Genome Biology

https://doi.org/10.1186/s13059-025-03928-5

Article in Press

SDrecall: a sensitive approach for variant detection in segmental duplications

Received: 03 Mar 2025

Accepted: 29 Dec 2025

Published online: 12 January 2026

Cite this article as: Yang, X., She, C., Zhang, C. *et al*. SDrecall: a sensitive approach for variant detection in segmental duplications. *Genome Biol* (2026). https://doi.org/10.1186/s13059-025-03928-5

Xing Yang, Chun She, CaiCai Zhang, Daniel Leung, Jing Yang, Koon-Wing Chan, Jaime Duque, Yu Lau & Wanling Yang

SDrecall: a sensitive approach for variant detection in segmental duplications

Xing Tian Yang[1]*, Chun Hing She[1]*, CaiCai Zhang[1], Daniel Leung[1], Jing Yang[1], Koon-Wing Chan[1], Jaime S Rosa Duque[1], Yu Lung Lau[1]†, Wanling Yang[1]†

[1] Department of Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

* These authors contributed equally to this work.

† Correspondence: lauylung@hku.hk; yangwl@hku.hk

Email addresses:

Xing Tian Yang: yangyxt@hku.hk

Chun Hing She: snakesch@connect.hku.hk

CaiCai Zhang: u3009162@connect.hku.hk

Daniel Leung: dan.leung@connect.hku.hk

Jing Yang: jingy09@hku.hk

Koon-Wing Chan: kwchan@hku.hk

Jaime S Rosa Duque: jsrduque@hku.hk

Yu Lung Lau: lauylung@hku.hk

Wanling Yang: yangwl@hku.hk

Abstract

Variant calling in segmental duplications is challenging for short-read sequencing because of ambiguous read origins. We present SDrecall, a method for sensitive variant detection in these regions. Upon constructing a network of homologous sequences, SDrecall realigns reads to each segmental duplication from its homologous counterparts. Realignments are phased and assembled into haplotypes via graph-based algorithms, followed by integer linear programming to retain the two most plausible haplotypes. Tested against long-read benchmarks, SDrecall achieved 95% sensitivity, while maintaining manageable false positives for short variants. SDrecall thus offers significant value for molecular diagnosis in terms of causal mutation detection within homologous regions.

Keywords

Segmental duplication; Molecular diagnosis; Variant Caller

Background

Next-generation sequencing (NGS) has revolutionized human genome analysis, yet certain genomic regions remain difficult to map accurately. Segmental duplications (SDs)—defined as genomic segments of at least 1 kb with high sequence identity (typically above 90%) to other regions(1-3)—are particularly challenging because short reads offer limited sequence information to determine their true origins. In modern NGS workflows, DNA is typically fragmented into pieces ranging from 250 to 500 bp, and paired-end sequencing generates two 150 bp reads from both ends of each fragment. Consequently, if a fragment originates entirely within an SD, its read pair may align to multiple similar regions in the reference genome, leading to high mapping ambiguity according to modern mapping algorithms(4-6). As the size of SD grows, more fragments enveloped by the SD are likely to be mapped with undetermined origins (Figure 1A). As a proof of concept, data from Genome in a Bottle(7) project (GIAB) sample HG002 show that mapping quality (MAPQ)—a metric inversely related to mapping ambiguity—of the reads originated from the SD decreases as the size of SD grows(Figure 1B).

According to the whole-genome assembly comparison (WGAC)(8), which is considered the gold standard for SD detection(9-12), GRCh37 and GRCh38 contain approximately 144 Mbp and 162 Mbp of SDs, respectively—about 5% of the genome—that overlap with roughly 2,400 genes(13) (around 6% of protein-coding regions). The telomere-to-telomere (T2T)-CHM13 assembly, however, reveals an expanded SD landscape (approximately 218 Mbp, or 6.6% of the genome)(14, 15), indicating a bigger burden from SDs. This presents further challenges for the molecular diagnosis of Mendelian disorders associated with genes harboring

homologous sequences such as Chronic Granulomatous Disease (CGD, *NCF1*)(16), Spinal Muscular Atrophy (SMA, *SMN1/SMN2*)(17, 18), Congenital Adrenal Hyperplasia (CAH, *CYP21A2*)(19), and Gaucher Disease (GD, *GBA*)(20) . According to the Online Mendelian Inheritance in Man (OMIM)(21), we summarized the SD-overlapping situation of 200 selected disease causal genes to a map in Additional file 1: Fig. S3.

The mapping ambiguity within SDs can be naturally overcome by long-read sequencing—commonly referred to as Third Generation Sequencing (TGS) since long reads are generally larger than most SDs and spans unique genomic sequences. Despite the rapid development of TGS, its cost remains substantially higher than that of conventional NGS(22), leading to limited application in large-scale clinical practice. In addition, given the huge amount of legacy NGS data from patients accumulated in the last two decades, revisiting them with advances in variants detection within SDs might lead to significant novel diagnoses and insights. Therefore, improving the analysis of NGS data for sensitive variant detection within SDs remains a valuable advancement in genomic diagnostics. While previous methods offered some solutions(23, 24), they have typically focused on limited genomic regions tied to specific disorders, and cannot be generalized to other SD regions in human genome.

The mapping ambiguity and variant detection within segmental duplications (SDs) cannot be completely resolved due to the inherently limited information carried by short reads, thus rendering the simultaneous achievement of high sensitivity and precision impossible. In the molecular diagnosis of Mendelian diseases, sensitivity is of utmost importance, whereas reduced precision can be mitigated to a certain extent through further downstream analysis. In this context, detected variants are

typically ranked to identify disease causality(25, 26), and most false positives (FPs)—whether functionally irrelevant or common in the population—do not withstand the causality evaluation.

Ebbert et al. proposed a method(27) to improve the mapping in low-mappability regions by realigning reads from all related homologous regions. This work, as a proof-of-concept, managed to recover numerous SD variants that might explain part of missing heritability of Alzheimer's Disease. However, their method was not benchmarked against gold standard callsets to quantify the improved variant detection sensitivity. In addition, this work lacks rigorous measures to control the false positives introduced by realignments, which might lead to excessive noises for downstream analysis.

Here we introduce SDrecall, a novel approach for sensitive detection of single nucleotide variants (SNVs) and small indels (< 30bp) within SDs. SDrecall offers small variant detection complementary to the traditional variant callers like GATK(28) and DeepVariant(29). According to benchmarks with golden callsets derived from long-read sequencing data in the GIAB project, SDrecall improves the variant detection sensitivity to approximately 95% in comparison to the benchmark callsets from the GIAB project while managed to remove 88% false positives (FPs) introduced by read realignments. To the best of our knowledge, SDrecall is the first comprehensive tool designed to detect small variants in SDs with high sensitivity while stringently controlling false positives based on NGS data. It offers full inspection of genomic regions camouflaged by ambiguous alignments while minimizing relevant FPs clouding the causal variant identification in molecular diagnosis. This tool is poised to play a crucial role in significantly enhancing the

molecular diagnosis rate of Mendelian disorders and has already helped capture causal variants in three CGD patients.

Results

To demonstrate the efficacy of SDrecall in terms of enhancing the detection sensitivity of small variants (SNPs and indels) within segmental duplications (SDs), we need to first briefly introduce the general scheme of the workflow. With user-defined regions of interest, which are the protein coding regions by default, SDrecall identifies the SDs overlapping with these regions, as well as their counterpart SDs genome-wide to form groups of SDs sharing homologous sequences. Subsequently, within each group, SDrecall recruits all the overlapping reads, which are potentially misaligned due to homology, and re-aligns them respectively to each SD that overlaps with protein coding regions (protein coding SD, pcSD) in the same group. This homology-guided read recruitment and re-mapping improves variant calling to near perfect (approximately 99%) sensitivity within SDs, indicating a strong competence of previous homologous counterpart identification for all SDs.

Subsequently, to remove the excessive false positives introduced by the realignments, SDrecall phases and assembles the realigned reads into longer micro-haplotypes via a graph-based phasing and assembly process. Given the diploid nature of human genome, SDrecall further adopts a binary integer linear constraint model to identify misaligned reads introduced by re-mapping, effectively reducing the number of false positives while maintaining the improved detection sensitivity in SDs. The resulting SDrecall variants are then merged into the callset from a traditional caller to complement their variant detection within SDs for downstream analysis. A symbolic scheme of the workflow is provided in Figure 1C.

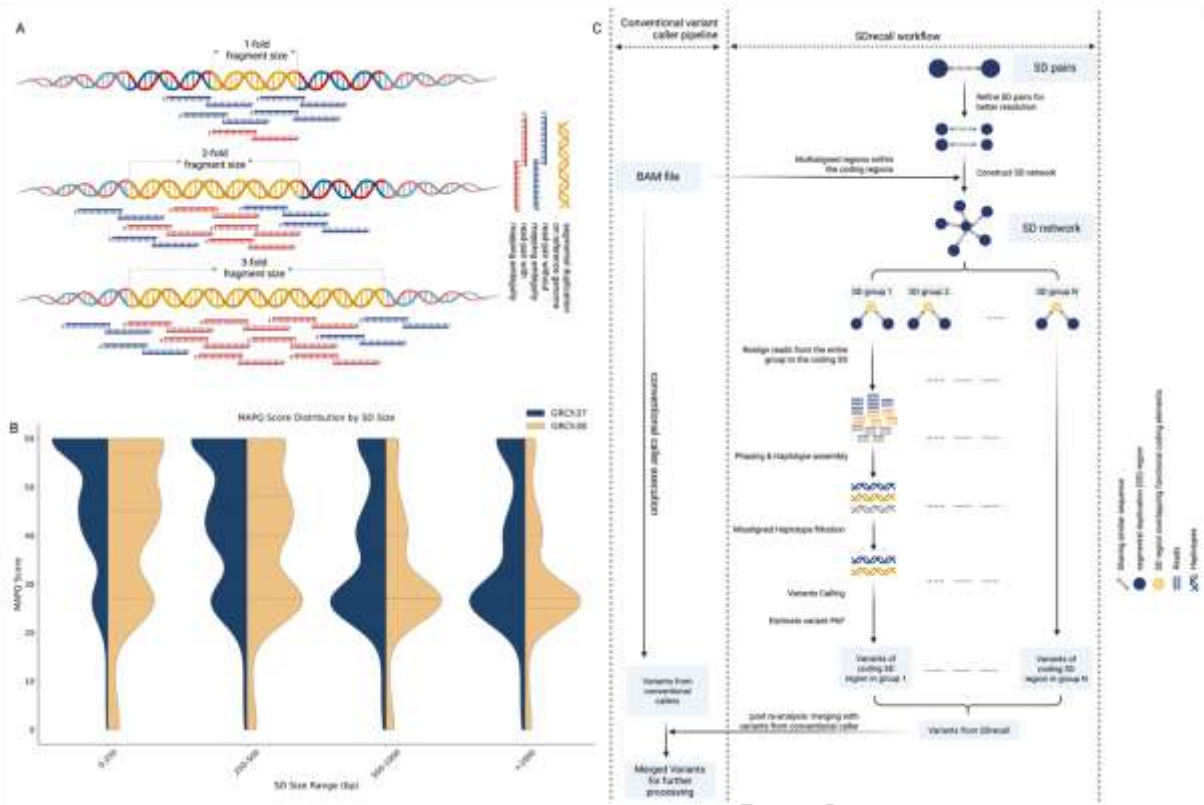

Fig 1. A. Schematic illustrating how paired-end reads, particularly those with insert sizes smaller than the SD length, can map ambiguously to multiple homologous SD regions. B. Mapping quality (MAPQ) score distributions for reads within SDs, stratified by SD size range, in GRCh37 (light blue) and GRCh38 (light orange) assemblies. C. Implementation of SDrecall. Size of each vertex is proportional to the size of SD region it represents. All paired SDs from WGAC are first refined to retain only paired subsegments with high similarity. Refined pairs overlapped with the multi-aligned regions from the input alignments are then used to construct a network of SDs which helps identify SDs grouped by their sharing homologous sequence. For each SD overlapped with coding regions, all reads aligned to the SDs from the same group are recruited and remapped to the coding SD. Realigned reads are then phased and assembled into longer haplotypes, which are then identified and filtered for misalignment. The remained reads are used to detect variants within coding SDs

and the total callset across all coding SDs are then merged with variants generated by traditional variant callers like GATK and DeepVariant.

Benchmarking against variants detected by Long Reads

To assess the sensitivity of variant detection across all pcSDs (which span approximately 30Mb) in the human genome, we benchmarked the performance of SDrecall using 6 samples from the Genome In a Bottle (GIAB) project. These samples, indexed HG002 through HG007, include two trio sets, the Ashkenazi trio (HG002, HG003, HG004) and the Chinese trio (HG005, HG006, HG007). On both GRCh37 and GRCh38 assemblies, GIAB provided comprehensive benchmark callsets derived from long-read sequencing data, including the v4.2.1 callset for all 6 samples and the Challenging Medical Relevant Genes (CMRG) benchmark callset for HG002. The GIAB CMRG callset includes benchmark variants in 273 challenging, medically relevant genes(30). We compared the combined callset of SDrecall and GATK/DeepVariant with these two benchmark callsets on both the GRCh37 and GRCh38 reference genomes. Our benchmarking test focused only on regions with sufficient coverage (regardless of MAPQs), low mappability, and high-confidence variants defined by GIAB. Detailed procedures for this benchmarking process are provided in Additional file 1: section 8.

In addition to benchmarking on GRCh37 and GRCh38, we also compared the merged callset of SDrecall and GATK/DeepVariant with CMRG benchmark callsets on the latest T2T-CHM13 assembly for sample HG002. The detailed methods(31, 32) and performance is included in Additional file 1: section 8.5. The detailed preparation process of this benchmarking is described in Additional file 1: section 2.

The performance of SDrecall is evaluated by three metrics, including variant detection sensitivity, variant detection precision and genotype accuracy within the benchmarking regions. The detection sensitivity is calculated as the fraction of true positives (TPs) according to the golden callset within the benchmark region that are successfully detected. The detection precision is calculated as the fraction of detected variants being TPs within the benchmark region. The genotype accuracy is calculated as the fraction of detected TPs having accurate zygosity.

Improved Detection Sensitivity

Using a callset combined from the variants detected by GATK/DeepVariant and the variants detected by SDrecall on unfiltered realignments, we detected around 99% true positives (TPs) on GRCh37 and 98.9% on GRCh38. In comparison, GATK alone captured only 77.4% TPs on GRCh37 and 73.8% TPs on GRCh38, while DeepVariant alone captured only 73.6% TPs on GRCh37 and 69.5% TPs on GRCh38. Although read realignment improved SD variant detection, it also introduced a substantial number of false positives (FPs), with precision rates at just 7.4%. To mitigate this, SDrecall implemented phasing and misalignment identification, significantly improving precision to around 40%, while maintaining sensitivities to around 95% on both GRCh37 and GRCh38 (Figure 2A). The detailed benchmarking performance data from the merged callset after misalignment elimination is displayed in Table 1. Additionally, in the combined callset between SDrecall and GATK, 90.0% of detected TPs on GRCh37 and 89.7% of detected TPs on GRCh38 have accurate genotypes. In the combined callset between SDrecall and DeepVariant, 91.1% of detected TPs were identified with accurate genotypes on both GRCh37 and GRCh38. When compared with the GIAB CMRG callset, SDrecall also performed similarly in

terms of sensitivity, precision and genotype accuracy. The detailed benchmarking data in terms of sensitivity and precision are provided in Table 1. As to the genotype accuracy, 100% of detected TPs on GRCh37 and 94.6% of detected TPs on GRCh38 have accurate genotypes.

To further test the performance of SDrecall, we mark the true positives detected by SDrecall exclusively and the true positives detected by either GATK or DeepVariant (Figure 2B) within the benchmark regions in 6 samples from GIAB. SDrecall exclusively captured TPs missed by traditional variant callers in 102 and 128 SD-overlapping protein coding genes on GRCh37 and GRCh38, respectively. A detailed list of variants is provided in the Additional file 2: Table S1.

Distinguishing rare and common variants to aid molecular diagnosis

A typical application of NGS in molecular diagnosis leverages population allele frequencies (PAFs) obtained from public databases to help filter out common variants that are unlikely to be disease causal (Figure 2C). However, most state-of-the-art databases, including gnomAD(33), 1000 Genome(34), and TOPMed(35), do not provide reliable PAF estimates for variants in SD regions. To achieve more accurate PAF estimations for variants from the SD regions, SDrecall provided a statistical framework that allows such estimation using in-house control cohorts. Since in-house cohorts are typically small and have insufficient statistical power compared to public population databases, we adopted a one-tailed binomial test (Figure 2E) to control the risk of overestimation, preserving the causal variant detection sensitivity of SDrecall while mitigating the risk of mis-classifying rare

variants as common. The detailed introduction of the binomial test is included in the Method section. With this step, SDrecall can identify common variants detected within SDs with limited support from public databases, which help us estimate the impact of the FPs introduced by SDrecall.

Reduced False Positives

In comparison with the v4.2.1 benchmark callset across 6 GIAB samples, the read realignment introduced 12710 and 15652 false positives (FPs) in the merged callset between SDrecall and GATK on GRCh37 and GRCh38, respectively. After applying phasing and misalignment identification (detailed in Methods), SDrecall successfully eliminated 10911 and 13261 FPs on GRCh37 and GRCh38, respectively. Given the limited sequence information captured by short paired-reads, even with our comprehensive misalignment elimination, it is impossible to eliminate all the misalignments. However, despite the precision of SDrecall cannot be comparable to the traditional callers like GATK and DeepVariant, the significantly reduced FP number still help eliminate most of the noise for final causal candidate selection in modern molecular diagnosis workflow (Figure 2C). To assess the impact of the FPs after the misalignment elimination on causal variant identification, we evaluated the number of rare (PAF < 0.01) and potentially deleterious (CADD(36) phred-scale score ≥ 20) FPs.

To supplement the regional gaps of population allele frequencies (PAFs) in SD regions in gnomAD (v2 exome dataset for GRCh37; v4 genome dataset for GRCh38), we constructed a callset from our in-house control cohort consisting of 498 samples. We applied SDrecall to these samples and merge the variants with BCFTools to a single multi-sample VCF file, providing comprehensive allele count

data for variants within SDs. Using this in-house callset and the aforementioned binomial test(Figure 2E), we classified 12434 (GRCh37) and 15423 (GRCh38) FPs as common variants. Additionally, 12066 (GRCh37) and 14495 (GRCh38) FPs were considered potentially neutral according to their CADD phred-scale scores (< 20). A Venn diagram illustrating the breakdown of these FPs for both assemblies is shown in Figure 2D.

Crucially, after identifying common and likely neutral FPs, only eight rare and potentially deleterious FPs remained on GRCh37, and five on GRCh38, for further expert review of their potential pathogenicity. This minimal number of clinically relevant FPs underscores the effectiveness of our filtering approach and highlights SDrecall's ability to significantly improve variant detection in SD regions while keeping the false positive burden low for downstream clinical interpretation.

Benchmark Callset	Assembly	Conventional caller	Sample	Conventional + SDrecall Sensitivity (%)	Conventional + SDrecall Precision(%)	Conventional Sensitivity (%)	Conventional Precision (%)	Total TPs (N)	Conventional + SDrecall rare pLoF FPs (N)	Benchmark Region Size (bp)
GIAB v4.2.1	GRCh38	GATK	HG002	93.9068	32.3058	75.448	85.9470	559	2	494255
			HG003	96.2672	36.3501	78.389	84.0000	509	1	480967
			HG004	94.3396	33.5796	74.5283	83.3333	527	2	475192
			HG005	94.3431	46.1607	72.2628	84.4350	548	1	375411
			HG006	96.0912	37.5318	73.127	81.7851	615	0	512867
			HG007	93.5433	35.9347	71.0236	78.1629	637	1	488962
		DV	HG002	93.2021	32.5625	70.8408	95.4327	559	1	494255
			HG003	96.6601	37.1321	73.0845	94.8980	509	0	480967
			HG004	93.7381	33.5826	67.9317	94.2105	527	0	475192
			HG005	94.3431	46.745	70.8029	92.1615	548	0	375411
			HG006	95.935	38.0155	67.4797	93.6795	615	0	512867
			HG007	92.9356	36.5658	66.876	92.6087	637	0	488962
	GRCh37	GATK	HG002	95.7143	41.875	79.3878	82.9424	491	3	398174
			HG003	93.9394	40.556	76.5152	80.4781	529	4	392668
			HG004	95.1579	36.6586	78.5263	79.7009	475	1	399150
			HG005	92.7239	50.611	78.1716	83.0677	536	3	305230
			HG006	95.1807	43.3046	76.2478	79.3165	581	1	378065
			HG007	95.3052	41.0931	77.6995	80.6846	427	0	319214
		DV	HG002	95.5193	42.8311	75.7637	93.7028	491	1	398174
			HG003	94.3289	41.7923	72.4008	94.8020	529	0	392668
			HG004	95.1579	36.9885	73.0526	91.7989	475	1	399150
			HG005	93.097	51.3374	76.6791	89.8901	536	2	305230
			HG006	94.4923	43.9552	71.6007	93.6652	581	0	378065
			HG007	94.8478	41.5811	72.1311	93.8650	427	0	319214
GIAB CMRG	GRCh38	GATK	HG002	97.2973	47.3684	70.2703	83.8710	37	0	35973
		DV		97.2973	47.3684	78.3784	90.6250	37	0	35973
	GRCh37	GATK		96.2963	40	70.3704	70.3704	27	1	29566
		DV		96.2963	41.2698	81.4815	84.6154	27	0	29566

Table 1. The benchmark results for the merged callset between SDrecall and GATK/DeepVariant(DV). SDrecall callset is merged with one conventional variant caller (GATK/DeepVariant) to generate a merged callset. Variants from the merged callset and the benchmark callset are sliced to select the ones located within the benchmark region, which is generated according to Additional file 1: section 8.4.

Rare and pLoF FPs are false positives that are neither identified as common variants nor identified as neutral/benign (CADD phred score < 20). GIAB stands for the Genome In A Bottle project and CMRG stands for Challenging Medically Relevant Genes. The displayed precision and recall rates are the benchmarking performance achieved after the misalignment filtration.

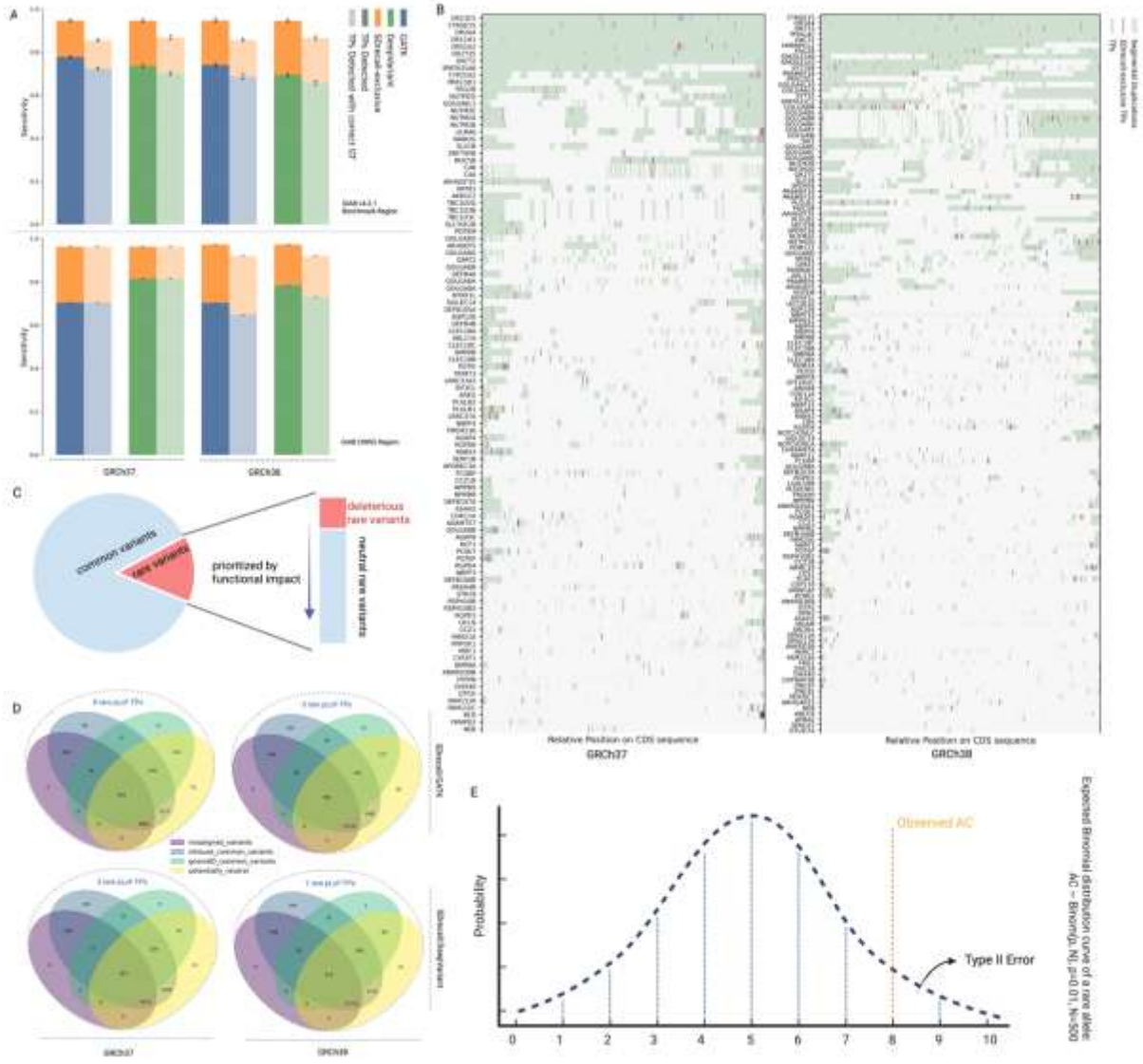

Fig 2. A. Recall rates of GATK/DeepVariant and GATK/DeepVariant–SDrecall callsets evaluated on six GIAB samples. The top panel compares calls to the GIAB v4.2.1 benchmark, and the bottom panel shows comparisons to the GIAB Challenging Medically Relevant Genes (CMRG) callset (available only for HG002).

Darker-shaded portions of the stacked bars represent true positives (TPs) detected without considering genotype (GT) accuracy, while lighter-shaded portions denote TPs with accurate GT calls.

B. Distribution of detected TPs along the coding sequences (CDS) of selected genes. Each horizontal bar spans a gene's entire CDS, with darker segments marking regions covered by segmental duplications. Blue vertical lines indicate TPs common to both callsets, and red vertical lines indicate TPs exclusively detected by SDrecall. Genes are selected for having SDrecall exclusively detected TPs located within the CDS of their canonical transcripts.

C. Schematic of the variant interpretation workflow: rare variants are prioritized by effect size and functional relevance using bioinformatic tools, and the final candidate list is manually reviewed to identify the causal variant(s).

D. Breakdown of false positives (FPs) across all six GIAB samples. FPs were categorized based on evidence of misalignment, designation as common (using an in-house control cohort or gnomAD), and prediction of neutrality by CADD; only the rare, deleterious FPs remained for expert review, with different counts observed for GATK–SDrecall and DeepVariant–SDrecall on GRCh37 and GRCh38.

E. Binomial test for distinguishing rare from common variants: for a variant covered by n haplotypes, the expected minor allele count is modeled by a Binom(n, 0.01) distribution; variants with observed counts significantly exceeding expectation ($p < 0.01$) are classified as common.

Application of SDrecall in real cases

We tested the performance of SDrecall in identifying disease-causing mutations in SD regions on three CGD patients. All three patients carry a

homozygous GT deletion at 75-76 bases in the coding sequence of *NCF1* gene(37), which leads to a frameshift effect with two consecutive altered codons. The new reading frame starts at the 26th amino acid residue of the protein sequence, where tyrosine is replaced by a histidine and is immediately followed by a premature stop codon, leading to a protein product with lost function, which further causes CGD. *NCF1* is known for its high average sequence similarity (> 98%) with its two paralogous pseudogenes, *NCF1B* and *NCF1C*. Notably, this mutation was missed by GATK in all three patients while it was later detected and confirmed by the GeneScan (Applied Biosystems) analysis. Using the same alignment files as input to GATK, SDrecall successfully detected the disease-causal deletion in all three patients. As a proof of concept, we highlighted the ability of the tool to capture clinically significant variants in challenging genomic duplications.

Computational Performance

We evaluated SDrecall on six GIAB samples, with detailed coverage depths and fragment sizes provided in Additional file 1: Table S1. Each analysis took approximately 4–6 hours to complete, using 10 CPUs when targeting the entire human exome. If runtimes from GATK/DeepVariant is considered, additional 6-7 hours are needed to process the mapped read data across the entire exome. In practical settings, phenotypic analysis often narrows the focus to a defined list of candidate genes, thereby substantially reducing the computational burden. In our practice, when targeting SD-overlapping exons of known causal genes for primary immunodeficiencies (around 500) (38), the runtime was significantly reduced to around 10 minutes per sample. The primary computational bottleneck is the identification of non-overlapping maximal cliques during phasing. This process has

been optimized using numba(39). All the codes are available at https://github.com/snakesch/SDrecall under the BSD-3 Clause license.

Discussion

In this work, we developed SDrecall, an NGS-based variant caller designed to sensitively detect SNVs and small indels (small variants) within segmental duplications (SDs). For any given SD, reads derived from it may be misassigned to its homologous regions and subsequently discarded by downstream variant callers due to high mapping ambiguity—thereby concealing genuine variants. SDrecall overcomes this limitation by retrieving all reads originating from the query SD, regardless of their initial alignment, thereby recovering the variants that would otherwise be missed.

To accomplish this, SDrecall first identifies all homologous counterparts (HCs) for any query SD in the human reference genome with high sensitivity and accuracy. Although WGAC provides a sensitive mapping of SD pairs, its binary output cannot disentangle overlapping SDs to reveal all HCs for a given SD. To overcome this limitation, SDrecall constructs a multiplex network that encodes both homologous and overlapping relationships among SDs. By traversing this network, SDrecall efficiently recovers all HCs for any query SD. While BLAT(40) can perform a similar search, it is optimized for short sequence queries and is at least 100× slower for batch queries of sequences ≥1 kb compared to our network-based approach. Similarly, tools like CORA(41) can track homologous counterparts by constructing a homology table with k-mers. However, this table is primarily designed as an intermediate step for accelerating read mapping, making it less suitable for our application. As evidence of our method’s comprehensiveness, SDrecall improves

small variant detection sensitivity to approximately 99% within protein-coding elements overlapping SDs (pcSDs) by realigning reads from all identified HCs to the corresponding pcSDs.

To reduce the FPs introduced by realigned reads originated elsewhere, SDrecall further phased and assembled realigned reads into distinct haplotypes based on the sharing variants among different pairs of reads. Because misaligned haplotypes are generally more similar to their true origin than to their aligned region and tend to carry an excess of variants relative to other overlapping haplotypes, we use binary integer linear constraint (BILC) programming to distinguish and filter out these misaligned haplotypes. By enforcing constraints that at most two haplotypes are correctly aligned in any well-covered region due to diploidy of human germline genome, our approach reliably selects the correctly aligned haplotypes for downstream analysis. Although there are scenarios that somatic mutations may occur in early development, resulting in more than two haplotypes among the covered reads, they are not sufficiently common to be considered by SDrecall as our goal primarily lies in rescuing germline variants within SDs. This strategy effectively removed 88% of FPs introduced by read realignment, pushing the precision rate from around 7% to 40% while maintaining the detection sensitivity at around 95%.

From a practical standpoint, SDrecall operates at a sensitivity-focused setting, which results in lower precision than traditional callers—a consequence of prioritizing variant recovery in segmental duplications (Figure 2A). This trade-off constrains its applicability to scenarios where sensitivity is paramount, such as rare disease diagnostics, where there is a strong prior expectation of a rare, high-impact variant and comprehensive variant ranking is routine. In this context (Figure 2C), the additional rare variants recovered in SDs increase the likelihood of identifying a

causal allele, while most false positives are deprioritized by automated filters based on population frequency and predicted functional effect. Consequently, the residual FP burden after filtering is minimal, enabling thorough inspection of low-mappability regions without substantially complicating causal variant determination. Nevertheless, any causal candidate identified by SDrecall should be confirmed using orthogonal assays such as long-range PCR(42), GeneScan fragment analysis(43), or—where feasible—long-read sequencing. SDrecall is particularly useful for patients whose phenotype implicates multiple candidate loci including SD-rich regions, because it is more scalable than locus-specific wet-lab assays and substantially more cost-effective than genome-wide long read sequencing. In such workflows, SDrecall serves as a pre-screening tool to nominate a small set of plausible causal variants for targeted validation, supporting both large-scale reanalysis of legacy NGS datasets and routine molecular diagnostics for incoming patients.

From a gene-level perspective, we tallied the number of additional variants called by SDrecall across all genes in six GIAB samples (Figure 2B). On average, exonic variants were detected in 201 functional coding genes on GRCh37 and in 233 functional coding genes on GRCh38. Upon reviewing the ClinVar data released on July 24, 2024, we identified 14 genes on both GRCh37 and GRCh38 that are reported as pathogenic for certain Mendelian disorders. These genes include *CYP21A2*(44, 45), *NCF1*(46), *NEB*(47), *PMS2*(48), *TTN*(49), *CBS*(50), *KCNE1*(51), *NUTM2B*(52), *OTOA*(53), *HBA2*(54), *RBFOX3*(55), *PI4KA*(56), *RAB43*(57), *SIK1*(58) and *STRC*(59). A detailed list of variants exclusively detected by SDrecall and their overlapping genes, as well as the diseases from ClinVar are recorded in Additional file 2: Table S1.

SDrecall's effectiveness depends on both the read depth and fragment size of the input alignments. In our benchmarks, for example, no variants were recovered in *SMN1* or *SMN2* (genes associated with spinal muscular atrophy(18)) because their exons did not have sufficient coverage in any of the six GIAB samples. Moreover, the unusually large average fragment size (≈600 bp; see Additional file 1: Table S1) in these samples increases the likelihood that read pairs span the junctions between SDs and unique genomic regions, thereby reducing mapping ambiguities. Despite these factors—which may underestimate SDrecall's true potential—we observed that nearly 800 genes, including *SMN1* and *SMN2*, exhibited a marked shift toward higher MAPQ scores after realignment. This MAPQ distribution shift underscores the broader benefit of SDrecall for improving variant detection, as detailed in Additional file 1: section 9.

Although SDrecall enables sensitive variant detection in SD regions, its effectiveness and accuracy are affected by several limitations. First, the realignment-based variant recovery approach appears to be less effective in more challenging genomic contexts such as low-complexity regions, long stretches of tandem repeats, and overlapping duplications. During development, we observed that a few realigned reads are having MAPQs lower than 40 and they are mainly observed in regions listed above. A common feature shared by these regions is that similar sequences are often found adjacent to or even overlapping with each other. Therefore, realigned reads may still have multiple similar matches within the unmasked region, leading to high mapping ambiguity. A symbolic diagram is provided to illustrate the mechanism (Additional file 1: Fig. S11). For example, in the 1q21.1 region on chromosome 1, enriched with *NBPF* gene family duplications, adjacent or overlapping duplicated sequences hinder unique realignment(60). Second, although SDrecall achieves an

average genotype accuracy of around 90%, some variants still are detected with underestimated dosage, leading to an underestimation of their effect sizes. Further improvements are needed to enhance genotype detection accuracy.

Conclusions

To summarize, SDrecall represents a significant advancement in variant detection within SDs for NGS data. This accomplishment has remained elusive for many years despite numerous efforts. It is the first tool to construct a systematic and accurate SD network of human reference assemblies while addressing complexities such as nested duplications, allowing for efficient and extensive extraction of groups of regions sharing homologous sequences for further analysis. It significantly improved the variant detection sensitivity to 95% across all SDs while minimizing the count of relevant FPs left for causal candidate selection. All these benefits have been demonstrated in real CGD patients by SDrecall, capturing causal variants previously undetected by conventional NGS analysis. SDrecall addresses what was once considered an insurmountable challenge in NGS data analysis, filling a critical gap in variant detection and molecular diagnosis of Mendelian Disease patients. It can be a valuable asset for both individual molecular diagnosis and legacy data re-analysis, significantly enhancing our ability to detect and revisit genetic variations in segmental duplications.

Methods

Establishment of a robust SD network

To accurately identify regions of SDs on the human reference genome assemblies, a comprehensive paired SD map was constructed using whole-genome assembly comparison (WGAC)(8). However, not all segments in SD pairs are sufficiently similar to cause mapping ambiguity. To extract the segments with sufficiently high degrees of similarity, we employed minimap2(5) to compare each SD against its paired counterpart, retaining only the sub-segments with a local mismatch rate (including gaps) below 10% (Figure 3A, Additional file 1: Fig. S2). To improve computational efficiency, the alignment file is filtered to select only SDs that overlap coding regions (or user-defined regions of interest), exceed the average fragment length, and have a read depth ≥ 3. This filtering process creates a tailored set of SDs based on each input alignment file for subsequent analysis.

A simple pairwise map of SDs is insufficient to efficiently identify all homologous counterparts of a given SD region. To accurately capture the complex relationships among SDs, we constructed a multiplex network from the refined SD segments (Figure 3B). This network incorporates two types of connections: sequence similarity (SS) edges, which connect SD regions that share a high degree of sequence identity, and physical overlap (PO) edges, which connect SD regions that overlap on the reference genome. Analysis of short-read data from the HG002 sample (GIAB) revealed over 10k SD regions interconnected by 7k SS edges and 65k PO edges (Additional file 1: section 3), highlighting the extensive and non-binary nature of SD relationships. The subnetwork surrounding the *NCF1* gene and its paralogs *NCF1B* and *NCF1C* (Figure 3C) exemplifies this complexity.

To efficiently extract all homologous counterparts (HCs) of any query SD, we developed an in-house method based on Dijkstra's algorithm(61), implemented via the graph-tool Python interface(62) (Additional file 1: Fig. S4). This customized algorithm enables the extraction of all HCs for each SD, forming groups of SDs sharing homologous sequences for subsequent realignment of the mapped reads. The detailed methodology(63-65) is described in Additional file 1: section 3.

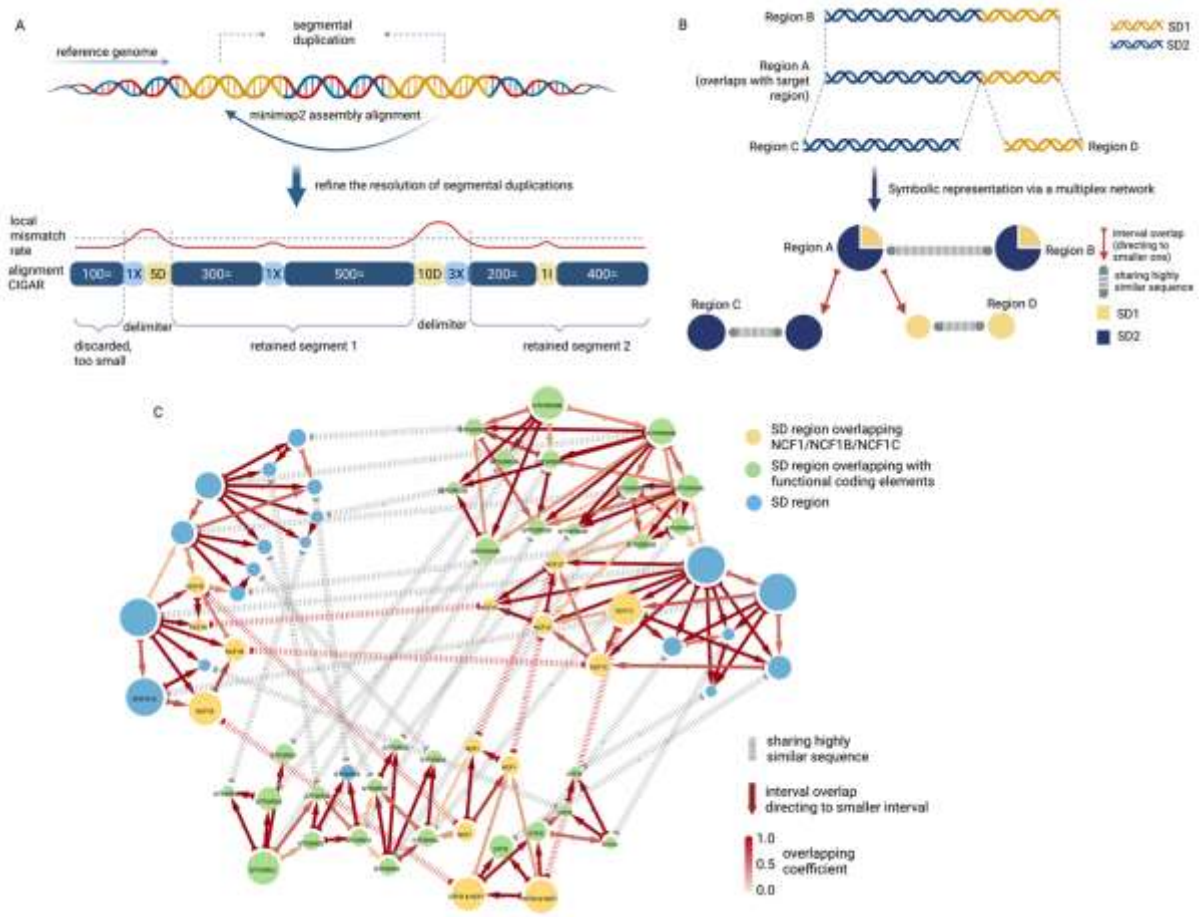

Fig 3. A. Initial SD pairs, defined by whole-genome assembly comparison, are refined using minimap2 to identify sub-segments of high sequence similarity. The alignment between SD regions is represented by a CIGAR string, which encodes matches (=), mismatches (X), insertions (I), and deletions (D) in the alignment.

B. Symbolic representation of convoluted relationships between SD regions (A, B, C, and D). Pie charts for two vertices show the fraction of the region covered by two distinct homologous sequences.

C. An *NCF1/NCF1B/NCF1C*-specific subnetwork extracted from the complete SD network, highlighting the relationships among SD regions overlapping this gene family. The red dashed edges highlighted the sequence sharing relationship among *NCF1/NCF1B/NCF1C*-overlapping SD regions. Red arrows stand for the overlapping relationship between SD regions and are pointed to the smaller one between the two.

Realignment

Leveraging the SD network, SDrecall identifies groups of SDs sharing homologous sequences. Within each group, SDrecall recruits reads from all group members and realign them to each SD that overlaps with protein coding regions (pcSD) in the same group, respectively (Figure 4A). Since multiple SD regions within a group may overlap with protein coding regions, each SD region can act as either a source or a target during multiple reciprocal realignment processes. These realignments ensure that every pcSD is fully covered by all reads that may originate from it, thereby capturing all variants within the pcSD. In the subsequent variant calling using BCFtools(66), the realignments enabled the detection of, on average, approximately 99% of the true positives (TPs) within the benchmark regions (as defined in the Benchmark section of the Results). This high capture rate was observed across six GIAB samples for both the GRCh37 and GRCh38 assemblies, when compared to their respective benchmark callsets. This near-perfect variant detection sensitivity proves that the multiplex SD network accurately identifies all the

HCs for each pcSD, which offers sufficient read coverage for downstream variant detection.

While this realignment strategy maximizes variant detection sensitivity, implementing stringent false positive (FP) control can further enhance its effectiveness in molecular diagnosis. Since SDrecall calls variants based on reads collected from pcSDs and their corresponding HCs, a large number of FPs can result, either from variants originated from the HCs or the natural sequence differences among paralogous regions on the reference genome assembly (paralogous sequence variants, PSVs). To limit the number of PSVs, reference sequences of HCs were aligned to the pcSD (Figure 4B) to create a set of intrinsic alignments, which are used in downstream FP elimination (Additional file 1: Fig. S7). A detailed description of the entire process above is provided in Additional file 1: section 5.

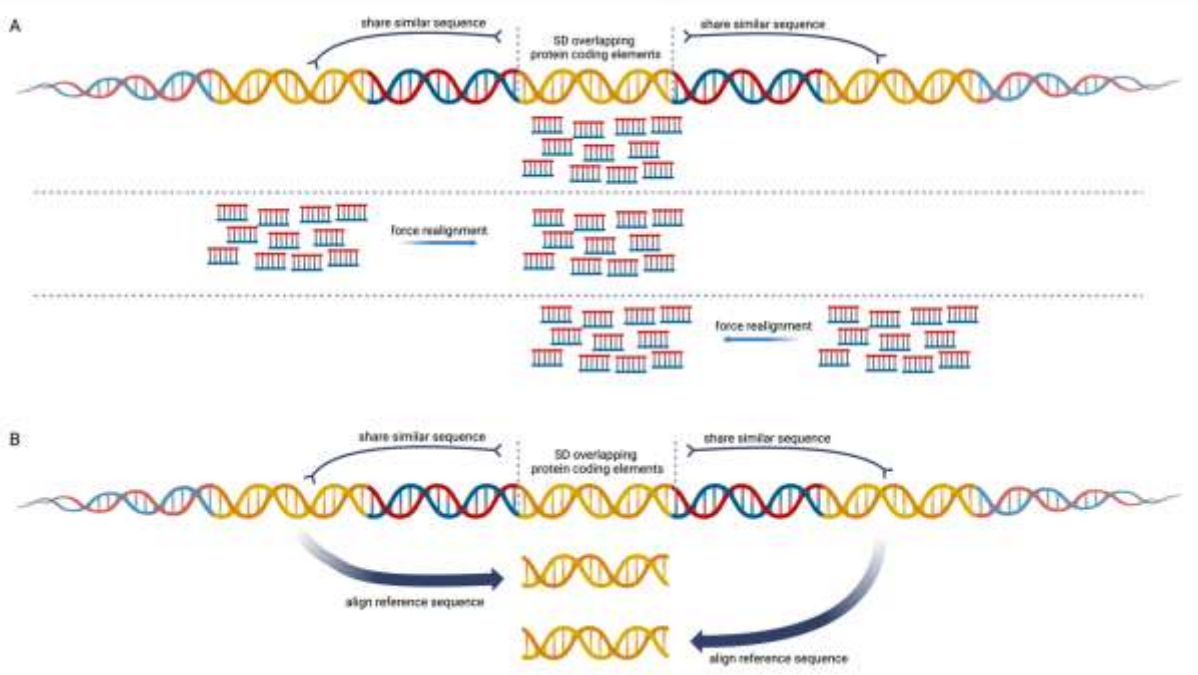

Fig 4. A. Overview of the realignment step in SDrecall. Using default settings, SDrecall recovers variants in segmental duplications encapsulating functional coding

elements. SDrecall pools reads mapped to the functional coding SD and its homologous counterparts and realigns all reads against the functional coding SD. B. Realignment of reference sequences to identify false positives. SDrecall aligns reference sequences of homologous counterparts to the functional coding SD to identify variants originated from reference sequences of homologous counterparts.

Phasing of realigned reads

To reduce false positives (FPs), SDrecall employs graph-based phasing to group realigned reads into haplotypes. Paired-end reads (termed "fragments" hereafter as paired reads are sequenced from short DNA fragments) are represented as vertices in a graph, with edges connecting fragments potentially originating from the same haplotype. Each edge requires two connected fragments to share identical sequences within their overlaps. Given that SNV density within SDs is 1.47 SNVs per kbp(67), haplotypes within SD regions contain much less non-reference alleles than reference alleles, therefore larger number of variants (SNVs and indels) on the shared sequence between two fragments indicates a higher likelihood that they originate from the same haplotype (Figure 5A). Therefore, the number of shared variants, as well as the overlap region size are used for calculating edge weights, which is used to indicate the likelihood that two overlapping fragments originate from the same haplotype.

For the fragments that do not overlap with each other, there is no observed evidence to determine whether they can possibly originate from the same haplotype. Therefore, we connect them by edges with zero weights, indicating zero confidence for originating from the same or distinct haplotypes. As a result, only pairs of fragments with high quality mismatches on the sequences within their overlapping

region remain unconnected due to the observed evidence to reject the possibility that they originate from the same haplotype (Figure 5A). To group fragments into distinct haplotypes, we identify groups of fragments where no evidence is observed to reject potential haplotype sharing between any pair of fragments within the group. This strategy is implemented by an in-house developed algorithm named Greedy Clique Expansion, which is based on the “seed-and-expand” heuristics for clique identifications(68), to identify disjoint maximal cliques (a maximal clique means a group of vertices where all pair of vertices are connected, and the group cannot be further expanded to include any more adjacent vertex, disjoint emphasizes that multiple maximal cliques that do not overlap with each other) within a graph (Figure 5B). As a result, most realigned reads are grouped and assembled into longer consensus sequences representing distinct micro-haplotypes for downstream analysis. A detailed explanation of this process(39, 69-71) is provided in Additional file 1: section 6.

Misalignment identification

Considering the diploid nature of the human germline genome, at most two haplotypes are expected in a region with sufficient coverage. In general, haplotypes carrying more variants are more prone to incorrect alignment due to paralogous sequence variants (PSVs) and variants originated from homologous counterparts (HCs). To address this, in a well-covered region, the ideal approach would be to select the two haplotypes with the fewest variants, presuming they represent the true haplotypes. However, directly comparing all assembled haplotypes within a fixed-size window is impractical because haplotypes vary considerably in their lengths and genomic spans. Therefore, we adopt a sliding window to scan through the well-

covered region and compare haplotypes encompassing the localized region at each position of the window. For each haplotype, we calculate the likelihood of misalignment considering all the haplotype comparisons it is involved in. The calculated likelihoods are then integrated to a BILC programming model(72) (Equation 1), which is used to identify the optimal combination of haplotypes considered as correctly aligned.

Given a total number of N haplotypes assembled by the remapped reads, the BILC sets a binary variable $X_i \in \{X_1, X_2, \dots, X_N\}$ to each haplotype i, indicating whether it is correctly aligned (X_i = 1) or misaligned (X_i = 0). Additionally, each X_i is linked to a positive coefficient ε_i, which indicates the likelihood of haplotype i being misaligned. In general, BILC tries to find a combination of X_i that minimizes the output of the linear combination, hence minimizing the total likelihood of misalignment for the determined correctly aligned haplotypes. As previously mentioned, we adopt a sliding window to scan through the well-covered regions. At each location of the sliding window, we set a constraint that only two haplotypes enclosing the window are correctly aligned (sum of X_i equals 2). Given that each haplotype can enclose the window at multiple stops, every X_i is bound to several constraints (Figure 5C), which helps avoid the trivial solution (all X_i are set to 0) to the optimization. The BILC was solved by the efficient integer linear programming solver HiGHs(72). A detailed explanation of the BILC, especially to the calculation of coefficients are provided in Additional file 1: section 7.

$$\min_{X_i \in \{0,1\}} \sum_{i=1}^{N} \varepsilon_i X_i$$

Equation 1. The Binary linear integer formula where $X_i \in \{0, 1\}$ indicates whether haplotype i is correctly mapped (1) or mismapped (0), N is the total number of

haplotypes, ε_i is the coefficient of each binary integer which is proportional to the likelihood of misalignment.

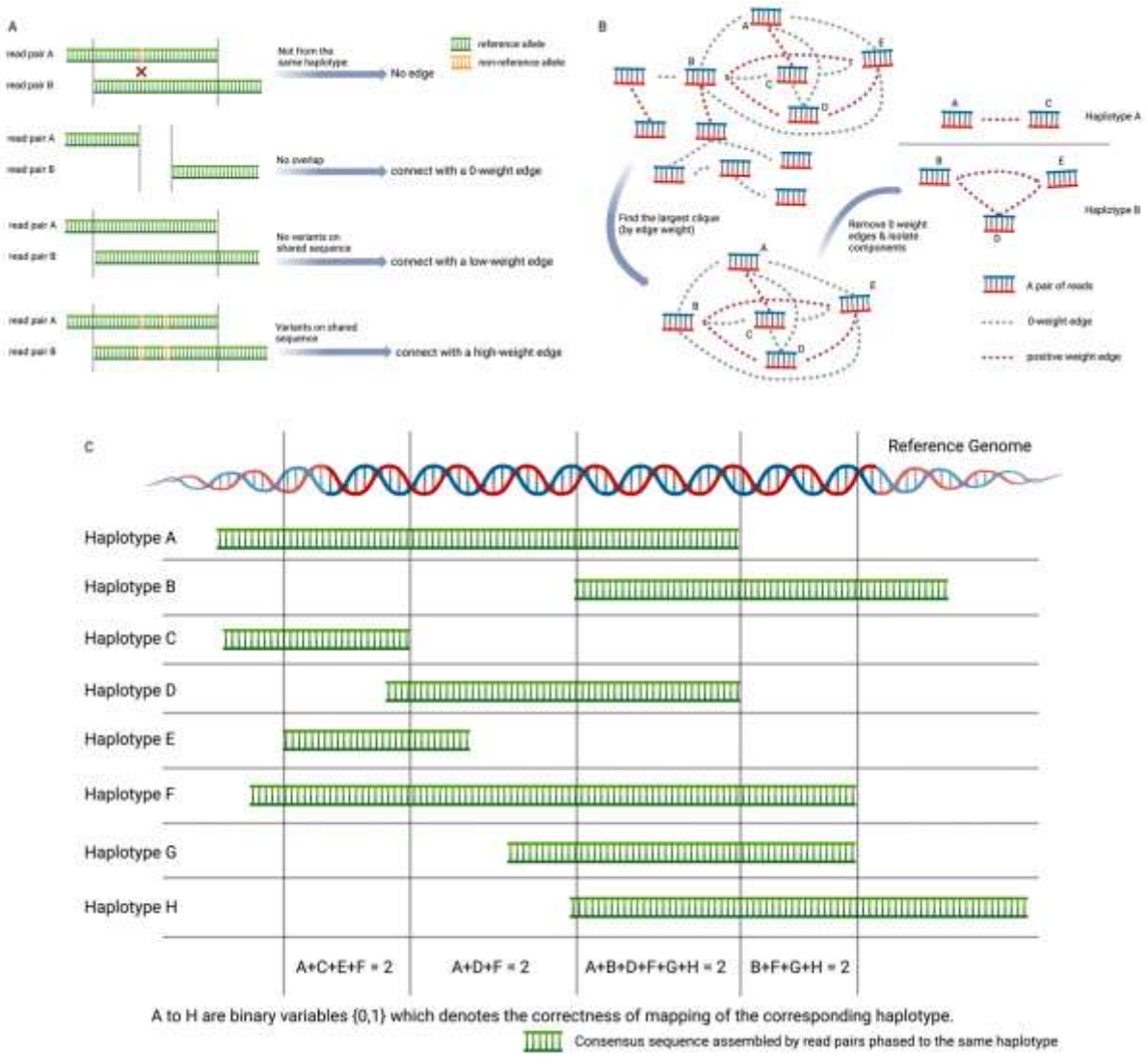

Fig 5. A. Stratified edge weight determination across different overlap situations between two read pairs. From top to bottom, (1) presence of high-quality mismatches on the sequences between two read pairs within their overlaps, (2) No overlaps between two read pairs, (3) identical sequence (all reference alleles) shared between two read pairs within their overlaps, (4) identical sequence (with non-reference alleles) shared between two read pairs within their overlaps.

B. Maximal clique search algorithm for haplotype phasing. Read pairs with strong evidence to reject the haplotype sharing are disconnected in the constructed network.

SDrecall first identifies subgraphs of interconnected read pairs (cliques), preferentially with larger aggregated edge weights (maximal cliques). Within each identified clique, SDrecall next removed zero-weight edges and isolated graph components with nonzero-weight edges. The isolated components represent the most likely haplotypes supported by the observations from the alignments.

C. Symbolic scheme of constraint establishment for the binary integer linear constraint programming. The diploid nature of the human genome asserted that only two haplotypes could be correctly aligned in any well-covered region. Based on this fact, as the window sliding through, we set up a constraint within each position of the window that the total summation of all X_i corresponding to the haplotypes enclosing the window equals to 2.

Variant Calling

Following the misalignment elimination, SDrecall used BCFtools for efficient variant detection across all regions piled with filtered realigned reads. Though followed by the misalignment elimination, the realignment process still inevitably introduces misaligned reads that barely harbor non-reference alleles, which might dilute allele depth of variant alleles at certain sites, leading underestimation of the variant dosage. To address this issue, SDrecall adjusts the genotype calls for certain variants based on the number of haplotypes supporting the variant allele, as determined from the phased alignments. The detailed process is described in Additional file 1: section 8.2.

Since SDrecall is intended to be supplementary to traditional variant callers, variants called by SDrecall were merged into those called by GATK HaplotypeCaller(73) or DeepVariant for downstream benchmarking, where the

GATK/DeepVariant callset is directly acquired from the input alignments before any realignment manipulation. Given the superior genotype accuracy of GATK/DeepVariant, priority is given to the GATK/DeepVariant callset whenever there is a conflict in genotype calling during the merging process. The detailed merging process is described in Additional file 1: section 8.3.

Identify Common Variants within Segmental Duplications

As previously introduced in the Results section, the variants within SDs are usually not well covered in public databases since the population variants are all derived from NGS data. In this study, we applied SDrecall to an inhouse control cohort and classify the variants as common/rare based on the allele frequency data from the limited cohort sample. Assuming variant allele count (AC) at a given site follows a binomial distribution with n as the total allele number of alleles and p as the PAF(Equation 2). SDrecall sets the null hypothesis that the query variant has a PAF smaller than 0.01, a typical and empirical cutoff between rare and common variants(74-76). Next, we tested the null hypothesis to see whether the observed AC was large enough to reject the null. The significance level is set at 0.01 to establish an upper bound for the tolerable probability of Type II error (Figure 2E). Therefore, any variants with an observed AC that fall within the bottom 99% of the distribution are classified as rare, allowing for conservative retention of rare variants to ensure sensitivity in identifying potential causal candidates.

$$AC \sim Binom(n, p)$$

Equation 2. Variant allele count is modeled by a binomial distribution with a total of n alleles and a probability of the true population allele frequency of the variant allele in the general population.

Declarations

Ethics approval and consent to participate

This study used both publicly available sequencing datasets and in-house whole-genome sequencing data from patients with immunological disorders. Use of the in-house samples was approved by the Institutional Review Board of the University of Hong Kong / Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) under the protocol “Database and Patient Registry on Immunological Disorders” (IRB reference UW 08-301). Written informed consent for participation, and for storage and research use of clinical and genetic data, was obtained from all participants or from their parents/legal guardians in the case of minors, in accordance with this approved protocol. All procedures involving human participants were conducted in accordance with the ethical standards of the HKU/HA HKW IRB, the Hospital Authority, and the University of Hong Kong, and with the principles of the **Declaration of Helsinki** and its later amendments. Analyses of public benchmark datasets (e.g. Genome in a Bottle and related resources) were performed on fully de-identified data available from open repositories and did not require additional institutional ethics approval.

Consent for publication

For the in-house samples, written informed consent for publication of anonymised research data, including genomic data, was obtained from all participants or from their parents or legal guardians in the case of minors, as required by HKU/HA HKW IRB protocol **UW 08-301**. This manuscript does not

contain any individual person's identifiable data (such as photographs, detailed pedigrees, or case descriptions that could enable re-identification). For the public benchmark datasets, consent for publication was obtained by the original data generators, and only de-identified data were used in this study.

Availability of Data and Materials

Most datasets used in this study were obtained from public repositories. Mapped whole-genome sequencing data for the six Genome in a Bottle (GIAB) samples (HG002, HG003, HG004, HG005, HG006, HG007) used for SDrecall benchmarking were downloaded from the GIAB project; accession numbers and direct download URLs for the original BAM files are listed in Additional file 1: Section 2. Segmental duplication annotations for different human genome assemblies were obtained from the UCSC Genome Browser annotation tracks, and the download URLs are provided in Additional file 1: Section 1. Download URLs for the benchmark region BED files and benchmark VCF files for the six GIAB samples are listed in Additional file 1: Sections 8.4 and 8.5.

SDrecall is available as open-source software under the BSD 3-Clause License at GitHub (https://github.com/snakesch/SDrecall)(77). The exact version of the code used in this manuscript has been archived on Zenodo (DOI: 10.5281/zenodo.17667346)(78) under the CC BY 4.0 license. The supplementary dataset, including the cohort VCF file containing 498 in-house control samples used in this study for allele frequency estimation, has been deposited on Zenodo (DOI: 10.5281/zenodo.17669242)(79) under the CC BY 4.0 licence.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by grants from the Hong Kong Society for the Relief of Disabled Children and Jeffrey Modell Foundation (YLL). WY thanks support from Health and Medical Research Fund (HMRF) of Hong Kong Government (Project No.: 10212696 and PR-HKU-7).

Author Contributions

Xing Tian Yang and Chun Hing She are responsible for the development of SDrecall. The concept of SDrecall was designed by Wanling Yang, Xing Tian Yang, and Yu Lung Lau. Xing Tian Yang led the design and implementation of algorithms used in SDrecall, conducted the benchmarking analyses, and drafted the initial manuscript. Chun Hing She contributed to the design and implementation of the algorithms, assisted with manuscript preparation, and led the packaging of SDrecall scripts for future distribution. Caicai Zhang supported the implementation of the integer linear constraint programming. Daniel Leung, Jing Yang, Koon-Wing Chan, Jaime S Rosa Duque, and Yu Lung Lau contributed to SDrecall's application in real clinical cases by gathering data, performing phenotype diagnosis, and GeneScan analysis. Wanling Yang and Yu Lung Lau supervised the project and provided important feedback on data interpretation and manuscript preparation. They also secured the funding for the project and corresponded with the journal. All authors read and approved the final manuscript.

Acknowledgements

All figures are assembled with BioRender.com. While Figure 1A/1C, Figure 2C/2E, Figure 3A/3B, Figure 4, Figure 5, and Additional file 1: Figure S4-S8 are created with BioRender.com

Authors' information

Not applicable

Footnotes

Not applicable

Additional files

Additional file 1: Supplementary material (DOCX)

Contains detailed supplementary methods and most supplementary figures (Fig. S1–S9) with their legends. It also includes Supplementary Tables S1 and S2 with their captions.

Additional file 2: Table S1 (XLSX)

Comprehensive list of variants exclusively detected by SDrecall in segmental duplications, including overlapping genes and associated diseases annotated from ClinVar.

Additional file 3: Fig. S1 (PDF)

Shifts in mapping quality (MAPQ) values for reads overlapping genes on GRCh37 before and after SDrecall realignment, shown per gene across the genome.

Additional file 4: Fig. S1 (PDF)

Shifts in mapping quality (MAPQ) values for reads overlapping genes on GRCh38 before and after SDrecall realignment, shown per gene across the genome.

Peer review information

Yafei Mao and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

The peer-review history is available in the online version of this article.

References

1. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. TRENDS in Genetics. 2001;17(11):661-9.
2. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297(5583):1003-7.
3. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome research. 2001;11(6):1005-17.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
5. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9.
7. Genome in a bottle—a human DNA standard. Nature Biotechnology. 2015;33(7):675-.
8. Rouchka EC, Gish W, States DJ. Comparison of whole genome assemblies of the human genome. Nucleic acids research. 2002;30(22):5004-14.
9. Išerić H, Alkan C, Hach F, Numanagić I. Fast characterization of segmental duplication structure in multiple genome assemblies. Algorithms for Molecular Biology. 2022;17(1):4.
10. Dallery J-F, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, et al. Gapless genome assembly of Colletotrichum higginsianum reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. BMC genomics. 2017;18:1-22.
11. Delehelle F, Cussat-Blanc S, Alliot J-M, Luga H, Balaresque P. ASGART: fast and parallel genome scale segmental duplications mapping. Bioinformatics. 2018;34(16):2708-14.
12. Pu L, Lin Y, Pevzner PA. Detection and analysis of ancient segmental duplications in mammalian genomes. Genome research. 2018;28(6):901-9.
13. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330(6004):641-6.
14. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science. 2022;376(6588):eabl3533.
15. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.
16. Roos D, Kuhns DB, Maddalena A, Roesler J, Lopez JA, Ariga T, et al. Hematologically important mutations: X-linked chronic granulomatous disease (third update). Blood Cells, Molecules, and Diseases. 2010;45(3):246-65.

17. Lefebvre S, Bürglen L, Reboullet S, Clermont O, Burlet P, Viollet L, et al. Identification and characterization of a spinal muscular atrophy-determining gene. Cell. 1995;80(1):155-65.
18. Wirth B. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). Human mutation. 2000;15(3):228-37.
19. White PC, Speiser PW. Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Endocrine reviews. 2000;21(3):245-91.
20. Hruska KS, LaMarca ME, Scott CR, Sidransky E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Human mutation. 2008;29(5):567-83.
21. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005;33(suppl_1):D514-D7.
22. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024:110842.
23. Feng Y, Ge X, Meng L, Scull J, Li J, Tian X, et al. The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic SMN1 copy-number and sequence variant analysis by massively parallel sequencing. Genetics in Medicine. 2017;19(8):936-44.
24. Trabucco SE, Gowen K, Maund SL, Sanford E, Fabrizio DA, Hall MJ, et al. A novel next-generation sequencing approach to detecting microsatellite instability and pan-tumor characterization of 1000 microsatellite instability–high cases in 67,000 patient samples. The Journal of Molecular Diagnostics. 2019;21(6):1053-66.
25. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics. 2017;18(10):599-612.
26. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine. 2015;17(5):405-23.
27. Ebbert MT, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome biology. 2019;20:1-23.
28. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 0 1- 0 33.
29. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-7.
30. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nature biotechnology. 2022;40(5):672-80.
31. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
32. O'Connell KA, Yosufzai ZB, Campbell RA, Lobb CJ, Engelken HT, Gorrell LM, et al. Accelerating genomic workflows using NVIDIA Parabricks. BMC Bioinformatics. 2023;24(1):221.

33. Koch L. Exploring human genomic diversity with gnomAD. Nature Reviews Genetics. 2020;21(8):448-.
34. Siva N. 1000 Genomes project. Nature biotechnology. 2008;26(3):256-7.
35. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.
36. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2019;47(D1):D886-D94.
37. Iwata M, Nunoi H, Yamazaki H, Nakano T, Niwa H, Tsuruta S, et al. Homologous dinucleotide (GT or TG) deletion in Japanese patients with chronic granulomatous disease with p47-phox deficiency. Biochemical and biophysical research communications. 1994;199(3):1372-7.
38. Bousfiha A, Moundir A, Tangye SG, Picard C, Jeddane L, Al-Herz W, et al. The 2022 update of IUIS phenotypical classification for human inborn errors of immunity. Journal of Clinical Immunology. 2022;42(7):1508-20.
39. Lam SK, Pitrou A, Seibert S, editors. Numba: A llvm-based python jit compiler. Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC; 2015.
40. Kent WJ. BLAT—the BLAST-like alignment tool. Genome research. 2002;12(4):656-64.
41. Yorukoglu D, Yu YW, Peng J, Berger B. Compressive mapping for next-generation sequencing. Nature biotechnology. 2016;34(4):374-6.
42. Davies PA, Gray G. Long-range PCR. PCR Mutation Detection Protocols: Springer; 2002. p. 51-5.
43. Kulkarni M, Desai M, Gupta M, Dalvi A, Taur P, Terrance A, et al. Clinical, immunological, and molecular findings of patients with p47phox defect chronic granulomatous disease (CGD) in Indian families. Journal of clinical immunology. 2016;36(8):774-84.
44. Witchel SF. Congenital adrenal hyperplasia. Journal of pediatric and adolescent gynecology. 2017;30(5):520-34.
45. Speiser PW, Azziz R, Baskin LS, Ghizzoni L, Hensle TW, Merke DP, et al. Congenital adrenal hyperplasia due to steroid 21-hydroxylase deficiency: an Endocrine Society clinical practice guideline. The Journal of Clinical Endocrinology & Metabolism. 2010;95(9):4133-60.
46. Roos D, de Boer M, Köker MY, Dekker J, Singh-Gupta V, Åhlin A, et al. Chronic granulomatous disease caused by mutations other than the common GT deletion in NCF1, the gene encoding the p47phox component of the phagocyte NADPH oxidase. Human mutation. 2006;27(12):1218-29.
47. Lehtokari VL, Kiiski K, Sandaradura SA, Laporte J, Repo P, Frey JA, et al. Mutation update: the spectra of nebulin variants and associated myopathies. Human mutation. 2014;35(12):1418-26.
48. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. Gastroenterology. 2008;135(2):419-28. e1.
49. Gerull B, Gramlich M, Atherton J, McNabb M, Trombitás K, Sasse-Klaassen S, et al. Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. Nature genetics. 2002;30(2):201-4.
50. Kraus JP, Janošík M, Kožich V, Mandell R, Shih V, Sperandeo M, et al. Cystathionine β-synthase mutations in homocystinuria. Human Mutation. 1999;13(5):362-75.

51. Splawski I, Shen J, Timothy KW, Lehmann MH, Priori S, Robinson JL, et al. Spectrum of mutations in long-QT syndrome genes: KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. Circulation. 2000;102(10):1178-85.
52. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. Nature genetics. 2019;51(8):1222-32.
53. Sugiyama K, Moteki H, Kitajiri S-i, Kitano T, Nishio S-y, Yamaguchi T, et al. Mid-frequency hearing loss is characteristic clinical feature of OTOA-associated hearing loss. Genes. 2019;10(9):715.
54. Garrick D, De Gobbi M, Samara V, Rugless M, Holland M, Ayyub H, et al. The role of the polycomb complex in silencing α-globin gene expression in nonerythroid cells. Blood, The Journal of the American Society of Hematology. 2008;112(9):3889-99.
55. Lal D, Reinthaler EM, Altmüller J, Toliat MR, Thiele H, Nürnberg P, et al. RBFOX1 and RBFOX3 mutations in rolandic epilepsy. PloS one. 2013;8(9):e73323.
56. Pagnamenta AT, Howard MF, Wisniewski E, Popitsch N, Knight SJ, Keays DA, et al. Germline recessive mutations in PI4KA are associated with perisylvian polymicrogyria, cerebellar hypoplasia and arthrogryposis. Human Molecular Genetics. 2015;24(13):3732-41.
57. Bellucci A, Longhena F, Spillantini MG. The role of Rab proteins in Parkinson's disease synaptopathy. Biomedicines. 2022;10(8):1941.
58. Hansen J, Snow C, Tuttle E, Ghoneim DH, Yang C-S, Spencer A, et al. De novo mutations in SIK1 cause a spectrum of developmental epilepsies. The American Journal of Human Genetics. 2015;96(4):682-90.
59. Moteki H, Azaiez H, Sloan-Heggen CM, Booth K, Nishio S-y, Wakui K, et al. Detection and confirmation of deafness-causing copy number variations in the STRC gene by massively parallel sequencing and comparative genomic hybridization. Annals of Otology, Rhinology & Laryngology. 2016;125(11):918-23.
60. Vandepoele K, Van Roy N, Staes K, Speleman F, Van Roy F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. Molecular Biology and Evolution. 2005;22(11):2265-74.
61. Johnson DB. A note on Dijkstra's shortest path algorithm. Journal of the ACM (JACM). 1973;20(3):385-8.
62. P. Peixoto T. The graph-tool python library. figshare; 2017.
63. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431-2.
64. Kobourov SG. Spring embedders and force directed graph drawing algorithms. arXiv preprint arXiv:12013011. 2012.
65. Aho AV, Kernighan BW, Weinberger PJ. Awk—a pattern scanning and processing language. Software: Practice and Experience. 1979;9(4):267-79.
66. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
67. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science. 2022;376(6588):eabj6965.
68. Bomze IM, Budinich M, Pardalos PM, Pelillo M. The maximum clique problem. Handbook of Combinatorial Optimization: Supplement Volume A: Springer; 1999. p. 1-74.

69. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261-72.
70. Alekseyenko AV, Lee CJ. Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases. Bioinformatics. 2007;23(11):1386-93.
71. Stovner EB, Sætrom P. PyRanges: efficient comparison of genomic intervals in Python. Bioinformatics. 2020;36(3):918-9.
72. Milano M, Trick M. Constraint and integer programming. Constraint and integer programming: toward a unified methodology Berlin: Springer. 2004.
73. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv. 2017:201178.
74. Nicolas G, Charbonnier C, Campion D. From common to rare variants: the genetic component of Alzheimer disease. Human heredity. 2017;81(3):129-41.
75. Christophersen IE, Rienstra M, Roselli C, Yin X, Geelhoed B, Barnard J, et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nature genetics. 2017;49(6):946-52.
76. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
77. Yang XT, She CH, Zhang C, Leung D, Yang J, Chan K-W, Rosa Duque JS, Lau YL, Yang W. SDrecall: a sensitive approach for variant detection in segmental duplications. Github. https://github.com/snakesch/SDrecall (2024).
78. Yang XT, She CH, Zhang C, Leung D, Yang J, Chan K-W, Rosa Duque JS, Lau YL, Yang W. SDrecall: a sensitive approach for variant detection in segmental duplications. Zenodo. https://zenodo.org/records/17667346 (2025).
79. Yang XT, She CH, Zhang C, Leung D, Yang J, Chan K-W, Rosa Duque JS, Lau YL, Yang W. SDrecall: a sensitive approach for variant detection in segmental duplications. Zenodo. https://zenodo.org/records/17669242 (2025).