Robust and accurate genome analysis with SOPHiA DDM™ WGS solutions

#1878

Izabela Matyszczak, David Fast, Tommaso Coletta, Tomas Zimmermann, Bita Khalili, Dmitri Ivanov, Loc Tran, Amjad Alkodsi, Yuanlong Liu, Yohann Nedelec, Jaume Bonet Martinez, Mark Laver, Christian Pozzorini , Alex Tuck, Zhenyu Xu
SOPHiA GENETICS, Rolle, Switzerland
Conflicts of interest: All authors are employees of SOPHiA GENETICS.

1 Highlights

The SOPHiA DDM™ lpWGS and 30xWGS solutions demonstrate high sensitivity and precision across all variant types, surpassing established benchmarks for germline variant detection.

The solutions offer a complete and streamlined workflow, from sequencing to variant interpretation and reporting, enabling comprehensive and actionable genomic insights for rare and inherited disease testing.

2 Background

Accurate identification of genetic variants is essential for characterizing inherited and rare diseases. Whole-genome sequencing (WGS) has emerged as a powerful approach, offering comprehensive insight into all classes of genomic variation in a single assay. However, transforming WGS into a routine clinical tool requires robust, high-performing analytic solutions that can reliably detect variants. SOPHiA GENETICS has developed two complementary WGS approaches within the SOPHiA DDM™ platform: a low-pass WGS (lpWGS) solution optimized for CNV and RoH detection, and a high-depth 30xWGS solution designed for full-spectrum variant analysis. These solutions are engineered to deliver high accuracy and scalability for diverse genomic applications.

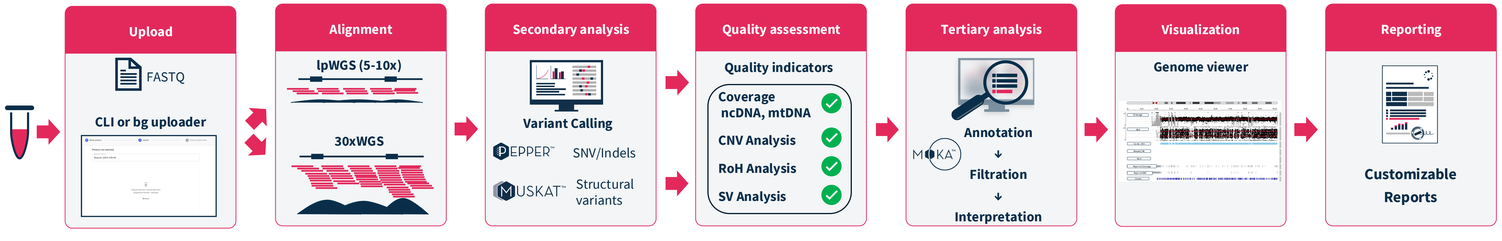
3 Aims

1. **Demonstrate a complete SOPHiA DDM™ workflow, from sequencing data to variant interpretation.**
2. **Assess analytical performance of the lpWGS and 30xWGS solutions across key germline variant types.**
3. **Benchmark the solutions against established tools and public reference datasets using real and simulated data.**

4 SOPHiA DDM™ WGS Workflow

The SOPHiA DDM™ platform supports two WGS workflows tailored to different analytical needs:

- The **lpWGS** workflow (5–10×) enables detection of CNVs, RoH, and mitochondrial variants offering a cost-effective alternative to chromosomal microarrays
- The **30xWGS** workflow (~30×) supports full-spectrum variant analysis, including SNVs/Indels, CNV, RoH, balanced structural variants (insertions, inversions, translocations), and mitochondrial variants



5 Analytical Performance and Benchmarking

MUSKAT™ Structural Variants

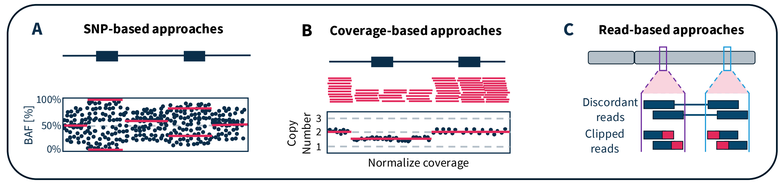
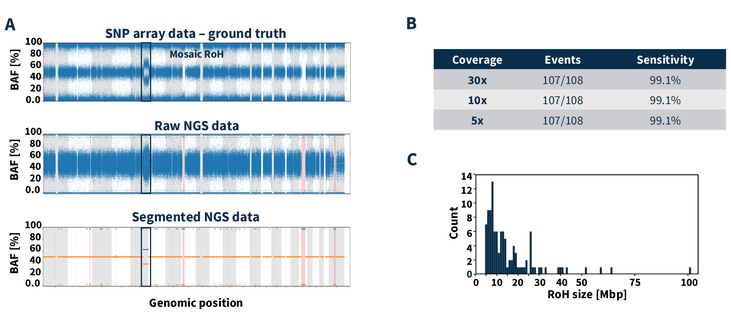


Figure 1. Overview of the approach. The SOPHiA DDM™ MUSKAT™ algorithm detects a broad range of structural variants by integrating three complementary signals from WGS data. **A)** SNP-based approach uses B-allele frequency (BAF) that captures allelic imbalance, enabling detection of regions of homozygosity (RoH) and copy number variants (CNVs). **B)** Coverage-based approaches identify CNVs through coverage deviations. **C)** Read-based approach uses split and discordant reads to detect balanced and complex events with precise breakpoint resolution. The lpWGS leverages signals A and B, while the higher coverage of 30xWGS (~30×) enables integration of all three signals, unlocking more precise and comprehensive structural variant detection.

ROH



Coverage	Events	Sensitivity
30x	107/108	99.1%
10x	107/108	99.1%
5x	107/108	99.1%

Figure 2. RoH detection. A) RoH are identified by evaluating BAF at polymorphic sites across the genome. Extended regions with BAF deviating from the expected 50% heterozygous state may indicate constitutional or mosaic RoH. Coverage data is integrated to distinguish true homozygosity from BAF shifts caused by copy number alterations. **B)** The approach was benchmarked using 20 samples from the 1000 Genomes Project, previously characterized by SNP arrays. Out of 108 RoH events ≥5 Mbp, 107 were correctly detected, yielding a sensitivity of 99.1% even at 5× coverage. **C)** Detected events span a broad range of sizes.

CNV

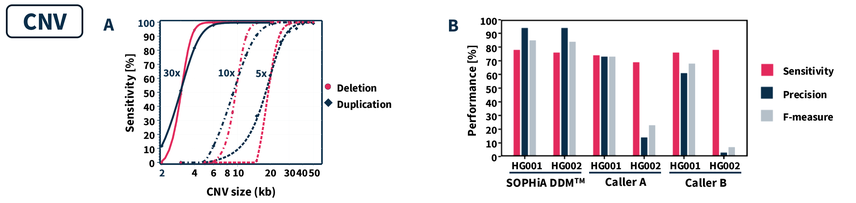


Figure 3. CNV detection. A) CNV detection was performed by analyzing deviations in normalized coverage across the genome. Performance was assessed using simulated CNVs of varying sizes (5–100 kb, in 5 kb increments) and copy number states (1 for deletions, 3 for duplications), injected into 15 TCGA WGS samples. For each size and state, 5 events were introduced and repeated across 5 replicates, generating 75 datasets and 750 CNVs per bin. Sensitivity was evaluated at 30×, 10×, and 5× coverage, demonstrating robust detection of events ≥10 kb. **B)** CNV detection was also performed using split and discordant read signals. Analytical performance was evaluated on the Genome in a Bottle (GIAB) reference sample HG001 and HG002 with characterized CNV events using 30× coverage data. Results were benchmarked against leading CNV detection tools, showing higher sensitivity, precision, and F1-score for the evaluated approach, underscoring the effectiveness of read-based evidence in CNV detection.

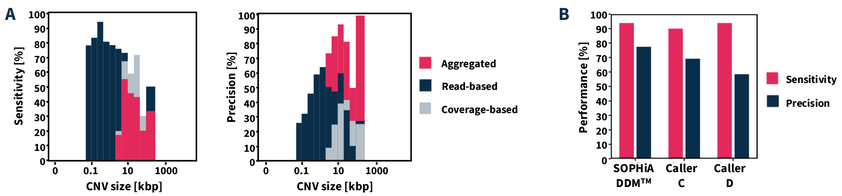


Figure 4. Impact of signal integration on CNV detection. A) Sensitivity and precision were evaluated across CNV size bins using individual signal types (coverage-based and read-based) and an aggregated approach. Coverage-based detection showed higher sensitivity for larger events, while read-based detection achieved higher precision for smaller CNVs. Aggregating both signals substantially improved precision for events ≥10 kb. Balancing the strengths of each method leads to increased overall performance. **B)** Benchmarking on the GIAB reference sample HG002 confirmed that the integrated approach outperforms leading CNV callers in both sensitivity and precision, highlighting the value of multi-signal integration for accurate structural variant detection.

PEPPER™ SNV/Indels

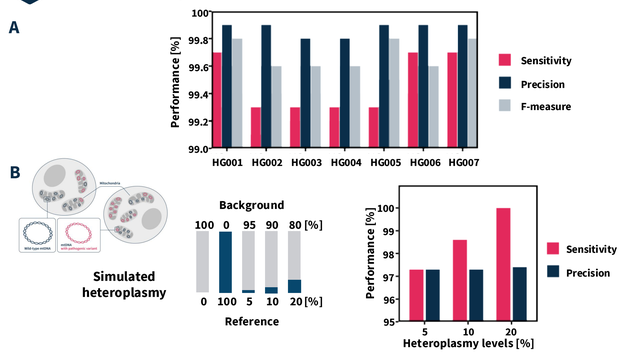


Figure 5. SNV/Indel detection in nuclear and mitochondrial genomes. A) Detection performance of SNV/Indels in the nuclear genome was evaluated across GIAB reference samples HG001–HG007. Results show consistently high sensitivity, precision, and F1-scores, demonstrating accurate variant calling across diverse genomic backgrounds. **B)** Mitochondrial variant detection was assessed using synthetic mixtures simulating heteroplasmic variants at 5%, 10%, and 20% levels, with GIAB reference samples used as backgrounds. High performance was maintained even at low heteroplasmy, supporting robust detection of mitochondrial variants across a range of allele fractions.

6 Conclusions

- Multi-signal integration enables accurate detection of structural variants.
- RoH detection is robust using combined BAF and coverage signals.
- CNV detection improves with integrated coverage and read evidence.
- SNV/Indel calling remains accurate in mtDNA even at low heteroplasmy.

Presented at the European Human Genetics Conference, ESHG, Milan, Italy, May 24-27, 2025