Comprehensive Structural Variant Detection: From Mosaic to Population-Level

Moritz Smolka[1], Luis F. Paulin[1], Christopher M. Grochowski[2], Medhat Mahmoud[1,2], Sairam Behera[1], Mira Gandhi[3], Karl Hong[4], Davut Pehlivan[2,5], Sonja W. Scholz[6,7], Claudia M.B. Carvalho[2,3], Christos Proukakis[8], Fritz J Sedlazeck[1,9]

1: Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
2: Department of Molecular and Human Genetics, Baylor College of Medicine, TX, USA
3: Pacific Northwest Research Institute (PNRI), Seattle, USA
4: Bionano Genomics, 9540 Towne Centre Dr #100, San Diego, USA
5: Division of Neurology and Developmental Neuroscience, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
6: Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA
7: Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD, USA
8: Department of Clinical and Movement Neurosciences, Royal Free Campus, Queen Square Institute of Neurology, University College London, London, UK
9: Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, USA

Abstract

Long-read Structural Variation (SV) calling remains a challenging but highly accurate way to identify complex genomic rearrangements. Here, we present Sniffles2, which is faster and more accurate than state-of-the-art SV caller across different coverages, sequencing technologies, and SV types. Furthermore, Sniffles2 solves the problem of family to population-level SV calling to produce fully genotyped VCF files by introducing a gVCF file concept. Across 31 Mendelian samples, we accurately identified causative SVs around *MECP2*, including highly complex alleles with three overlapping SVs. Sniffles2 also enables the detection of mosaic SVs in bulk long-read data. This way, we were able to identify multiple mosaic SVs across a multiple system atrophy patient brain. The identified SV showed a remarkable diversity within the cingulate cortex, impacting both genes involved in neuron function and repetitive elements. In summary, we demonstrate the utility and versatility of Sniffles2 to identify SVs from the mosaic to population levels.

Introduction

The role and biological impact of Structural Variation (SV) have become evident[1,2]. SVs are loosely defined as 50 bp or larger genomic alterations that fall into five types (insertions, inversions, deletions, duplications, and translocations) or a combination of these types[1]. Given that this type of variant impacts the greatest number of nucleotides in a genome, it is not surprising that evidence is mounting regarding their importance across all categories of life. This starts e.g., with important speciation events[3] and impacts plants[4,5], but goes further across human diseases (Mendelian[6,7] as well as complex diseases[8–10]) to cancer development[11–13] (e.g., HLA loss, oncogene amplification). Despite the importance of SVs, we are still struggling to detect germline vs. somatic SVs or even robustly identify *de novo* SVs[14–16]. The least often studied and thus challenging SVs are insertions that, as many studies showed, amount to half of all SVs found in a human genome[17–19]. The latter can either be recovered by long-read mapping methods or *de novo* assemblies followed by a genomic alignment[1,20].

Long-read sequencing came a long way over the past years from a novelty to a population/production scale mechanism to study SV[21,22]. The error rate of Oxford Nanopore and PacBio HiFi are both ever decreasing, soon reaching levels of Illumina-like errors along the genome[23,24]. Most recently, Oxford Nanopore Technologies (ONT) provided an insight into the upcoming chemistry update to produce Q20+ reads, which further seem to reduce the error rates (~2%)[25]. Indeed, several studies have now started to sequence larger and larger data sets or even medical applications using PacBio HiFi or ONT[21,26]. This trend started with GENCODE [22], but is ever increasing to other projects (e.g., All of Us initiative, CARD) and is currently peaking in the G42 endeavor to sequence multiple hundreds of thousands genomes. This trend also requires more efficient software to not just detect SVs, but also to merge and produce a fully genotyped VCF file[27,28]. The degrees of error and cost for long-read are also starting to promote applications in medical or clinical space[29,30]. This is needed as several genes or regions of the genome remain a “dark matter”[20,31]. Here, recent studies showed 386 medically relevant genes that are still escaping the analysis of standard clinical Illumina WGS[31]. Most of these genes (~70%) can be assessed using long-read technologies, but several challenges remain.

Furthermore, there are more complex SVs beyond simple deletions, duplications, inversions, insertions, and translocations that can lead to a Mendelian disease[6]. The genomic locus including the dosage-sensitive gene *MECP2* at Xq28 is particularly susceptible to such genomic instability due to nearby inverted and direct orientation low-copy repeats (LCRs)[32–34]. The protein encoded by the *MECP2* gene, Methyl-CpG binding protein 2 (MeCP2), is critical for brain function by acting as an epigenetic regulator[35]. Copy-number variation spanning the gene causes *MECP2* Duplication Syndrome (MDS) (MIM:300260) in males with 100% penetrance[36]. The most prevalent clinical features of MDS are infantile hypotonia, developmental delay, intellectual disability, frequent respiratory infections, and refractory epilepsy[37]. One of the frequent complex allele presentations is an allele constituted by an inverted triplication flanked by duplications (DUP-TRP/INV-DUP). This aberration is generated by a given pair of inverted low-copy repeats telomeric to *MECP2*, being responsible for 20% to 30% of the MDS cases[6]. When generated, this structure includes two breakpoint junctions (Jct) connecting the end of the duplication to the end of the triplication (Jct1) and the beginning of the triplication to the beginning of the duplication (Jct2). Given the nature of this event, we lack the ability not only to detect this, but also how to describe this in a standardized VCF format. Part of the complexity originates as the reads themselves only partially indicate the allele, e.g., highlighting a shorter inversion[27].

In addition to complex variants, multiple studies have shown that there are mosaic or low-frequency SVs that are likely causal across neurological diseases or other diseases[9]. As an example, single-cell studies show us that there can be variable CNVs across multiple cells in the brain[9]. However, their true frequency is unknown, with around 12% of healthy cortical neurons having Mb-scale CNVs[38]. A possible role in neurodegenerative disease[39] has not been adequately explored. In synucleinopathies, which include Parkinson's disease and Multiple System Atrophy[40] (MSA), somatic CNVs of the highly relevant *SNCA* gene have been reported [41,42], and scWGS in MSA has shown Mb-scale CNVs in ~30% of cells[42]. Still, these CNVs studies lack resolution as breakpoints are defined within +/- multiple kbp and only very large ~1Mbp+ CNV events are reported[38,43,44]. An identification of complex SVs arising in neurodevelopment was so far only possible with WGS of clonally-expanded precursors [9,42]. Thus, so far, we struggle to identify the underlying alleles even for large already reported CNVs along the human genome.

In this paper, we present Sniffles2, which we extended not only for germline SVs but to further solve the problem of population-scale SV calling for long-reads. In addition, Sniffles2 now enables the detection of low-frequency SVs across data sets, which again opens the field of cell heterogeneity for long-read applications. Sniffles2 is a redesign of the popular SV caller Sniffles and is thus more accurate and faster than the previous implementation. We first highlight Sniffles2 performance over multiple benchmark sets. We further investigate how the new population or family mode for SV calling improves the accuracy and performance across Mendelian disease probands with ONT. Here we can showcase the boundaries of long-read SV calling by assessing highly complex SVs around *MECP2.* Lastly, we investigate Sniffles2 abilities to identify low-frequency/mosaic SV across an MSA brain sample and compare its performance to Illumina sequencing and Bionano optical genome mapping. Overall, Sniffles2 pushes the boundaries of long-read based SV calling and thus demonstrates the utility of such an approach further than any existing approach. Sniffles2 remains an open source (MIT license) and is available at: https://github.com/fritzsedlazeck/Sniffles

Results

Accurate detection of complex structural variations at scale

Sniffles2 is a complete redesign and extension of the popular SV caller Sniffles. **Figure 1** gives an overview of its main components. Sniffles2 improves germline SV calling (**Figure 1A**) but further enables family and population SV calling at scale and ease (**Figure 1B**) and implements novel methods to identify mosaic SVs (**Figure 1C**). A detailed description of Sniffles2 can be found in the methods section.

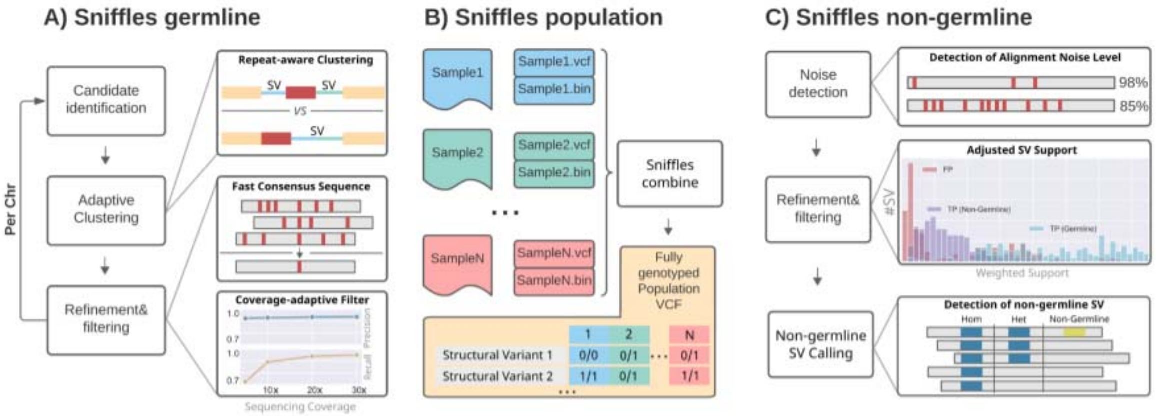

Figure 1: *Overview of Sniffles2* ***A)*** *The germline SV detection has been improved through inclusion of repeat aware clustering. In addition, the fast consensus sequence and coverage-adaptive filtering result in higher accuracy of the SV calls.* ***B)*** *One key limitation of current SV calling is the generation of fully genotyped population VCF. Sniffles2 implements a concept similar to a gVCF file where single sample calling is only done once and thus improves accuracy and reduces runtime multiple-fold.* ***C)*** *Non-germline SV detection is enabled by improved detection and filtering of low variant allele frequency SV across a bulk sample. This is enabled over additional noise detection methodology as well as refinement and filtering approaches that we developed.*

Figure 1A shows a summary of the most important steps applied by Sniffles2 to identify germline SVs. In brief, we use a fast yet high-resolution clustering approach, which identifies SVs in three key steps. First, putative SV events are extracted from read alignments (split reads and inline insertion or deletion events) and allocated to high-resolution bins (default: 100bp) based on their genomic coordinates and putative SV type. Second, neighboring SV candidate bins are subsequently merged based on a standard deviation measure of SV starting positions within each growing bin. Using optional tandem repeat annotations, Sniffles2 dynamically adapts clustering parameters during SV calling, allowing it to detect single SVs that have been scattered because of alignment artifacts. Finally, identified clusters are separately reanalyzed and split based on putative SV length. Final SV candidates are subjected to quality control based on read support, breakpoint variance and expected coverage changes.

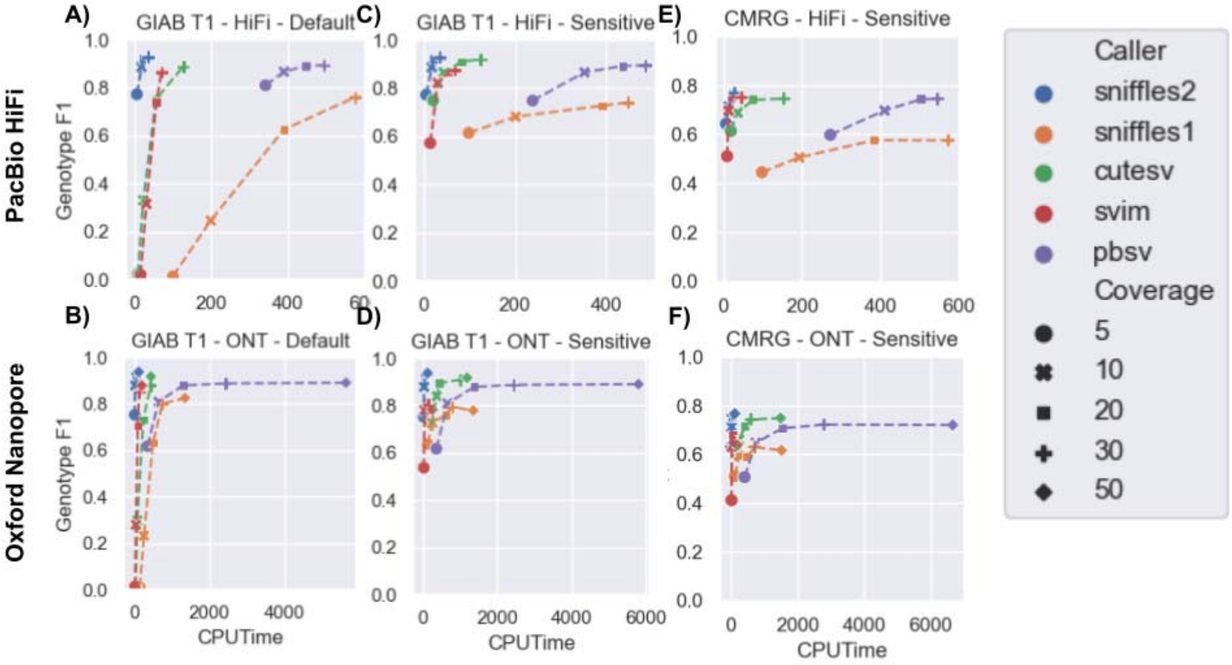

***Figure 2**: Performance assessment of Sniffles2. Performance metrics for correctly identifying and genotyping SVs across PacBio HiFi (upper column) and Oxford Nanopore (lower column) (see **Supplementary Table 1** for details). **A+B:** comparison across Tier1 GIAB genome-wide SV (y-axis) across different coverages (symbols) and SV caller (colors) with respect to the CPU time (x-axis). Across the programs the suggested/default parameters were used. **C,D:** Similar comparison as A+B across Tier1 GIAB SV call set, but this time with maximum sensitivity of different SV callers. E+F: GIAB challenging medical gene benchmark for SV*.

We assessed the performance of Sniffles2 with respect to Sniffles[27] (v1.12), cuteSV[45] (v1.0.11), PBSV[46] (v2.6.2) and SVIM[47] (v1.4.2) using Truvari[48] and the GIAB recommended parameters[49]. **Figure2** shows the results across different GIAB benchmarks. Across the default coverage (30x for HiFi, 50x for ONT), Sniffles2 shows the best performance with respect to correctly identified and genotyped insertions (HiFi: F-score 0.918, ONT: F-score 0.929) and deletions (HiFi: F-score 0.942, ONT: F-score 0.948) (see **Supplementary Tables 1 & 2** for details). Sniffles2 achieves a better result in a fraction of the time across data sets compared to Sniffles (v1.12), being over 16 times (HiFi) and 11 times (ONT) faster in processing a 30x coverage data set, respectively. **Figure 2A+B** shows the results for PacBio HiFi and ONT, respectively. In addition, Sniffles2 is also the fastest method overall, requiring 34.16 CPU minutes for processing a 30x coverage HiFi dataset (real time using 8 processor cores: 14.37 minutes), which was twice as fast as SVIM, the 2nd fastest method. For a 30x coverage ONT dataset, Sniffles2 was also close to twice as fast (1.98x) as the second fastest caller (SVIM), while also having an over 9.6% higher F-score. Considering Sniffles2 multi-processing capability (not supported by SVIM), the speedup increases even further, to more than 5.4-fold and 7.5-fold for HiFi 30x, ONT 30x data sets, respectively. When reducing the coverage from 30x to 10x we observe only a slight reduction in F-score for Sniffles2 (HiFi: reduction F-score 0.042, ONT: reduction F-score 0.054). This is in stark contrast to other programs such as cuteSV, where using default parameters, F-score dropped by an average of over 60% (HiFi: reduction F-score 0.56, ONT: reduction F-score 0.58). Even when using only 5x for Sniffles2, we still observe a high accuracy ONT (F-score: 0.75) and HiFi (F-score: 0.77). This is achieved as Sniffles2 includes an automated parameter selection for filtering of SV candidates based on the available coverage. In contrast, other SV callers rely on manual adjustment of these parameters to retrieve acceptable results across coverages and sequencing technologies. **Figure 2A+B** shows this clearly as all other SV callers show a decreased performance across lower coverage. Even when tuning the parameters for other SV callers (**Figure 2C+D**), Sniffles2 remains the highest accuracy (see **Supplementary Table 1** for details). **Supplementary Table 3** also shows the evaluation with respect to Tier2, a more challenging region of the GIAB benchmark set. Again, Sniffles2 even increases the performance difference compared to other SV caller. Lastly, we benchmarked Sniffles2 across a more challenging SV data set across 386 medically relevant, but highly polymorphic/challenging genes[31]. GIAB has recently released this call set of ~200 SV covering around 70% of these genes[31]. **Figure 2E+F** shows the results. Again, Sniffles2 outperforms the other SV callers in terms of accuracy and speed using default parameters. The next best performing SV caller (pbsv for HiFi, cuteSV for ONT) both achieved 2.1% and 1.8% lower genotyping accuracy even at 30x coverage. **Supplementary Table 4** contains the detailed results across all SV callers. Overall, Sniffles2 outperforms other state-of-the-art SV caller across the entire genome including the most challenging regions/genes. Sniffles2 improves insertion identification through two additional methods: First, the consensus module corrects sequencing-related errors in the recovered insertion sequences using a fast pseudo-alignment-based approach. This allows Sniffles2 to attain the second highest mean sequence identity of (HiFi: 0.948, ONT: 0.939), after pbsv (HiFi: 0.953, ONT: 0.949), while Sniffles2 is over 14x (HiFi) and 36x (ONT) faster (see **Supplementary Figure 1, Supplementary Table 5** containing insertion sequence accuracy across all callers) at 30x coverage. Second, Sniffles2 increases the sensitivity for the detection of large insertions by recording additional supporting alignment signals in the affected regions (see **Supplementary Figure 2, Supplementary Table 6**) at much higher speed than pbsv, the only SV caller with a comparable accuracy for long insertions.

Lastly, GIAB only represents one individual benchmarked across most studies (HG002). Thus next, we used Dipcall[50] together with three T2T assemblies (HG01243, HG02055, HG02080) to further assess the performance of Sniffles2. Clearly, we give Dipcall the benefit of the doubt, knowing that the accuracy will be lower than the GIAB vetted benchmark set. Overall, Sniffles2 performs the best across all samples having on average a F-Score of 0.80 at 30x coverage ONT and HiFi compared to 0.77 F-score for the next best SVcaller (cuteSV), at nearly 3.5 times the speed (73.72 CPU minutes versus 256.65 CPU minutes on average) for default parameters. **Supplementary Table 7** contains the detailed results for each benchmarked program across these three samples. Besides the here benchmarked insertions and deletions, we also benchmarked Sniffles2 on duplications, inversions and translocations using simulated data as no benchmark exists. Overall, Sniffles2 again outperformed all other methods in speed and accuracy (see **Supplementary Figure 3 and Supplementary Table 8**) (see methods for details).

Given all these comparisons across different ethnicities, coverage levels and sequencing technologies, we conclude that Sniffles2 improves the detection of SVs in terms of accuracy and speed compared to other state-of-the-art methods.

Enabling family to large cohort studies to discover the impact of complex Structural Variation

Over the past years, an uptake of ever larger studies utilizing long-reads is foreshadowing a trend in genomics to utilize long-reads more often than ever[21]. To promote this, Sniffles2 is fast and efficient, but further implements a strategy to obtain a fully genotyped population VCF. Traditionally this is a multi-stage process of calling, merging, genotyping, and re-merging[21,51,52]. This is clearly inefficient as the bam/cram alignment files need to be assessed twice. Even so, this process can only be achieved by using a few of the existing methods (SVJedi[53], Sniffles[27], CuteSV[45]). Sniffles2 strategy only requires an initial calling and merging to obtain a fully genotyped population-level VCF. **Figure 1B** illustrates the principle. The calling can be done independently per sample and thus allows to scale to large data sets. Each sample run produces a single germline VCF file accompanied with a binary file that serializes every single candidate SV down to a single read support. Next, both files per sample are provided as a list to Sniffles2 merge, which combines the SV across the samples and fills the missing information utilizing the binary files per sample. These files typically range from 75 to 250Mb per ~30x ONT sample. This process is extremely efficient as it scales linearly with the number of samples and allows the samples to be analyzed in parallel and independent of each other (see **Supplementary Table 9, Supplementary Figure 4**). In addition, it solves the “n+1” problem to include a batch of samples at a later stage of a project.

To assess the performance of this approach, we measured the Mendelian inconsistency rate (see methods)[54]. Here we counted in how many of the called SV, the genotypes of the proband do not concord with the genotype of the parents (e.g., F 0/0, M 0/0, P 1/1) or the other way around (e.g., F 1/1, M 0/0, P 0/0). For Sniffles2, we obtained a low Mendelian inconsistency rate of 5.65% with a 0.31% missing genotype rate (**Figure 3A**). The latter is driven by a user parameter where Sniffles2 does not report a genotype where only 5 or fewer reads are present. In comparison, cuteSV with a simple merge (SURVIVOR[55]) presented a mendelian inconsistency of 3.07% with a much higher missingness of 33.60% of all genotypes compared to the 0.31% of Sniffles2. When we apply a re-genotyping and re-merging of the cuteSV results, we obtain a Mendelian inconsistency rate of 6.02% with a missingness 1.97%, both higher than Sniffles2. Furthermore, the cuteSV approach took more than 43 hours CPU hours (43:10:05, **Supplementary Table 10**) in contrast to 77.28 CPU seconds for Sniffles2 for a given trio. Thus, rendering it impractical for larger cohorts. Note for cuteSV this is even ~2.8 times slower on a single trio than Sniffles2 producing a fully genotyped VCF file across 768 samples.

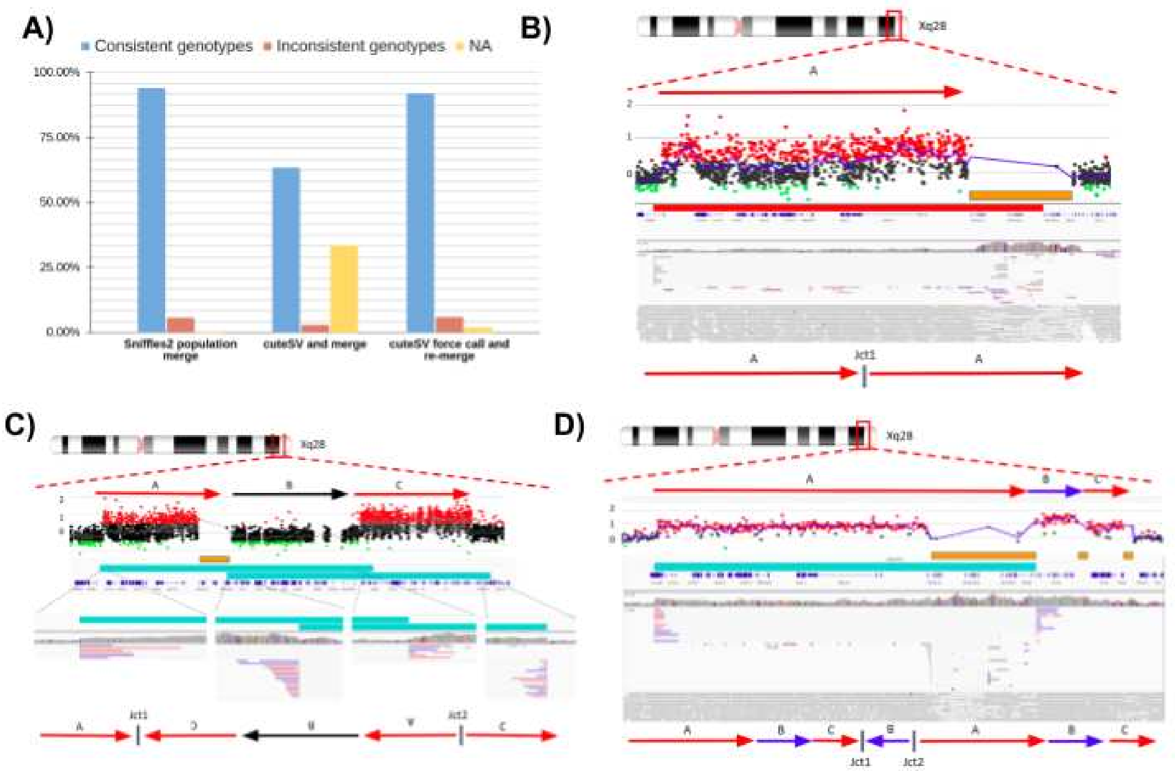

***Figure 3**: Sniffles2 population approach and application to Mendelian disease. **A** Comparison of the proportion of consistent, inconsistent, and missing (./.) genotypes across HG002/3/4 computed by the bcftools Mendelian plugin for Sniffles2 population merge and cuteSV. To achieve similar results (right side), cuteSV requires more than 2,000x the time. **B)** Tandem duplication that was fully resolved by Sniffles2 in one of the patients (BH14233_1). Sniffles2 was able to identify and map the junction of the duplication within a segmental duplication region where array data does not provide information. For the array, dots represent genomic positions being assayed. Black dots represent a log ratio between -0.35 and 0.35, red dots represent a log ratio above 0.35 and green dots represent a ratio below -0.35. Consistent (at least 3 consecutive probes) log ratios above 0.35 represent a region of copy number gain and below -0.35 represent copy number loss. Orange bars indicate segmental duplication (SegDups) where no probe can be designed for the array. **C)** Detailed aCGH view of a complex duplication-normal-duplication (DUP-NML-DUP) with breakpoints within* SegDup *(SegDups) or low-copy repeats (LCRs) region (orange bar) where Sniffles2 is indicating two overlapping inversions in IGV (teal bars) forming junctions 1 and 2 (Jct1 and Jct2). Top arrows indicate the reference orientation (duplications in red, neutral regions in black) of each genomic fragment and bottom arrows indicate the resolved rearrangement structure (DUP-NML-INV/DUP) including Jct1 and Jct2. **D)** Complex duplication-triplication-duplication structure as highlighted in aCGH data with* SegDups *and LCRs highlighted (orange bars). Sniffles2 identifies the inversion breakpoint at Jct2 (teal bar) but cannot fully resolve the entire allele including Jct1 as it's also not possible to be reported in the VCF standard. Red arrows indicate duplicated regions and blue arrows show triplicated portions. The context of the resolved structure is a DUP-TRP/INV-DUP where the triplication is inverted forming Jct1 and Jct2*[33].

ID	Sex	Inheritance	Pathogenic CGR	Coordinates CGR (aCGH)						Coordinates (Sniffles2)			
				CNV	Chr	Start	SegDUP	End	SegDUP	SV	Chr	Start	End
BH14233_1	M	Maternal	Tandem Duplication	DUP	X	**153084841**	-	**153414342**	Yes	DUP	X	**153084620**	**153483866**
BH13948_1	M	Maternal	Tandem Duplication	DUP	X	**152877325**	-	**153414342**	Yes	DUP	X	**152808716**	**153487348**
BH15642_1	F	*de novo*	Tandem Duplication	DUP	X	**153289589**	-	**153399165**	-	DUP	X	**153289208**	**153386550**
BH13947_1	M	Maternal	DUP-NML-INV/DU	DUP1	X	**153106533**	-	**153414342**	Yes	INV	X	**153106249**	**153937616**
				DUP2	X	**153938964**	-	**154293950**	-	INV	X	**153492860**	**154294604**
BH15700_1	M	Maternal	DUP-TRP/ INV-DUP	DUP1	X	**153131406**	-	153409337	-	INV	X	**153131086**	153520844
				TRP	X	**153523170**	Yes	153565901	Yes				
				DUP2	X	153575989	-	153623000	Yes				
BH15701_1	M	Maternal	DUP-TRP/ INV-DUP	DUP1	X	**153189181**	-	153420198	Yes	INV	X	**153188685**	153499734
				TRP	X	**153505485**	Yes	153565901	Yes				
				DUP2	X	153575989	-	153623000	Yes				
BH15646_1	M	Maternal	Terminal DUP/ Recombinant	DUP	X	**147326287**	-	Telomere	-	INV	X	**1406919**	**147326058**
				DEL	X	Telomere	-	**1405994**	-				
BH15692_1	M	*de novo*	Terminal DUP/ Translocation Y	DUP	X	**151905254**	Yes	Telomere	-	BND	X	**151904176**	**N]Y:23243741]**
				DUP	Y	**23243948**	-	23655166	Yes				
				DEL	Y	24095954	Yes	Telomere	-				
BH15696_1	M	*de novo*	Terminal DUP/ Translocation Y	DUP1	X	**148351663**	-	148384182	-	BND	X	**148351433**	**]Y:28389311]N**
				DUP2	X	**148706667**	-	Telomere		BND	X	**148384577**	[Y:25210061[N
				DEL	Y	**28458870**	Yes	Telomere		BND	X	**148705972**	N]Y:25654822]
BH14229_1	M	Maternal	Terminal DUP/Unknown structure	DUP	X	**151893933**	Yes	Telomere	-	INV	X	**151919987**	155251615
BH13949_1	M	Maternal	Terminal DUP/Unknown structure	DUP1	X	**144057799**	-	144066387	-	DUP	X	**144056099**	150063756
				TRP	X	144067901	-	144101282	-	INV	X	144069177	150064223
				DUP2	X	144101282	-	Telomere	-	BND	X	144086954	[10:42527789[N

Table 1: *Table across all the probands assessed here and highlighting in bold which junctions could be resolved using Sniffles2. Highlighted in green are the results discussed in the main text.*

Next, we applied this population/family approach of Sniffles2 across 31 Oxford Nanopore data sets that represented cases of Mendelian disorders in probands. The CPU runtime for merging the individual samples was roughly 28 minutes (28:27) to produce a fully genotyped population VCF file. Across the seven complete families (proband, mother, father) we measured an average of 3.65% Mendelian inconsistency rate and 0.90% of missingness (see methods, **Supplementary Table 11, Supplementary Figure 5**). The probands for sequencing were selected based on a Mendelian disease that often is caused by SVs impacting the *MECP2* gene at Xq28 locus. As described in the introduction this is a severe neurodevelopmental disorder that is often caused by extreme complex alleles in this region. We were interested if Sniffles2 together with ONT data can resolve the breakpoints which were not always solvable using array data and if we were able to fully explain the entire allele or just partially solve the junctions. To address this, we filtered SV based on ChrX together with their size (10kbp) and filtered for SV only being *de novo* or inherited from the mother.

Within this cohort, Sniffles2 achieves a high rate of detection across junctions, but sometimes struggles to recapitulate the entire allele that contains complex SVs. **Table 1** shows the details per proband. In samples harboring a tandem duplication, Sniffles2 was able to properly detect the aberration and fully resolve its architecture. In our cohort, these duplications span the dosage sensitive gene (*MECP2*) and form a single breakpoint junction (Jct1), confirming a tandem duplication structure. As highlighted in sample BH14233_1, although aCGH broadly defines the genomic interval of the duplicated region, Sniffles2 can properly give positional context of genomic fragment defining at nucleotide-level resolution to be a tandem duplication on the allele even though the end of the duplication is within a segmental duplication region (orange bar) (**Figure 3B**).

Interestingly, a portion of the inversions that Sniffles2 was able to detect were not simple genomic inversions, but instead part of more complex structures that could not be fully resolved using current bioinformatic tools. A more complex allele is detected in sample BH13947_1, which consists of a duplication-normal-duplication (DUP-NML-INV/DUP) with breakpoints spanning segmental duplications (SegDups) (**Figure 3C**). Here Sniffles2 indicates two overlapping inversions which form junctions 1 and 2 (Jct1 and Jct2) generating a DUP-NML-DUP/INV structure. In sample BH15646_1, the inversion called by Sniffles2 spanning nearly the entire X chromosome (~148 Mb) represents the breakpoint junction of a recombinant chromosome. In the sample, aCGH data shows a short-arm deletion and a long-arm duplication, i.e. DEL-NML-DUP structure. Sniffles2 is able to positionally connect the beginning of the duplication to the end of the deletion forming Jct1 (**Supplementary Figure 7**). This aberration is generated *de novo* as the result of a maternal meiotic recombination between heterozygous homologous X-chromosomes for a pericentric inversion[56].

Another example is represented by an apparent 311kb inversion detected in sample BH15700_1. This inversion is part of a DUP-TRP/INV-DUP structure (**Figure 3D**), which is generated by a given pair of inverted SegDups and produces an inverted triplication flanked by duplications[33]. When generated, this structure includes two breakpoint junctions (Jct) connecting the end of the duplication to the end of the triplication (Jct1) and the beginning of the triplication to the beginning of the duplication (Jct2). While Sniffles2 can properly detect the inverted breakpoint generating Jct2, it is not able to fully resolve the context of the larger structure due to Jct1 being embedded within a pair of inverted SegDups with 99.9% sequence similarity.

In this cohort, Sniffles2 can correctly detect with nucleotide-level resolution the precise breakpoints defining a genomic interval in patients carrying complex genomic rearrangements (CGRs). Importantly, a large portion of the CGRs in this cohort have at least one of the breakpoint junctions mapping to SegDups; those can be fully resolved by Sniffles2 together with copy-number information. Nevertheless, CGRs that have two breaks mapping to pairs of highly identical SegDups, such as in the DUP-TRP/INV-DUP events, still represent an important challenge for SV callers and are also too complex to appropriately report within the limits of a VCF standard. Additionally, Sniffles2 infers positional connections that help resolve a given complex allele architecture with information that aCGH alone cannot provide. Thus, overall this highlights the benefit of Sniffles2 family/population mode to enable these types of comparisons.

Identification of mosaic SVs reveals new insight in diversity

We have shown that Sniffles2 accurately identifies SVs across the entire genome and that it enables better scaling and accuracy across families and even larger population levels. Nevertheless, as we know from many studies, germline variants are not the only source of structural variation. Often somatic/mosaic variants are important. This has been indicated in e.g. cancer and neurological disorders [9,12]. Thus, Sniffles2 is equipped with a non-germline mode to identify mosaic and low-frequency SVs across a single sequenced sample. **Figure 1C** shows the principal steps where the main innovation is to weigh the support of each read taking into consideration its edit distance as a confidence measure. To circumvent the impact of sequencing error rates on mosaic SV detect we filter out SV where the average edit distance of reads supporting exceeds a threshold, which is estimated per data set to account for different sequencing error levels (see methods).

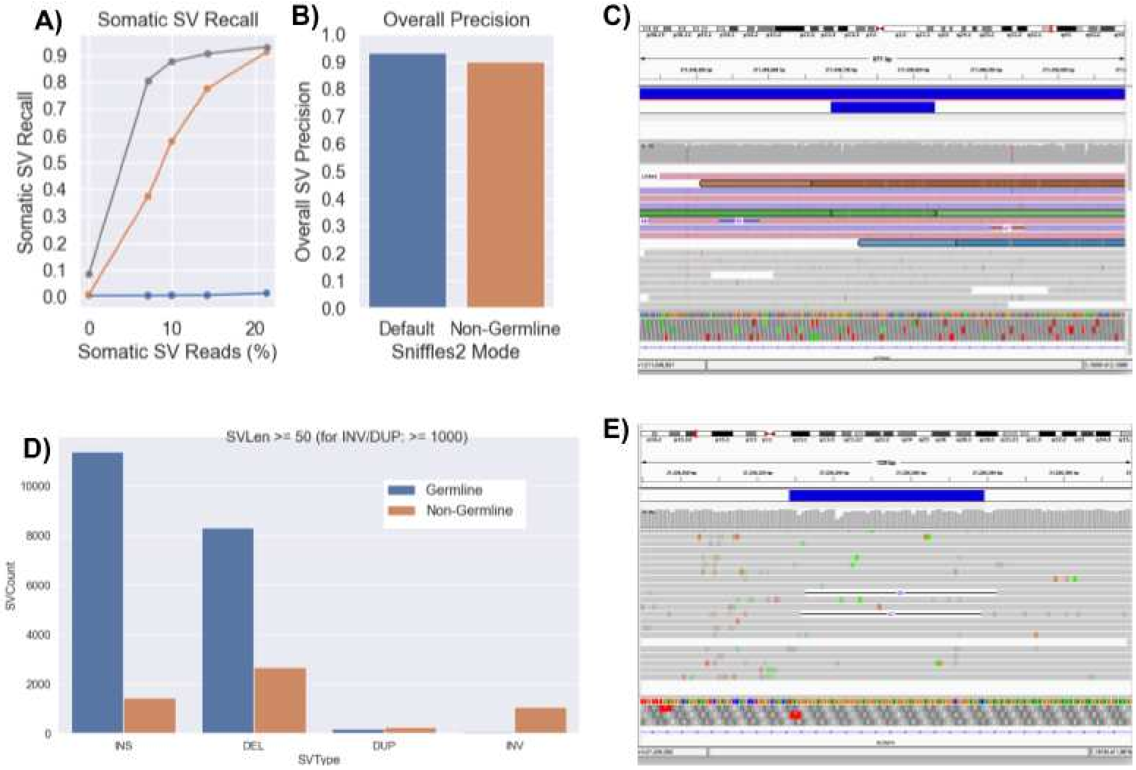

***Figure 4**: Recovery of somatic SVs using the Sniffles2 non-germline mode **A+B)** benchmark of mixtures of HG002 with HG004. We spiked HG002 in various concentrations and measured the recall (**A**) and precision (**B**) of Sniffles2 at different variant allele frequencies. **C)** Example of a false positive inversion (small blue bar on top) caused by likely chimeric reads. The three supporting INV reads are highlighted in brown, green, blue. Other likely chimeric reads are shown in light blue and red. **D)** Overview of SV types identified as germline (blue) and non-germline events in the MSA patient brain sample. **E)** Example of mosaic repeat contraction detected by Sniffles2.*

To assess the performance of Sniffles2 across mosaic SVs, we first synthetically merged HG002 (at low concentrations: 5x, 7x, 10x or 15x) with the 55-65x coverage read data from the sample of the mother (HG004). This yielded multiple, synthetic samples with constant total coverage of ~70x, but varying concentrations of HG002 in them. We only assessed this for ONT data as this technology often is sequenced at higher sequence depth than PacBio Hifi. The latter can still be used but is not benchmarked here. **Figure 4 A&B** highlights the results across this synthetic data set across different concentrations of HG002 (x-axis, **Supplementary Table 12**). **Figure 4A** shows the recall of SVs across the different concentrations. In blue, we highlight the performance of Sniffles2 germline-mode, which shows a 0% recall as expected for the somatic mutations. Orange highlights the performance of non-germline mode for Sniffles2. Sniffles2 achieves high recalls at around 10% (57.89% recall), 15% (77.45% recall) and 20% (91.18%) variant allele frequency (VAF) across the reads, while maintaining a constant high average precision (~88%) (**Figure 4B**). Nevertheless, we observe a reduced recall at 7% VAF (ie. ~7% of reads supporting an SV) of 33% recall. This is due to the lack of supporting reads as at 7% VAF there is only on average 1-3 reads to indicate a SV. The gray curve (**Figure 4A**) shows the force calling/genotyping performance of Sniffles2, another mode that was implemented. Here, Sniffles2 knows the individual SV (via a provided VCF file) and identifies a SV if only a single read is supporting the allele (i.e., genotyping). Thus, with only 70x coverage we can identify somatic SV down to around 7% VAF.

Next, we applied Sniffles2 non-germline mode to an affected brain region (cingulate cortex) of an MSA patient at 55x coverage using ONT. Here we are interested in all types of SVs, including rearrangements (INV & DUP). In this case, however, we need to be alert for the possibility that chimeras can form inversions or other duplications and as such contribute to the overall apparent somatic SV calls. In this sample, the initial call set included 9,290 inversion calls, most of them below 1 kbp (89.26%). On visual inspection it was clear that several SV are supported by two or more reads which is also consistent with an overall ~2% chimeric read frequency for this ONT data set. **Figure 4C** shows an example of an inversion identified likely because of chimeric reads. Therefore, we filtered INV and DUP that are smaller than 1 kbp to avoid chimera-based SV calls. **Figure 4D** shows the overall number of SV and their type after this filtering for germline and non-germline SV. After applying these filtering steps, the final set of non-germline SVs included 1,465 insertion, 2,667 deletion, 267 duplication and 1,061 inversion events. In **Figure 4E**, an example for one of the mosaic deletions that was identified by Sniffles2 is shown that appears to impact a short tandem repeat directly. The deletion overlaps the neuronal *KNCIP4* gene, which encodes for an interactor of voltage-gated potassium channels[57]. This serves as an example that non-germline SVs identified by Sniffles2 can even include potential direct disruptions of genes.

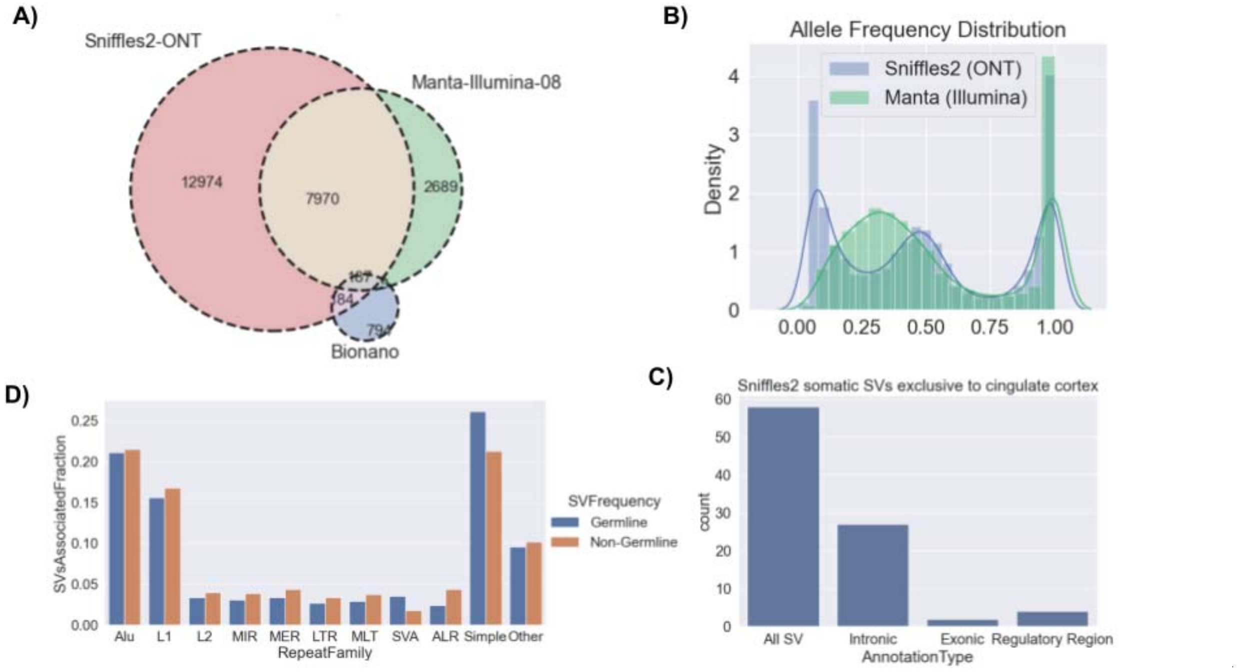

***Figure 5:** Insights into somatic SVs in the MSA patient brain sample. **A)** Overall comparison of SVs detected in ONT (Sniffles2), Illumina (Manta) and Bionano data sets. **B)** Distribution of*

variant allele frequencies (ie. the percent of read support) for SVs identified by Sniffles2 and Manta. ***C)*** *Association of Sniffles2 germline and non-germline SVs with repeat elements.* ***D)*** *Mosaic SVs detected in the cingulate cortex only and their overlap with gene structure.*

Next, we compared the different technologies to the Sniffles2 results. The same brain region was also sequenced by Illumina short reads (90x) and analyzed by optical genome mapping (Bionano 690x) (see methods). **Figure 5A** highlights the agreement of all SVs detected in the same sample by the three technologies for a minimum length of 50bp and excluding translocations. Overall, 38.27% of all 21,315 SVs (germline and non-germline) detected by Sniffles2 were also identified by either Manta[58] (Illumina), Bionano, or both. At the same time, Bionano showed a significantly higher overlap with Sniffles2 (31.60% of its calls) than with Illumina (16.70% of its calls). The overlap between Sniffles2 and the other technologies was also much higher for deletions (53.10%) than for insertions (31.44%). These differences are likely explained by the individual difficulty in detecting larger (i.e., Illumina) and smaller (i.e., Bionano, threshold approximately 1kbp) SV, respectively. Next, we took a closer look at putatively somatic SVs detected by Sniffles2. For this, we separated Sniffles2 detected SVs by their reported variant allele frequencies (VAF) into germline (VAF>0.3) and likely non-germline (VAF<0.3) calls. Only 13.14% of 4,537 non-germline SVs reported by Sniffles2 could be detected by either Illumina or Bionano, highlighting the difficulty in identifying rare SVs. For deletions, only 384 (18.43%) of SVs were also found by either Bionano and/or Illumina. Only 142 (11.87%) non-germline insertions reported by Sniffles2 were detected by the other methods. For duplications and inversions, only 28 (10.81%) and 42 (4.21%) of Sniffles2 non-germline SVs were identified in the Illumina or Bionano data sets, respectively. For somatic calls (VAF<0.3) that were only supported by Sniffles2, we selected 23 SV that are larger than 1kbp and manually genotyped them using Bionano high coverage data. Out of these four were directly confirmed and five SV were overlapping but had incongruent sizes between the two calls. For five SV we did not obtain enough evidence in the Bionano data sets to reject as there was a lack of Bionano label. Thus only 39.13% of SV from Sniffles2 only calls could not be confirmed further after manual inspection of Bionano, which most of them were smaller SV calls below 3kbp.

We further noted a shift in the allele frequencies across the Manta calls compared to the Sniffles2 calls **(Figure 5B)**. As expected, for Sniffles2, we observe a multimodal distribution with three peaks, representing homozygous, heterozygous, and non-germline SVs, respectively. In contrast, Manta shows two main peaks in their allele frequency distribution. A homozygous (~0.9-1 AF) and a broad peak around 0.3 AF, which would be below the typical expected heterozygous AF peak. For Sniffles2, we furthermore observe ~ 75% increase in the area under the curve in the putative non-germline range of allele frequencies (0.0-0.5) when compared to the area under the curve between the heterozygous and homozygous allele frequencies (0.5-0.1).

Next, we focused on the non-germline SVs exclusive to the cingulate cortex brain region. For this, we also sequenced the neighboring cingulate white matter from the same patient using Illumina. We used SVTyper[59] to genotype Sniffles2 SVs (only deletions, duplications, and inversions) that were not initially identified by Manta against the aligned Illumina reads from both brain regions. This way we identified 389 SV that initially were not identified in Illumina but were genotyped as present in either cingulate cortex or cingulate white matter. Of these, we identified 58 (14.91%) Sniffles2 non-germline SVs that had read support unique to the cingulate cortex region. 29 of these SV overlapped with genes including *GRIN2A*, which encodes for *NMDA* receptor subunit 2a. Of these 29 SVs, 27 overlapped introns and 2 affected at least one exon. Furthermore, 4 SVs disturb regulatory regions associated with at least one gene, including *PEX26*, *DLL1* and *ABCA2*. This further highlights Sniffles2's ability to not only detect SVs that cannot be identified using Illumina data alone but also the unique presence of a subset of these calls within a brain region. Their distribution across functional areas of the genome in a brain affected by multiple system atrophy merits further study.

A significant fraction of germline SVs is known to be associated with genomic repeats such as *Alu* and L1 elements. To understand the differences between germline and non-germline SVs in this regard, we separately analyzed the association of Sniffles2 germline and non-germline SV calls with different repeat families (**Figure 5D**). In summary, we found that a higher fraction of overall non-germline SVs was associated with repeat elements than for germline SVs across most repeat families. Interestingly, the patterns of repeat association significantly differed between the individual SV types (**Supplementary Figure 7**). For duplications, the fraction of non-germline SVs associated with LINE1 and simple repeats showed the highest difference over germline SVs. For inversions, close to 40% of non-germline SVs were associated with Alu-elements, in comparison to less than 20% of germline SVs. The fraction of non-germline inversions associated with other repeat elements, including LINE2, MIR, MER, LTR and simple repeats was increased as well, however in contrast to duplications not LINE1 elements. This highlights the different ways in which repeat elements can cause new somatic structural variants to form. **Supplementary Figure 8** shows read alignments for a subset of non-germline duplications and inversions that were only called by Sniffles2 and subsequently manually curated, together with their relation to nearby repeat elements. For deletions and insertions, we observe slightly lower fractions of non-germline SVs associated with most repeat types, except for centromeric satellite (ALR/alpha) repeats. For this repeat family, especially non-germline insertion SVs had a higher fraction associated. Repetitive elements may be associated with neurodegenerative disorders, through increased expression and / or de novo somatic genomic integration[39]. The observance of a higher fraction of non-germline insertion and deletion SVs being associated with these repeats could suggest a further correlation for this on the level of an individual brain region. Overall, this also highlights the differences between repeat families in their effects on somatic SV generation[60].

Discussion

In this paper, we present a new version of the highly popular SV caller: Sniffles. Sniffles2 is a significant improvement in terms of accuracy and runtime not only compared to Sniffles v1, but also to all other commonly used long-read based SV callers (see **Figure 2**). We show higher accuracy across different coverages (5-50x) using different sequencing technologies (PacBio HiFi and ONT) and even across all SV types. This is achieved by an automatic parameter optimization that is part of Sniffles2 compared to all other SV caller that require manual adjustments. Besides this Sniffles2 is also able to genotype SV and leverage phased reads (using HP and PS tags) as input to provide phased SV in a VCF file. Nevertheless, Sniffles2 is more than a simple extension (**Figure 1**). For the first time, we demonstrate a gVCF concept for SV calling and implemented a working version in Sniffles2. This instantaneously halves the requirements of computing and storage for population/family SV or even tumor vs. normal SV calling. Thus, resolving the ever-larger demands of long-read data sets[21]. Furthermore, it solves the n+1 problems when a new sample is added on a later stage of the project. We demonstrated the new merge concept across the GIAB family where Sniffles2 produced a fully genotyped VCF file within minutes resulting in a lower Mendelian error rate. We could demonstrate the utility across the 31 ONT Mendelian samples, where Sniffles2 resolved SVs mapping to complex regions of the genome with a direct impact to disease, supporting copy-number data. This clearly illustrates the benefit of this novel approach that can easily scale to new population long-read challenges.

Another novelty of Sniffles2 is the non-germline mode that enables the detection of low-frequency/mosaic SVs with a standard sequencing run, while maintaining a high precision. We have demonstrated this novel approach using a synthetic data set of HG002 and his mother HG004. This showed the accuracy and recall of Sniffles2 while depending on only 2-3 reads overall to distinguish SV from noise. Here we identified that chimeric reads lead to wrongly called Inversions and in rare cases Duplications (**Figure 4**). We then turned our attention to MSA, a rare sporadic neurodegenerative disease related to Parkinson's, with negligible heritability (<7%)[61]. We performed ONT WGS on an affected brain region from one patient, where Sniffles2 was able to identify somatic SV and showcased great performance compared to Illumina and optical genome mapping approaches, thus overall highlighting the fact that Sniffles2 is highly versatile and accurate. While thresholding on the variant allele frequencies (here 0.3 AF and below) for the identification of potential somatic variants is straightforward, it is of course not generalizable. For multiple SV we saw a continuum in VAF which suggests that some SV with apparent AF <0.3 may also be germline (see **Figure 5B**). Thus, the comparison to population data or to different tissues is still favorable (e.g., tumor vs. normal). The possible role of somatic SV's in MSA is under investigation[42], although further validation data from more cases and controls would be required to allow interpretation of the present findings.

Despite all the novelties and solving central problems of SV calling at scale and accuracy for long-reads many challenges remain. We still lack high quality benchmarks for non-insertion and deletion calls including complex rearrangements. Achieving this will boost the field of SV detection further and promote novel methodologies. In this study, we could only evaluate other SV types (inversions, duplications, and translocations) via simulated data using an established pipeline. **Supplementary Figure 3** summarizes these results. While this is unsatisfactory, it remains the only way as benchmark sets focus on insertion and deletions only. In addition, Sniffles2 does not yet solve the issues with highly rearranged regions where SVs can be overlapping to each other. This remains a near future goal of Sniffles2 and will also require improved benchmark sets and even standards to report these events as the VCF standard does not provide a clear recommendation. Nevertheless, this is clearly needed as our experiments on the Mendelian cohort shows. Another highly important factor that Sniffles2 improves is tumor normal comparisons. Here both new modes, the somatic to detect low variant frequency SV which are common in cancer and the merging mode to compare tumor vs. normal, will improve the detection of cancer driving mutations. We have currently not included these in the manuscript, but it should be obvious how they will improve SV calling and prioritization also in this field.

Overall, this paper reports the innovations across Sniffles2 and highlights them across Mendelian cases and an MSA patient. We believe that our advancements (merging and non-germline mode) will spark novel findings across human diseases and diversity with respect to SV. Furthermore, we believe that these will also be important for other species and not only human experiments. Even though the genotype model for Sniffles2 is designed for diploid organisms, Sniffles2 is capable of also detecting SV in haploid or polyploid organisms. For higher ploidy levels we would suggest running the non-germline mode as otherwise the genotype caller will penalize true SV. Thus, again highlighting Sniffles2 as a highly accurate and versatile method to detect SV of any kind from mosaic to population level.

Online Methods

Patient Enrollment:
The 31 individuals (proband and parents) included in this study were enrolled into research protocols approved by the Institutional Review Board (IRB) at Baylor College of Medicine and the Pacific Northwest Research Institute (H-29697 and H-47127, WIRB#20202158).

Sniffles2 methodology

Sniffles2: Germline calling

An overview of the steps involved in Sniffles2 germline SV detection algorithm is shown in **Supplementary Figure 9**.

Sniffles2 germline mode accepts aligned long-reads as input (BAM or CRAM format, sorted by genomic coordinate and indexed). First, read alignments are parsed and pre-filtered based on minimum mapping quality (default: 25), minimum alignment length (default: 1kb), and maximum number of split alignments (default: $3 + 0.1\ ReadLengthKb$). Split alignments are analyzed to extract SV signals for insertions, deletions, duplications, inversions, and breakends. Next to analyze splits, inline alignments are scanned for insertion and deletion signals. Sniffles2 does not merge nearby inline insertion and deletion events at this point. SV signals that fulfill a minimum length threshold (default: $0.9\ MinSVLength$) are subsequently recorded in high-resolution genomic bins. Start and end positions of alignments are recorded in a separate data structure for facilitating later coverage computation without requiring reopening of alignment files.

Sniffles2 employs a three-phase clustering process to translate individual SV signals into putative SV candidates. First, SV signals extracted from reads in the preprocessing step are clustered based on their indicated SV type and genomic start position. Second, insertion and deletion sequences in each cluster stemming from the same read are merged to correct for alignment errors in highly repetitive regions. Third, preliminary clusters are re-split to represent different supported SV lengths.

The first clustering phase constitutes a fast pass over all bins (default bin size: 100bp) containing SV signals extracted from alignments in the preprocessing step. Bins are traversed from chromosome start to end separately for each of the five basic SV types. Neighboring bins are merged if the inner distance between them is smaller than a threshold calculated based on the minimum standard deviation of the genomic SV start positions within each bin. The inner distance threshold d_n is calculated as: $d_n = r \cdot min(\sigma_{StartA}\ , \sigma_{StartB})$, where r is a constant (default: 2.5), and σ_{StartA} , σ_{StartB} refer to the standard deviation of indicated SV start positions in the two neighboring bins, respectively. In regions spanning tandem repeats, a more relaxed clustering criterion is applied: Neighboring bins are also clustered when their outer distance falls below a threshold defined based on the indicated average SV length of the SV signals stored in the neighboring bins. This threshold d_r is calculated as: $d_r = min(h_{mas}, h \cdot [x_A + x_B])$, where h and h_{max} are constants (default: 1.5 and 1kb, respectively) and x_A,x_B refer to the mean indicated SV length in the two neighboring bins. Whenever two neighboring bins have been merged, the clustering is restarted at the bin preceding the merged pair, facilitating the growth of SV clusters in both upstream and downstream directions. The first clustering phase is completed as soon as the last bin in the chromosome has been reached.

The second clustering phase constitutes merging of insertion and deletion events stemming from the same read that have been placed within the same initial cluster. Events with an inner distance closer than the set threshold (default: 150bp) are merged. In areas of tandem repeats, the distance threshold is set to the size of the initial cluster itself.

In the third phase, clusters are split by indicated SV length of the contained SV signals and subsequently re-merged, which leads to the final separation of SVs that share a start position on the reference but have different lengths. Bins are traversed from those containing small to large SV signals and merged in a similar fashion to phase one, based on the relative difference in SV length between neighboring bins being no larger than a given threshold (default: 0.33). In clusters overlapping tandem repeats, Sniffles2 does not perform re-splitting.

Differentiated clustering parameters are applied to breakend-type SVs, since no length is available as a metric to drive clustering.

At the beginning of postprocessing, SV candidates are generated from the final clusters resulting at the end of the last stage. Start coordinates and SV length are determined based on the median of the most common values supported by the reads. Standard deviations are calculated for the trimmed distribution of indicated SV start position and lengths. The quality value is summarized as the mean mapping quality of supporting reads. SVs are labeled as precise when the sum of SV start and length standard deviation is less than the set threshold (25bp).

SV candidates are filtered based on absolute and relative (compared to the SV length) standard deviation of their coordinates. In addition, type-specific coverage filtering is applied to deletions and duplications, requiring central coverage changes consistent with the detected variant. Instead of requiring users to settle for a predefined, static minimum read support threshold, Sniffles2 dynamically adjusts the minimum support value based on estimates of global and regional sequencing coverage. By default, the minimum read support threshold is calculated as $MinSupport = \alpha \cdot ([1-\lambda]\, Cglobal + \lambda\, Clocal)$. Where Cglobal and Clocal refer to average chromosomal and SV surrounding coverage, respectively. The parameters are set as α=0.1 and λ=0.75, by default. For insertion and deletion SVs, support from inline alignments and split alignments is output separately. Additionally indicated support from soft-clipped reads is additionally recorded for insertion SVs.

Genotypes are determined using a maximum-likelihood approach. The genotype quality is calculated based on the likelihood ratio of the second most likely to the output genotype: $Q = -10\, log_{10}(\, L_2/L_1)$, whereas L_1 and L_2 refer to the likelihood of the most likely genotype and second most likely genotype, respectively. Genotype likelihoods are computed for a binomial distribution for the observed number of variant and reference reads. Genotype likelihoods are set as 1.0-ß for 1/1, 0.5 for 0/1, and ß for 0/0, whereas ß represents the genotype error introduced through sequencing and alignment artifacts and is set to ß=0.05 by default.

For insertion SVs, sequencing and read aligner errors are corrected using a fast kmer-based pseudo-alignment method. Through this, Sniffles2 generates a consensus sequence in two steps: In the first step, the best possible starting sequence is chosen from the supporting read with the smallest distance in SV start position and length to the final reported SV coordinates. K-mers (default length: 6bp) are enumerated for this read's supported insertion sequence and a taboo set of repetitive k-mers, which occur more than once in the sequence, is built. Simultaneously, the positions of non-repetitive k-mers are stored in an anchor table to facilitate pseudo-alignment of the other reads. In the second phase, k-mers from other reads insertion sequences are enumerated. When a k-mer is present in the anchor table, the corresponding position in both the initial insertion sequence and the current read are stored. After all reads have had their k-mers anchored, sequences between anchored k-mers are extracted from the pseudo-aligned reads. These sequences from between the anchored k-mers constitute the parts of each read's insertion sequence in disagreement with the initial sequence. Finally, coordinates of the initial sequence are traversed, and the consensus is generated as the most common base at the respective position throughout all pseudo-aligned reads. Long insertions (i.e. multiple kbp) are often difficult to detect even in long-read data because reads often do not span the full insertion sequence. To improve detection of long insertions, Sniffles2 records these clipped read events as additional support for presence of a large insertion. This enables Sniffles2 to accurately detect large insertions even when the SV is fully covered by just by a single read.

Post processed and annotated SV calls that passed quality control checks are written to the output VCF file. Quality control filters applied to SV candidates by default include absolute and relative standard deviation of the SV breakpoints, coverage change for copy number variants and minimum coverage in the surrounding genomic region. Additionally, all unfiltered SV candidates and genome-wide coverage information are written to a specified output SNF file, which may be consecutively used as input for multi-sample calling (see Section 1b). Using the --qc-output-all option, all unfiltered candidates (except for the minimum SV length filter) can also be directly written to the VCF output file complete with the respective reasons for why they would have been filtered by default.

Full parallelization across chromosomes is applied through all key steps in Sniffles2, including preprocessing, clustering and postprocessing. The final SV calls are written to a sorted VCF output file. Alternatively, Sniffles2 also supports direct output to a sorted, bgzipped and tabix-indexed VCF file.

Sniffles2: Combined Calling (Population Mode)

Sniffles2 produces a fully genotyped population VCF file by introducing a specialized mode (*Sniffles2 combine*) for both family- and population-level SV calling. *Sniffles2 combine* is built around a new specialized binary file format (SNF), designed to store a complete snapshot of structural variation and sequencing coverage for a single sample. Mergeable SNF files for later population-level calling are designed to be easily produced as a side-product of regular single-sample SV calling using Sniffles2, by using the optional --snf output argument. Based on individual use case requirements, Sniffles2 can simultaneously produce SNF files and/or regular VCF files in a single run of processing an individual sample.

SNF files consist of a JSON-based index followed by a series of multiple gzip-compressed blocks (separated by genomic coordinates). Each block stores all putative SV candidates, separated by SV type, for a single sample's respective genomic region. This includes candidates only supported by e.g. a single read, that would normally be ignored. Each block furthermore stores sequencing coverage information (500bp resolution by default). All stored SV candidates contain a compressed form of all the information of the final SV calls, as they would be output in a single-sample VCF file, such as start, end positions, standard deviation, and alternative alleles. SNF blocks span a genomic region of 100kb by default. This small block size comparison to a typical mammal genome allows Sniffles2 to combine a high number of samples simultaneously while keeping a manageable memory footprint.

SNF files, once generated, can then be used as input for the *Sniffles2 combine* mode, producing a final, fully genotyped population-level VCF file within seconds. SNF files may also be reused in the combine step, e.g., when the population is later on extended, when individual samples need to be re-run, or when querying whether a later newly identified SV is present in a population. These use cases would not be possible without costly reprocessing of all samples with the currently prevalent method of forced calling. A schematic of SNF file structure can be found in **Supplementary Figure 10.**

When presented with multiple SNF files as input, Sniffles2 combines them through a single pass over chromosomal region. For each region, the respective SNF blocks overlapping it are loaded, including all SV candidates and coverage information from each sample. In the following step, Sniffles2 groups the loaded SV candidates based on SV type and coordinate-based matching criteria. For each SV candidate, Sniffles2 first checks if there is an already existing, matching group. An SV candidate matches a group if it has the same SV type and the sum of start position and length deviation is less than $M \cdot \sqrt{min(SVLength, GroupSVLength)}$, where M is set to 500bp by default (user-adjustable). The start position and SV length of a group are defined as the arithmetic mean of all SVs currently contained in it. In case there are one or more groups that fulfill the matching criteria for the current SV candidate, the group with the smallest deviation metric is chosen and the SV candidate placed therein. The coordinates of the selected group are then subsequently updated to represent the new average position of length of the contained candidates. If there are no matches, a new SV group is created. By default, Sniffles2 allows for matching multiple SVs from the same sample within a group (can be disabled using *a* dedicated parameter).

This partition of SNF files into individually loadable blocks keeps Sniffles2 memory footprint manageable even when processing a high number of samples and/or samples with high-coverage. Sniffles2 further implements a dynamic binning strategy for accelerating the grouping phase. Sniffles2 first assigns all loaded SV candidates from the current chromosomal region to bins based on SV type and start position. Bins are then traversed from low to high coordinate within the current block, while collecting encountered SV candidates. When the number of SV candidates exceeds a certain threshold (default: $PopulationSize * 0.5$), the collected SV candidates are grouped as described above. Triggering the grouping stage only when a set number of SV candidates is reached allows for the highest possible accuracy in matching SVs from different samples in regions with low complexity, while keeping the runtime manageable even in regions with a high density of SV candidates. To avoid edge effects, the final resolving of SV groups with genomic coordinates close to the ends (default: <2.5kb) of the respective bin are carried over and finally resolved in conjunction with the grouping of the next bins. The same strategy is applied to SV groups close to the genomic start or end coordinate of the currently processed SNF block.

By default, Sniffles2 combine mode will output all resulting SV groups in the population that meet at least one of two criteria: A. The SV has been detected with high-confidence (i.e. passes all quality control checks) in at least one sample and/or B (default: high-confidence call in at least one sample). The SV is present in a sufficiently high number of individual samples, even though it may not have passed individual quality-control checks (default: present in at least $max(0.2\ PopulationSize, 2)$ samples). These parameters are also user-adjustable and can be adjusted or disabled without having to re-generate the SNF files for the individual samples.

Each final SV group that passes the above criteria is output as an SV in the final population-level VCF file, including the genotypes from all samples. For samples that did not have a SV candidate that could be matched to the group, Sniffles2 first uses the coverage information stored in the SNF file of the respective sample to determine if the sequencing depth at around the group's genomic location was sufficiently high (default value: 5x). If it is, the sample genotype for that SV is output as 0/0 if there is no evidence, and otherwise as missing (./.). For all SVs, the number of reads supporting the SV and supporting the reference are output for all samples, allowing for differentiation between true biological and technically induced absence of each SV from a sample.

Sniffles2 combine is fully parallelized, allowing leveraging multi-core CPU systems not just for calling individual samples but also the final combination step. This, in conjunction with the separation of SNF files into blocks and dynamic binning strategy, together enables Sniffles2 to perform scalable population-level SV calling.

Sniffles2: Non-Germline calling

In the non-germline mode, a reduced default minimum support multiplier is applied (default: 0.025) to increase sensitivity for low-frequency SVs. At coverage levels of 30x to 50x, this leads to a minimum read support of 2-4 reads for the detection of non-germline SVs. To balance out the increased influence of sequencing and alignment artifacts at this lowered read support threshold, additional filtering based on alignment quality is applied. In the preprocessing steps, the length-weighted number of mismatches is recorded for all SV signals, excluding insertions and deletions. After calling, SVs with an average weighted mismatch ratio of larger than a threshold $t = c * a$, where a is the average length-weighted mismatch number for all reads and c is a constant (default: 1.66) are filtered. The additional, coverage-based filtering steps for CNVs applied in the germline mode are not applied in non-germline mode, as coverage changes induced by somatic SVs are not reliably measurable.

Benchmarking Methodology

Benchmarking SV callers on GIAB, 1000 Genomes and CMRG

Reads were mapped using minimap2[62] (v2.17-r941) technology specific preset parameters. The -Y option was supplied to disable hard clipping (required by pbsv) and generate the –MD tag (required by Sniffles1) and the PacBio / ONT presets were used respectively. Resulting alignments were converted to BAM format, sorted and indexed using samtools (v1.13).
As measure of coverage across all benchmarked data sets, we used the mapping coverage as reported by mosdepth[63] (version 0.3.2), which was averaged across all autosomes.

Besides HG002 we also benchmarked SV on three assemblies of the 1000genomes (HG01243, HG02055, HG02080). Here we leveraged the phased HiFi assemblies provided at https://github.com/human-pangenomics/hpgp-data and the corresponding long-reads. The benchmark set was derived from a dipcall[50] (version 0.2) alignment against GRCh38 reference. This result was used together with the corresponding bed files for benchmarking.

We used Truvari[48] (version 2.1) for benchmarking the accuracy of all SV callers across datasets. For benchmarking, we used the *--passonly* parameter to include only those SVs from caller and gold standard that are not marked as filtered. For the GIAB benchmarks, we additionally used the *--giabreport* parameter to generate the benchmark-specific detailed report.

As included regions Tier 1 regions were used unless otherwise specified. For all other parameters, default values were used.

Callers were first benchmarked using default parameters and callers other than Sniffles2 were separately benchmarked on GIAB by manually setting the minimum read support parameter to 2 (sensitive).

SVIM[47] (v1.4.2) does not include filtering steps in its main pipeline, which caused it to perform poorly (F-measure) in most benchmarks, and we were not able to identify a recommended default cutoff for the quality value that SVIM outputs along with its SV calls. Therefore, in line with previous SV caller benchmarks, we filtered the output of SVIM to include only calls with a minimum read support of 10 by default (equal to the default of cuteSV and Sniffles1) or 2 (sensitive).

For benchmarking Sniffles2 (build rc11_a973e), we only used the default parameters with the exception of non-germline SVs, where the *--non-germline* option was supplied. For Sniffles[27] (v1.12), default parameters were used. For cuteSV[45] (v1.0.11), we used the additional parameters recommended by the authors for use with HiFi / ONT datasets in their GitHub documentation, as well as the --genotype option. For pbsv[46] (v2.6.2), we supplied the --ccs option for analyzing HiFi data, as recommended by the authors. Both pbsv and Sniffles2 support the use of tandem repeat annotations for improving SV calling in repetitive regions. For pbsv and Sniffles2, we therefore supplied the tandem repeat annotations for GRCh37 / GRCh38 which we obtained from the pbsv repository on GitHub: https://github.com/PacificBiosciences/pbsv .

For all SV callers that have an option for specifying the number of multiprocessing threads, we set the number of threads as 8. We measure and report the total CPU time and wall clock time using the UNIX time command. For the benchmarks including only insertions and deletions, we used SnpSift [64](v4.3t) to filter the output of all SV callers to include only those types of structural variants. To prepare SV caller output for benchmarking, VCF files were sorted using bedtools, compressed and indexed using bgzip and tabix. For SVIM, SVs labeled as INS:NOVEL were relabeled to INS, in order to be able to be matched to insertions in the benchmark sets by Truvari.

Simulation of different SV types using SURVIVOR

SURVIVOR [55](v1.0.7) was used to simulate SV types not covered by the GIAB and other benchmarks. For this benchmark, 3000 duplications, inversions and translocations were each simulated within a length range of 500bp to 30kb on the human reference genome hg19 in diploid mode. A total sequencing depth of 30x was simulated for ONT reads, with the error profile obtained using the SURVIVOR scanreads command from the HG002 ONT Q20+ data set. SVs were called using each SV caller for the simulated reads using the default parameters and postprocessing steps also used in the GIAB and other benchmarks (see respective methods subsection). The SURVIVOR eval command was used (matching threshold: 500bp) to obtain TP, FN and FP counts for each caller and simulated SV type from which precision, recall and F-measure were calculated.

Measurement of Insertion Sequence accuracy

Accuracy of insertion sequences recovered by the SV callers was measured using Biopython's[65] (v1.79) pairwise2 global alignment function. First, the true positive calls from all investigated SV callers on the data set were intersected, to establish a common set of calls to benchmark. Next, the gold standard and reported insertion nucleotide sequences were aligned and the resulting score was normalized by length of the gold standard sequence to compute the alignment identity. We measured sequence accuracy separately for the GIAB HiFi and ONT data sets (30x coverage). Results are shown in **Supplementary Figure 1**. The respective script is made available in the supplementary materials.

Simulation of low-frequency SVs

Low-frequency SVs were simulated by combining varying coverage titrations of HG002 and HG004 into synthetic samples with different levels of mosaicism. Recovery of SVs unique to HG002 (based on being absent from HG004 according to the GIAB gold standard) were then used as benchmark to measure recall for low-frequency SVs. For benchmarking the ability of Sniffles2 to detect low-frequency SVs, we simulated synthetic data sets with 65x/5x, 63x/7x, 60x/10x and 55x/15x, where the coverage refers to HG004 and the second one to HG002. Next, we subsampled the GIAB Tier 1 benchmark set of SV calls to exclude those present in the mother (HG004), i.e. those SVs either heterozygous or homozygous in HG004. To measure recall for low-frequency SVs, we ran Sniffles2 in non-germline mode on the synthetic samples and used Truvari as described in the methods section on GIAB benchmarks to compute the recall for the rare HG002 SVs introduced into each HG004 data set. As in all the other GIAB benchmarks, analysis was limited to insertion and deletion SVs. In order to measure the precision of Sniffles2 non-germline mode, we separately benchmarked it on the 70x coverage HG002 data set and compared it to the default Sniffles2 germline mode.

MSA patient analysis

Optical mapping data on MSA patient brain

Ultra-high molecular weight (UHMW) DNA was isolated from frozen human brain tissues using a Bionano Prep SP Tissue and Tumor DNA Isolation kit (#80038) according to the Bionano prep SP Brain Tissue Isolation Tech Note (#3400). In short, approximately 20mg frozen tissue was homogenized using a Qiagen TissueRuptor (9002755), passed through a 40um filter, and treated sequentially with Qiagen protease (catalog #19155), proteinase K, and RNAse A in lysis and binding buffer. The homogenate was then treated with PMSF to de-activate the Protease and Proteinase K, washed, and eluted. The extracted DNA was mixed using an end-over-end rotator for 1 hour at 5rpm and allowed to rest at room temperature until homogenous (approximately 1 week). 750ng purified UHMW DNA was fluorescently labeled at the recognition site CTTAAG with the enzyme DLE-1 and subsequently counter-stained using a Bionano Prep DLS Labeling Kit (#80005) following manufacturer's instructions (Bionano Prep Direct Label and Stain (DLS) Protocol #30206). Optical genome mapping was performed using a Saphyr Gen2 platform for a final effective coverage of 894X for the pons and 754X for the cingulate. Effective coverage is defined as the total raw coverage of molecules ≥ 150kbp in length multiplied by the proportion of molecules which align to the reference genome.

Calling of low allele frequency structural variants was performed using the rare variant analysis pipeline (Bionano Solve version 3.6) on molecules ≥ 150kbp in length. De novo assembly was performed using the longest 250X molecules of each dataset. The variant annotation pipeline (Solve 3.7) was used to detect which structural variant calls in the cingulate are present in the pons structural variant calls and/or molecules. See the Bionano Solve Theory of Operations for more details.

MSA samples comparison

Illumina reads were mapped to the human genome Grch38 using bwa[66] mem (version 0.7.17-r1188) with default parameters including -M to mark split reads as secondary alignments. Subsequently we identified SV using manta[58] (version 1.6.0).

For ONT, reads were mapped using minimap2 [62] (version 2.17-r941) with present parameters for ONT. Subsequently we identified SV using Sniffles2 with the default parameters for non-germline mode.
The Bionano data smap file was converted by SURVIVOR smaptovcf into a VCF file.

To compare SVs called by Sniffles2, Manta (Illumina), and Bionano we used SURVIVOR merge using a 10kb threshold, matching SV type and ignoring reported SV strand. We extended it to 10kbp after testing 500, 1kbp and 5kp thresholds and observed that the accuracy of the breakpoints from Bionano required the larger parameter.
The genotype columns in the SURVIVOR merge output were compared for each SV to determine presence or absence in the results reported by the respective method.
Subsequently, to further investigate SVs absent from the Manta call sets, we additionally genotyped the respective Sniffles2 calls against the raw Illumina read alignments for the same (cingulate cortex) as well as a different brain region (cingulate white matter) using svtyper (version: 0.7.1).[59] SVs reported as having at least one supporting read by svtyper were considered as present in a sample.

Mendelian inconsistency benchmark in population mode

Mendelian benchmark/ inconsistency

To assess the performance of Sniffles2 population mode, we used the Ashkenazim family trio. We called SV using Sniffles2 and cuteSV. For Sniffles2 we used a minimum SV length of 50 and with the output being the SNF binary file that contains the unfiltered SV candidates and genome-wide coverage information (see below Sniffles2-pop) to then merge with Sniffles2 population-level calling providing the reference genome to obtain the sequences of the deletions. For the case of cuteSV we used v1.0.11 with recommended parameters for Oxford Nanopore data (see below cuteSV-merge). Then, we merged the results of cuteSV using SURVIVOR v1.0.7 with a maximum distance between breakpoints of 1kb, a minimum support of one and taking into account the SV type (see below cuteSV-merge). Next, we performed force-calling with cuteSV using as input the merged SV from SURVIVOR. Finally, we performed a second merge with SURVIVOR with identical parameters (see below cuteSV-merge).

We then tested the mendelian inconsistency of the genotypes using the BCFtools v1.14 mendelian plugin[54]. The mendelian plugin denotes a genotype consistent when the proband genotype is in concordance with the parental genotypes (e.g F 0/0, M 0/1, P 0/0), inconsistent when the proband and parental genotypes do not match (e.g. F 0/1, M 1/1, P 0/0) and NA when the proband has a missing genotype (./.). For all analysis time was measured utilizing the linux time command.

```
# Sniffles2-pop
/usr/bin/time -v -o sniffles2.timelog Sniffles2 \
 --input HG00[2|3|4].bam \
 --snf HG00[2|3|4]-sniffles2.snf
 --vcf HG00[2|3|4]-sniffles2.vcf
 --threads 8
 --minsvlen 50
 --sample-id HG00[2|3|4]

/usr/bin/time -v -o sniffles2.timelog Sniffles2 --input \
 HG002-sniffles2.snf \
 HG003-sniffles2.snf \
 HG004-sniffles2.snf \
 --vcf HG002-trio-sniffles2-merge.vcf \
 --threads 8 \
 --reference hs37d5.fa


# cuteSV-merge
/usr/bin/time -v -o cutesv.timelog cuteSV HG00[2|3|4].bam hs37d5.fa \
 HG00[2|3|4]-cutesv.vcf cutesv-tmp-workdir \
 --max_cluster_bias_INS 100 \
 --diff_ratio_merging_INS 0.3 \
 --max_cluster_bias_DEL 100 \
 --diff_ratio_merging_DEL 0.3 \
 --threads 16 \
 --genotype

# list files for SURVIVOR
ls HG00*-cutesv.vcf > cutesv-files-to-metge.txt
CUTESV_MERGE_LIST="cutesv-files-to-metge.txt"
MAX_DISTANCE_BREAKPOINTS="1000"
MIN_NUM_SUPPORT_CALL="1"
USE_TYPE="1"
USE_STRAND="0"
PARAM_DISABLED="0"
MIN_SV_SIZE="50"
CUTESV_MERGE_VCF="HG002-trio-cutesv-merge.vcf"

/usr/bin/time -v -o survivor.timelog survivor merge \
 ${CUTESV_MERGE_LIST} \
 ${MAX_DISTANCE_BREAKPOINTS} \
 ${MIN_NUM_SUPPORT_CALL} \
 ${USE_TYPE} \
 ${USE_STRAND} \
 ${PARAM_DISABLED} \
 ${MIN_SV_SIZE} \
 ${CUTESV_MERGE_VCF}

/usr/bin/time -v -o cutesv.timelog cuteSV HG00[2|3|4].bam hs37d5.fa \
 HG00[2|3|4]-cutesv-forcecall.vcf cutesv-tmp-workdir \
 --max_cluster_bias_INS 100 \
 --diff_ratio_merging_INS 0.3 \
 --max_cluster_bias_DEL 100 \
 --diff_ratio_merging_DEL 0.3 \
 --threads ${PPN} \
 -Ivcf HG002-trio-cutesv-merge.vcf \
 --genotype

ls HG00*-cutesv-forcecall.vcf > cutesv-forcecall-files-to-metge.txt
CUTESV_FORCECALL_MERGE_LIST="cutesv-forcecall-files-to-metge.txt"
CUTESV_FORCECALL_MERGE_VCF="HG002-trio-cutesv-forcecall-merge.vcf"


/usr/bin/time -v -o survivor.timelog survivor merge \
 ${CUTESV_FORCECALL_MERGE_LIST} \
 ${MAX_DISTANCE_BREAKPOINTS} \
 ${MIN_NUM_SUPPORT_CALL} \
 ${USE_TYPE} \
 ${USE_STRAND} \
 ${PARAM_DISABLED} \
 ${MIN_SV_SIZE} \
 ${CUTESV_FORCECALL_MERGE_VCF}
```

8. Chromosome X disorder patient analysis

Sniffles2 population mode was used to analyze 31 Oxford Nanopore samples that represented cases of Mendelian disorders in probands. We obtained the bam files by running PRINCESS[28] (version 1.0) using the default parameters and "ont" flag. PRINCESS implicitly calls Minimap2[62] (version 2.17) with the following parameters "-ax map-ont -Y --MD"; Later, we sorted the output using samtools[54] (version 1.9). For all samples, unfiltered SV candidates and genome-wide coverage information are written to a specified output SNF file and then merged with Sniffles2 population-level calling. General statistics, such as SV sizes and composition (proportion of each SV type) were computed using a custom python script (cat population.vcf | sniffles2_vcf_parser.py parsesv).

Given the nature of the dataset, only the SV calls from chromosome X were used. Additionally, for specific individuals (BH14379, BH14413) SV from chromosome Y were analyzed given that Both aCGH and Sniffles2 called translocations to chromosome Y.

Then, all SVs that were less than 10kb were filtered, as aCGH data showed large events were involved. Finally, we filtered out SV that occurred in the father, as this disorder is fully penetrant in males (cat population.vcf | sniffles2_vcf_parser.py population --ont-31).

Data availability

GIAB HG002 PacBio HiFi data is hosted at the github server: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/

ONT HG002: https://labs.epi2me.io/gm24385_q20_2021.10/
ONT HG004: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/UCSC_Ultralong_OxfordNanopore_Promethion/

GIAB benchmark sets:
Genome wide: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/
Medical regions: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/

The 1000 genomes data sets of the three genomes were downloaded from : https://github.com/human-pangenomics/hpgp-data The dipcall results that we leveraged as benchmark are deposited at https://github.com/smolkmo/Sniffles2-Supplement

The other data sets will be made available over SRA.

Code availability

Source code for Sniffles2 is available at https://github.com/fritzsedlazeck/Sniffles the auxiliary scripts are available at https://github.com/smolkmo/Sniffles2-Supplement

Acknowledgments

This research was supported in part by the National Institute of General Medical Sciences (R01 GM132589) to CMBC & FJS. This research was supported in part by the Intramural Research Program of the National Institutes of Health (National Institute of Neurological Disorders and Stroke; project number: 1ZIANS003154). This research was supported in part by the MSA Trust to CP. FJS, MS, MM LP are partially supported by NIH grants (UM1HG008898, 1U01HG011758-01, U01 AG058589 and PO 75N95021P00215)

Contributions

MS implemented the software. MS, FJS designed the study. CMBC, CP, MG, DP, SS & FJS generated the data. MS, LFP, CMG, SB, KH, MM, FJS contributed to data interpretation. All the authors reviewed and edited the manuscript.

Competing interests

FJS receives research support from PacBio and Oxford Nanopore. S.W.S. is a member of the Scientific Advisory Council of the Lewy Body Dementia Association and an editorial board member of JAMA Neurology, Journal of Parkinson's Disease, and the Neurodegeneration Specialty Section for Frontiers of Neurology, Frontiers of Psychiatry, and Frontiers of Neuroscience. LFP is sponsored by Genentech, Inc. KH is an employee of Bionano Genomics.

References

1. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
2. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
3. Weissensteiner, M. H. *et al.* Discovery and population genomics of structural variation in a songbird genus. *Nature Communications* vol. 11 (2020).
4. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145–161.e23 (2020).
5. Soyk, S. *et al.* Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. *Nat Plants* **5**, 471–479 (2019).
6. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
7. Beck, C. R. *et al.* Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. *Cell* **176**, 1310–1324.e10 (2019).
8. Leija-Salazar, M. *et al.* Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol Genet Genomic Med* **7**, e564 (2019).
9. Sekar, S. *et al.* Complex mosaic structural variations in human fetal brains. *Genome Res.* **30**, 1695–1704 (2020).
10. Schmidt, K., Noureen, A., Kronenberg, F. & Utermann, G. Structure, function, and genetics of lipoprotein (a). *Journal of Lipid Research* vol. 57 1339–1359 (2016).

11. Baslan, T. *et al.* High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab812.
12. Aganezov, S. *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
13. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259–1271.e11 (2017).
14. Layer, R. M., Sedlazeck, F. J., Pedersen, B. S. & Quinlan, A. R. Mining Thousands of Genomes to Classify Somatic and Pathogenic Structural Variants. doi:10.1101/2021.04.21.440844.
15. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
16. Belyeu, J. R. *et al.* De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021).
17. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
18. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* vol. 3 160025 (2016).
19. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
20. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
21. Coster, W. D., De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nature Reviews Genetics* vol. 22 572–587 (2021).
22. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).

23. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

24. October 2021 GM24385 Q20+ Simplex Dataset Release. https://labs.epi2me.io/gm24385_q20_2021.10/ (2021).

25. October 2021 GM24385 Q20+ Simplex Dataset Release. https://labs.epi2me.io/gm24385_q20_2021.10/ (2021).

26. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).

27. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

28. Mahmoud, M., Doddapaneni, H., Timp, W. & Sedlazeck, F. J. PRINCESS: comprehensive detection of haplotype resolved SNVs, SVs, and methylation. *Genome Biol.* **22**, 268 (2021).

29. Gorzynski, J. E. *et al.* Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *N. Engl. J. Med.* **386**, 700–702 (2022).

30. Goenka, S. D. *et al.* Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01221-5.

31. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotechnology* (2022) doi:10.1038/s41587-021-01158-1.

32. Carvalho, C. M. B. *et al.* Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum. Mol. Genet.* **18**, 2188–2203 (2009).

33. Carvalho, C. M. B. *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).

34. Liu, P., Carvalho, C. M. B., Hastings, P. J. & Lupski, J. R. Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.* **22**, 211–220 (2012).

35. Guy, J., Cheval, H., Selfridge, J. & Bird, A. The role of MeCP2 in the brain. *Annu. Rev. Cell*

Dev. Biol. **27**, 631–652 (2011).

36. del Gaudio, D. *et al.* Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males. *Genet. Med.* **8**, 784–792 (2006).
37. Ramocki, M. B., Tavyev, Y. J. & Peters, S. U. The MECP2 duplication syndrome. *Am. J. Med. Genet. A* **152A**, 1079–1088 (2010).
38. Chronister, W. D. *et al.* Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep.* **26**, 825–835.e7 (2019).
39. Proukakis, C. Somatic mutations in neurodegeneration: An update. *Neurobiol. Dis.* **144**, 105021 (2020).
40. Fanciulli, A. & Wenning, G. K. Multiple-system atrophy. *N. Engl. J. Med.* **372**, 249–263 (2015).
41. Mokretar, K. *et al.* Somatic copy number gains of α-synuclein (SNCA) in Parkinson's disease and multiple system atrophy brains. *Brain* **141**, 2419–2431 (2018).
42. Perez-Rodriguez, D. *et al.* Investigation of somatic CNVs in brains of synucleinopathy cases using targeted SNCA analysis and single cell sequencing. *Acta Neuropathologica Communications* **7**, 1–22 (2019).
43. Knouse, K. A., Wu, J. & Amon, A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res.* **26**, (2016).
44. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. *PLoS Comput. Biol.* **16**, e1008012 (2020).
45. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
46. PacificBiosciences. GitHub - PacificBiosciences/pbsv: pbsv - PacBio structural variant (SV) calling and analysis tools. *GitHub* https://github.com/PacificBiosciences/pbsv.
47. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads.

Bioinformatics **35**, 2907–2915 (2019).

48. English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity. doi:10.1101/2022.02.21.481353.
49. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
50. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
51. Tusso, S. *et al.* Ancestral Admixture Is the Main Determinant of Global Biodiversity in Fission Yeast. *Mol. Biol. Evol.* **36**, 1975–1989 (2019).
52. Chander, V., Gibbs, R. A. & Sedlazeck, F. J. Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* **8**, (2019).
53. Lecompte, L., Peterlongo, P., Lavenier, D. & Lemaitre, C. SVJedi: genotyping structural variations with long reads. *Bioinformatics* **36**, 4568–4575 (2020).
54. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
55. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
56. Pettersson, M. *et al.* Cytogenetically visible inversions are formed by multiple molecular mechanisms. *Hum. Mutat.* **41**, 1979–1998 (2020).
57. Burgoyne, R. D. Neuronal calcium sensor proteins: generating diversity in neuronal Ca2+ signalling. *Nat. Rev. Neurosci.* **8**, (2007).
58. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
59. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
60. Pascarella, G. *et al.* Recombination of repeat elements generates somatic complexity in human genomes. *bioRxiv* (2020) doi:10.1101/2020.07.02.163816.

61. Federoff, M. *et al.* Genome-wide estimate of the heritability of Multiple System Atrophy. *Parkinsonism Relat. Disord.* **22**, 35–41 (2016).
62. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab705.
63. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
64. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).
65. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
66. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).