# SVIM-asm: Structural variant detection from haploid and diploid genome assemblies

David Heller [1] and Martin Vingron [1*]

[1]Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Berlin, 14195, Germany
*To whom correspondence should be addressed.

**Abstract**

**Motivation:** With the availability of new sequencing technologies, the generation of haplotype-resolved genome assemblies up to chromosome scale has become feasible. These assemblies capture the complete genetic information of both parental haplotypes, increase structural variant (SV) calling sensitivity and enable direct genotyping and phasing of SVs. Yet, existing SV callers are designed for haploid genome assemblies only, do not support genotyping or detect only a limited set of SV classes.
**Results:** We introduce our method SVIM-asm for the detection and genotyping of six common classes of SVs from haploid and diploid genome assemblies. Compared against the only other existing SV caller for diploid assemblies, DipCall, SVIM-asm detects more SV classes and reached higher F1 scores on two recently published assemblies of the HG002 individual.
**Availability and Implementation:** SVIM-asm has been implemented in Python and can be easily installed via bioconda. Its source code is available at github.com/eldariont/svim-asm.
**Contact:** vingron@molgen.mpg.de
**Supplementary information:** Supplementary data are available online.

## 1 Introduction

As one of the main classes of genomic variation, structural variants (SVs) comprise a diverse range of genomic rearrangements with sizes larger than 50 bps. Although there are considerably less SVs than Single Nucleotide Variants (SNVs) or small indels in an average human genome, SVs affect more base-pairs (1000 Genomes Project Consortium, 2015). Consequently, SVs have a strong effect both on the healthy human phenotype and human disease.

Due to the availability of affordable and accurate next-generation sequencing (NGS) technology, SVs are now commonly detected by the analysis of sequencing reads. Typically, the reads from a genome under investigation are aligned to an existing reference genome to reveal differences between both genomes. This

alignment-based approach works well in most genomic regions but suffers from three major problems: (1) low mappability in repetitive regions of the genome, (2) the reference bias working against reads differing from the reference, and (3) the problem that long SVs cannot be completely covered by a single read.

De novo assembly presents itself as a viable alternative for the detection of SVs from sequencing reads. Using sequence overlaps between reads, longer genomic fragments, called contigs, can be reconstructed computationally. Then, these assembly contigs can be aligned to a reference or comparison genome. Although the task of genome-genome alignment remains challenging, above mentioned problems of read alignment are alleviated by the alignment of long assembly contigs.

While a growing number of software tools detect SVs from the alignments of short and long reads (Kosugi et al., 2019), only few tools are available for the detection of SVs from genome-genome alignments. Three such tools, AsmVar, Assemblytics and SyRy, have been developed recently but only for the analysis of haploid genome assemblies (Goel et al., 2019).

Until recently, genome assembly methods usually collapsed the two parental haplotypes in a diploid genome into a haploid genome representation. With the availability of longer sequencing reads and complementary sequencing technologies like Hi-C and Strand-Seq, however, the routine production of haplotype-resolved human genome assemblies has become feasible (Garg et al., 2019; Nurk et al., 2020). These haplotype-resolved assemblies capture the complete genetic information of both parental haplotypes, can increase variant calling sensitivity (Chaisson et al., 2019) and enable direct genotyping and phasing of variants. Within the next few years, the routine generation of diploid genome assemblies from different ethnic backgrounds will offer an unprecedented view into the variation landscape of the human genome. Yet, only one method, DipCall, has been published so far for the detection of large insertions and deletions from haplotype-resolved genome assemblies (Li et al., 2018). In this study, we introduce our method SVIM-asm for the detection and genotyping of six common classes of SVs, including insertions and deletions, from haploid and diploid genome assemblies.

## 2 Materials and Methods

SVIM-asm (Structural Variant Identification Method for Assemblies) is based on our previous method SVIM that detects SVs in long-read alignments (Heller and Vingron, 2019). Although SVIM-asm follows a similar workflow as SVIM, several adaptions have been made to consider the unique properties of assembly alignments compared to read alignments (see Figure S1). To enable the analysis of both haploid and diploid genome assemblies, SVIM-asm implements two different pipelines (see Figure S2).

Diploid genome assemblies consist of two sets of contigs, one for each parental haplotype. In the first step of the pipeline (COLLECT), SV signatures are extracted separately for each haplotype from discordant alignments of individual

contigs to the reference. The discordancies fall into two categories: a) long alignment gaps within alignment segments (intra-alignment discordancies) and b) discordant positions and orientations between alignment segments (inter-alignment discordancies).

In the second step of the pipeline (PAIR), signatures from opposite haplotypes are compared and paired up if sufficiently similar. To measure the similarity of two signatures, the edit distance (Levenshtein distance) between their haplotype sequences is computed with the library edlib (Šošić and Šikić, 2017). Based on the computed distances, very similar signatures (i.e. signatures with similar haplotype sequences) from different haplotypes are merged.

In the third step of the pipeline (GENOTYPE), paired signatures from the two opposite haplotypes are merged into homozygous SV candidates while variants without a partner on the other haplotype are called as heterozygous SV candidates. Finally, the genotyped SVs are written out in VCF as members of one of six SV classes (OUTPUT).

In contrast to their diploid counterparts, haploid assemblies consist of only a single set of contigs. For diploid organisms, this set often represents a mixture of the two haplotypes. Due to the missing second haplotype, it is not possible to estimate genotypes from haploid genome assemblies. After the same first step (COLLECT) is applied to the assembly alignments, the PAIR and GENOTYPE steps are skipped for haploid assemblies and the detected SV signatures are written out immediately (OUTPUT).

# 3 Results

We compared our tool, SVIM-asm (v1.0.0), to the DipCall pipeline (v0.1). For the evaluation we chose two publicly available diploid genome assemblies of the HG002 individual from Wenger et al. (Assembly A) and Garg et al. (Assembly B) (see Supplementary Methods) (Wenger et al., 2019; Garg et al., 2019). We aligned the fragments separately for each haplotype using minimap2 (v2.17-r941) and produced genotyped SV calls using SVIM-asm and DipCall, respectively (Li, 2018).

When compared against the GIAB SV benchmark set (Zook et al., 2020) using truvari (v2.0.1), both methods reached F1 scores above 90% when only the variant calls were evaluated (see Figure S3). SVIM-asm performed slightly better than DipCall with F1 scores of 93.2% (Assembly A) and 93.7% (Assembly B) compared to 91.7% and 92.5%, respectively. When additionally requiring matching genotypes, the values for both tools decreased but SVIM-asm outperformed DipCall by a wide margin (see Figure 1). While SVIM-asm reached F1 scores of 62.4% and 67.0%, DipCall only reached 48.3% and 54.0%, respectively. When measuring precision and recall across variant lengths, we observed that SVIM-asm reached a higher recall particularly for large deletions and insertions (see Figure S4). Unlike DipCall, SVIM-asm is able to detect inversions, two types of duplications, and translocation breakpoints. From the Assemblies A and B it detected 14475 / 14001 insertions, 9479 / 9217 deletions, 90 / 73 inver-

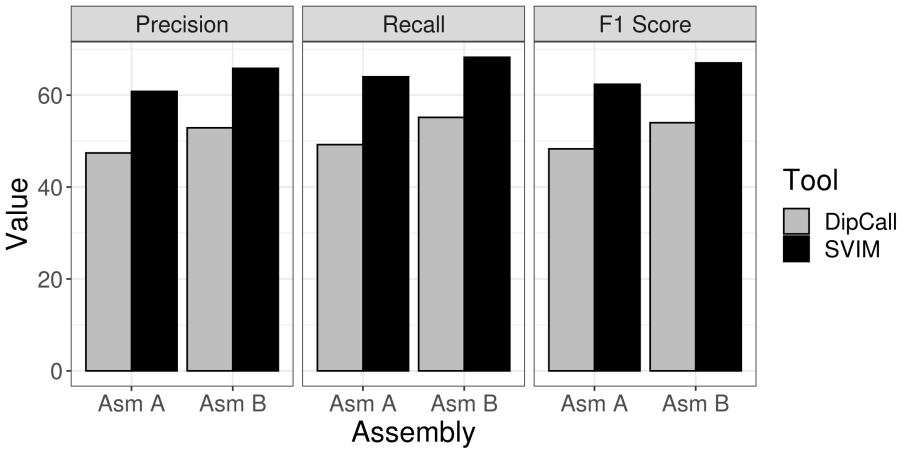


Figure 1: Comparison of SV detection and genotyping performance of DipCall and SVIM-asm on two diploid genome assemblies (x-axis, Asm A by Wenger et al. and Asm B by Garg et al.). Plotted are the precision, recall and F1 score (y-axis) measured by comparing against the GIAB SV benchmark set. Only SV calls with identical genotypes in the benchmark set were classified as true positives.

sions, 124 / 101 tandem duplications, 2 / 4 interspersed duplications, and 3110 / 4028 translocation breakpoints, respectively (see Figures S5 and S6).

# 4 Discussion

De novo genome assembly provides a good alternative to read-based analyses because it can overcome some of the problems of read alignment, such as mappability issues, reference bias and poor sensitivity for large variants. In this context, SVIM-asm is an important tool to leverage the growing number of high-quality haplotype-resolved genome assemblies for the comprehensive characterization of SVs. Compared to existing tools for assembly-based SV detection, SVIM-asm supports more SV types and reached a higher SV calling performance in our benchmarks, particularly for the detection of variant calls with correct genotypes.

# Funding

This work was supported by the German Bundesministerium für Bildung und Forschung (031L0169A) and the doctoral program of the International Max Planck Research School for Biology and Computation.

# References


1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015.

Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, 10(1): 1–16, 2019.

Shilpa Garg, Arkarachai Arkarachai Fungtammasan, Andrew Carroll, Mike Chou, Anthony Schmitt, Xiang Zhou, Stephen Mac, Paul Peluso, Emily Hatas, Jay Ghurye, et al. Efficient chromosome-scale haplotype-resolved assembly of human genomes. *bioRxiv*, 2019. doi: 10.1101/810341.

Manish Goel, Hequan Sun, Wen-Biao Jiao, and Korbinian Schneeberger. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.*, 20 (1):1–13, 2019.

David Heller and Martin Vingron. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35(17):2907–2915, 2019.

Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, 20(1):117, 2019.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 1:7, 2018.

Heng Li, Jonathan M Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods*, 15(8):595–597, 2018.

Sergey Nurk, Brian P Walenz, Arang Rhie, Mitchell R Vollger, Glennis A Logsdon, Robert Grothe, Karen H Miga, Evan E Eichler, Adam M Phillippy, and Sergey Koren. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *BioRxiv*, 2020.

Martin Šošić and Mile Šikić. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.

Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162, 2019.

Justin M Zook, Nancy F Hansen, Nathan D Olson, Lesley Chapman, James C Mullikin, Chunlin Xiao, Stephen Sherry, Sergey Koren, Adam M Phillippy, Paul C Boutros, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.*, pages 1–9, 2020.