Genome in a Bottle (GIAB) and Telomere-to-Telomere (T2T) consortia, demonstrates the high accuracy achieved by this integrated approach.

This white paper presents the specific results of specific studies with specific samples. The results, comparisons, and trends identified in this paper may not be applicable or generalizable to other samples, other studies, other platforms, or other bioinformatics analyses.

DeepVariant

Date of first publication: March, 2026 | Publication date of this version: August, 2026

Data on File.
For Research Use Only. Not for use in diagnostic procedures.

1.1 Overview of SBX

SBX is a high-throughput sequencing technology designed to increase the scale, efficiency, and speed of genomic analysis. A primary characteristic of SBX technology is the utilization of a biochemical conversion process that encodes the sequence of target nucleic acids into Xpandomers, which are surrogate polymers extending over 50 times longer than the original DNA templates. These Xpandomers are engineered with high signal-to-noise reporter codes, thereby facilitating accurate nanopore sequencing. The synthesis of Xpandomers involves specialized molecular structures, chemistries, and enzymes which facilitate precise, template-dependent synthesis. This approach enables high-quality sequencing data, achieving a >99% basecalling accuracy rate. It also enables near real-time data processing with high scalability (high-throughput parallel processing) through the use of CMOS-based technology. This compatibility with parallel architectures makes SBX highly suitable for large-scale applications in genomics.[1]

SBX Duplex (SBX-D) is a high-accuracy sequencing method with a simple library prep workflow and fast turnaround time for the sequencing step. Standard approaches for achieving high accuracy can come at the expense of throughput, because the resolution of errors without sacrificing yield can be a significant challenge. SBX-D overcomes these challenges by linking both strands of the target DNA in a single sequencing read (Figure 1). Single-strand errors that arise on separate strands of the DNA molecule can thus be resolved to achieve higher accuracy. The SBX-D workflow involves: **(a)** target dsDNA ligation to a hairpin and sequencing adapters, **(b)** amplification of the library, **(c)** automated Xpandomer synthesis, and **(d)** single-molecule sequencing using a highly parallel CMOS-based sensor array. Potential errors can be detected post sequencing via the formation of intramolecular consensus reads, with discordant positions marked as low-confidence calls. This method can sequence homopolymer, repeat, and complex regions by utilizing the structural integrity provided by hairpins. With a maximum throughput of over 500 M bases per second, SBX-D can generate billions of high-accuracy duplex reads from single-molecule DNA fragments.[1]

SBX-Fast is a rapid version of SBX-D that may be used when input DNA is not limiting and amplification is not essential. The simplified library preparation workflow offers faster turnaround times. Eliminating amplification reduces homopolymer errors, leading to improved indel calling accuracy, particularly in long homopolymers and tandem repeats.

1.2 Benchmarking of SBX quality metrics

High-quality sequencing, indicated by metrics such as high coverage uniformity, longer reads, and low error rates (particularly in long homopolymers) improves the accuracy of germline variant calling.

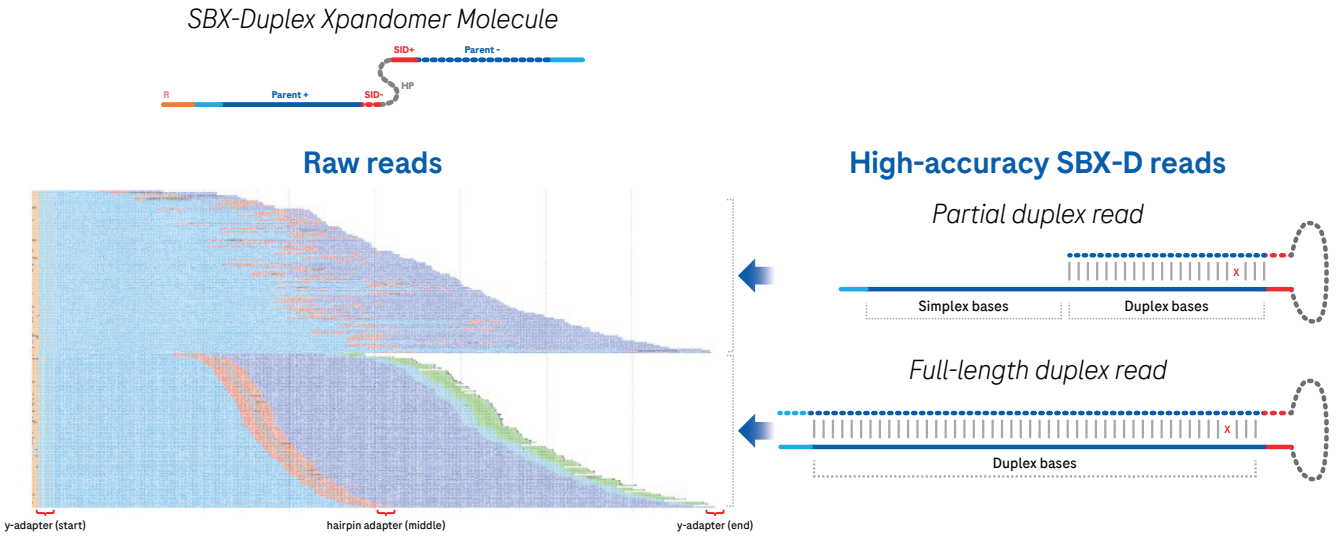


Figure 1. Schematic of SBX duplex sequencing read structure. Full length duplex reads encompass sequences where both the forward and reverse strands are fully sequenced, while partial duplex reads contain a partial sequence of the reverse (or forward) strand. Both full and partial duplex reads can be used in data analysis; however, the utilization of the simplex bases in a single-stranded tail of a partial duplex read in variant calling is application-dependent. The proportions of single strand bases vary from library to library and sample to sample, with Genome in a Bottle (GIAB) samples returning duplex fractions of approximately 75%.

- **High coverage uniformity** across the genome reduces the risk of missing variants and enables confident variant detection by removing biases caused by variable coverage.
- **Longer reads** provide more contiguous information that reduces alignment ambiguity and improves resolution of complex regions and structural variants.
- **Low error rates** reduce false positives and negatives, including in homopolymeric regions where errors are more likely to occur due to their repetitive nature. More accurate representation of these regions increases reliability of variant detection.

To evaluate potential sequencing bias, we assessed the uniformity of SBX-D concordant duplex coverage across various genomic regions. As illustrated in Figure 2, SBX WGS shows 99.9% of the genome achieving at least 10x concordant duplex coverage when sequenced to a median depth of 30x concordant duplex coverage (deduped).

Longer DNA inserts facilitate the mapping of reads to highly similar regions of the genome. Shorter reads can lead to incorrect variant identification in challenging genomic areas like long repeats or segmental duplications due to multimapping issues. Despite variable lengths, mean insert lengths of approximately 240 bp for SBX-D and approximately 190 bp for SBX-Fast are adequate for precise mapping and germline variant calling (Figure 3).

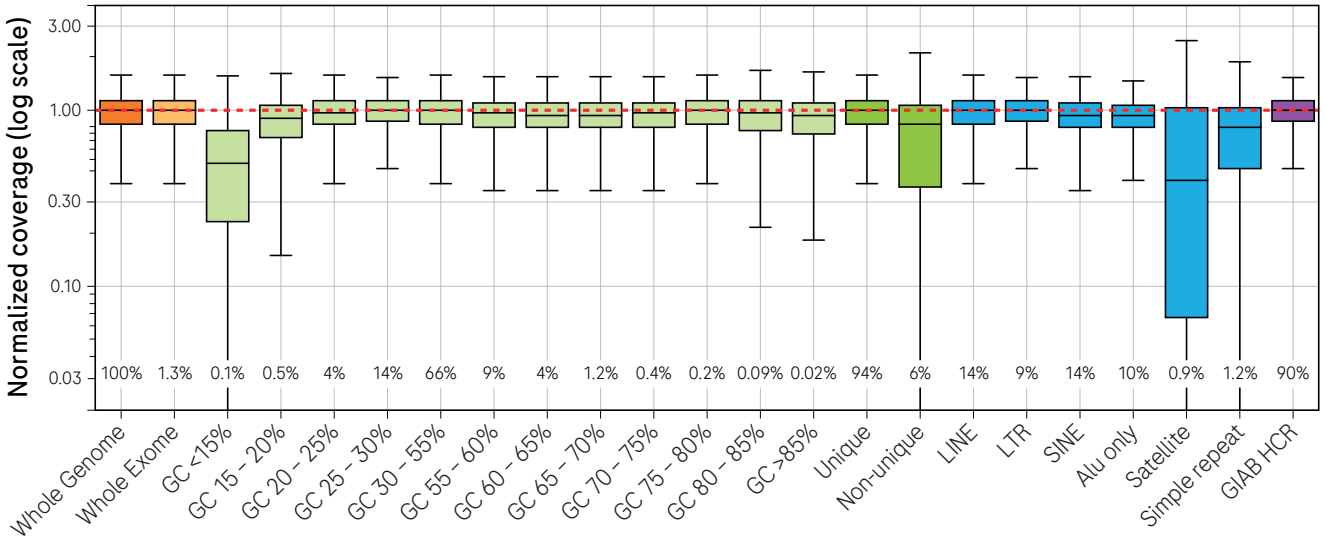


Figure 2. WGS (deduped) concordant duplex coverage uniformity achieved with SBX, stratified by genomic context. LINE: Long Interspersed Nuclear Elements, LTR: long terminal repeat; SINE: short interspersed nuclear elements, GIAB HCR: Genome in a Bottle high confidence regions. Percentage statistics reported below boxplots represent % of the genome covered by each genomic region.

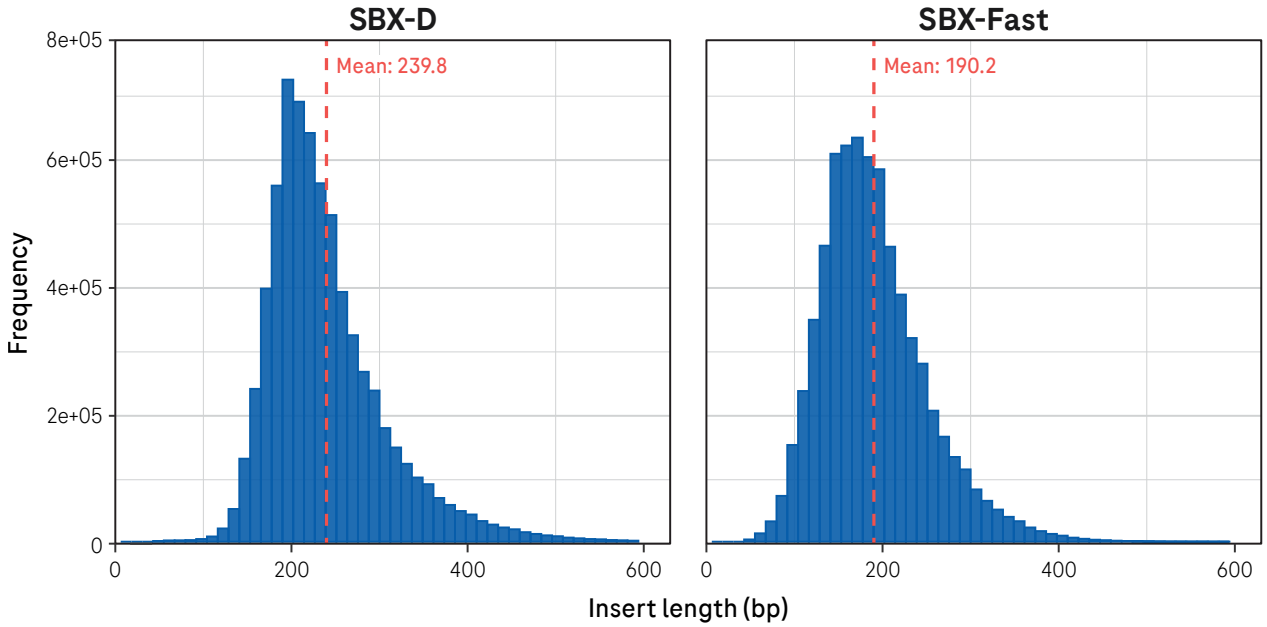


Figure 3. Mapped insert length distribution for SBX-D (left) and SBX-Fast (right), obtained from WGS of the HG001 sample.

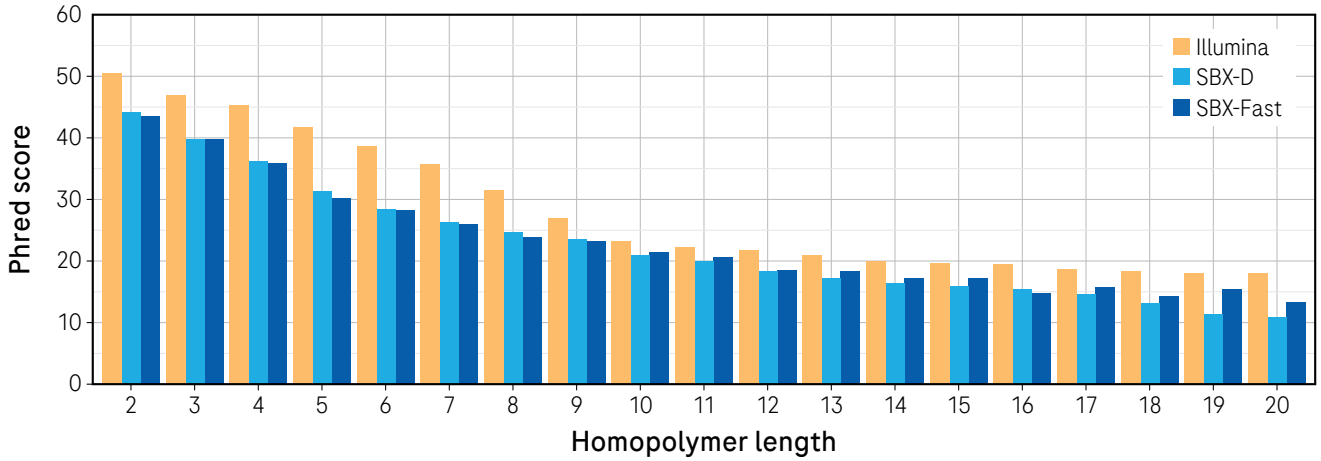


Figure 4. Comparison of homopolymer accuracy for WGS of HG001, performed using an Illumina PCR-free workflow (NovaSeq 6000 (without DRAGEN)), SBX-D, and SBX-Fast. Accuracy measurements are based on linear reference alignment using bwa mem.

Note that while duplex coverage for SBX-D is always expressed in terms of concordant duplex bases, simplex bases of partial duplex reads and both discordant and concordant duplex bases are included in SBX insert length statistics because all of these base types can be used by the aligner when reads are mapped to the reference genome. In addition, simplex bases may be utilized for variant calling in some applications. Thus, by utilizing these simplex bases—which are generated as a natural part of any SBX-D run—in certain applications an SBX-D run with 30x duplex coverage may have a higher total base coverage than a 30x coverage run on other technologies (see Table 1 below for per sample total coverages).

Library preparation or sequencing errors can lead to incorrect variant calling, particularly in challenging genomic regions such as homopolymers and tandem repeats. While genome-wide accuracy for SBX duplex bases achieve Phred quality scores ~Q40, non-duplex single-strand tails exhibit lower scores in the ~Q22 range. Evaluation of SBX homopolymer accuracy indicates sustained accuracy, even in homopolymers up to 20 bp. Although both SBX-D and SBX-Fast demonstrate comparable homopolymer accuracy, SBX-Fast shows slightly higher Phred scores compared to SBX-D in longer homopolymer regions as a result of its amplification-free workflow (Figure 4). Accuracy was measured on bwa-mem aligned BAM files using SBX-Optimized Open Source ("XOOS") tools or "BEST" software as the Giraffe pangenome aligner does not perform left alignment and is subject to SNP errors in difficult regions due to higher sensitivity.

Table 1. Total SBX-D coverage (including both concordant duplex bases and simplex bases of partial duplex reads) for HG001 – HG007

Sample	HG001	HG002	HG003	HG004	HG005	HG006	HG007
Total coverage (median)	39x	38x	38x	39x	39x	38x	39x



Figure 5. SBX WGS small variant calling workflow.

2. Accurate SBX germline variant calling workflow

The workflow for germline variant calling from SBX-based WGS data utilizes the high-level procedures identified in Figure 5. First, consensus reads are demultiplexed according to each sample index and subsequently aligned to the reference genome. Duplicate reads generated during the amplification or sequencing processes are subsequently designated as such based on the spatial proximity of the start and end positions of individual reads. Finally, variant callers are employed to identify germline variants within the duplicate marked BAM/CRAM files, thereby generating variant calls in Variant Call Format (VCF).

2.1 Leveraging pangenome aligner to improve the mapping of SBX reads

Traditional read mapping aligns reads to a single, linear reference genome. While this approach is widely used, it presents certain limitations. A linear reference cannot take into account the genetic variability present within a species, and its accuracy can degrade in the presence of this variation. This leads to biased mapping and loss of information, particularly in highly polymorphic regions and regions containing structural variation.

Pangenomes are alternative reference structures that represent the known genomic variation found within a species (Figure 6). This information can be incorporated into the read mapping process, resulting in reduced reference bias and more accurate downstream analyses. Pangenome-informed mapping is surjected on GRCh38, thereby translating the benefits of this strategy to commonly used genome resources.

We have developed a read mapping pipeline for SBX reads based on the pangenome mapping tool Giraffe, which is open-source and community-developed.[2] Giraffe's development has been closely coordinated with the Human Pangenome Reference Consortium (HPRC), allowing it to take advantage of invaluable pangenomic data from public research. Like in other sequencing technologies[4] for SBX data, we see substantial improvements in accuracy of downstream analyses when using these pangenome-based methods, as opposed to traditional linear reference aligners.

We have also achieved additional gains over the standard best practices pipeline for read mapping with Giraffe.

- We have created custom versions of the HPRC's pangenomes that mask out known errors in the GRCh38 reference build that have previously been published by the Genome Reference Consortium. These errors include contaminant sequences and false duplications that can create the false appearance of high mapping uncertainty.
- We take advantage of recent methodological advances in "personalizing" the pangenome to select a subset of haplotypes that most closely match the sample genome.[4]
- We use the reads' alignments in the pangenome to disambiguate discordant base calls in duplex reads. This essentially amounts to casting a tie-breaking vote for the discordant base call based on the presence or absence of that sequence in the pangenome.

As the HPRC releases more comprehensive genome assemblies, we expect substantial improvements in SBX read alignment to pangenome reference, especially in under-represented alleles.

As noted above, all SBX-D bases (including simplex bases of partial duplex reads, and concordant and discordant duplex bases) can be used in the alignment step.

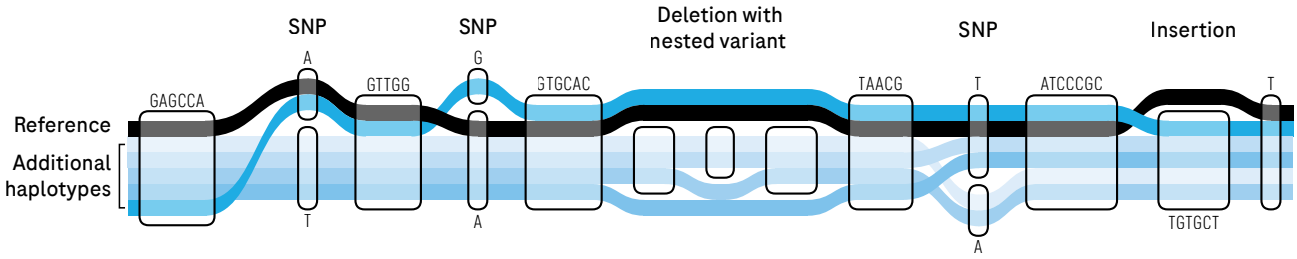


Figure 6. Schematic representation of a pangenome graph in which additional haplotypes augment the reference with known variation, creating a graph structure in which sequences can diverge and reconverge around sites of variation. Image generated using sequenceTubeMap.[5]

2.2 *Detecting amplification (sequencing) duplicates among SBX reads*

In many DNA sequencing workflows, duplicate marking is a critical preprocessing step used to identify and flag duplicates. These duplicates can skew downstream analyses, such as variant calling, by artificially inflating read depth and introducing bias. In some approaches—in tools like Picard's MarkDuplicates, for example—reads that likely come from the same original molecule are grouped and all except the best read are flagged as duplicates. Duplicate groups are defined by 5' start position and orientation within a library/read group, using both the 5' coordinates of both reads for paired-end analysis.

While SBX chemistry utilizes a PCR-free library preparation workflow, it remains important to identify and remove duplicates caused by the process of linear amplification or as a result of the sequencing itself because they can impact variant calling performance. Given that SBX is a single-end sequencing methodology, duplicate marking employs the precise matching of both start and end coordinates of the reads. This is achieved using specialized duplicate marking software developed at Roche. It should be noted that conventional duplicate marking tools such as Picard are not appropriate for use with SBX's single-end data. The use of such tools may result in an inaccurately elevated duplication rate when applied to SBX data.

2.3 *DeepVariant for accurate germline variant calling*

DeepVariant is a tool utilizing a deep learning-based solution designed and developed to meet a growing demand for highly accurate variant calling among scientists.

Developed by Google, DeepVariant uses convolutional neural networks (CNNs) to detect single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) from next-generation sequencing (NGS) data. Unlike traditional variant callers that rely on hand-crafted statistical models, DeepVariant takes a novel approach: it converts sequencing read data into image-like representations and applies a trained neural network to predict the presence of genetic variants.[6]

This deep learning architecture allows DeepVariant to achieve high accuracy, often outperforming established tools in benchmark tests, especially in complex or error-prone regions of the genome. It is compatible with multiple sequencing platforms, including Illumina, PacBio, and Oxford Nanopore, making it highly adaptable to diverse datasets and research needs.

DeepVariant has been widely adopted due to its robustness, reproducibility, and ease of use. The tool also includes pre-trained models optimized for common use cases, including genome or exome analysis, trio analysis, or joint analysis of large research cohorts, facilitating integration into a variety of workflows.

As genomic technologies continue to advance, tools like DeepVariant are essential for maximizing the accuracy and utility of sequencing. By combining advanced deep learning architectures such as CNNs with genomic science, DeepVariant significantly improves variant calling accuracy, and therefore our ability to interpret and understand variation in the human genome.[6]

DeepVariant for SBX-D and SBX-Fast builds on the foundation of the pangenome-aware DeepVariant models trained for short-read, paired-end data (latest GitHub release is v1.9). This model was further optimized for SBX-D and SBX-Fast data in the following ways: DeepVariant uses a De-Bruijn graph to identify potential haplotypes that are used to realign reads. To prevent the creation of too many paths through this graph, we add a filter which requires any 1- or 2-base pair insertion to have more than 8% evidence in the reads. This is needed as SBX data contain discordant insertions since a base is chosen in lieu of having a gap during the consensus generation step. In addition, the ws_min_base_quality, which is used in the process of realignment to determine evidence for reference, is raised from the default of 20 to 25 for SBX data. In addition to the SBX-specific improvements, the version of DeepVariant in this paper contains general improvements, which will be part of the DeepVariant release v1.10.

2.4 *Model Training Data and Process for Evaluated Models*

The DeepVariant models used in this study have been trained on BAM files from GIAB samples with deduplicated, concordant duplex coverage exceeding median 30x. Various downsampling ratios were applied during the training process, leading to the creation of more image examples from different downsampled depths. This approach ensures that the models are well-equipped to accurately call variants even in low-coverage scenarios, mitigating potential issues that might arise from insufficient read depth.

In order to test our DeepVariant approach, we trained SBX models separately for each of the seven HG001 – HG007 samples, leaving the one to be tested out of the training set to allow for evaluation of said model using the excluded sample. These models were trained with both SBX-D and SBX-Fast data. In addition, chromosomes 20, 21, and 22 were excluded in every model for every GIAB sample. Evaluations on chromosomes 21 and 22 were subsequently used to select a model, whereas chromosome 20 was used solely for testing. Additionally, the Challenging Medically Relevant Genes (CMRG) regions[7] were excluded from all training in all models.

For the GIAB v4.2.1 high confidence regions, whole-genome evaluations are presented for the excluded GIAB sample from each model's training. For the HG002 T2T v1.1 truth set,[8] accuracy is reported only on chromosome 20, the holdout chromosome, and training was performed using HG002 chromosomes 1 – 19, as well as the rest of the GIAB samples. This approach was taken as the T2T truth set is exclusively available for the HG002 sample, and its inclusion in training significantly improves accuracy. Of the seven trained models, the model that excluded HG001 in training is made available.

For comparison, evaluations with Illumina DeepVariant models used the DeepVariant v1.9 production models available via GitHub. The training data for these models are described at: https://github.com/google/deepvariant/blob/r1.9/docs/deepvariant-details-training-data.md.

In these models, we excluded HG003 from training. Chromosomes 20, 21, and 22 were also excluded from all training in the same manner described for SBX data.

3. Germline small variant calling performance

3.1 Germline variant calling evaluation using GIAB NIST benchmarking (high confidence) regions

GIAB samples are an invaluable resource for benchmarking germline small variant calling performance using any novel sequencing technology, including SBX. These highly curated benchmarking samples provide gold standards for evaluating variant calling accuracy as they contain a comprehensive catalog of known variants, enabling rigorous assessment of sensitivity and specificity. Moreover, GIAB's detailed documentation of variants allows for in-depth analysis of sequencing performance in challenging genomic regions such as homopolymers and GC-rich regions.[9]

For benchmarking the germline variant calling against multiple truth sets, Illumina WGS PCR free deduped 30x (mapped) coverage data, downloaded from[10], was remapped. Note, the Illumina HG001 – HG005 data used was generated on a NovaSeq 6000 and the Illumina HG006 and HG007 data used was generated on a HiSeq X to ensure consistent, high accuracy across all samples. For SBX-D and SBX-Fast 30x deduped (mapped) concordant duplex coverage was used. Note, SBX-Fast data was originally presented at the 2025 European Human Genetics Conference[11] and SBX-D data was made available within the Google Brain Genomics Public Repository (gs://brain-genomics-public/research/sbx/2025/full_bam/). Subsequently, germline small variant calling was performed using DeepVariant v1.9 [Docker version: google/deepvariant:pangenome_aware_deepvariant-sbx https://github.com/google/deepvariant/blob/r1.9/docs/roche-sbx-case-study.md] on bam files including the simplex tails [https://github.com/google/deepvariant/releases/tag/v1.9.0], and variant calling performance (F1 scores) was evaluated using hap.py v0.3.12 against the NIST v.4.2.1 truth set.[12,13]

DeepVariant performed well with all three datasets, with respect to both SNP and indel calling. As expected, SBX-D consistently produced higher SNP F1 scores compared to SBX-Fast as a result of better mappability from longer SBX-D insert lengths. In contrast, SBX-Fast demonstrated higher indel F1 scores due to higher (long) homopolymer accuracy (Figure 7).

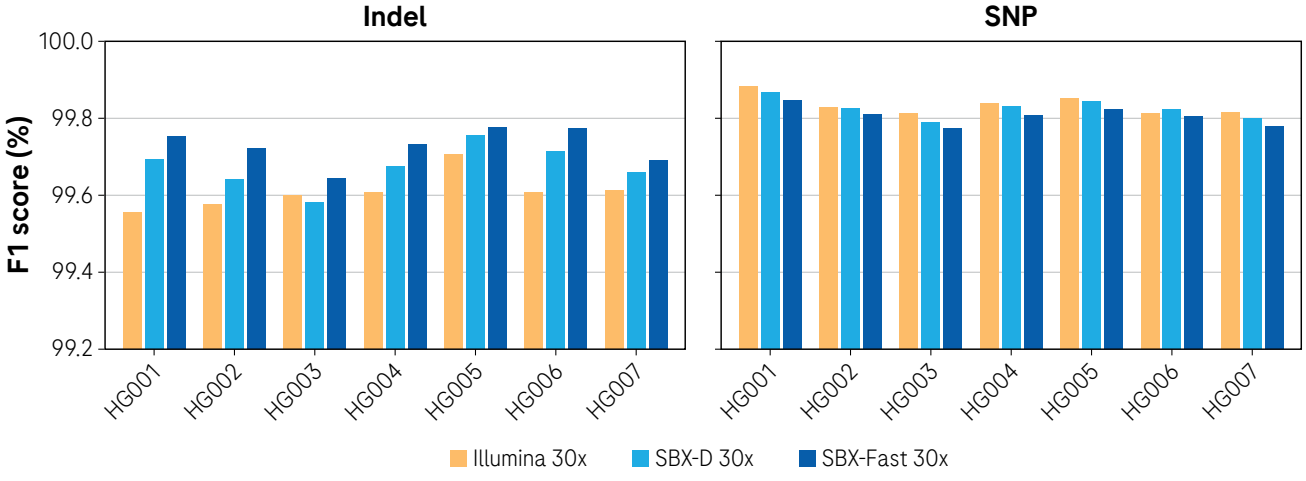


Figure 7. Benchmarking germline small variant calling performance on GIAB samples. Indel and SNP calling F1 scores for 30x Illumina PCR free, SBX-D and SBX-Fast are shown.

It is important to note the total coverage from SBX data can vary depending on the fraction of bases from partial duplex reads with only single-strand support (simplex bases), and that DeepVariant can use evidence from both duplex and simplex portions of partial duplex SBX reads for variant calling. In this study, simplex bases were excluded from concordant duplex coverage calculations. If included in the coverage calculation, these bases—which are generated by the SBX-D workflow in any event—would have bolstered coverage by an average of ~10x. By utilizing these additional single-strand bases for variant calling, SBX achieved better indel F1 scores in some cases than those obtained from the NovaSeq 6000 and HiSeq X data (analysis via DRAGEN was not performed), despite slightly lower homopolymer accuracy.

Long (>12 bp) homopolymer regions have a high concentration of inherited insertions and deletions, despite making up less than 1% of the genome. For example, over 20% of indels in HG001 are found in such regions. Sequencing platforms generally show a higher error rate in longer homopolymer regions. Analyzing DeepVariant's performance across different homopolymer lengths revealed consistently high F1 scores for both SBX-D and SBX-Fast. Notably, SBX-Fast performed slightly better relative to SBX-D in homopolymers longer than 18 bp (Figure 8).

3.2 Germline variant calling performance stratified by genomic context

Accurate germline variant calling is challenging in many genomic regions beyond homopolymers, including tandem repeats, segmental duplications, hard-to-map areas, and regions with extreme GC content. We benchmarked DeepVariant's performance in these difficult regions using SBX-D, SBX-Fast, and Illumina PCR-free WGS data (Genome Stratification v3.6[9]).

DeepVariant performed well across all genomic regions in all three datasets. However, lower F1 scores were observed in all three datasets for regions with low GC content, segmental duplications, and regions with low mappability (Figure 9). Longer inserts/sequencing reads, more diverse pangenome references, and improved coverage uniformity (determined with concordant duplex coverage) in low GC regions could contribute to improved germline variant calling in these regions.

3.3 Evaluation of SBX germline variant calling performance against the telomere-to-telomere (T2T) draft benchmark

We compared DeepVariant's performance on the T2T-CHM13 assembly,[8] using SBX-D, SBX-Fast, and Illumina PCR-free WGS data. The T2T-CHM13 assembly fills in previously missing regions—centromeres, large satellite arrays, ribosomal DNA, and other GC-rich or hard-to-map repeats—thereby expanding the landscape in which variant callers must operate. Unlike the GRCh38 reference assembly, which contains approximately 150 million base pairs of gaps distributed throughout the genome, underrepresents repetitive sequences, and comprises a mosaic of haplotypes due to its construction from multiple individuals, the T2T-CHM13 provides a more complete and accurate representation of the human genome. This exposes genomic contexts that were previously inaccessible, offers a more stringent benchmark for read alignment and variant calling, and necessitates continued refinement of analysis pipelines as reference genomes become increasingly comprehensive. High accuracy SNP and indel calling was achieved from both SBX-D and SBX-Fast data with this challenging T2T benchmark (Figure 10).

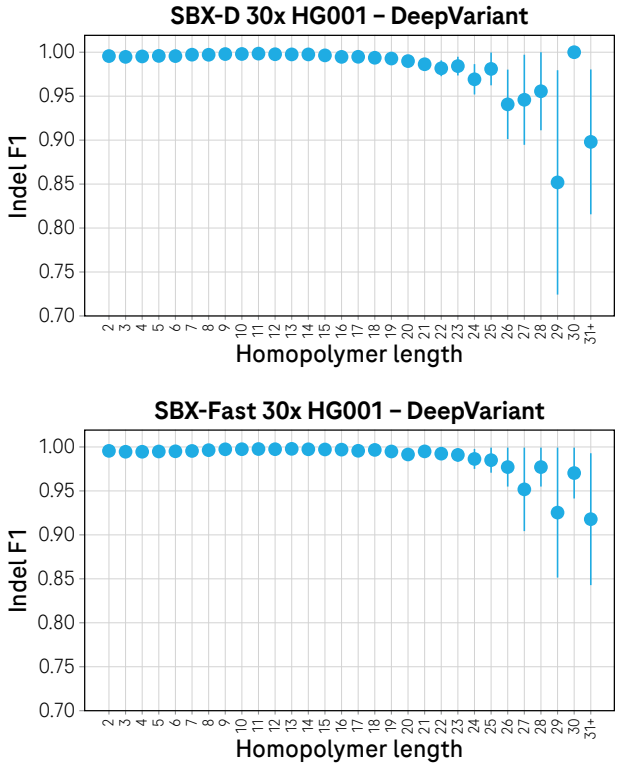


Figure 8. Breakdown of DeepVariant germline small variant calling performance based on homopolymer length. DeepVariant indel calling F1 scores remained high for both SBX-D and SBX-Fast, even for homopolymers as long as 25 bp. The error bars denote the 95% confidence interval, estimated as F1 +/- 2 * sqrt(F1 * (1 - F1) / N), where N is the total number of variant calls in the callset.

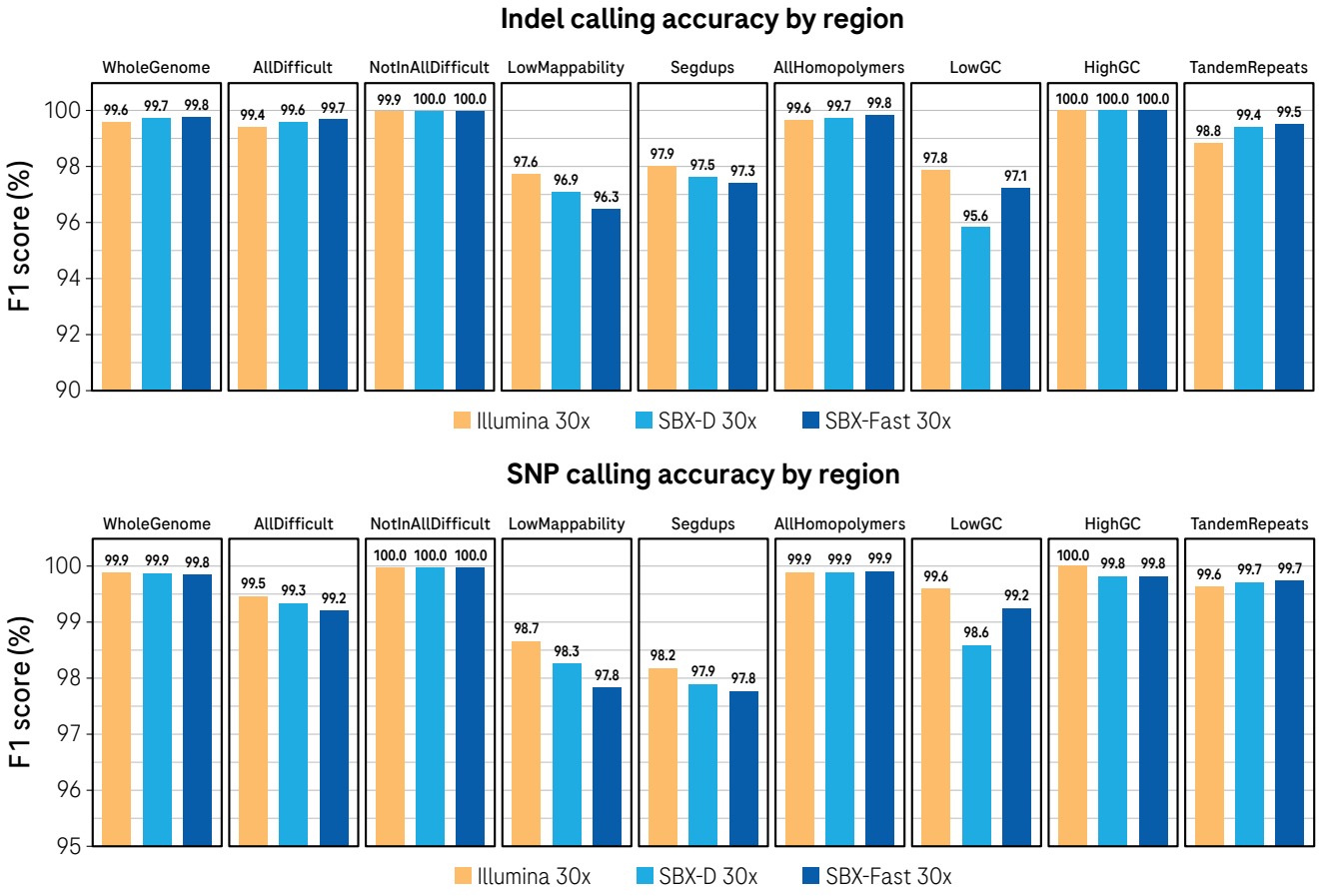


Figure 9. Breakdown of DeepVariant germline small variant calling performance in challenging genomic regions intersected by GIAB benchmarking (high confidence) regions in HG001. As noted, Illumina's NovoSeq 6000 reads were used for HG001 – HG005 and HiSeqX reads were used for HG006 and HG007.

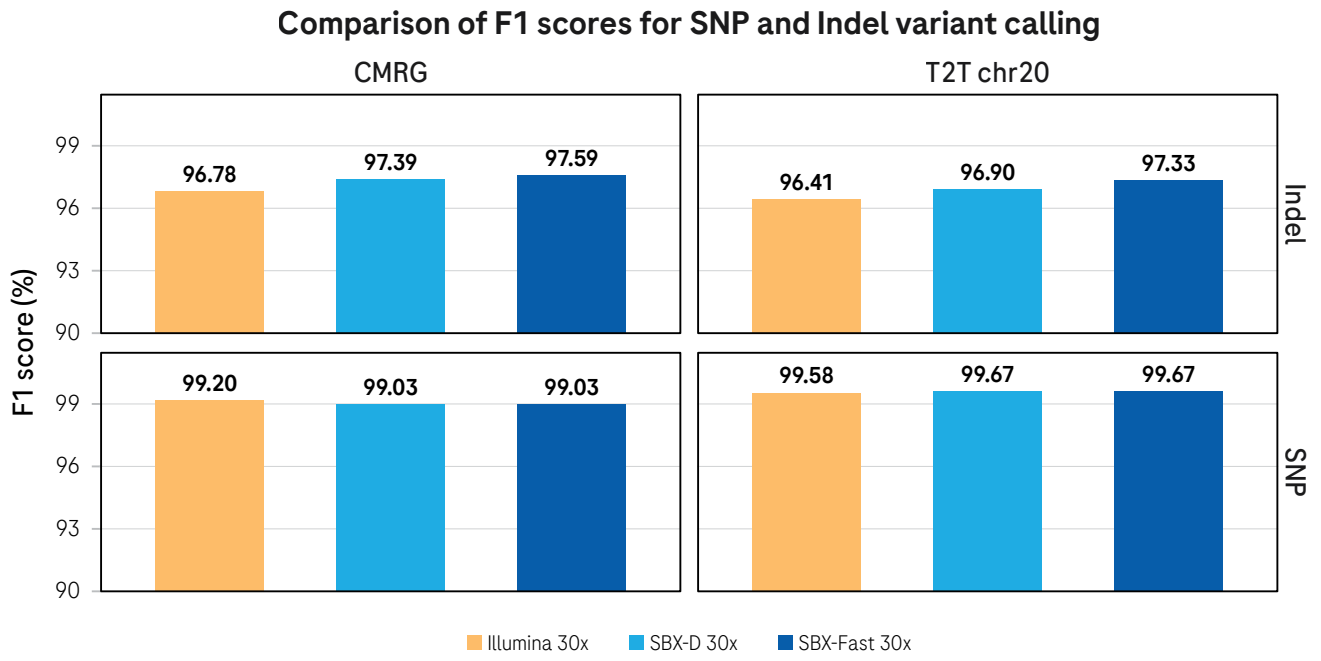


Figure 10. Benchmarking of HG002 germline variant calling on the T2T truth set (chr20) and 273 CMRGs.

3.4 Germline variant calling performance against the CMRG benchmark

Next, we evaluated the performance of germline small variant calling against the CMRG benchmark.[7] This benchmark comprises 273 highly complex, medically significant genes that meet stringent criteria: their full sequence—plus 20 kb flanking regions—must be present in a single assembled contig and align without breaks to both GRCh37 and GRCh38 reference genomes (though overlap with segmental duplications is permitted). A significant portion (at least 15% of the gene body in 99% of the 273 genes) poses sequencing or variant detection challenges due to low mappability and repetitive elements. Furthermore, 11% of indels within these genes are larger than 15 bp, complicating their accurate detection and impacting the precision and recall of analysis tools. Consequently, the 273 CMRGs in the benchmark dataset represent difficult-to-analyze genomic regions crucial for accurate variant identification, thereby challenging existing methodologies across sequencing, alignment, variant calling, and representation.

Figure 10 illustrates the performance of DeepVariant on the HG002 T2T benchmarking truth set (HG002-T2TQ100-v1.1) and in the CMRG regions (v1.00). SBX-D and SBX-Fast were highly competitive under the conditions in these experiments with respect to SNP and indel calling accuracy across both benchmarking regions. Combining DeepVariant with other specialized variant callers, such as those optimized for segmental duplications or pseudogenes/homologues, is expected to further improve variant detection in challenging, medically relevant genes.

4. Conclusion and future directions

In this white paper we have presented key SBX-D and SBX-Fast sequencing quality metrics, such as coverage uniformity, insert length distribution, and error rates, resulting from the germline variant calling workflow in the experiments presented here (demultiplexing, alignment with a pangenome aligner, duplicate marking using specialized tools for single-end data, and variant calling with DeepVariant). Benchmarking against GIAB and T2T truth sets, confirmed high germline variant calling accuracy for both SNPs and indels, especially in challenging genomic regions like homopolymers and repeats. With high sequencing accuracy and insert lengths longer than typical short-read technology, SBX supports impressive variant calling performance when combined with advanced bioinformatics tools such as a pangenome aligner and DeepVariant.

We are actively pursuing continuous improvements to germline small variant calling with the DeepVariant SBX workflow. To this end, future SBX chemistry development is focused on improved sequencing quality with even longer reads, reduced SNP and indel error rates, and more uniform coverage. While homopolymer and tandem repeat accuracy are crucial for indel calling, current limitations in both SNP and indel calling arise from difficulties in aligning reads within highly homologous genomic regions. Beyond longer reads, bioinformatics strategies such as including a more comprehensive pangenome reference (ongoing HPRC efforts) and decoy contigs to mask or target unknown false duplications, are being contemplated to enhance accuracy in these complex regions.

5. Final notes on training DeepVariant

A pre-trained SBX model is available for download [gs://brain-genomics-public/research/sbx/2025/model] and a case study for Deepvariant for SBX is available [https://github.com/google/deepvariant/blob/r1.9/docs/roche-sbx-case-study.md].

Note, the SBX technology and analysis tools are currently in development. The content of this material reflects current study results and/or design goals. As SBX chemistry advances, state-of-the-art models will be trained, updated, and made available to users. Currently, DeepVariant pre-trained models are exclusively available for WGS, but future releases will include support for other applications such as exome sequencing.

6. References

1. Kokoris, Mark, et al. Sequencing by Expansion (SBX)–a novel, high-throughput single-molecule sequencing technology. *bioRxiv* (2025): 2025 – 02.
2. Sirén, Jouni, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374.6574 (2021): abg8871.
3. Olson, N.D., Wagner, J., Dwarshuis, N. et al. "Variant calling and benchmarking in an era of complete human genome sequences." *Nat Rev Genet* 24, 464 - 483 (2023)
4. Sirén, Jouni, et al. Personalized pangenome references. *Nature Methods* 21.11 (2024): 2017-2023.
5. Beyer, Wolfgang, et al. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* 35.24 (2019): 5318 – 5320.
6. Poplin, Ryan et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* vol. 36,10 (2018): 983-987. doi:10.1038/nbt.4235
7. Wagner, Justin, et al. "Curated variation benchmarks for challenging medically relevant autosomal genes." *Nature biotechnology* 40.5 (2022): 672 – 680.
8. Nurk, Sergey et al. The complete sequence of a human genome. *Science* (New York, N.Y.) vol. 376,6588 (2022): 44 – 53. doi:10.1126/science.abj6987
9. Dwarshuis, Nathan, et al. The GIAB genomic stratifications resource for human reference genomes. *Nature communications* 15.1 (2024): 9029.
10. Baid, Gunjan, et al. An extensive sequence dataset of gold-standard samples for benchmarking and development. *BioRxiv* (2020): 2020 – 12.
11. "Enabling rare disease research with rapid workflows by SBX technology and the AVENIO Edge automated KAPA HyperExome V2 solution", European Human Genetics Conference in Milan, Italy, May 2025
12. Wagner, Justin, et al. Benchmarking challenging small variants with linked and long reads. *Cell genomics* 2.5 (2022).
13. Krusche, Peter, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology* 37.5 (2019): 555 – 560.

7. Contributors

Special appreciation and sincere gratitude are given to the following contributors from the Google Research and Google Cloud teams for their collaboration and significant contributions to the studies performed and the results demonstrated within this white paper.

Pi-Chuan Chang
Experimental design, error analysis, model training

Alexey Kolesnikov
De Bruijn graph modifications, error analysis

Lucas Brambrink
Model representation, small model training

Ali Jafari
Cloud TPU training

Omar Bayoumi
Cloud infrastructure support, resource management

Daniel E. Cook
Code optimizations for TPU training

Kishwar Shafin
Experimental design, error analysis, model training

Andrew Carroll
Experimental design, error analysis

Published by:

Roche Sequencing Solutions, Inc.
4300 Hacienda Drive
Pleasanton, CA 94588

sequencing.roche.com

For Research Use Only. Not for use in diagnostic procedures.
The content of this material reflects current study results or design goals.
All product names and trademarks are the property of their respective owners.
© 2026 Roche Sequencing Solutions, Inc. All rights reserved.

MC-US-17767 08/26