Material and Methods

Genomic data

We re-used Pacbio HiFi long reads produced by the Darwin Tree of Life project (The Darwin Tree of Life Project Consortium et al. 2022). Raw long reads were downloaded from the SRA database with fastq-dump from the sra-tools toolkit (« sra-tools » 2025), then merged in a single fastq with samtools (Danecek et al. 2021; Li et al. 2009). The reference genome and assembly reports produced from these reads were downloaded from the NCBI genome assembly database with the datasets toolkit (O'Leary et al. 2024). The quality of raw long reads was assessed with fastqc and nanoplot before filtering (De Coster et Rademakers 2023; Andrews 2010). Then long reads were filtered with chopper (De Coster et Rademakers 2023), removing reads longer than 20 Mb, shorter than 500 bp and with an average Phred-scaled quality score lower than 10.Then the quality of filtered reads was assessed a second time with nanoplot.

Structural Variants calling and genotyping

We implemented an automatic and reproducible pipeline coded in Snakemake to map reads on the reference genome, call and genotype SVs and check the quality at each critical step (Mölder et al. 2021). All the settings for the analyses were coded in a yaml config file for reproducibility.

We used two different aligners, minimap2 and ngmlr, to map filtered long reads on the reference genome (Sedlazeck et al. 2018; Li 2018). Minimap2 was used with the map-hifi preset for the Pacbio HiFi sequencing technology. Similarly ngmlr was used with the pacbio preset and alignments with an identity lower than 0.65 were discarded. Alignements were sorted and indexed with samtools. Mapping quality was checked with samtools stats and plot-bamstats (Danecek et al. 2021; Li et al. 2009). The mappability and callability of each alignment was assessed and reported in bed files for post-processing and filtering purposes. The mappability was evaluated with genmap (Pockrandt et al. 2020) with default parameters by computing the uniqueness of k-mers for each position in the genome. The callability was evaluated based on depth calculation along the genome with mosdepth (Pedersen et Quinlan 2018). The mean coverage was computed on the aligned bam files with mosdepth summary. With mosdepth quantize we divided the genome in four callability categories (no coverage, low coverage, callable and high coverage). Regions were labelled as low coverage if the mean coverage was less than 10. On the contrary, they were labelled as high coverage if they had a coverage higher than three times the genome wide average coverage.

Four SV callers were used with each of the two alignments, resulting in an ensemble of eight different callsets. This allowed us to capture the most exhaustive number of SVs and to evaluate the uncertainty of each call and the dataset specific performance of each combination aligner-caller in downstream statistical analyses. The four SV callers, SVIM, Sniffles2, CuteSV and Debreak, were among the top SV callers in recent benchmarks (Ahsan et al. 2023; Liu et al. 2024; Smolka et al. 2024; Jiang et al. 2020; Chen et al. 2023; Heller et Vingron 2019). All SV callers were set with a required minimum coverage of 10 and a minimum SV length to detect of 30 bp. SVIM (Heller et Vingron 2019) was set with a minimum mapping quality of 20, a maximum tolerated gap between adjacent alignment segments of 10, a maximum tolerated overlap between adjacent alignment of 5, a minimum depth of 10, a minimum quality score of 10 and a minimum number of reads supporting the SV of 10. Sniffles2 (Smolka et al. 2024) was set with a minimum number of supporting reads automatically chosen based on coverage, a minimum SV length screen ratio 0f 0.9, a mapping quality of 25 and a cluster binsize of 100 bp. In CuteSV (Jiang et al. 2020), we set a maximum cluster bias of 1000. For insertions, the difference ratio for merging breakpoints was 0.3, while for deletions it was 0.5. DeBreak (Chen et al. 2023) was set with the '–rescue_DUP', '–rescue_large_ins' and '–poa' options.

We used the sniffles2-plot package (Smolka et al. 2024) to generate a set of QC summary plots for each combination aligner-caller. We removed BND (break ends) and pre-processed each callset with JasmineSV (Kirsche et al. 2023) and custom python scripts using pysam (« Pysam » 2018) in order to properly convert duplications to insertions for genotyping with SVjedi-graph (Romain et Lemaitre 2023). We genotyped SVs with SVjedi-graph (Romain et Lemaitre 2023) independently for each aligner-caller combination, with a relaxed minimum number of alignments of 1. After genotyping, we filtered the callsets consistently with bcftools (Danecek et al. 2021) to remove SVs larger than 30 Mb, with a minimum total depth of 10 (or 5 for the alternative allele), and a maximum depth of 500.

The eight ensemble callsets were merged with JasmineSV with the –ignore_strand option and maximum distance between SV breakpoints of 1 Mb. Then the merged callset was re-genotyped with SVjedi-graph with the same settings as before. Sex chromosomes and chloroplastic, mitochondrial and unplaced chromosomes were excluded from the vcf. A final vcf merging the eight aligner-caller vcf files and the jasmine merged vcf was produced with bcftools.

The mappability at the breakpoints of called SVs was evaluated with truvari anno grm (English et al. 2022; 2024) with default parameters by creating a kmer over the upstream and downstream reference and alternate breakpoints and then remapping that kmer to the reference genome.

For the final assessment of the results, 100 random insertions, deletions and inversions were plotted with samplot (Belyeu et al. 2021) for visual inspection and a custom Rmarkdown report was generated (R Core Team 2023). During this report we output diagnostic plots and computed performance scores for each aligner-caller. Since the reference genome assembly and SV calling were done from the same reads coming from a single individual, we assumed that called SVs can only be heterozygous. Hence we considered that SV called, not filtered out and genotyped as heterozygotes were true positive calls. On the contrary, SV genotyped as homozygotes were considered as false positive calls. SV not called by an aligner-caller but considered as a true positive for at least one other aligner-caller were counted as false negative calls. From these counts we were able to compute a precision, recall and F1 score for each aligner-caller. These tools- and dataset-specific scores were used to estimate an ensemble score for each call to propagate calling uncertainty in downstream analyses.

References

Ahsan, Mian Umair, Qian Liu, Jonathan Elliot Perdomo, Li Fang, et Kai Wang. 2023. « A Survey of Algorithms for the Detection of Genomic Structural Variants from Long-Read Sequencing Data ». *Nature Methods* 20 (8): 1143‑58. https://doi.org/10.1038/s41592-023-01932-w.

Andrews, Simon. 2010. « FastQC: a quality control tool for high throughput sequence data ». Babraham, UK. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Belyeu, Jonathan R., Murad Chowdhury, Joseph Brown, Brent S. Pedersen, Michael J. Cormier, Aaron R. Quinlan, et Ryan M. Layer. 2021. « Samplot: A Platform for Structural Variant Visual Validation and Automated Filtering ». *Genome Biology* 22 (1): 161. https://doi.org/10.1186/s13059-021-02380-5.

Chen, Yu, Amy Y. Wang, Courtney A. Barkley, Yixin Zhang, Xinyang Zhao, Min Gao, Mick D. Edmonds, et Zechen Chong. 2023. « Deciphering the Exact Breakpoints of Structural Variations Using Long Sequencing Reads with DeBreak ». *Nature Communications* 14 (1): 283. https://doi.org/10.1038/s41467-023-35996-1.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. « Twelve years of SAMtools and BCFtools ». *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

De Coster, Wouter, et Rosa Rademakers. 2023. « NanoPack2: Population-Scale Evaluation of Long-Read Sequencing Data ». Édité par Can Alkan. *Bioinformatics* 39 (5): btad311. https://doi.org/10.1093/bioinformatics/btad311.

English, Adam C., Egor Dolzhenko, Helyaneh Ziaei Jam, Sean K. McKenzie, Nathan D. Olson, Wouter De Coster, Jonghun Park, et al. 2024. « Analysis and Benchmarking of Small and Large Genomic Variants across Tandem Repeats ». *Nature Biotechnology*, avril. https://doi.org/10.1038/s41587-024-02225-z.

English, Adam C., Vipin K. Menon, Richard A. Gibbs, Ginger A. Metcalf, et Fritz J. Sedlazeck. 2022. « Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity ». *Genome Biology* 23 (1): 271. https://doi.org/10.1186/s13059-022-02840-6.

Heller, David, et Martin Vingron. 2019. « SVIM: Structural Variant Identification Using Mapped Long Reads ». Édité par Inanc Birol. *Bioinformatics* 35 (17): 2907‑15. https://doi.org/10.1093/bioinformatics/btz041.

Jiang, Tao, Yongzhuang Liu, Yue Jiang, Junyi Li, Yan Gao, Zhe Cui, Yadong Liu, Bo Liu, et Yadong Wang. 2020. « Long-Read-Based Human Genomic Structural Variation Detection with cuteSV ». *Genome Biology* 21 (1): 189. https://doi.org/10.1186/s13059-020-02107-y.

Kirsche, Melanie, Gautam Prabhu, Rachel Sherman, Bohan Ni, Alexis Battle, Sergey Aganezov, et Michael C. Schatz. 2023. « Jasmine and Iris: Population-Scale Structural Variant Comparison and Analysis ». *Nature Methods* 20 (3): 408‑17. https://doi.org/10.1038/s41592-022-01753-3.

Li, Heng. 2018. « Minimap2: Pairwise Alignment for Nucleotide Sequences ». Édité par Inanc Birol. *Bioinformatics* 34 (18): 3094‑3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et 1000 Genome Project Data Processing Subgroup. 2009. « The Sequence Alignment/Map Format and SAMtools ». *Bioinformatics* 25 (16): 2078‑79. https://doi.org/10.1093/bioinformatics/btp352.

Liu, Yichen Henry, Can Luo, Staunton G. Golding, Jacob B. Ioffe, et Xin Maizie Zhou. 2024. « Tradeoffs in Alignment and Assembly-Based Methods for Structural Variant Detection with Long-Read Sequencing Data ». *Nature Communications* 15 (1): 2447. https://doi.org/10.1038/s41467-024-46614-z.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. « Sustainable Data Analysis with Snakemake ». *F1000Research* 10 (janvier):33.

https://doi.org/10.12688/f1000research.29032.1.

O'Leary, Nuala A., Eric Cox, J. Bradley Holmes, W. Ray Anderson, Robert Falk, Vichet Hem, Mirian T. N. Tsuchiya, et al. 2024. « Exploring and Retrieving Sequence and Metadata for Species across the Tree of Life with NCBI Datasets ». *Scientific Data* 11 (1): 732. https://doi.org/10.1038/s41597-024-03571-y.

Pedersen, Brent S, et Aaron R Quinlan. 2018. « Mosdepth: Quick Coverage Calculation for Genomes and Exomes ». Édité par John Hancock. *Bioinformatics* 34 (5): 867‑68. https://doi.org/10.1093/bioinformatics/btx699.

Pockrandt, Christopher, Mai Alzamel, Costas S Iliopoulos, et Knut Reinert. 2020. « GenMap: Ultra-Fast Computation of Genome Mappability ». Édité par Jinbo Xu. *Bioinformatics* 36 (12): 3687‑92. https://doi.org/10.1093/bioinformatics/btaa222.

« Pysam ». 2018. https://github.com/pysam-developers/pysam.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Romain, Sandra, et Claire Lemaitre. 2023. « SVJedi-Graph: Improving the Genotyping of Close and Overlapping Structural Variants with Long Reads Using a Variation Graph ». *Bioinformatics* 39 (Supplement_1): i270‑78. https://doi.org/10.1093/bioinformatics/btad237.

Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, et Michael C. Schatz. 2018. « Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing ». *Nature Methods* 15 (6): 461‑68. https://doi.org/10.1038/s41592-018-0001-7.

Smolka, Moritz, Luis F. Paulin, Christopher M. Grochowski, Dominic W. Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, et al. 2024. « Detection of Mosaic and Population-Level Structural Variants with Sniffles2 ». *Nature Biotechnology*, janvier. https://doi.org/10.1038/s41587-023-02024-y.

« sra-tools ». 2025. https://github.com/ncbi/sra-tools.

The Darwin Tree of Life Project Consortium, Mark Blaxter, Nova Mieszkowska, Federica Di Palma, Peter Holland, Richard Durbin, Thomas Richards, et al. 2022. « Sequence Locally, Think Globally: The Darwin Tree of Life Project ». *Proceedings of the National Academy of Sciences* 119 (4): e2115642118. https://doi.org/10.1073/pnas.2115642118.