

SOFTWARE

Open Access



# TRsv: simultaneous detection of tandem repeat variations, structural variations, and short indels using long read sequencing data

Shunichi Kosugi<sup>1,2,3,4\*</sup>  and Chikashi Terao<sup>3,4,5</sup>

\*Correspondence:  
shunichi.kosugi@nig.ac.jp

<sup>1</sup> Center for Genome Informatics, Joint Support-Center for Data Science Research, Research Organization of Information and Systems Center for Genome Informatics, 1111, Yata, Mishima, Shizuoka 411-8540, Japan

<sup>2</sup> Advanced Genomics Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

<sup>3</sup> Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan

<sup>4</sup> Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-Cho, Tsurumi-Ku, Yokohama, Kanagawa 230-0045, Japan  
<sup>5</sup> The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan

## Abstract

Tandem repeat copy number variations (TR-CNVs), structural variations (SVs), and short indels have been responsible for many diseases and traits, but no tools exist to distinguish and detect these variants. In this study, we developed a computational tool, TRsv, to distinguish and detect TR-CNVs, SVs, and short indels using long reads. In evaluation with simulated and real datasets, TRsv outperformed existing tools for detection of TR-CNVs and indels and performed equally well for detection of SVs. We demonstrated genome-wide detection of TR-CNVs, including variants associated with gene expression, disease, and quantitative traits, using 160 long-read whole genome sequencing data and TRsv.

**Keywords:** Tandem repeat, Tandem repeat expansion, TR, STR, VNTR, Structural variation, SV, Indel, Long read, WGS

## Background

Genomic structural variations (SVs) of at least 50 bp and short indels of less than 50 bp are the primary determinant of genomic variation and have much greater potential to alter gene function and gene regulation than single nucleotide variants (SNVs). On the other hand, tandem repeat variations within tandem repeat regions are copy number variation of their repeat units (referred to as TR-CNVs in this study). Although TR-CNV is a kind of SV and short indel, they are often treated confusedly.

Tandem repeats (TRs) are tandem repeats of short motifs or DNA segments. Short tandem repeats (STRs) are tandemly repeat sequences with motifs ranging from 1 to 6 bp. STRs are also called microsatellites whereas tandem repeats with motif length > 6 bp are called minisatellites or variable number of tandem repeats (VNTRs). In this study, all tandem repeats with  $\geq 1$  bp motif were considered TRs. Although tandem repeats represent only about 3% of the human genome [1], many insertions (INs) and deletions



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

(DELs) are observed in tandem repeat regions, corresponding approximately to repeat expansion and contraction, respectively. This is because tandem repeats are order of magnitude more mutable than mutations in non-repeat regions [1, 2]. The high mutation rate of tandem repeats is explained by mechanisms based on repeat structure, such as strand misalignment by strand slippage and secondary structure formation during replication [3, 4]. Recurrent mutations at high mutation rates generate multiple alleles with different numbers of repeats at many tandem repeat loci, a unique property of tandem repeat polymorphisms that is rarely observed in other types of variants [5].

TR-CNVs, especially tandem repeat expansions, are associated with many diseases and traits [6–10]. Repeat expansion disorders, consisting of more than 50 neurodegenerative and muscular diseases, are known to be caused by rare tandem repeat expansions with repeat unit copy sizes larger than a certain threshold copy number, primarily in the coding or 5' UTR regions of a subset of specific genes [8, 9]. Most of the repeat units involved in the repeat expansion diseases are trinucleotide units, such as GCA and GGC, and the expansion repeats are likely to be translated as homo amino acid polymers by a non-AUG RAN translation mechanism or as fusion proteins with the corresponding native proteins, which may exhibit some toxicity in the nervous system [11, 12]. Recent studies have shown that there is a higher burden of tandem repeat expansions in patients with autism [13, 14] and schizophrenia [15] and in cancer genomes [16, 17]. Apart from the disease-causing function of rare repeat expansion variants, common TR-CNVs in non-coding and coding regions regulate the expression of neighboring genes and the protein function, respectively, and are associated with a number of quantitative traits dependent on copy number of repeat units [18–23]. This suggests that TR-CNVs may be a resource to fill in the missing heritability not captured by traditional SNP-based genome-wide association studies.

The current standard method for the genome-wide detection of TR-CNVs and SVs is to use short read WGS data and associated bioinformatics tools [10, 24]. Detection of TR-CNVs and SVs is more difficult than detecting SNVs because the alignment of short reads to repetitive regions is often ambiguous. Furthermore, TR-CNV/SV detection tools using short reads have difficulty detecting variants beyond read length (100~200 bp) or paired-end read library size (200~500 bp). Long reads overcome the limitation of short read-based detection since long reads span most of the tandem repeat regions. Although there was a problem with error-prone sequencing of long reads, the sequencing accuracy of PacBio HiFi reads reaches 99.9% and that of recently improved ONT Nanopore reads about 98%.

Genome-wide detection and genotyping of TR-CNVs with long read data can be performed using specific tools, such as PacMonSTR [25], RepeatHMM [26], Straglr [27], tandem-genotypes [28], LongTR [29], TRGT [30], and TREAT [31], which use information from long reads aligned to reference tandem repeat regions. TR-CNVs can also be detected using long read-based structural variation (SV) detection tools. However, INs and DELs called from SV detection tools are often detected at multiple sites within a single tandem repeat region, which is due to redundant or overly specific read alignments observed in repetitive regions. Furthermore, these TR-CNV/SV detection tools cannot detect tandem repeat insertions in non-repeat or undefined tandem repeat regions and may not distinguish between repeat and non-repeat insertions within TR regions.

To overcome the current limitations in TR-CNV/SV detection, we developed a computational tool, TRsv, to detect and genotype TR-CNVs, SVs, and indels using long read alignment data. TRsv calculates the copies of repeat units in the INS and DEL sequences detected in tandem repeat regions for each haplotype and distinguishes non-TR variations within TR regions. In addition, TRsv detects tandem repeat INSs in non-repeat or undefined repeat regions by examining the INS sequences, as well as SVs and indels outside TR regions. TRsv also annotates INSs with duplication and mobile elements. TRsv using 138 publicly available PacBio HiFi WGS data detected a median of approximately 17,275 ( $\geq 50$  bp) and 254,159 ( $\geq 3$  bp) TR-CNV alleles per individual, representing approximately 60% of the SVs/short indels present in the human genome, respectively. Tandem repeat sites, especially high allele frequency TR-CNV sites, were highly enriched in Chip-seq peaks of DNA damage checkpoint kinases (ATM/ATR) and DNA replication origins, suggesting that tandem repeat regions are susceptible to DNA mutation and actively repaired. Analysis with RNA-seq data from 59 matched samples identified many TR-CNV expression quantitative loci (eTR-CNVs) that were significantly enriched for genes associated with specific diseases and traits, including schizophrenia, coronary artery disease, and refraction disorders.

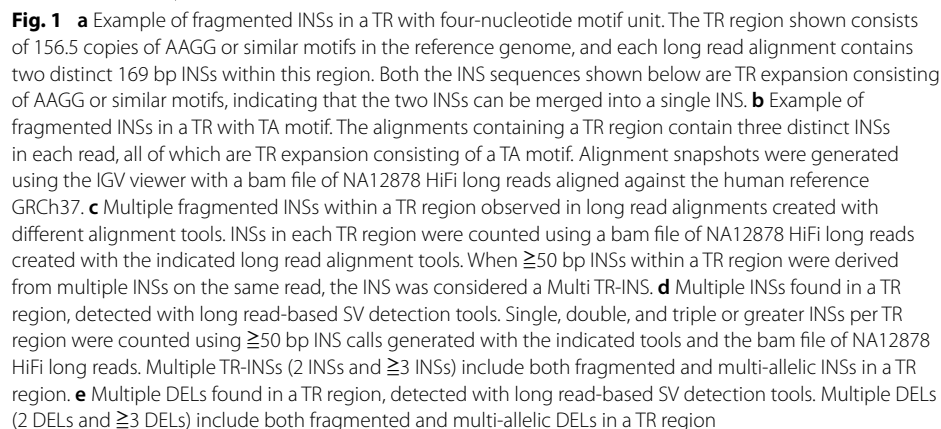
## Results

### Problems in detecting TR variations

Accurate calling of TR-CNVs from TR regions is challenging due to two main problems: fragmented INSs/DELs and non-TR-INSs. Alignment of long reads to reference genomes often generates fragmented INS or DEL alignments within a tandem repeat region, resulting in multiple INS and/or DEL calls with a haplotype within a single tandem repeat region even if the sequence corresponding to the INS or DEL is composed of a repeat unit of the tandem repeat region. For two STR regions with AAGG and TA repeat unit, we show examples of fragmented INSs composed of the corresponding repeat unit in a single NA12878 PacBio HiFi long read alignment (Fig. 1a and 1b). The fragmented INS alignments were observed in all HiFi long read alignment data generated by alignment tools commonly used for long read alignment (Minimap2 [32], NGM-LR [33], and Winowmap2 [34]), with fragmented INSs ranging from 14 to 28% of the INSs observed in the TR regions (Fig. 1c). As a result, all existing long read-based SV detection tools, including cuteSV [35], dysgu [36], Nanovar [37], pbsv, Sniffles2 [38], SVDSS [39], and SVIM [40], detect multiple DELs and INSs in a single TR region, ranging from 7 to 19% (Fig. 1d) and 6% to 24% (Fig. 1e) of DELs and INSs called in the TR regions, respectively. Another problem is the presence of non-TR-INSs, i.e., sequences within the TR region that are not composed of a TR repeat unit. This class of INS consists of two types: one is INSs containing irrelevant sequences with no or few copies of a repeat unit (Additional file 1: Fig. S1a and b), and the other is INSs of mobile elements (MEs), such as Alu, L1, and SVA (Additional file 1: Fig. S1c and d). We found that about 5% of INSs in the TR region were non-TR-INSs.

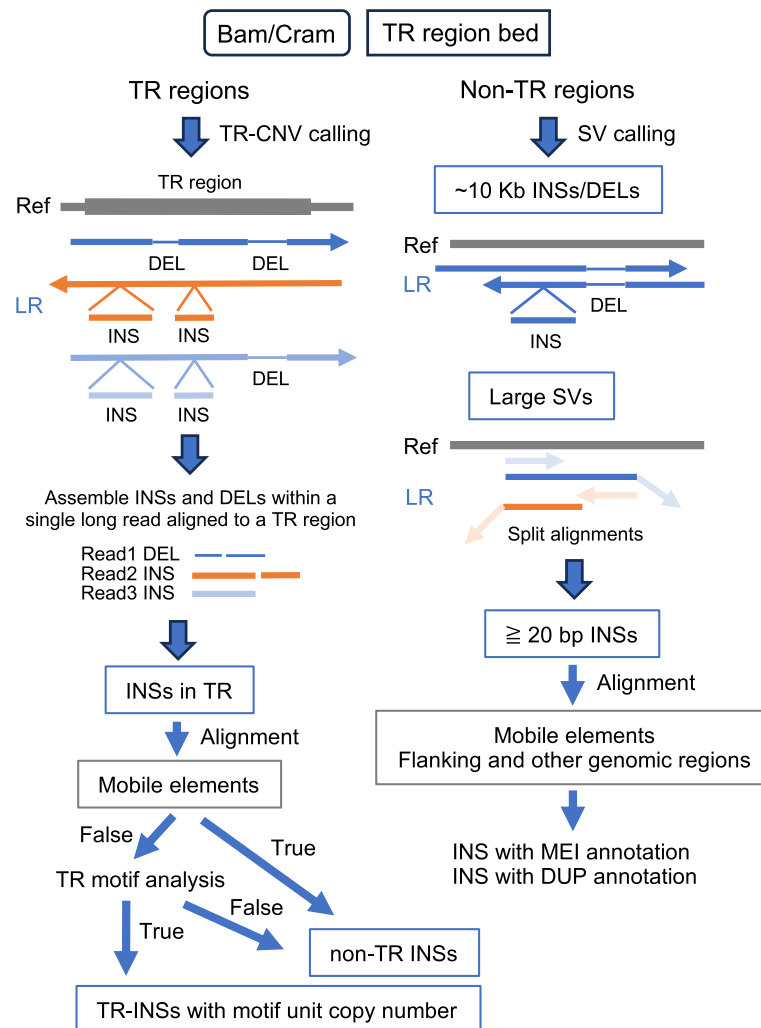
### Development of TRsv to distinguish and detect TR-CNVs, SVs, and indels inside and outside the TR regions

To address these issues related to TR-CNV calling, we developed TRsv, a computational tool that detects TR-CNVs within and around the TR regions in a motif/sequence-resolved manner, as well as SVs outside the TR regions. Although any custom TR data



can be used for TRsv, in this study we used a combination of TR data from the UCSC and HipSTR sites [41], in which TR regions with  $\geq 30\%$  reciprocal overlap were merged without redundancy and restricted to a size range from 20 bp to 10 Kb (see Methods in details). TRsv first examined the similarity of the INS sequences detected within and around the TR regions to MEs to detect ME INSs within TR regions (Fig. 2). When more than half of an INS sequence consisted of a TR motif or its similar motifs, the INS

was assigned as a TR-CNV in the TR region. Multiple INSs in one TR region aligned by the same long read were merged into one TR-INS, as well as multiple DELs in one TR region. When TR-INSs or -DELs of different sizes were aligned by different long reads, these INSs or DELs were assigned as TR-CNVs of distinct haplotypes. TRsv also detected SVs (DEL, INS, DUP, INV, and TRA) and short indels outside the TR region, including short SVs/indels embedded within read alignments and large SVs supported



**Fig. 2** Overview of the TRsv algorithm. TRsv calls TR-CNVs within the TR regions and SVs outside the TR region separately, using a bam/cram file of long read alignment data and a bed file defining the TR regions. For TR-CNV calling, when a single long read aligned to a TR region contains multiple INSs and/or DELs, these variants are concatenated as a single variant, assuming that they are copies of the TR repeat unit in the first stage. When the variants derived from multiple long reads assigned within the TR region are of different types or sizes, the variants are defined as two different alleles depending on the number of supporting reads and the length ratio. Furthermore, TR-INSs are defined by sequence alignment and motif analysis to determine whether the INS contain copies of the TR repeat unit or unrelated sequences such as mobile elements. For SV calling outside the TR regions, INSs and DELs are detected using evidence of long-read alignments spanning these variants or split alignments of long reads that indicate the breakpoints of large SVs. In addition, sequence alignment and motif analysis are performed for  $\geq 20$  bp INSs outside the TR regions to determine whether these INSs are mobile element INSs, tandem repeat INSs, or DUP-containing INSs. Variants called from multiple samples are joint-called to create a single vcf file

by split read alignments. The INSs in the non-TR region were annotated as follows: check for similarity to ME in the INS sequence to see if the INS is a ME insertion; check for the other genomic sequences of the INS to see if the INS is a tandem/interspersed DUP. To identify INSs containing tandem repeats outside the TR regions, INSs in non-TR regions were further checked using the Tandem Repeat Finder (TRF) tool to see if the INS sequences contained TR copies of a short motif. Low-quality SVs from non-HiFi reads were filtered using a machine learning-based method. TRsv also had the ability to merge (joint call) multiple vcf files generated from multiple samples and annotate SVs that overlap with gene regions.

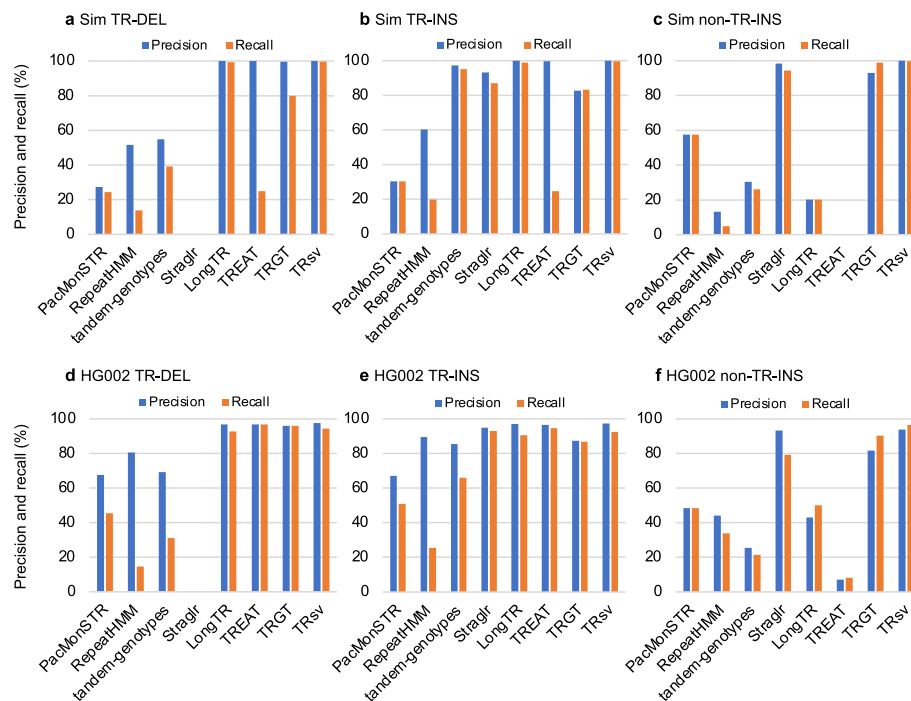
**Evaluation of TR-CNVs called using TRsv and other existing tools**

To evaluate the accuracy of TR-CNV detection of TRsv and existing TR-CNV detection tools, we used simulated and real datasets of TR-DEL, TR-INS, and non-TR-INS. We selected existing tools for genome-wide detection of TR-CNVs using long read WGS data: tandem-genotypes [28], RepeatHMM-Scan [26], PacMonSTR [25], Straglr [27], LongTR [29], TRGT [30], TREAT [31] (Table 1). TRiCoLoR [42] could not be executed in our computing environment due to a system error. For the evaluation using simulated data, we created three simulated datasets by introducing approximately 1000 of artificial TR-DELS, TR-INSs, and non-TR-INSs into existing TR sites on diploid chromosome 1 of the human GRCh38 reference genome, followed by creation of artificial 20 × PacBio HiFi long reads using the simulated genome (see Methods). The simulated non-TR-INSs were introduced into existing TR sites, but half of them were MEs and the other half were tandem repeats with repeat units unrelated to those of the introduced TR sites. TR-CNVs called using the simulated datasets were evaluated for each tool (Fig. 3, Additional file 1: Table S2 for numerical data). Only TRsv showed nearly 100% precision and recall across the three datasets (Fig. 3a–c). LongTR also exhibited almost 100% precision and recall for both TR-DEL and TR-INS data, but only 20.2% precision and 20.1% recall for non-TR-INS data. For the evaluation using real data, we used a curated TR-CNV data selected from the GIAB HG002 TR-CNV catalog (see Methods). This GRCh38-based reference dataset for HG002 consists of 3109 TR-INSs,

**Table 1** Summary of TR-CNV detection tools

| Tool              | Repeat unit <sup>a</sup> | Input data       | Runtime <sup>b</sup> |
|-------------------|--------------------------|------------------|----------------------|
| adVNTR            | 6–100                    | bam/cram         | 32,856               |
| PacMonSTR         | No limit                 | bam              | 20,457               |
| RepeatHMM-Scan    | 1–15                     | bam              | 166                  |
| Straglr           | No limit                 | bam              | 21.2                 |
| Tandem-genotypes  | No limit                 | maf <sup>c</sup> | 1.9                  |
| LongTR            | No limit                 | bam/cram         | 8.6                  |
| TREAT             | No limit                 | bam              | 15.8                 |
| TRGT              | No limit                 | bam              | 8.2                  |
| TRsv <sup>d</sup> | No limit                 | bam/cram         | 11.9                 |

<sup>a</sup> Unit size range of tandem repeat that can be used  
<sup>b</sup> Execution time (minutes) per 10,000 tandem repeats per CPU core  
<sup>c</sup> maf alignment data generated with the last alignment tool  
<sup>d</sup> TR-CNVs and SVs with ≥3 bp were detected with 6 CPU cores



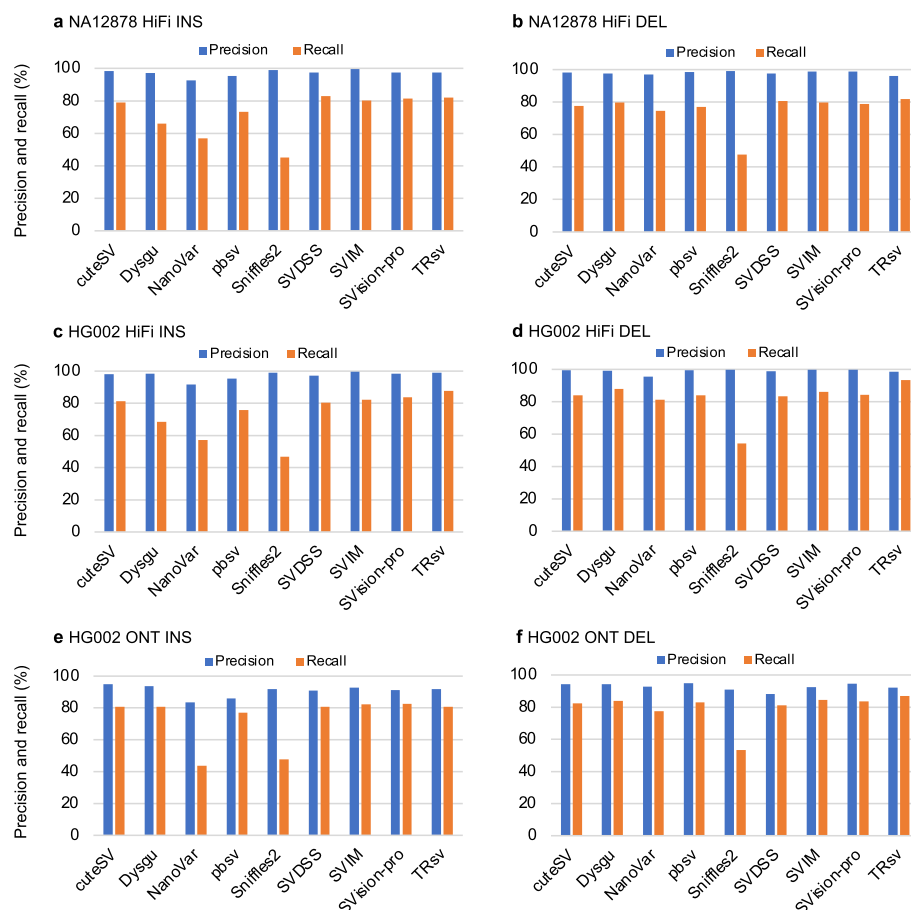
**Fig. 3** Evaluation of TR-CNV detection tools using simulated data (a–c) and real data (d–f). Simulated data consisted of three sets of simulated PacBio HiFi long reads created using the diploid chr1 of the GRCh38 reference, in which approximately 1000 of artificial TR-DELs, TR-INSs, or non-TR-INSs were incorporated into the existing TR sites of chr1. Non-TR-INSs are insertions that do not consist of repeat units of the incorporated TR sites. For real data, HG002 HiFi long read WGS data and reference TR-CNVs selected from the GIAB HG002 TR-CNV catalog were used. Evaluation results using simulated TR-DEL (a), TR-INS (b), and non-TR-INS (c) data. Evaluation results using the real HG002 data of TR-DEL (d), TR-INS (e), and non-TR-INS (f). Straglr is not designed to detect TR-DELs. In the evaluation for non-TR-INS, INSs called as standard INSs (non-TR-INSs) were evaluated as TP for the TRsv tool, while for the other tools TPs were considered if the called INS sequence was found to contain only less than 50% reference repeat content. For LongTR, variants with INEXACT\_ALLELE tag in the output vcf file were considered as TP for non-TR-INS data. Precision (%) and recall (%) are indicated with blue and orange bars, respectively

2416 TR-DELs, and 144 non-TR-INSs (all  $\geq 50$  bp) at 4746 TR sites. TR-CNV detection tools were run on the 4746 reference TR sites using HG002 PacBio HiFi WGS data. The evaluation of the called TR-CNVs showed that only TRsv had high precision and recall for all three types, TR-DEL (97.5% precision and 94.4% recall), TR-INS (97.4% precision and 92.5% recall), and non-TR-INS (93.9% precision and 96.5% recall). LongTR, TREAT, and TRGT also showed high precision (96.7%, 96.9%, and 96.1%, respectively) and recall (92.6%, 96.9%, and 96.0%, respectively) for TR-DELs but not for the other data, and Straglr, LongTR, and TREAT also showed high precision (94.8%, 96.9%, and 96.4%, respectively) and recall (92.9%, 90.6%, and 94.5%, respectively) for TR-INSs but not for the other data (Fig. 3d, e). These results indicate that TRsv can not only accurately detect TR-CNVs, but also discriminate between non-TR-INSs and TR-INSs in TR regions. We determined the runtime per 10,000 tandem repeats per CPU core for each tool using HG002 HiFi data and found that TRsv was one of the tools showing the shortest runtime, despite that TRsv simultaneously detects SVs outside the TR regions (Table 1). Tandem-genotype had the shortest runtime, but it often took several days or more to create the input maf alignment file with the LAST tool. These results show that

TRsv outperforms the existing TR-CNV detection tools tested in terms of both precision, sensitivity, and runtime.

#### Evaluation of SVs and short indels called outside TR regions

To compare the SV detection performance between TRsv and other eight long read-based SV detection tools, SVs outside the TR regions called using three long read WGS datasets (NA12878 HiFi, HG002 HiFi, and HG002 ONT) were evaluated using the corresponding reference SV datasets (3717 DELs and 4195 INSs for NA12878, 3663 DELs and 4656 INSs for HG002) (see Methods). TRsv had the highest recall (81.3–93.3%) of DEL among the nine tools for the three datasets, and the recall of INS was as high (81.9–87.8%) as the other tools (Fig. 4, Additional file 1: Table S2 for numerical data). In terms of precision, all tools showed high precision (90–99%) for both INS and DEL on all three datasets (Fig. 4). The evaluation of TRsv false positive calls by manual visual inspection (see Methods) showed that TRsv could detect INSs and DELs with nearly 100% precision (Additional file 1: Table S3). TRsv used a machine learning-based filtering strategy (see Methods for details) for SV calls from non-HiFi reads, thereby improving precision,



**Fig. 4** SVs called outside TR regions with long read-based SV detection tools. Precision (blue bars) and recall (orange bars) of INSs (a, c, e) and DELs (b, d, f) called with long read-based SV detection tools in NA12878 HiFi (a, b), HG002 HiFi (c, d), and HG002 ONT (e, f) WGS data. SV calls of  $\geq 50$  bp that were outside the TR regions were used for the evaluation using the reference SV sets for NA12878 and HG002 (Methods for details)

especially for INS calls in PacBio CLR data (Table 2). For DUP calling, TRsv detected the highest number of  $\geq 1$  Kb TP DUPs outside the TR regions in both NA12878 and HG002 HiFi data, with precision comparable to the other tools (Additional file 1: Fig. S2). The efficient DUP detection of TRsv is likely due to its algorithm that integrates read depth and split-read statistics. In the evaluation of genotyping, all the tools except Dysgu, SVDSS, SVIM, and NanoVar showed high ( $>96\%$ ) precision for all datasets (Additional file 1: Fig. S3).

TRsv detected short indels as short as 1 bp outside the TR regions using HiFi long read alignment data. We evaluated the accuracy of indel detection using TRsv, PEPPER-Mergin-DeepVariant (PEPPER) [43], and GATK4 [44] by analyzing shared and unique calls between TRsv and PEPPER or GATK4 (see Methods). TRsv called more indels for both insertions (ins) and deletions (del) than the other tools with nearly 100% precision for unshared unique calls (Fig. 5a and b) The precision of the unshared unique calls of PEPPER was also nearly 100% with over 97% of the PEPPER indel calls shared with the TRsv indel calls in all size ranges (Fig. 5b and d), suggesting the highest recall of TRsv. The number of indel calls shared between TRsv and PEPPER or GATK4 decreased as the indel size increased (Fig. 5c–e), suggesting that TRsv had higher recall in larger indels than the other tools.

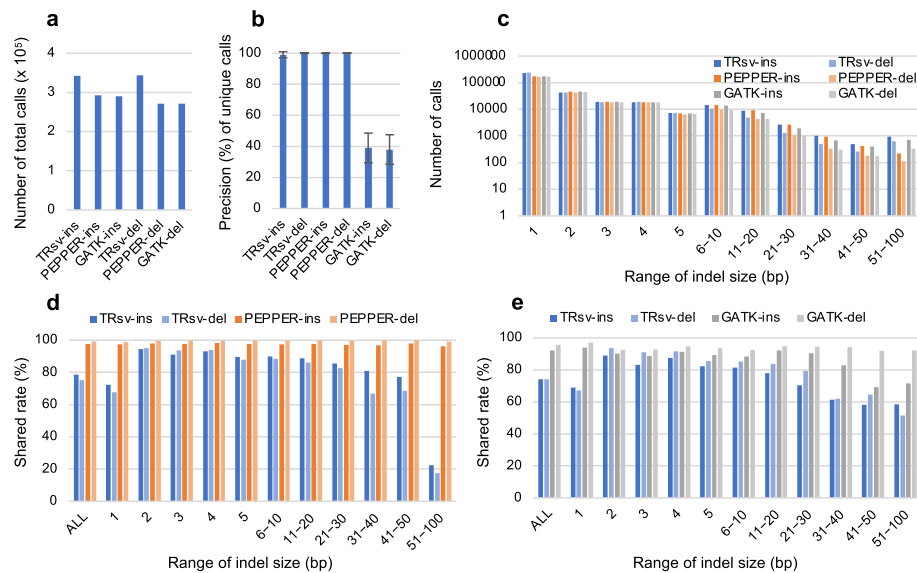
**TR-CNVs, SVs, and short indels detected from 138 PacBio HiFi WGS dataset using TRsv**

We collected 138 publicly available, high coverage, parentage-free PacBio HiFi WGS data from diverse populations, mainly obtained from the Human Pangenome Reference Consortium (HPRC), the Human Genome Structural Variation Consortium (HGSVC) in the 1000 Genomes Project (1KGP), and the Genome in a Bottle Consortium (GIAB). From the 138 HiFi WGS datasets, TRsv detected 60,840 TR-CNV and 78,064 SV sites outside the TR regions for  $\geq 50$  bp variants (HiFi-var50) and 804,415 TR-CNV and 1,599,476 SV/

**Table 2** Precision and recall of SVs before and after machine learning-based filtering of TRsv

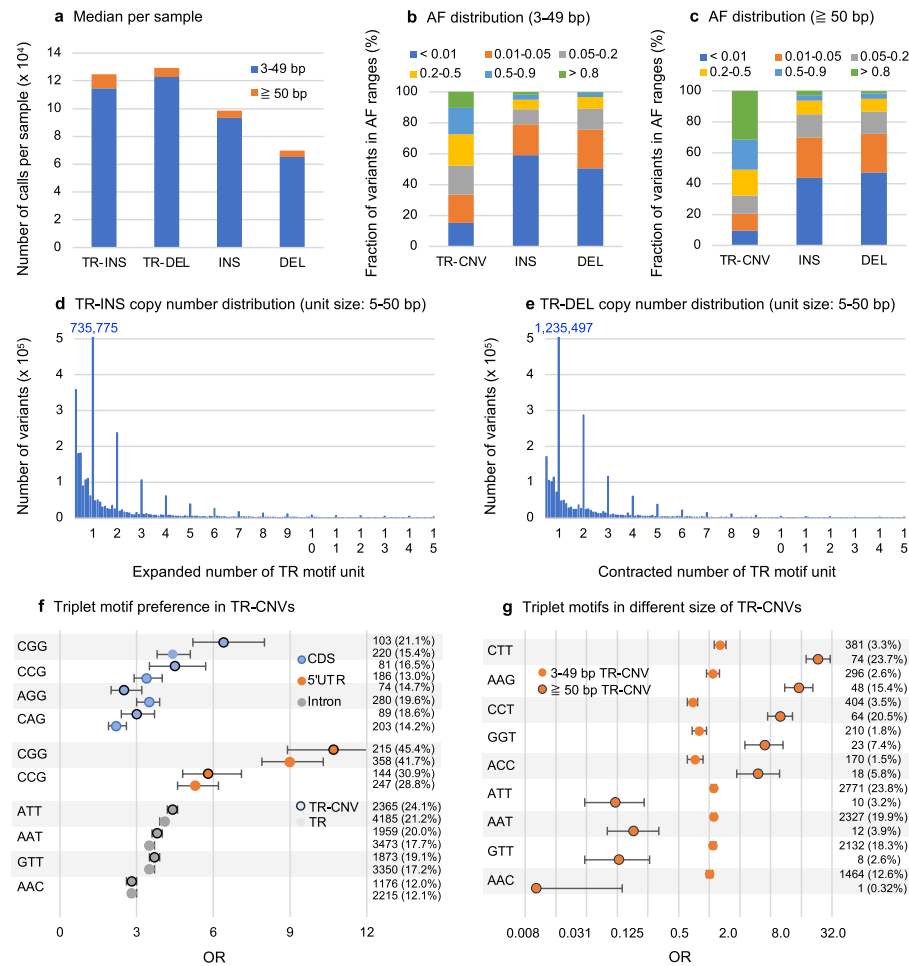
| Sample  | Read       | SV type | ML-Filtering <sup>a</sup> | Number of calls | Precision (%) | Recall (%) |
|---------|------------|---------|---------------------------|-----------------|---------------|------------|
| NA12878 | PacBio CLR | INS     | —                         | 4949            | 70.1          | 77.0       |
|         |            |         | +                         | 3841            | 88.6          | 75.6       |
|         |            | DEL     | —                         | 2842            | 94.3          | 72.1       |
|         |            |         | +                         | 2800            | 95.1          | 71.6       |
|         | Nanopore   | INS     | —                         | 4179            | 88.4          | 80.4       |
|         |            |         | +                         | 3744            | 94.4          | 78.5       |
|         |            | DEL     | —                         | 3529            | 84.7          | 81.9       |
|         |            |         | +                         | 3271            | 90.0          | 79.2       |
| HG002   | PacBio CLR | INS     | —                         | 5686            | 67.6          | 82.3       |
|         |            |         | +                         | 4054            | 92.3          | 74.1       |
|         |            | DEL     | —                         | 3269            | 94.1          | 84.0       |
|         |            |         | +                         | 3217            | 95.2          | 77.2       |
|         | Nanopore   | INS     | —                         | 4767            | 83.3          | 84.9       |
|         |            |         | +                         | 4081            | 92.7          | 81.1       |
|         |            | DEL     | —                         | 3753            | 86.3          | 88.3       |
|         |            |         | +                         | 3460            | 92.1          | 87.0       |

<sup>a</sup> Machine learning-based filtering was applied (+) or not applied (—) for SV calls



**Fig. 5** Short indels outside TR regions. **a** Total number of indel calls for TRsv, PEPPER, and GATK4. We called ins and del in the size range 1–100 bp with NA12878 PacBio HiFi WGS data for TRsv and PEPPER and with NA12878 Illumina short read WGS data for GATK4. Indels outside TR regions were used for analysis. **b** Precision of unique indels not shared between tools. Among the unique indels not shared between TRsv and PEPPER or GATK4, 200 indels each of ins and del were randomly selected, and the precision of the selected indels was estimated by manual visual inspection with the IGV viewer using the WGS data for both the HiFi long read and Illumina short reads. The black thin bars on the bars show the confidence intervals obtained using a binomial test. **c** Number of indel calls across indel size range. In TRsv, the number of ins and del calls in the indicated indel size range is indicated by blue and light blue bars, respectively. The number of PEPPER ins and del calls is indicated by orange and light orange bars, respectively, and the number of GATK4 ins and del calls is indicated by gray and light gray bars, respectively. The number of indel calls is given in a base-10 log scale. **d** Percentage of indels shared between TRsv and PEPPER. The percentage (shared rate) of ins and del calls matched in the indicated indel size range between TRsv and PEPPER calls is shown as blue and pale blue bars for TRsv and orange and pale orange bars for PEPPER, respectively. **e** Percentage of indels shared between TRsv and GATK4. The percentage of ins and del calls matched in the indicated indel size range between TRsv and GATK4 calls is shown as blue and pale blue bars for TRsv and gray and pale gray bars for GATK4, respectively

indel sites for  $\geq 3$  bp variants (HiFi-var3) (Additional file 1: Table S4). The median number of TR-CNV alleles per sample was 17,275 for  $\geq 50$  bp and 254,159 for  $\geq 3$  bp, and the median number of SVs/indels outside the TR regions was 10,131 for  $\geq 50$  bp and 168,492 for  $\geq 3$  bp (Fig. 6a and Additional file 1: Table S4). Of the TR-CNV sites, 65% in HiFi-var3 and 57% in HiFi-var50 were multi-allelic sites with different copies of expanded and/or contracted repeat units, and in each individual, about 19% in HiFi-var3 and 23% in HiFi-var50 were multi-allelic sites (Additional file 1: Table S4). Of 39,339 INSs in HiFi-var50, 19,947 (50.7%), 11,135 (28.3%), and 2790 (7.1%) were ME-INSs with homology to ME, DUP-INSs with homology to other genomic regions, and TR-INSs with tandem repeats outside the reference TR regions, respectively. Of the TR sites with INS, 1795 (5.2%) sites had non-TR-INSs in HiFi-var50 (Additional file 1: Table S4). Allele frequencies (AFs) of TR-CNVs were significantly greater than those of SVs (Fig. 6b and c). The trend toward higher AF was more pronounced for  $\geq 50$  bp TR-CNVs than for TR-CNVs in the 3–49 bp range. TRsv determines the increase or decrease in the copy number of repeat units of TR-CNVs to the last digit. We measured the frequency of increased or decreased repeat unit copy number of TR-INSs and TR-DELS in different repeat unit size ranges (Fig. 6d



**Fig. 6** Properties of TR-CNVs and SVs/indels detected from 138 human HiFi WGS datasets. **a** Median number of TR-CNVs and SVs per individual. Variants in the ranges 3–49 bp and  $\geq 50$  bp are indicated by blue and orange bars, respectively. **b, c** Percentage of variants stratified by AF. The percentages of variants in the AF ranges of  $< 0.01$ ,  $0.01-0.05$ ,  $0.05-0.2$ ,  $0.2-0.5$ ,  $0.5-0.8$ , and  $> 0.8$  for TR-CNVs, INSs (outside TR regions), and DELs (outside TR regions) in the ranges of 3–49 bp (**b**) and  $\geq 50$  bp (**c**) are shown as blue, orange, gray, yellow, cyan, and green bars, respectively. **d** Copy number distribution of repeat units in TR-INSs with unit sizes from 5 to 50 bp. The frequency of TR-INSs corresponding to the indicated number of unit copies determined to one decimal place is indicated by blue bars. The number of TR-INSs with one unit copy is shown in blue. **e** Copy number distribution of repeat units in TR-DELs with unit sizes from 5 to 50 bp. The frequency of TR unit copy number is shown as in **d**. **f** Triplet motifs of TRs and TR-CNVs abundant in gene regions. The enrichment of TR and TR-CNV triplet motif sets present in the CDS, 5'-UTRs, and introns of protein-coding genes was tested as a control for the 3-mer frequency of the corresponding gene region to calculate odds ratios (OR). The motifs in the opposite-strand genes were reverse complemented. ORs in the CDSs, 5'-UTRs, and introns of the triplet motif sets are shown by light blue, orange, and gray circles, respectively, with confidence intervals. The ODs of TR and TR-CNV are shown by circles without and with black borders, respectively. The motif on the left represents one of three overlapping motifs with 1 bp shifts on that repeat (e.g., CGG contains two other overlapping motifs, GGC and GCG). On the right side, the number of corresponding TRs and TR-CNVs and their percentages of the total triplet TRs/TR-CNVs are shown. **g** Triplet motifs enriched or depleted in 3–49 bp TR-CNVs and  $\geq 50$  bp TR-CNVs. Enrichment of the TR-CNV triplet motif sets in either CDSs, 5'-UTRs, or introns of the protein-coding genes was tested against TR triplet motif sets. ORs for TR-CNVs  $\geq 50$  bp and 3–49 bp are shown by orange circles with and without black borders, respectively. The motifs were represented as in **g**. On the right, the number of corresponding TR-CNVs and their percentages in all triplet TR-CNVs are shown

and e for 5–50 bp unit size range, Additional file 1: Fig. S4 for the other unit size ranges). For both TR-INSs and TR-DELs, clear peaks of integer copy number frequency were observed in any unit size ranges (81% of TR-CNVs were integer copy number in HiFi-var3), indicating that repeat unit-based expansion and contraction generate INSs and DELs within the TR regions.

### **Triplet motifs observed in repeat expansion disorders are enriched in both TRs and TR-CNVs within CDSs and 5′-UTRs**

TR expansion in gene regions, especially those with a repeat unit size of 3–6 bp, is responsible for many diseases. We examined which of 1–10 bp unit sizes of TRs and TR-CNVs are biased in the CDS, 5′UTR, and intron of protein-coding and non-coding genes. As expected, the frequencies of TRs and TR-CNVs of unit sizes other than 3 bp, 6 bp, and 9 bp were highly constrained in the CDSs (Additional file 1: Fig. S5a and e), suggesting negative selection against frameshift translation. In the non-coding gene exons, TRs and TR-CNVs of 3 bp, 6 bp, and 9 bp units were moderately enriched (Additional file 1: Fig. S5h and j), suggesting some non-coding genes may have been translated. In the 5′-UTRs, both TRs and TR-CNVs with unit sizes of 3 bp, 6 bp, and 9 bp were positively biased (Additional file 1: Fig. S5b and f), suggesting that some 5′-UTRs may be translated, as observed in the disease-related repeat expansions in the 5′-UTRs. In contrast, no significant unit size enrichment or constrain was observed in the introns of the protein-coding and non-coding genes (Additional file 1: Fig. S5g and k). The enrichment of 3 bp, 6 bp, and 9 bp units in the CDS/exon and 5′-UTR was greater for TR-CNVs than TRs, suggesting that there may be translational regulation via copy number changes in the TR region. We then focused on the most significantly enriched 3 bp (triplet) motifs to determine which triplet motifs are more abundant among all the triplet motifs located in CDSs, 5′-UTRs, or introns. Four triplet motif sets with 1 bp shifts, CGG/GGC/GCG, CCG/CGC/GCC, AGG/GGA/GAG, and CAG/AGC/GCA, were more common in CDSs than the other 16 motif sets for both TRs and TR-CNVs, accounting for 71% of all TR-CNV triplet occurrences in CDSs (Fig. 6f). The top two sets of CDSs, CGG/GGC/GCG and CCG/GCC/CGC, were also the top two sets of 5′-UTRs, representing 76% of all TR-CNV triplets present in the 5′-UTRs. In particular, the CGG/GGC/GCG GC-rich triplets, involved in repeat expansion diseases, were present at a higher rate in TR-CNVs than in TRs. On the other hand, the most abundant triplets in introns were AT-rich triplets such as ATT and AAT (Fig. 6f). We examined the extent to which each triplet motif set in 3–49 bp TR-CNVs and  $\geq 50$  bp TR-CNVs is biased compared to those in TRs in gene regions containing transcribed regions and their 1 Kb flanking regions. Five triplet sets, including CTT/TTC/TCT and AGG/GGA/GAG, were present at high rates in  $\geq 50$  bp TR-CNVs but not in 3–49 bp TR-CNVs, whereas the four sets that were abundant in introns were present at low rates in  $\geq 50$  bp TR-CNVs (Fig. 6g). The results suggest that the degree of copy number change of the triplet motifs in the TRs depends on the sequence of the motif and the region in which the TR is present.

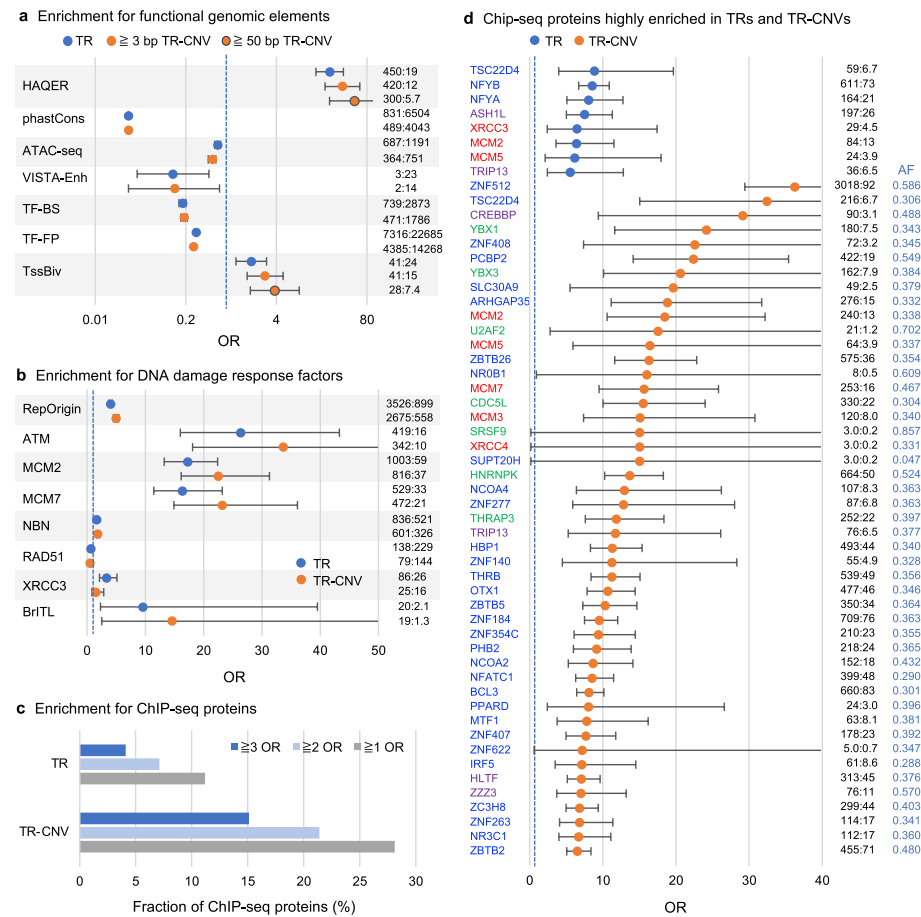
### **TR-CNVs and TRs are enriched in HAQERs and in DNA repair sites**

Preferentially occurring sites of TRs and TR-CNVs in the size range 100–5000 bp were examined for functional genomic regions using permutation tests (see Methods). The

enriched genomic regions included the Human Ancestor Quickly Evolved Region (HAQER) [45] and the promoter regions in bivalent chromatin states (TssBiv) (Fig. 7a). HAQERs are human genomic regions highly divergent from the human-chimpanzee ancestor and evolved under elevated mutation rates and positive selection. HAQERs have also been shown to be enriched in TssBiv regions [45]. About 29.5% and 27.5% of HAQERs ( $N=1526$ ) overlapped with TRs and TR-CNVs, respectively. The enrichment of TR-CNVs was more pronounced for  $\geq 50$  bp TR-CNVs than  $\geq 3$  bp TR-CNVs (Fig. 7a). TR-CNVs with higher AFs were more enriched in HAQER than TR-CNVs with lower AFs (Additional file 1: Fig. S6a and b). The most frequent TR size range for TR-CNVs overlapping HAQERs was 500–1000 bp (Additional file 1: Fig. S6c and d), which was consistent with the mean size of HAQERs (800 bp). In contrast, evolutionarily conserved regions were highly constrained for both TRs and TR-CNVs (Fig. 7a). Transcriptionally regulatory sites, including open chromatin regions (ATAC-seq), enhancers (VISTA-Enh), and transcription factor (TF) binding sites (TF-BS and TF-FP), were also constrained for both TRs and TR-CNVs (Fig. 7a). These results suggest that many transcriptional regulatory regions are largely composed of conserved regions lacking TRs.

(See figure on next page.)

**Fig. 7** Enrichment of TR-CNVs detected in 138 HiFi WGS datasets for functional genomic elements. **a** Enrichment for functional genomic elements and epigenetic regulatory sites. To examine the enrichment of the indicated genomic elements or epigenetic sites in TRs ( $N=113,383$ ) and TR-CNVs ( $N=70,822$  for  $\geq 3$  bp,  $N=34,818$  for  $\geq 50$  bp), we performed permutation tests using randomly selected elements from the GRCh37 human genome corresponding to the number of the genomic elements. TRs included 100–5000 bp of TR and TR-CNV sites. To fairly compare the enrichment of TRs and TR-CNVs, a fixed size of 100 bp from their centers was used to determine their overlap with the 200 bp center region of each genomic element. For fastCons and TF-BS, the data from only chr1 were used for analysis due to the large data size. ORs for TR,  $\geq 3$  bp TR-CNVs, and  $\geq 50$  bp TR-CNVs are shown by circles in blue, orange, and orange with black borders, respectively, with confidence intervals. A blue dot line indicates a border of OR 1.0. On the right, the former, separated by a colon, shows the number of TRs/TR-CNVs that overlap with the corresponding elements, and the latter the average of 1000 permutation tests. HAQER: human ancestor quickly evolved regions ( $N=1526$ ), phastCons: evolutionarily conserved regions ( $N=1,071,394$ ), ATAC-seq: open chromatin regions ( $N=1,071,394$ ), VISTA-Enh: experimentally determined enhancer regions ( $N=1927$ ), TF-BS: TF-binding sites (ChIP-seq data of 87 TFs for the NA12878 lymphoblastoid cell line) ( $N=263,442$ ), TF-FP: TF-binding sites (DNase I footprinting regions for hematopoietic cells) ( $N=2,006,091$ ), TssBiv: promoters in bivalent chromatin states ( $N=2047$ ). All the permutation tests showed  $p$  values of  $< 1 \times 10^{-7}$ , except for VISTA-Enh ( $1.1 \times 10^{-5}$  for TR and  $4.2 \times 10^{-4}$  for TR-CNVs) and TssBiv ( $1.8 \times 10^{-4}$  for TRs). **b** Enrichment for ChIP-seq data involved in DNA damage response. Permutation tests for the indicated ChIP-seq proteins to determine the enrichment of TRs and  $\geq 3$  bp TR-CNVs were performed as in **a**. RepOrigin: human core DNA replication origins ( $N=62,939$ ), ATM: checkpoint kinase activated upon DNA double-strand breaks ( $N=1209$ ), MCM2/MCM7: DNA helicase involved in DNA replication licensing ( $N=4299$  for MCM2, 2462 for MCM7), NBN: a factor involved in DNA repair through homologous recombination ( $N=39,807$ ), XRCC3: a factor involved in the homologous recombination repair pathway ( $N=1986$ ), BrtTL: DNA break sites upon ATR checkpoint kinase inhibition ( $N=166$ ). All the permutation tests showed  $p$  values of  $< 1 \times 10^{-7}$ , except for XRCC3 (0.015 for TR-CNVs). **c** Summary of enrichment for ChIP-538 data. TRs and TR-CNVs were limited to transcribed and 5 Kb flanking gene regions. Permutation tests for a total of 538 ChIP-Seq proteins to determine the enrichment of non-variable TRs and TR-CNVs were performed as in **a**. The percentage of Chip proteins enriched with ORs of  $\geq 1.0$ ,  $\geq 2.0$ , and  $\geq 3.0$  among the total Chip proteins tested for non-variable TRs ( $N=22,262$ ) and  $\geq 3$  bp TR-CNVs ( $N=32,162$ ) is shown by blue, light blue, and gray bars, respectively. **d** ChIP-538 proteins enriched in non-variable TRs and TR-CNVs with  $\geq 3.0$  OR. TRs and TR-CNVs were limited to transcribed and 5 Kb flanking gene regions. ORs of TRs and  $\geq 3$  bp TR-CNVs are indicated by blue and orange circles, respectively. The numbers separated by a colon on the right are shown as in **a**. The blue numbers on the right indicate the mean AFs of the overlapped TR-CNVs. Proteins from the ChIP-seq data involved in transcriptional regulation, chromatin modification, DNA replication and repair, and RNA processing are shown on the left in blue, purple, red, and green, respectively



**Fig. 7** (See legend on previous page.)

As highly mutable regions such as TR regions are constantly subject to DNA replication stall, mutation, and repair, we examined whether TRs and TR-CNVs are enriched in sites to which DNA damage response factors are recruited. DNA double-strand break checkpoint kinase (ATR) and DNA replication licensing factors (MCM2/MCM7), as well as ATM inhibition-induced double-strand break sites (BrITL [46]) and replication origins (RepOrigin), were enriched in TR regions with more prominently in TR-CNV regions (Fig. 7b). This observation suggests that DNA replication stress causes DNA replication forks to stall at TR regions, leading to expansion or contraction of the repeat unit. However, proteins involved in the DNA double-strand break-induced homologous recombination (HR) repair pathway (RAD51 and NBN) were not enriched in TR and TR-CNV regions, which suggests that replication stalled at TR regions collapses and is repaired possibly through non-HR repair pathway, including the micro-homology mediated end joining (MMEJ) pathway [47].

**Some DNA-binding proteins are enriched for binding to TR-CNV sites but not to non-variable TR sites**

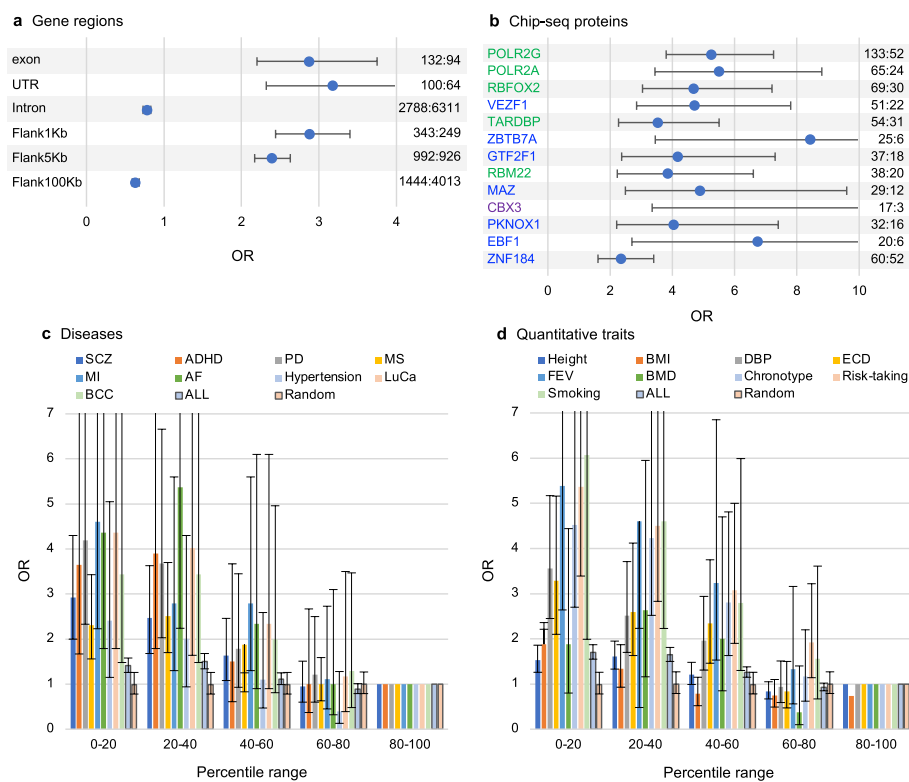
Although TF-binding sites were generally depleted in the TR and TR-CNV regions, some DNA-binding proteins may be abundant in the TR or TR-CNV regions, because

repeated TF-binding motifs generally increase the binding affinity of TF. We examined the enrichment of protein binding sites present in TRs/TR-CNVs using the ENCODE ChIP-seq data (Chip-538) for 538 DNA-binding proteins from GM12878 and K562 cell lines. Seventy-one (15.1%) ChIP proteins were enriched in TR-CNV regions with  $\geq 3$  OR while only 16 (4.1%) were enriched in non-variable TRs with  $\geq 3$  OR (Fig. 7c). These enriched ChIP proteins included proteins involved in DNA replication/repair (MCM2/3/5/7, XRCC3), chromatin modification (CREBBP, ASH1L), and transcriptional regulation (TSC22D4, NFYA/B, PCBP2) as well as those involved in RNA processing (YBX1/3, U2AF2, CDC5L, HNRNPK) (Fig. 7d). TR-CNV sites overlapping with the ChIP-seq data had higher levels of mean AF in most cases (Fig. 7d). These results suggest that some TR sites, especially mutation-prone TR-CNV sites, are both transcriptional regulatory sites and DNA damage susceptibility sites induced by replication and transcriptional stress [48].

#### **TR-CNVs associated with gene expression are enriched in gene regions involved in many diseases and traits**

To determine whether TR-CNVs are associated with gene expression, we used publicly available LCL-derived RNA-seq data from 59 samples, 39 of which were included in the 138 HiFi WGS data. We added PacBio CLR and HiFi long read WGS data for the remaining 20 samples and created a new vcf file (R59-var20) containing  $\geq 20$  bp TR-CNVs for the 59 samples using TRsv. To identify TR-CNVs associated with gene expression (eTR-CNVs), we tested the association between normalized gene expression of the LCL-derived RNA-seq data from 59 samples and TR-CNV unit copy number. A total of 24,514 genes (17,160 coding genes) were tested in a linear regression model against 59,959 TRs (55,211 TRs for coding genes). A total of 104 eTR-CNVs were identified that were controlled for significance with  $FDR < 0.05$ , and these eTR-CNVs positively or negatively affected the expression of various genes depending on their unit copy number (Additional file 1: Fig. S7). Due to the small number of FDR-based eTR-CNVs, the upper and lower percentiles of the eTR-CNVs sorted based on  $p$  values of the association tests were used for downstream analysis. Site-based eTR-CNVs with non-redundant TR sites, ranked in the top 10th percentile ( $N = 5988$ ,  $p$  values:  $9 \times 10^{-12}$  to  $2.8 \times 10^{-2}$ ), were more abundant in the exonic regions and the gene flanking regions within 1 Kb or 5 Kb of the terminal exons and less abundant in the introns and the 5–100 Kb flanking regions, compared to eTR-CNVs ranked in the bottom 20th percentile (Fig. 8a). We examined the enrichment of the top 10th percentile eTR-CNVs for DNA-binding factors using the ChIP-538 dataset. The top 10th percentile eTR-CNVs were significantly enriched in transcriptionally active regions ( $p$  values:  $< 1 \times 10^{-4}$  for enrichment), including 27 ChIP-seq proteins of RNA polymerase complex (POLR2G and POLR2A), RNA processing factors (RBFOX2, TARDBP, and RBM22), and TFs, compared to the bottom 20th percentile ones (Fig. 8b). In contrast, there were no depleted ChIP-seq proteins with  $< 1 \times 10^{-4}$   $p$  value in the top 10th percentile eTR-CNVs. These results suggest that the top 10th percentile eTR-CNVs are involved in transcriptional regulation.

To examine whether eTR-CNVs are involved in diseases and quantitative traits (QTs), 37 disease-associated and 109 QT-associated genes collected from the GWAS catalog were used for association tests. The gene-based eTR-CNVs with non-redundant genes



**Fig. 8** eTR-CNVs associated with gene expression. **a** Gene regions enriched in eTR-CNVs. We determined the enrichment of the top 10th percentile of site-based eTR-CNVs ( $N = 5799$ ) against the bottom 20th percentile eTR-CNVs ( $N = 11,657$ ) in each gene region shown at the left. ORs are indicated by blue circles with confidence intervals. Flank1Kb, Flank5Kb, and Flank100Kb indicate the regions within 1 Kb, 5Kb, and 100 Kb from the terminal exons, respectively. On the right side, the number of eTR-CNVs in the top 10th percentile and the bottom 20th percentile are shown separated by colons. **b** Chip-538 proteins enriched for eTR-CNVs. In the site-based eTR-CNVs overlapping with the Chip-seq peaks for the proteins shown on the left, the enrichment of the top 10th percentile eTR-CNVs against the bottom 20th percentile eTR-CNVs was tested. Proteins are classified by different colors, as in Fig. 7d. ORs and the number of eTR-CNVs are shown as in **a**. All the indicated associations showed  $p$  values of  $< 1 \times 10^{-6}$  in the test. **c** Disease-associated genes enriched in eTR-CNVs. In the gene regions (exons, UTRs, introns, and flank-100Kb) of disease-associated protein-coding genes from the GWAS catalog, the enrichment of gene-based eTR-CNVs ( $N = 17,160$ ) in the indicated percentile range against those in the bottom 20th percentile was tested. ORs are shown as colored bars with confidence intervals: blue: schizophrenia (SCZ), orange: attention deficit hyperactivity disorder (ADHD), gray: Parkinson disease (PD), yellow: multiple sclerosis (MS), cyan: myocardial infarction (MI), green: atrial fibrillation (AF), pale blue: hypertension, pink: lung carcinoma (LuCa), pale green: basal cell carcinoma (BCC), light gray with edging: average for all 37 diseases (ALL), and pink with edging: Random as control. The “Random” control represents the average OD from 1000 enrichment tests with 800 randomly selected protein-coding genes. **d** Quantitative traits-associated genes enriched in eTR-CNVs. In the gene regions of QT-associated protein-coding genes from the GWAS catalog, the enrichment of eTR-CNVs was tested as in **c**. ORs are shown as colored bars with confidence intervals: blue: body height, orange: body mass index (BMI), gray: diastolic blood pressure (DBP), yellow: electrocardiography (ECD), cyan: forced expiratory volume (FEV), green: bone density (BMD), light blue: chronotype measurement, pink: risk-taking behavior, light green: smoking behavior, light gray with edging: average for all 109 traits (ALL), and pink with edging: Random as control.

( $N = 17,160$ ) in the upper percentiles, such as the top 20th and 20–40th percentiles, were significantly enriched for genes associated with several common diseases (Fig. 8c), including schizophrenia (SCZ), attention deficit hyperactivity disorder (ADHD), Parkinson disease (PD), multiple sclerosis (MS), myocardial infarction (MI), atrial fibrillation (AF), hypertension, lung carcinoma (LuCa), and basal cell carcinoma (BCC), compared

to eTR-CNVs in the lower 20th percentile and those overlapping with randomly selected genes (Random) in permutation tests. For QTs, the eTR-CNVs in the upper percentiles were significantly enriched for genes associated with several QTs (Fig. 8d), including diastolic blood pressure (DBP), electrocardiography (ECD), forced expiratory volume (FEV), bone density (BMD), chronotype, risk-taking behavior, and smoking behavior, compared to those in the lower 20th percentile and the Random controls. These results suggest that the transcriptional regulation of several disease-associated and QT-associated genes is regulated by eTR-CNVs.

## Discussion

This study shows that long read alignments often generate multiple INSS and DELs in a TR region even though these variants are in the same allele and consist of the repeat unit copies of the TR region. In addition, INSS found in TR regions may contain sequences unrelated to the TR unit, such as mobile elements. TRsv overcomes these limitations by integrating multiple variants found in a single long read alignment and by carefully checking INS sequences. In evaluations using simulated and real long read datasets, TRsv showed higher accuracy and sensitivity in detecting TR-CNVs, especially non-TR-INSS, than other existing long read-based TR-CNV detection tools. TRsv showed the shortest runtime comparable to existing tools, with no limitation on the number of TR regions and unit size. Furthermore, in detecting variants outside the TR regions, TRsv was one of the most accurate and sensitive long read-based SV detection tools and detected a higher number of indels with almost 100% precision than PEPPER and GATK4.

High-quality TR-CNVs detected using TRsv from 138 PacBio HiFi long read WGS data showed many characteristics specific to tandem repeat regions. This is the first population-scale analysis for TR-CNVs using long read WGS data and should provide a more realistic view of TR-CNV structure than previous studies using short read data. TR-CNVs exhibited considerably higher AF than SVs outside the TR regions, with larger TR-CNVs in variant size having higher AFs than smaller ones. Clear peaks of integer copy number frequencies of TR repeat units were observed in any unit size ranges. These observations together with the observed enrichment of DNA damage response factors and DNA replication origins in TRs/TR-CNVs regions suggest that repeat unit-based expansion and contraction in the TR regions occur at a high frequency by the mechanism of DNA replication stress at the TR sites causing DNA repair-induced mutations possibly via non-HR repair pathway. Many of the high-frequency variations observed in the TR regions do not appear to be lost by negative selection and may serve as human-specific variations that distinguish us from other primates, as supported by our observation that HAQER is significantly enriched in the TR/TR-CNV regions. The GC-rich triplet motifs in TRs, especially TR-CNVs, were significantly enriched in the 5'-UTRs and the coding exons. This suggests that the biased presence of expanded GC-rich triplet motifs in CDSs and 5'-UTRs that cause repeat expansion diseases is due to the intrinsic localization property of the GC-rich triplet TRs rather than a selected effect of triplets. The unit size in TRs was significantly biased to 3 bp, 6 bp, and 9 bp in the 5'-UTRs and coding exons but not in introns and intergenic regions, suggesting peptide translation of TRs in the 5'-UTRs. These observations support a proposed hypothesis

that repeat expansion diseases involve aberrant translation of GC-rich triplets that are inherently abundant in 5'-UTR and CDS [8, 12].

Many TR and TR-CNV regions are unlikely to be involved in transcriptional regulation because TRs/TR-CNVs were depleted in TF-binding sites and transcriptional regulatory regions. However, given the enrichment of several specific TFs in the TR-CNV region, it is likely that some variable TR sites are involved in transcriptional regulation. The gene expression-associated eTR-CNVs in the upper percentile, identified in this study, are enriched in exons and gene proximal regions and in sites where mRNA transcription and processing factors are recruited. These observations suggest that highly significant eTR-CNVs are involved in transcriptional regulation. In particular, several eTR-CNVs likely regulate the expression of genes associated with several common diseases and quantitative trait. Thus, eTR-CNVs may be strong alternative candidates for causal variants other than SNVs and indels identified by GWAS for common disease and trait. The results of the TR-CNV analysis indicate that variants observed within TR regions should be analyzed in terms of repeat unit sequence and copy number rather than considering them as SVs and indels.

## Conclusions

TRsv accurately discriminates between TR-INSs and non-TR-INSs at any TR sites to identify TR-CNVs and SVs/indels simultaneously, and further annotates INSs with homology to ME and other genomic regions as well as those containing tandem repeats outside TR regions. The population-scale analysis for TR-CNVs using a total of 160 long read WGS data, including 138 HiFi WGS data, provided a realistic view of TR-CNV structure. These results showed that TRsv can indeed detect high-quality TR-CNVs in diverse data and can be used to identify variants associated with gene regulation and trait expression. In addition, variants identified using 138 HiFi long read WGS data from diverse populations can be used as a reference panel for TR-CNVs, SVs, and indels.

## Methods

### WGS datasets

A summary of WGS datasets used in this study is presented in Additional file 2: Table S1. Illumina short-read and PacBio HiFi long-read WGS datasets of NA12878 were obtained from 1KGP ([ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/1000G\\_2504\\_high\\_coverage/data](ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data)) and NCBI (SRR9001768-SRR9001773), respectively. WGS datasets for HG002 were obtained from GIAB ([https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data\\_indexes/AshkenazimTrio/sequence.index.AJtrio\\_Illumina300X\\_wgs\\_07292015\\_updated.HG002](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data_indexes/AshkenazimTrio/sequence.index.AJtrio_Illumina300X_wgs_07292015_updated.HG002)) for short reads and from HPRC (SRR13684284, etc.) for long reads. ONT Nanopore long read WGS data for HG002 (SRR13062493, SRR13062494) and NA12878 (SRR15058167) were obtained through the European Nucleotide Archive (ENA, <https://www.ebi.ac.uk/ena/browser/home>). The PacBio HiFi/CCS long read WGS datasets for HG002 and NA12878 had coverage of  $39.9\times$  and  $30.2\times$ , respectively, and N50 read lengths of 19.1 Kb and 10.0 Kb, respectively. The coverage of the Nanopore long read WGS data of HG002 and NA12878 were  $31.6\times$  and  $31.1\times$ , respectively, and the N50 read length was 50.9 Kb and 17.1 Kb, respectively. The Illumina short-read WGS datasets for NA12878 and HG002 are 150 bp and 148 bp paired-end reads with  $36.7\times$  and  $30\times$  coverage, respectively. Other

long read WGS datasets were obtained from HPRC (<https://humanpangenome.org/data/>), HGSC ([ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/HGSC2](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSC2) [49, 50]), and others through ENA. All short and long reads were aligned to GRCh37 (hs37d5) or GRCh38 (GRCh38\_full\_analysis\_set\_plus\_decoy\_hla.fa) using bwa mem (v0.7.17, <https://github.com/lh3/bwa>) for short reads and Minimap2 (v2.24) [32] with `-MD -ax map-hifi` (`map-ont` for Nanopore and `map-pb` for PacBio CLR) options for long reads.

#### Reference TR data

The reference TR dataset (refTR20-10000) used for the generation of HiFi-var3, HiFi-var50, and R59-var20 datasets was based on a TRF-based tandem repeat file (`simpleRepeat.txt.gz`) obtained from UCSC (<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/>) and a HipSTR reference bed file [41] (GRCh37.hipstr\_reference.bed.gz) obtained from <https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/>. TR regions ranging from 20 to 10,000 bp from both files were used. In the `simpleRepeat.txt` file, when the overlap length of adjacent repeat regions exceeded 30% of the length of either region, either overlapping site was deleted. STR sites of the HipSTR repeats were merged into the `simpleRepeat.txt` file if the overlap length with the `simpleRepeat` data is less than 20% of the length of HipSTR repeat region. The resulting reference TR regions consisted of 1,123,327 TRs, including 442,943 HipSTR repeats. To find TR sequences with similarity to MEs, the reference TR unit sequences of 30 bp or more were searched for the sequences of the MEs, including ALU, LINE1, SVA, and HERVK using `blastn` [51] and `yass` [52] aligners. ME-homologous TRs were selected when the repeat units exhibited  $\geq 90\%$  sequence identity and  $\geq 80\%$  alignment coverage of the repeat unit length of  $\geq 100$  bp ( $> 90\%$  coverage for repeat unit sizes of 30–49 bp and  $\geq 85\%$  coverage for 50–99 bp) in either alignment of `blast` or `yass`, yielding 5334 ALU, 353 LINE1, 2697 SVA, and 3 HERVK TRs. The homologous TE types were annotated at the corresponding site in the reference TR file. Similarly, the GRCh38-based reference TR dataset (refTR20-10000.b38) was created using a TRF-based tandem repeat file obtained from UCSC (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/simpleRepeat.txt.gz>) and a HipSTR reference bed file converted to GRCh38-based coordinates using `liftOver`.

#### Reference TR-CNV data

For the evaluation of the TR-CNV detection, we used the GIAB HG002 TR benchmark catalog (HG002\_GRCh38\_TandemRepeats\_v1.0.1.vcf.gz) [53] as real reference data. This catalog has been generated by determining repeat units of INs and DELs and their genomic regions, identified based on haplotype-resolved long read assemblies, using TRF. TRF may identify repeat motifs with different start and end sites and sequences, depending on its options specified and repeat contents, which may affect the evaluation results especially for TR-CNV detection tools that identify repeat motifs alone. To alleviate this issue, we extracted  $\geq 50$  bp TR-CNVs ( $N=6058$ ) from the GIAB HG002 data with `TRRepeat` tag matching the refTR20-10000 reference data and further selected those with the same TR start coordinates and repeat units between these two datasets. Since the GIAB HG002 TR data was derived from haplotype-resolved assemblies, it contained duplicated variants and ambiguous variants that were not detected by the

alignment-based method. One of the duplicates with a variant length ratio of 0.8–1.25 was removed from the selected TR-CNVs, and for each TR sites, variants that were only supported by less than half of the  $\geq 50$  bp calls detected by the TR-CNV detection tools used in this study were manually visually inspected (see the following section), resulting in the removal of 451 variants and correction of 346 TR sites. In addition, 37 TR-INSs whose INS sequences consisted of less than 50% of the TR repeat units were classified as non-TR-INS. additional 107 non-TR-INSs were selected from the GIAB HG002 TR variants, which do not have a TRRepeat tag and are located within the refTR20-10000 reference TR regions. The resulting HG002 reference TR-CNV data consists of 3109  $\geq 50$  bp TR-INSs, 2416  $\geq 50$  bp TR-DELs, and 144  $\geq 50$  bp non-TR-INSs at 4746 TR sites.

#### Reference datasets for SVs outside TR regions

The reference SVs outside TR regions for NA12878 were based on the long read-based haplotype-resolved HGSVC variant data [50] (GRCh38-based, variants\_freeze4\_sv\_ins-del\_alt.vcf.gz), which was obtained from [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/HGSVC2/release/v2.0/integrated\\_callset/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/). The variants corresponding to NA12878 were extracted from the vcf file. The reference SV dataset for HG002 was based on the GIAB Tier1 v0.6 benchmarked SV sets [54], which was obtained from [ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002\\_NA24385\\_son/NIST\\_SV\\_v0.6/](ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/). SVs of  $< 50$  bp were removed from the GRCh37-based HG002 SV set. Since the number of true SVs was still insufficient in these reference SV sets, we supplemented them with the high-confidence SVs that were shared among multiple long read-based detection tools. To do so, we detected SVs from GRCh37-based and GRCh38-based HiFi long-read WGS data using nine long read-based SV detection algorithms (cuteSV, dysgu, NanoVar, pbsv, Sniffles2, SVDSS, SVIM, SVision-pro, and TRsv), and selected SVs that were commonly detected by at least four tools. The selection of commonly called SVs was based on breakpoint (BP) distances of  $\leq 200$  bp and length ratio of 0.5–2.0 for INS and  $\geq 50\%$  reciprocal overlap for the other types. The SV genotype shared by the most tools at a given site was used as the reference genotype, but if there were two top genotypes shared by the same number of tools, the genotype was considered undefined. The high-confidence SVs selected for NA12878 and HG002 were merged without redundancy with the GRCh38-based HGSVC reference SV set and the GRCh37-based GIAB SV set, respectively. Variants within the TR regions (refTR20-10000 and  $> 10$  Kb TR regions of simpleRepeat.txt) were removed from the two reference SV sets, resulting in total SVs of 8224 for NA12878 and 8342 for HG002.

#### Simulated TR-CNV datasets

Three sets of simulated diploid genomes were created using GRCh38 chromosome 1 (chr1) to simulate TR-DELs, TR-INSs, and non-TR-INSs. A total of 1000 simulated TR-DELs of 60 bp, 100 bp, 200 bp, 500 bp, and 1000 bp were introduced on diploid chr1 in a 4:4:3:2:2 ratio. The simulated TR-DELs were made within randomly selected TR regions from refTR20-10000.b38, with a heterozygous to homozygous ratio of 2:1, a repeat unit size of 2–100, and a TR region size of  $\geq 100$  bp. Similarly, a total of 1000 simulated TR-INSs of 60 bp, 100 bp, 200 bp, 500 bp, 1000 bp, and 2000 bp were introduced on diploid chr1 in a 4:4:3:2:2:1 ratio. The simulated TR-INSs were made within randomly selected

TR regions from refTR20-10000.b38, with a heterozygous to homozygous ratio of 2:1, a repeat unit size of 2–50, and a TR region size of  $\geq 50$  bp. The sequences of TR-INSs consisted of copies of the genome sequences of the selected TR sites. To mimic multiple/split TR-CNV alignments often observed in real TR sites, 100 (10%) variants each of the simulated TR-DELS and TR-INSs were split into two segments, each half the size, and inserted into the selected TR sites. In the non-TR-INS simulation, 500 transposable element insertions and 500 tandem repeat insertions different from the repeat unit of the TR site being inserted were inserted into the diploid chr1. The retro-transposable elements (300 of  $\geq 200$  bp ALUs, 100 of  $\geq 1$  Kb LINE1s, and 100 of  $\geq 1$  Kb SVAs) were randomly selected from the RepeatMasker file obtained from UCSC (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz>). The other 500 non-TR-INSs were generated as in the same manner as the simulated TR-INSs, but their sequences were created by converting A, C, G, and T in the TR repeat unit of the inserted TR site to C, A, T, and G, respectively. Exceptionally, the dinucleotide repeat units of TG and CA were converted to AA and GG, respectively. In all the three simulated genomes, 0.1% of SNVs were introduced at randomly selected positions with a heterozygous to homozygous ratio of 2:1. Simulated PacBio HiFi long reads were generated with the created diploid chromosome 1 using a Badread long read simulator [55] with options, `-quantity 20x -length 20,000,2000 -error_model pacbio2021 -qscore_model pacbio2021 -identity 30,3`. Each simulated long read data were aligned against the GRCh38 chr1 using Mini-map2. Non-overlapping 984 TR-DELS, 992 TR-INSs, and 996 non-TR-INSs were used to evaluate the TR-CNV detection tools.

### Variant calling

The tools used to detect TR-CNVs, SVs, and indels and their commands, options, and filtering conditions are described in Additional file 3: Supplementary Note. As input alignment files, all tools used bam files aligned to the GRCh37 (hs37d5.fa) and GRCh38 (GRCh38\_full\_analysis\_set\_plus\_decoy\_hla.fa) references with decoy sequences. For tandem-genotypes (v1.9.2) [28], maf files created with the LAST alignment tool and long read data were used. The other TR-CNV detection tools, including TRsv, were run with simulated or HG002 GRCh38-based bam files using the corresponding TR reference file. For RepeatHMM-Scan (v2.0.3) [26], only TRs with a repeat unit size of 1–15 bp were used, meeting the execution requirements of RepeatHMM. TRsv was run with the default options, except that the minimum size of TR-CNV/SV options was 3 for HiFi-var3, 20 for R59-var20, or 50 for HiFi-var50. The main default options of TRsv (v1.1) are minimum mapping quality of 1, minimum number of TR-CNV-supporting reads of 3 (5 for non-HiFi reads), minimum number of SV-supporting reads of 2 (3 for non-HiFi reads), minimum ratio of TR-CNV-supporting reads to read depth of 0.15, and minimum ratio of SV-supporting reads to read depth of 0.05. SV calling and indel calling were performed using cuteSV (v1.0.13) [35], Dysgu (v1.3.10) [36], NanoVar (v1.4.1) [37], pbsv (v2.8.0, <https://github.com/PacificBiosciences/pbsv>), Sniffles2 (v2.2) [38], SVDSS (v1.0.5) [39], SVIM (v2.0.0) [40], SVision-pro (v2.4) [56], PEPPER (r0.8) [43], GATK4 (v4.1.2) [44], and TRsv (v1.1), and only the variants outside the TR regions were used for evaluation. To detect fragmented TR-CNVs, bam alignment files of PacBio HiFi reads

were created using Minimap2 (v2.24) [32], NGM-LR (v0.2.7) [33], and Winowmap2 (v2.03) [34].

#### **Evaluation of TR-CNV detection tools**

To evaluate the performance of TR-CNV detection tools, including tandem-genotypes (v1.9.2), RepeatHMM (v2.0.3), PacMonSTR (v1.0), Straglr (v1.5.3), LongTR (v1.1), TRGT (v1.5.1), TREAT (v1.0.0), and TRsv (v1.1), we used long-read simulation data and real HG002 long-read data, which contained  $\geq 50$  bp TR-CNVs and non-TR-INSs. The simulated PacBio HiFi long read data for the simulated TR-CNV-containing GRCh38 chr1 genomes were prepared as in the previous section. The TR-CNV detection tools were run with the simulated and real long read alignment data and the corresponding reference TR bed files, and the resulting TR-CNV calls were evaluated with the corresponding TR-CNV reference data. For each reference TR region, the call was considered as TP when the length ratio of the called TR-CNV (INS or DEL) to the matched reference TR-CNV (INS or DEL) ranged from 0.8 to 1.25 for the simulated data and 0.67 to 1.5 for the real data (the evaluation parameters were summarized in Additional file1: Table S5). In the evaluation for non-TR-INS data, INSs called as standard INSs (non-TR-INSs), which had no “TR:CNV” tag in the fifth column of the output vcf file, were evaluated as TP for the TRsv tool, while for the other tools, TPs were considered when the called INS sequence was found to contain only less than 50% reference repeat content. For LongTR, variants with INEXACT\_ALLELE tag in the output vcf file were considered as TP for non-TR-INS data. Precision and recall for each type of TR-CNV were determined by dividing the total number of TP calls by the total number of calls and by the reference variants, respectively.

#### **Evaluation of short indel detection tools**

To compare TRsv performance to PEPPER or GATK4,  $\leq 100$  bp indels shared between the tools were determined using PacBio HiFi long read and Illumina short read WGS data of NA12878. The indels tested were limited to those outside the TR regions of the reference TR data (refTR20-10000). Two insertions (inss) from two tools were considered shared when the distance between the two inss was within 10 bp and the length ratio of the two inss was between 0.5 and 2. For  $\leq 3$  bp inss, when the distance of two inss was within 3 bp and the difference in length was 0 bp or 1 bp, they were considered shared. Two deletions (dels) from the two tools were considered shared based on the criterion of  $\geq 50\%$  reciprocal overlap. For  $< 3$  bp dels, they were considered shared when the distance of the two dels was  $\leq 2$  bp. Unique indels not shared between the tools were extracted from each tool's call set. To estimate the precision of indels in the unique call sets, 100 inss and 100 dels were randomly selected from the unique call set and manually validated by visual inspection with the IGV viewer (see the following section). The number of TP calls determined in the validation test was used to estimate the precision of unique calls for each type.

#### **Evaluation of SV detection tools**

SV detection tools, including cuteSV, dysgu, NanoVar, pbsv, Sniffles2, SVDSS, SVIM, SVision-pro, and TRsv, were evaluated using the SV reference datasets of GRCh38-based

NA12878 and GRCh37-based HG002. Using PacBio HiFi long read WGS data of NA12878 and HG002 and Nanopore long read WGS data of HG002, INSs, DELs, and DUPs were called with each tool, and  $\geq 50$  bp SV calls with  $> 2$  supporting reads outside the TR regions (refTR20-10000 and  $> 10$  Kb TR regions of simpleRepeat.txt) were used for evaluation. When two SVs from the same SV call set were located within 200 bp for INSs (within  $0.1 \times$  length for DELs) and the length ratio of these SVs was between 0.9 and 1.11, only one of them was used for evaluation. INS calls were considered TP when the distance of an INS call and a reference INS was  $\leq 100$  bp and the length ratio of these two INSs was between 0.67 and 1.5 (the evaluation parameters were summarized in Additional file1: Table S5). INSs, especially small INSs, may involve tandem duplication, in which case the first or second BP of the DUP is assigned as the INS BP, depending on the tool. For this reason, the threshold of INS distance was set to 100 bp, which is relatively large for long read-based detection tools. The other types of SV calls were considered TP based on the criterion of  $\geq 50\%$  reciprocal overlap between the call and reference SVs. DUP calls were added to INS calls with the first BP of DUP as the BP of INS, since DUP is a type of INS and some long read-based SV detection tools do not or rarely call DUP. Precision and recall of each type of SV were determined by dividing the total number of TP calls by the total number of calls and by the number of the reference SV, respectively. Precision for SV genotype was determined using SVs that matched the reference SVs. When the genotype of the called SV was equal to the genotype of the matched reference SV, the called genotype was considered true. When the matched reference SV genotype was undefined (i.e.,/.), the called genotype was excluded from the genotype evaluation test. Genotype precision was determined by dividing the number of true genotypes by the number of tested SVs with matched reference SVs.

#### Validation of variants by manual visual inspection

In some analyses, the detected variants were validated by manually observing evidence supporting the presence of the variant in the HiFi long-read alignment using the IGV viewer (<https://igv.org>). For the curation of the HG002 reference TR-CNV data (see Reference TR-CNV data section in Methods), when the ratio of the length of TR-INS or TR-DEL in the data to the length of the variant observed in the long-read alignment (supported by at least three reads) was 0.67 and 1.5, the TR-CNV was considered as matched variant. When the alignment corresponding to a TR-INS or TR-DEL was fragmented within the TR region, the sum of the lengths of multiple INSs or DELs was considered the size of TR-CNV at the TR site. For the short indel call validation (Fig. 5), we validated 100 inss and 100 dels that were randomly selected from a non-shared unique call set between the two tools with NA12878 PacBio HiFi long-read alignment data. The variants observed in the alignment had to be supported by at least two reads. The criteria for determining TP were the same as the criteria for determining the variants shared between two short indel detection tools (see the previous section), and the confidence intervals were determined as in [57]. In the validation of the SV calls for TRsv (Additional file 1: Table S3), the variants observed in the alignment had to be supported by at least three reads. The criteria for determining TP were the same as for the evaluation of SV detection tools (see the previous section). In the validation of the DUP calls (Additional file 1: Fig. S2), for a DUP to be considered a TP, the read alignment corresponding

to a DUP call had to have read coverage of at least 1.2-fold the read coverage of the flanking regions and have at least two read alignments around the first and second BPs of the DUP, each with 5'-clipped and 3'-clipped ends, respectively. Alternatively, if an INS of 0.8 or more times the size of the DUP call was observed within the DUP region in the long-read alignments, the call was considered TP.

### TRsv algorithm

An overview of the TRsv algorithm is shown in Fig. 2. TRsv calls TR-CNVs within TR regions and SVs and short indels outside TR regions separately. The main input files include a bed file specifying TR regions and a BAM/CRAM alignment file of PacBio HiFi, PacBio CLR, and ONT Nanopore long reads. The absolute values described in this section are the default values, and many of them can be specified with the options of TRsv. A minimum size of variants is specified with "min\_str\_len" for TR-CNV and "min\_len" for indel/SV. Secondary alignment reads, reads with mapping quality 0, and reads with < 200 bp (500 bp for non-HiFi reads) mapping length were filtered by default. Read depths were calculated for every 50 bp, and average read depths per chromosome were also calculated. When the ratio of read depth to the mean read depth in an alignment region was greater than 15, the variants in that alignment region were removed. Based on the TR regions specified in a bed file, read alignment regions were separated into TR and non-TR regions, and TR-CNVs inside the TR regions and SVs/indels outside the TR regions were detected separately.

### TR-CNV data collection

INSs and DELs 0.8 times "min\_str\_len" within TR regions were collected from each read using the CIGAR strings in the alignments. When  $\geq 50\%$  length of the DEL overlapped a TR region, the DEL was considered a TR-DEL. An INS was considered a TR-INS when it was located within a TR region or its 20 bp flanking regions since INSs located in close proximity to a TR region were often repeat expansion of the TR unit. Fragmented variants were often observed within a TR region of a single read alignment. Therefore, when multiple INSs or DELs were present in the same read within a TR region, these fragmented variants were integrated into a single variant by summing the lengths of the variants. If both INS and DEL were present in the same read within a TR region, the sum of positive length of INSs and the negative length of DELs was calculated and the absolute value of the sum was taken as the size of the integrated variant. Because soft-clipped ends of read alignments near the boundaries of a TR region suggest the duplication of that entire TR region, those clipped-end points were also collected.

### TR-CNV allele definition

The INS and DEL data collected within TR regions were clustered based on their size: two variants of the same type were classified to the same cluster when the ratio of their lengths was between 0.91 and 1.1 for HiFi data (0.67 and 1.5 for non-HiFi data). The number of reads supporting the clustered variants was determined for each cluster as "READS" and the ratio of READS to total reads in the TR region as "VRR." A TR region was considered to contain the DUP of the entire TR region when there were 5'- and 3'-clipped ends with  $\geq 0.15$  VRR at the 5'- and 3'-boundaries of a  $\geq 200$  bp TR,

respectively, and the read depth of the TR region was >1.2-fold greater than those of the flanking regions. The TR-DUP copy number was calculated based on the ratio of the TR read depth to the flanking read depth, named “DPR,” and classified to an INS cluster whose INS length was the TR length multiplied by the DUP copy number. If the standard < 20 Kb DEL or < 20 Kb DUP spanned a TR region, a TR-DEL or TR-INS allele corresponding the TR region size was assigned. When a variant was shared between overlapping TR regions, it was moved to a matched cluster with the highest number of supporting reads in either TR region. Clustered variants with < 3 READS (< 5 READS for non-HiFi reads) or < 0.15 VRR were removed. If there was a single cluster in a TR region, that cluster was defined as a TR-CNV allele. If there were multiple clusters in a TR region, the top two clusters with the highest READS were defined as TR-CNV alleles.

#### ***TR-CNV classification and copy number determination***

To examine if the INSs detected within TR regions were INSs of mobile elements (MEs), the sequences of TR-INS alleles, excluding alleles defined as TR-DUP and their TRs with TE homology, were searched for ME sequences using the yass pairwise sequence alignment tool [52] with the command, `yass -O 10 -m 15`. If the INS was a sequence merged from fragmented INSs, each fragmented INS was searched for MEs. The ME reference set, which was provided as an external fasta file, included ALU, LINE1, and SVA. A TR-INS was classified as a standard INS with ME INS annotation when the length of the alignments between the INS and an ME was  $\geq 60\%$  of the INS length,  $\geq 160$  bp for ALU,  $\geq 800$  bp for SVA, and  $\geq 800$  bp for LINE1. The TR-INSs with no similarity to MEs were examined to see if the INS sequence was composed of TR repeat units. For TRs with < 5 bp unit sizes, we determined the sequence content in the INS sequences that matched the sequence patterns of the repeat unit exactly (e.g., AGC, GCA, and CAG for a AGC unit) and allowed 1 bp mismatch for unit sizes of 2–4 bp to match those sequence patterns. When the sequence of the matched regions was < 50% of the INS length or had < 70% identity, the INS was classified as a standard INS with TR annotation. For TRs with  $\geq 5$  bp unit sizes, the TRF tandem repeat finder [58] and Multalin pairwise aligner [59] were used to determine the sequence content in an INS sequence that had homology to the repeat unit sequence. TRF was conducted with the following command: `trf “INS_seq” 2 7 7 80 10 “min_score” 2000 -d -l 1 -h -ngs`, where “INS\_seq” is a query INS sequence and “min\_score” is 10, 20, 30, or 50 for < 30 bp, 30–49 bp, 50–99 bp, or > 100 bp INSs, respectively. When a repeat unit sequence from the TRF output contained the TR unit or vice versa, the content of the sequence corresponding to the repeat was determined. Multalin alignment was conducted between the INS sequence and the TR unit sequence. The sequence content in the INS sequence of the TR unit with  $\geq 70\%$  identity in the Multalin alignments was determined. The better result between TRF and Multalin search results was adopted. If the INS was a sequence derived from fragmented INSs, each fragmented INS was searched for the unit sequences and the sum of the lengths of the matched sequences was determined. When the TR unit content in the INS sequence was  $\geq 50\%$  of the INS sequence length, the TR unit copy number was calculated by dividing the INS length corresponding to the TR unit content by the TR unit size. If the content of TR unit in the INS sequence was < 50% of the INS sequence length or the TR

unit copy number was  $< 0.02$ , the TR-INS allele was classified as a standard INS with TR annotation.

### ***SVs and short indels outside TR regions***

#### **(1) INS detection.**

Proximal INSs outside TR regions aligned in the same read were merged using one of the parameters (conditions: 1–4) shown in Additional file1: Table S6, depending on the distance and length of the two INSs. The position of the merged INSs was assigned to that of the longest INS. The INS data from all reads were combined and INSs at the same position were assigned as the same INS allele when the length ratio of two INSs was between 0.59 and 1.7. In the combined INS data, adjacent INSs from different reads were integrated as the same allele using one of the parameters (conditions: 6–9) shown in Additional file1: Table S6, depending on the distance and length ratio of the two INSs. Large INSs supported only by clipped alignments were also detected. The 5′- and 3′-clipped ends of alignment reads located within 50 bp were clustered separately and the clustered 5′- and 3′-clipped ends were used as the 5′- and 3′-BPs of an INS, respectively. INSs with a distance of up to 10 Kb between the 5′- and 3′-BPs were allowed. INSs with a distance between their two BPs contained a duplication(s) of the sequence between the BPs with  $> 1$  DPR and were assigned as INSs with DUP annotations. When there was only one clip end support on either BP for a BP-supported INS, such INSs were excluded. When there was unbalanced clipped-end supports in a BP-supported INS, such that the ratio of 5′- and 3′-clipped ends was greater than threefold, such INSs were excluded. All INSs with  $< 2$  READS ( $< 3$  READS for non-HiFi reads) or  $< 0.05$  VRR were removed.

#### **(2) DEL detection.**

Proximal DELs outside TR regions aligned in the same read were merged when the distance between DELs was  $< 500$  bp and smaller than 0.5 times the length of both DELs. The DEL data from all reads were combined, and DELs at the same position were assigned as the same DEL allele when the length ratio of two DELs was between 0.59 and 1.7. In the combined DEL data, adjacent DELs from different reads were integrated as the same allele using one of the parameters (conditions 10–12) shown in Additional file1: Table S6, depending on the distance and length ratio of the two DELs. Large DELs supported by clipped supplementary alignments of the same reads were also detected. The 3′- and 5′-clipped read ends located within 100 bp (50 bp for  $< 1$  Kb DEL and 150 bp for  $> 10$  Kb DEL) were clustered separately and the clustered 3′- and 5′-clipped ends were used as the 5′- and 3′-BPs of a DEL, respectively, and the distance between the BPs was defined as the DEL length. When a BP-supported DEL had only one clip end support on either BP or the DPR exceeded 0.85, such DELs were excluded. All DELs with  $< 2$  READS ( $< 3$  READS for non-HiFi reads) or  $< 0.05$  VRR were removed.

#### **(3) DUP detection.**

DUPs were detected using both long read split alignment and read depth (DPR) information. DUPs outside TR regions that were supported by clipped supplementary alignments of the same reads were detected. The 5′- and 3′-clipped read ends located within 100 bp (50 bp for  $< 1$  Kb DUP and 150 bp for  $> 10$  Kb DUP) were clustered separately and

the clustered 5′- and 3′-clipped ends were used as the 5′- and 3′-BPs of a DUP, respectively, and the distance between the BPs was defined as the DUP length. DUPs with < 2 supporting clipped ends at the either BPs, with < 0.05 VRR, with < 1.1 DPR, with < 100 bp BP distance, or with > 10 Mb BP distance were excluded. When a DUP had only one clip end support on either BP, a DPR less than 1.1, or a VRR less than 0.05, such DUPs were excluded.

(4) INV detection.

INVs outside TR regions that were supported by clipped supplementary alignments of the same reads were detected. When two alignments of opposite strands of the same read had a distance between 5′(3′)-clipped ends, the two clipped ends were used as the BPs of an INV, and the distance between the BPs was defined as the INV length. The 5′- and 3′-clipped read ends located within 50 bp were clustered separately. When an INV had only one clip end support on either BP, when the VRR was less than 0.05, or when the BP distance was less than 100 bp or larger than 50 Mb, such INVs were excluded.

(5) Machine learning-based filtering of SVs from non-HiFi reads.

SV call sets from PacBio CLR or Nanopore long reads contained many more FP calls than the call set from PacBio HiFi reads due to the high error rate of these reads. To predict whether an SV called with non-HiFi long reads is true or false, we used a machine learning strategy that exploits several characteristics of called SVs. SV calls of 10 samples each for PacBio CLR and Nanopore were divided into TP and FP call by using the HiFi read SV calls of the same sample as reference true SVs. Seven SV characteristics were collected for each TP and FP SV call site, including READS, VRR, SV length, read depth, mapping quality, and standard deviations of SV position and size from each read derived from the called SV. The collected data was used as training datasets with its TP and FP categories for the gradient boosting machine learning. Machine learning models for classifying true and false SVs were created using the xgboost R package (<https://cran.r-project.org/web/packages/xgboost/index.html>) and the PacBio CLR and Nanopore training datasets. The SV types were limited to INS and DEL due to the limited amount of data used for training. When evaluated on the NA12878 test data and the training data excluding NA12878 data, the prediction accuracies for DEL and INS were 0.87 and 0.71 for Nanopore and 0.8 and 0.95 for PacBio CLR, respectively. SVs predicted as FP in the xgboost model were filtered out in the final stage of TRsv. The precision and recall of INSs and DELs for NA12878 and HG002 before and after the filtering are shown in Table 2.

### **Sequence annotation of INSs outside TR regions**

INSs outside TR regions were examined for ME insertion, duplication of the flanking regions, or simple tandem repeats. First, sequence similarity to ME sequences was examined for  $\geq 90$  bp INSs, as described in the section of TR-CNV classification. INSs with significant similarity to an ME sequence were annotated as ME INSs. Second, sequence similarity between  $\geq 20$  bp INSs and their INS flanking sequences was examined using the yass aligner with the command, `yass -O 10 -m 15`. Flanking sequences 1.2 times longer than an INS were taken from both sides of the INS position. INSs with significant similarity to the flanking sequences were annotated as INS-DUPs, along with the matched length and direction. For non-ME INSs, the tandem repeat content in the INS

sequences were tested using TRF, as described in the TR-CNV classification section. For INSs with  $\geq 50\%$  tandem repeat content, a 100 bp region around the INS location was examined for overlap with any of reference TR regions. If there was an overlap with a reference TR region and the INS repeat unit was similar to the repeat unit of the reference TR region, the INS was assigned as a TR-INS with the annotation of the matched TR ID. When there was no overlap with the reference TRs and the repeat content in the INS sequence was  $\geq 50\%$ , the INS was annotated as a tandem repeat INS, which was different from the reference TRs.

#### **Low confident variant annotation**

Two categorized low confident annotations, LowConf and LowQual, were added to the FILTER field of a final output vcf file. LowConf was annotated to a variant that overlapped with the low-confident regions provided by an external file. In this study, variants that overlapped with low-quality-mapping regions observed in  $> 10$  Kb TR regions were annotated as LowConf variants. LowQual was annotated to the variant with a high percentage of reads with mapping quality 0 (MQ0) supporting the variant. This class of variants may be low-quality variants in regions where read mapping is difficult although reads with MQ0 were not used for variant calling by default. When the ratio of MQ0 variants to the sum of MQ0 and non-MQ0 variants (SAR) exceeded 0.6, the corresponding variant was annotated as a LowQual variant.

#### **Gene annotation**

Information on gene regions that overlap with TR-CNV/SV sites can be added in the INFO field of the vcf file, as described in our previous study [60]. The gene information used in this study was based on Homo\_sapiens.GRCh37.87.gff3 obtained from the Ensembl site, which contained 20,356 protein-coding genes and 21,555 non-coding genes.

#### **Genotyping**

TR-CNVs and SVs were genotyped based on VRR and DPR. For INS and INV, variants with  $< 0.7$  and  $\geq 0.7$  VRR were considered heterozygous and homozygous alleles, respectively. For TR-CNV, multi-allelic variants were considered heterozygous, and variants with  $< 0.7$  and  $\geq 0.7$  VRR were considered heterozygous and homozygous alleles, respectively. For  $< 500$  bp DELs, DELs with  $\geq 0.7$  VRRs or  $\geq 100$  bp DELs with  $\leq 0.1$  DPR were considered homozygous allele; other DELs were considered heterozygous allele. For  $\geq 500$  bp DELs, DELs with  $< 0.2$  DPR or  $\geq 0.7$  VRR were considered homozygous allele, and DELs with  $0.2-0.4$  DPR or  $< 0.7$  VRR were considered heterozygous allele. For BP-supported DELs, DELs with  $\leq 0.1$  DPR or  $\geq 0.7$  VRR were considered homozygous allele, and other DELs were considered heterozygous allele. For DUPs with  $\leq 1.5$  DPR, DUPs with  $\geq 0.8$  VRR were considered homozygous allele, and all others were considered heterozygous allele. For DUPs with  $> 1.5$  DPR, DUPs with  $\geq 0.7$  VRR were considered homozygous allele, and all others were considered heterozygous allele.

### ***Joint calling of variant call sets from multiple samples***

Joint calling of multi-sample TR-CNV/SV vcf files was based on the method described in our previous study [60]. TR-CNVs and SVs were joint-called separately. Overlapping SV sites from multiple samples were clustered primarily by the following criteria:  $\geq 80\%$  reciprocal overlap for DELs, DUPs, and INVs, and  $\leq 200$  bp BP distance for INs. Clustering was done by stepwise merging of closer sites for each SV type. Two adjacent INs were assigned to the same cluster according to the length, length ratio, and distance of the two INs. For each IN cluster, when the ratio of maximum to minimum IN size exceeded 1.8, the IN cluster was further divided into several clusters based on size. When an IN of a sample was located within 50 bp of the BPs of a BP-supported IN of another sample and the ratio of the two IN lengths (or IN BP distance) was between 0.5 and 2, these two INs were assigned to the same cluster. When the BP of a BP-supported IN were located within 50 bp of either BP of a DUP cluster, the IN was assigned to the DUP cluster. When an IN was located between the two BPs of a DUP and the ratio of the IN and DUP lengths was between 0.8 and 1.2, the IN was assigned to the DUP cluster. Median values of SV position and size and mean values of READS, VRR, DPR, and SAR were added to the POS and INFO fields of the output vcf file. For TR-CNV, the minimum and maximum lengths of TR-INS and/or TR-DEL, as well as TR ID and TR unit size, were added to INFO fields of the output vcf file. The variant genotype, length, READS, VRR, and/or TR-CNV unit copy number for each sample were added to the FORMAT field of the output vcf file.

### ***Enrichment analysis of TR-CNVs in gene regions***

TR regions in the reference TR bed file and HiFi-var3/-var50 vcf files were annotated for overlapping regions (CDS/exon, UTR, intron, and flanking regions) of a total 20,356 protein-coding genes and 23,564 non-coding genes, excluding pseudogenes. When a TR overlapped with multiple regions of a gene, the region in the top hierarchy (CDS/exon > UTR > intron > flanking) was selected. The gene flanking regions, flank1Kb and flank5Kb, were annotated for TRs overlapping the regions within 1 Kb and 5 Kb from the terminal exons of a gene, respectively. The intergenic region was defined as the region that does not overlap any gene regions including flank5Kb. The OD for TR-INS and TR-DEL occurrences in a gene region was calculated for the occurrence of non-variable TRs in the corresponding gene region. Alternatively, ODs for the occurrence of non-variable TRs and TR-CNVs in CDSs, UTRs, or introns were calculated for the occurrence of TRs and TR-CNVs in intergenic regions. For the analysis for triplet motifs of TR-CNVs, the frequency of 20 triplet sets with 1 bp shifts (e.g., CCG and CGC, and GCC for a CCG set) overlapping with the CDS, 5'-UTR, and intron of the protein-coding genes was determined. The triplet motifs overlapping the gene regions on the opposite strand were reverse complemented. As a control, 3-mer frequencies in each CDS, 5'-UTR, and intron of the protein-coding genes were counted in a strand-dependent manner. The OD for the occurrence of the triplet motif set of TR-CNVs in each gene region was calculated for the frequency of the 3-mer corresponding to the triplet motif set in the corresponding gene region.

### Enrichment analysis of TR-CNVs in functional genomic regions

Functional genomic data used in this study included human ancestor quickly evolved regions (HAQER) [45], evolutionarily conserved regions (phastCons), open chromatin regions (ATAC-seq), enhancer sequences (VISAT-Enh), TF-binding sites (TF-BS and TF-FP), and promoters in bivalent chromatin states (TssBiv). The HAQER data was obtained at <https://www.vertgenlab.org/haqer2022.html>. phastCons and VISAT-Enh files were downloaded from <https://bds.mpi-cbg.de/hillerlab/120MammalAlignm ent/Human120way/data/conservation/> and <https://enhancer.lbl.gov>, respectively. The phastCons data was converted to hg19 coordinates with liftOver. The ATAC-seq data for the NA12878 lymphoblastoid cell line was obtained at the NCBI GEO (GSE47753: <https://www.ncbi.nlm.nih.gov/geo/>) [61]. The TF-BS data was derived from Chip-seq data of 87 TFs from the NA12878 lymphoblastoid cell line, which was obtained from GREAP (<http://www.liclab.net/Greap/view/index>) [62]. The TF-FP data was 4.6 M consensus TF footprints from 243 biosamples by genomic DNaseI footprinting experiments [63]. The TssBiv data was obtained at <http://www.roadmapepigenomics.org/>. The Chip-seq data included DNA replication origins (RepOrigin) [64], DNA break sites upon ATR checkpoint kinase inhibition (BrITL) [46], and DNA damage response factors (ATM, MCM2, MCM7, NBN, RAD51, and XRCC3). The RepOrigin and BrITL Chip-seq data were obtained at the NCBI GEO site (RepOrigin: GSE128477, BrITL: GSE115623), and the coordinates of RepOrigin were converted to hg19-based ones with liftOver. The ATM, MCM2, MCM7, NBN, RAD51, and XRCC3 Chip-seq data were obtained from ENCODE (<https://www.encodeproject.org>) and converted to hg19 coordinates with liftOver. The ATM data was derived from the Hep G2 human hepatoma cell line, and the MCM2/7 and XRCC3 data were derived from the K562 leukemia cell line. The NBN and RAD51 data were mixed data from NA12878 and K562 cell lines. Other Chip-seq data (Chip-538) for 538 proteins from NA12878 or K562 cell lines were also obtained from ENCODE and converted to hg19 coordinates with liftOver.

To determine the enrichment of TRs (including polymorphic TR-CNVs) and TR-CNVs that overlap with genomic elements, we performed permutation tests using simulated genomic elements. The TR regions were restricted to those with a TR size range of 100–5000 bp. The overlaps between TR regions and genomic elements were determined using central 100 bp regions of the TR regions and central 200 bp regions of the genomic elements to reduce the size bias of TR-CNVs, which had larger TR sizes than non-variable TRs. A simulated genomic element dataset for each genomic element data was created by sampling an equal number of genomic elements from random locations in the GRCh37 genome excluding gap regions. The overlaps between the 200 bp regions of the simulated genomic elements and the central 100 bp regions of TR regions were counted. The permutation tests were repeated 1000 times to obtain the mean overlap count for each experiment. In the enrichment test with the Chip-538 data, non-variable TRs and TR-CNVs that overlapped with the CDSs, UTRs, introns, and flank5K regions of protein-coding genes were used and 200 permutation tests were performed. ORs for the TRs/TR-CNVs overlapping genomic elements were calculated using TRs/TR-CNVs

overlapping the corresponding simulated elements, along with their confident intervals and  $p$  values in the Fisher exact test.

#### Identification of eTR-CNVs associated with gene expression

A total of 59 RNA-seq data with matched PacBio HiFi or PacBio CLR WGS data were obtained from HGSVC. The paired-end RNA-seq reads for each sample were aligned with STAR against the GRCh37 reference, and gene-based count data were obtained using RSEM (rsem-calculate-expression command with `-star` and `-paired-end` options) [65]. The gene count data from 59 samples were merged using a rsem-generate-data-matrix command of RSEM, and genes with a maximum count of  $< 10$  in the merged data were filtered out. The count data were normalized by trimmed mean of M values (TMM) using the edgeR R package [66]. The TMM counts were further normalized with a rank-based inverse-normal transformation (INT), which normalized the residuals obtained by regressing gene count data on sex and super population as covariates. To obtain TR-CNV data for the RNA-seq samples,  $\geq 20$  bp TR-CNVs were detected using TRsv from 59 bam files of the corresponding samples and annotated for overlapping genes, including exons, introns, and flanking regions within 100 Kb of the terminal exons. The rationale of selecting the 100 Kb flanking region was that most of the expression quantitative traits (eQTLs) are observed within 100 kb of the transcription start sites [67]. The TR sites where less than 5 samples had TR-CNVs were removed. The copy number of repeat unit of TR-CNVs was represented with positive and negative values for TR-INSs and TR-DELs, respectively. The correlation between the TR-CNV copy number and the normalized gene expression level was evaluated by linear regression test. A total of 59,959 TR-CNVs were tested, yielding 179,872 associations. Since many TR-CNV sites and genes had multiple associations with shared genes or shared TR-CNVs, two eTR-CNV sets, a site-based eTR-CNV set with non-redundant TR sites and a gene-based eTR-CNV set with non-redundant genes, were created by selecting the eTR-CNV with the lowest  $p$  value from shared TR sites or genes. eTR-CNVs were ranked based on  $p$  values determined by the linear regression tests and fractionated by 10% or 20%. The top 10th percentile of gene-based and site-based eTR-CNVs contained 2452 and 5988 eTR-CNVs, respectively.

#### Association test between eTR-CNVs and traits

A total of 37 common disease-associated and 109 quantitative trait (QT)-associated genes were collected from the GWAS catalog (gwas\_catalog\_v1.0.2-associations\_e109\_r2023-03-03.txt), obtained at <https://www.ebi.ac.uk/gwas/>. The GWAS-genes whose associated regions were only intergenic regions were excluded, and traits with at least 100 associated genes were selected. To examine the association of eTR-CNVs with trait-associated genes, 17,160 eTR-CNVs that overlapped the transcribed regions of the protein-coding genes and the flanking regions within 100 Kb of their terminal exons were selected from 24,512 gene-based TR-CNVs that were tested for association with gene expression. The 17,160 eTR-CNVs were sorted in descending order of  $p$  value and

fractionated by 20%. The ORs of eTR-CNVs in each percentile fraction overlapping with the trait-associated genes against the eTR-CNVs in the lower 20th percentile overlapping with the trait-associated genes were determined for each trait, along with the confidence intervals and *p* values in the Fisher exact test. As controls, all genes for 37 diseases or for 109 QTs were tested for the association with eTR-CNVs. As another control, a permutation test was performed on 800 non-redundant genes randomly selected from 20,356 protein-coding genes to test for their association with eTR-CNVs. The mean OR obtained by repeating the permutation test 1000 times was determined for each trait.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03718-z>.

Additional file 1. Fig. S1 – Non-TR INS in TR region. Fig. S2 – DUPs outside TR regions, called with long read-based SV detection tools. Fig. S3 – Precision of SV genotypes, called with long read-based SV detection tools. Fig. S4 – Copy number distribution of repeat units in TR-CNVs. Fig. S5 – Size preference for TR repeat units in gene regions. Fig. S6 – Enrichment of TRs and TR-CNVs for HAQER. Fig. S7 – Examples of the association between unit copy number of TR-CNVs and gene expression. Table S2 – Numerical data of the evaluation results for TR-CNV and SV. Table S3 – Results of the evaluation of  $\geq 50$  bp TRsv SV calls outside TR regions by manual visual inspection (MVI). Table S4 – TR-CNV, SVs, indels detected using long read WGS data in this study. Table S5 – Parameters used for the evaluation of TR-CNV, SV, and short indel. Table S6 – Parameters for integrating INSs or DELs outside TR regions in the TRsv algorithm.

Additional file 2. Table S1 – Long read WGS data used in this study.

Additional file 3. Supplementary Note – Variant calling procedures.

Additional file 4. Peer review history.

## Acknowledgements

We thank Dr. Takeshi Usui (Shizuoka General Hospital) and Dr. Hideki Noguchi (Joint Support-Center for Data Science Research) for providing the environment for data analysis.

## Authors' contributions

S.K. conceived and designed the project. S.K. performed the analyses and created the TRsv code. S.K. and C.T. wrote the manuscript. All authors reviewed the paper and approved the final version.

## Funding

This work is supported by Japan Society for the Promotion of Science KAKENHI Grant JP17K07264 and JP21K06130. Japan Society for the Promotion of Science, JP17K07264, Shunichi Kosugi, JP21K06130, Shunichi Kosugi

## Data availability

TRsv code and its related data, including the reference TR datasets, are available at GitHub (<https://github.com/stat-lab/TRsv>) [68], and also available at Zenodo (<https://doi.org/10.5281/zenodo.16009888>) [69]. TR-CNV/SV vcf files (TRsv-v1.1.HiFi-var3.GRCh37.vcf.gz, TRsv-v1.1.HiFi-var50.GRCh37.vcf.gz, TRsv-v1.1.HiFi-var3.GRCh38.vcf.gz, and TRsv-v1.1.HiFi-var50.GRCh38.vcf.gz) generated using TRsv with 138 PacBio HiFi WGS data are available at Zenodo (<https://doi.org/10.5281/zenodo.16022190>) [70]. The result of the eTR-CNV association tests obtained using 59 RNA-seq data and its variant data (TRsv-v1.1.R59-var20.GRCh37.vcf.gz) is available at Zenodo (<https://doi.org/10.5281/zenodo.16022190>) [70]. The reference TR-CNV and SV data used for the tools' evaluation are also available at Zenodo (<https://doi.org/10.5281/zenodo.16024610>) [71]. The source of the long read and short read WGS data used in this study is presented in Additional file 2: Table S1, and all fastq read files are available from HGSCV ([ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/HGSCV2](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSCV2)), GIAB ([https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data\\_indexes](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data_indexes)), or ENA (<https://www.ebi.ac.uk/ena/browser/home>).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 27 February 2024 Accepted: 25 July 2025

Published online: 20 August 2025

## References

- Fan H, Chu JY. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics*. 2007;5:7–14.
- Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y, Genomes Project Chromosome YG, Erlich Y. Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am J Hum Genet*. 2016;98:919–33.
- Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T. Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *J Evol Biol*. 2023;36:321–36.
- Ghahsouran J, Hosseinzadeh H, Ghafouri-Fard S, Taheri M, Rezazadeh M. STRs: ancient architectures of the genome beyond the sequence. *J Mol Neurosci*. 2021;71:2441–55.
- Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. 2018;19:286–98.
- Zhou ZD, Jankovic J, Ashizawa T, Tan EK. Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. *Nat Rev Neurol*. 2022;18:145–57.
- Zhang S, Shen L, Jiao B. Cognitive dysfunction in repeat expansion diseases: a review. *Front Aging Neurosci*. 2022;14:841711.
- Depienne C, Mandel JL. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am J Hum Genet*. 2021;108:764–85.
- Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun*. 2021;9:98.
- Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet*. 2024. <https://doi.org/10.1038/s41576-024-00692-3>.
- Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, Margolis J, Peterson M, Markowski TW, Ingram MA, et al. Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci U S A*. 2011;108:260–5.
- Castelli LM, Huang WP, Lin YH, Chang KY, Hautbergue GM. Mechanisms of repeat-associated non-AUG translation in neurological microsatellite expansion disorders. *Biochem Soc Trans*. 2021;49:775–92.
- Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, Shleizer-Burko S, Lohmueller KE, Gymrek M. Patterns of de novo tandem repeat mutations and their role in autism. *Nature*. 2021;589:246–50.
- Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, Mirceta M, Mojarad BA, Yin Y, Dov A, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*. 2020;586:80–6.
- Mojarad BA, Engchuan W, Trost B, Backstrom I, Yin Y, Thiruvahindrapuram B, Pallotto L, Mitina A, Khan M, Pellicchia G, et al. Genome-wide tandem repeat expansions contribute to schizophrenia risk. *Mol Psychiatry*. 2022;27:3692–8.
- Erwin GS, Gursoy G, Al-Abri R, Suriyaprakash A, Dolzhenko E, Zhu K, Hoerner CR, White SM, Ramirez L, Vadlakonda A, et al. Recurrent repeat expansions in human cancer genomes. *Nature*. 2023;613:96–102.
- Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Hujoel MLA, McCarroll SA, Loh PR. Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. *Cell*. 2023;186(3659–3673): e3623.
- Margoliash J, Fuchs S, Li Y, Zhang X, Massarat A, Goren A, Gymrek M. Polymorphic short tandem repeats make wide-spread contributions to blood and serum traits. *Cell Genom*. 2023;3: 100458.
- Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, Loh PR. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science*. 2021;373:1499–505.
- Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. The impact of short tandem repeat variation on gene expression. *Nat Genet*. 2019;51:1652–9.
- Garg P, Martin-Trujillo A, Rodriguez OL, Gies SJ, Hadelia E, Jadhav B, Jain M, Paten B, Sharp AJ. Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am J Hum Genet*. 2021;108:809–24.
- Lu TY, Smaruj PN, Fudenberg G, Mancuso N, Chaisson MJP. The motif composition of variable number tandem repeats impacts gene expression. *Genome Res*. 2023;33:511–24.
- Hamanaka K, Yamauchi D, Koshimizu E, Watase K, Mogushi K, Ishikawa K, Mizusawa H, Tsuchida N, Uchiyama Y, Fujita A, et al. Genome-wide identification of tandem repeats associated with splicing variation across 49 tissues in humans. *Genome Res*. 2023;33:435–47.
- Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019;20:117.
- Ummat A, Bashir A. Resolving complex tandem repeats with long reads. *Bioinformatics*. 2014;30:3491–8.
- Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med*. 2017;9:65.
- Chiu R, Rajan-Babu IS, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol*. 2021;22:224.
- Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, Oma Y, Kino Y, Mitsuhashi H, Matsumoto N. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol*. 2019;20:58.
- Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol*. 2024;25:176.
- Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol*. 2024;42:1606–14.
- Tesi N, Salazar A, Zhang Y, van der Lee S, Hulsman M, Knoop L, Wijesekera S, Krizova J, Schneider AF, Pennings M, et al. Characterizing tandem repeat complexities across long-read sequencing platforms with TREAT and otter. *Genome Res*. 2024;34:1942–53.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.

33. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018;15:461–8.
34. Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods*. 2022;19:705–10.
35. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020;21:189.
36. Cleal K, Baird DM. Dysgu: efficient structural variant calling using short or long reads. *Nucleic Acids Res*. 2022;50: e53.
37. Tham CY, Tirado-Magallanes R, Goh Y, Fullwood MJ, Koh BTH, Wang W, Ng CH, Chng WJ, Thiery A, Tenen DG, Benoukraf T. Nanovar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol*. 2020;21:56.
38. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*. 2024. <https://doi.org/10.1038/s41587-023-02024-y>.
39. Denti L, Khorsand P, Bonizzoni P, Hormozdiari F, Chikhi R. SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nat Methods*. 2023;20:550–8.
40. Heller D, Vingron M. SVM: structural variant identification using mapped long reads. *Bioinformatics*. 2019;35:2907–15.
41. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14:590–2.
42. Bolognini D, Magi A, Benes V, Korbel JO, Rausch T. TRiCoLoR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience*. 2020. <https://doi.org/10.1093/gigascience/giaa101>.
43. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*. 2021;18:1322–32.
44. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
45. Mangan RJ, Alsina FC, Mosti F, Sotelo-Fonseca JE, Snellings DA, Au EH, Carvalho J, Sathyan L, Johnson GD, Reddy TE, et al. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell*. 2022;185(4587–4603): e4523.
46. Shastri N, Tsai YC, Hile S, Jordan D, Powell B, Chen J, Maloney D, Dose M, Lo Y, Anastasiadis T, et al. Genome-wide identification of structure-forming repeats as principal sites of fork collapse upon ATR inhibition. *Mol Cell*. 2018;72(222–238): e211.
47. Belan O, Anand R, Boulton SJ. Mechanism of mitotic recombination: insights from *C. elegans*. *Curr Opin Genet Dev*. 2021;71:10–8.
48. Lans H, Hoeijmakers JHJ, Vermeulen W, Marteijn JA. The DNA damage response to transcription stress. *Nat Rev Mol Cell Biol*. 2019;20:766–84.
49. Porubsky D, Hops W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggolini FA, Harvey WT, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*. 2022;185(1986–2005): e1926.
50. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021. <https://doi.org/10.1126/science.abf7117>.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
52. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res*. 2005;33:W540–543.
53. English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, Park J, Gu B, Wagner J, Eberle MA, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol*. 2025;43:431–42.
54. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. 2020;38:1347–55.
55. Wick RR. Badread: simulation of error-prone long reads. *J Open Source Softw*. 2019;4:1316.
56. Wang S, Lin J, Jia P, Xu T, Li X, Liu Y, Xu D, Bush SJ, Meng D, Ye K. De novo and somatic structural variant discovery with SVision-pro. *Nat Biotechnol*. 2025;43:181–5.
57. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var*. 2024;11:18.
58. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27:573–80.
59. Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. 1988;16:10881–90.
60. Kosugi S, Kamatani Y, Harada K, Tomizuka K, Momozawa Y, Morisaki T, BioBank Japan P, Terao C. Detection of trait-associated structural variations using short-read sequencing. *Cell Genom*. 2023;3: 100328.
61. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10:1213–8.
62. Yang Y, Qian F, Li X, Li Y, Zhou L, Wang Q, Zhou X, Zhang J, Song C, Yu Z, et al. GREAP: a comprehensive enrichment analysis software for human genomic regions. *Brief Bioinform*. 2022. <https://doi.org/10.1093/bib/bbac329>.
63. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al. Global reference mapping of human transcription factor footprints. *Nature*. 2020;583:729–36.
64. Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, Derelle R, Smith G, Rodriguez-Martinez M, Romano M, et al. A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat Commun*. 2020;11:4826.
65. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12: 323.

66. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.
67. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 2008;24:408–15.
68. Kosugi S, Terao C. TRsv. GitHub; 2025. <https://github.com/stat-lab/TRsv>.
69. Kosugi S, Terao C. TRsv. Zenodo. 2025. <https://doi.org/10.5281/zenodo.16009888>.
70. Kosugi S, Terao C. TR-CNVs, SVs, and short indels detected from 138 PacBio HiFi WGS data using TRsv. Zenodo. 2025. <https://doi.org/10.5281/zenodo.16022190>.
71. Kosugi S, Terao C. Reference TR-CNV and SV variants. Zenodo. 2025. <https://doi.org/10.5281/zenodo.16024610>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.