Supplementary Materials

A comprehensive assessment of tandem repeat genotyping methods for Nanopore long-read genomes

Elbay Aliyev[1*], Akshay Avvaru[1*], Wouter De Coster[2,3], Garrison M. Arner[1,4], Denis M. Nyaga[5], Sophia B. Gibson[6], Ben Weisburd[7], Bida Gu[8], Claudia Gonzaga-Jauregui[9], 1000 Genomes Long-Read Sequencing Consortium, Mark J.P. Chaisson[8,10], Danny E. Miller[6,11,12,13], Elizabeth Ostrowski[5,14], and Harriet Dashnow[1 ✉]

* These authors contributed equally
✉ Corresponding author

Supplementary Note

Algorithmic basis of tool performance

The differences in overall accuracy between tools can be traced to their distinct strategies across four key steps: read selection, repeat identification, haplogroup assignment, and consensus generation.

In read selection, tools apply different filtering criteria that directly affect sensitivity. STRkit applies a stringent mapping quality threshold, filtering out many informative reads and resulting in a high rate of uncalled loci. LongTR applies a high mean Phred quality threshold, which prevents it from making calls on R9.4.1 data where the mean basecall quality falls below this threshold. For repeat identification within reads, most tools rely on the original read alignment to define the repeat boundaries. LongTR and STRdust are exceptions, performing local realignment before determining the repeat sequence within each read. STRkit takes a motif-centric approach, aligning reads to perfect iterations of the input motif sequence.

Haplogroup assignment is arguably where tools diverge most substantially. LongTR iteratively builds candidate haplotypes from the reads and assigns reads by aligning them to these candidates. ATaRVa selects the most informative, high-quality SNPs from the reads or applies k-means clustering based on repeat lengths. Vamos segregates reads by max-cut heuristic that initializes partitions with the two reads with the most disagreeing heterozygous SNVs and assigns the remaining reads to a partition based on shared SNVs. STRkit supports multiple phasing strategies, including input haplotags, an external SNV VCF, de novo identification of informative SNVs combined with repeat length, or a Gaussian mixture model approach based

solely on repeat lengths. Straglr applies a Gaussian mixture model to segregate reads into haplotype groups based on repeat length distributions. STRdust uses haplotag information from alignment files by default, or performs hierarchical clustering based on pairwise read alignments.

Finally, consensus generation relies on partial order alignment (POA) across all tools that report allele sequences (except Straglr). Specifically, ATaRVa, Medaka Tandem, and vamos use the Python distribution of absolute banded partial order alignment implementation (abPOA, https://github.com/yangao07/abPOA/tree/main [1]), while LongTR, STRkit, and STRdust use C++ and Rust distributions of SIMD-accelerated partial order alignment (sPOA, https://github.com/rvaser/spoa [2]). The tools also slightly differ in the choice of reads within a haplogroup for consensus generation. STRdust identifies and excludes outlier reads before consensus generation. LongTR additionally applies a scoring model that aligns reads across different haplotype combinations to determine the final genotype. While all tools follow the same broad workflow of read filtering, repeat identification, haplogroup assignment, and consensus generation, it is the specific combination of strategies at each step that drives the observed differences in accuracy across tools and locus types.

Pathogenic locus genotyping presents unique challenges in tandem repeat analysis, as these loci exhibit error profiles and instability not typically observed at other genomic loci. To maximize tool sensitivity at these loci, we applied more lenient thresholds to ensure all available read information was considered. However, accurately calling expansions at pathogenic loci requires appropriately weighting longer alleles, as sequencing coverage of expanded alleles decreases with increasing length, increasing the risk of missed calls. Additionally, the inherent instability of pathogenic loci introduces higher allelic variance, requiring tools to weigh reads appropriately to accurately resolve the expanded allele.

Supplementary Note References

1. Gao Y, Liu Y, Ma Y, Liu B, Wang Y, Xing Y. abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. Bioinformatics. 2021;37:2209–11.

2. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.

Supplementary Figures

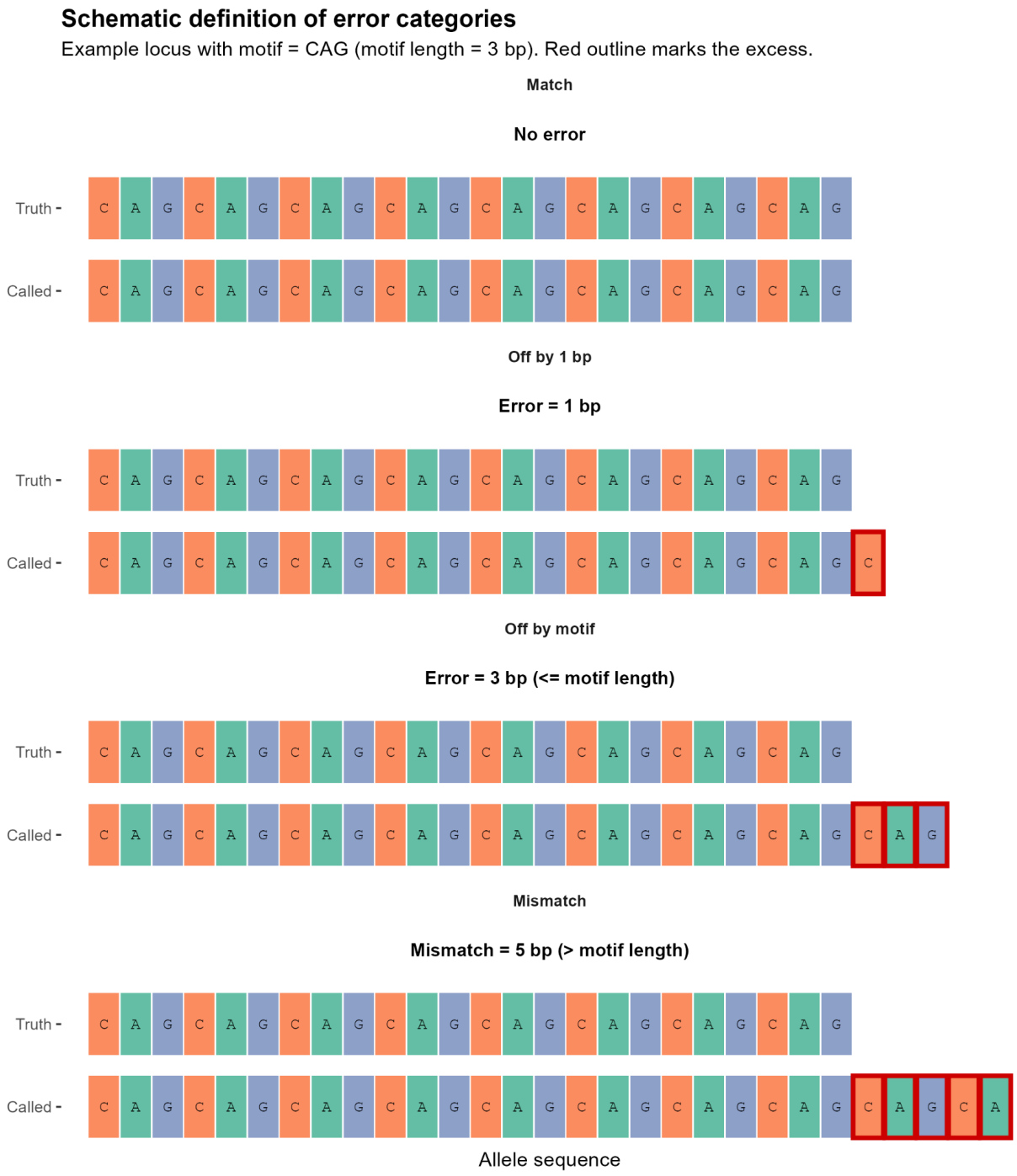

Fig. S1: **Schematic definition of error categories.** An example tandem repeat locus with motif CAG (motif length = 3 bp) is shown to illustrate how allele-length differences were classified. Match indicates no difference between the truth and called allele lengths. Off by 1 bp indicates an absolute error of exactly 1 bp. Off by motif indicates an absolute error greater than 1 bp but

less than or equal to the motif length of the locus. Error indicates an absolute error greater than the motif length. The red outline illustrates excess sequence in the called allele relative to the truth allele.

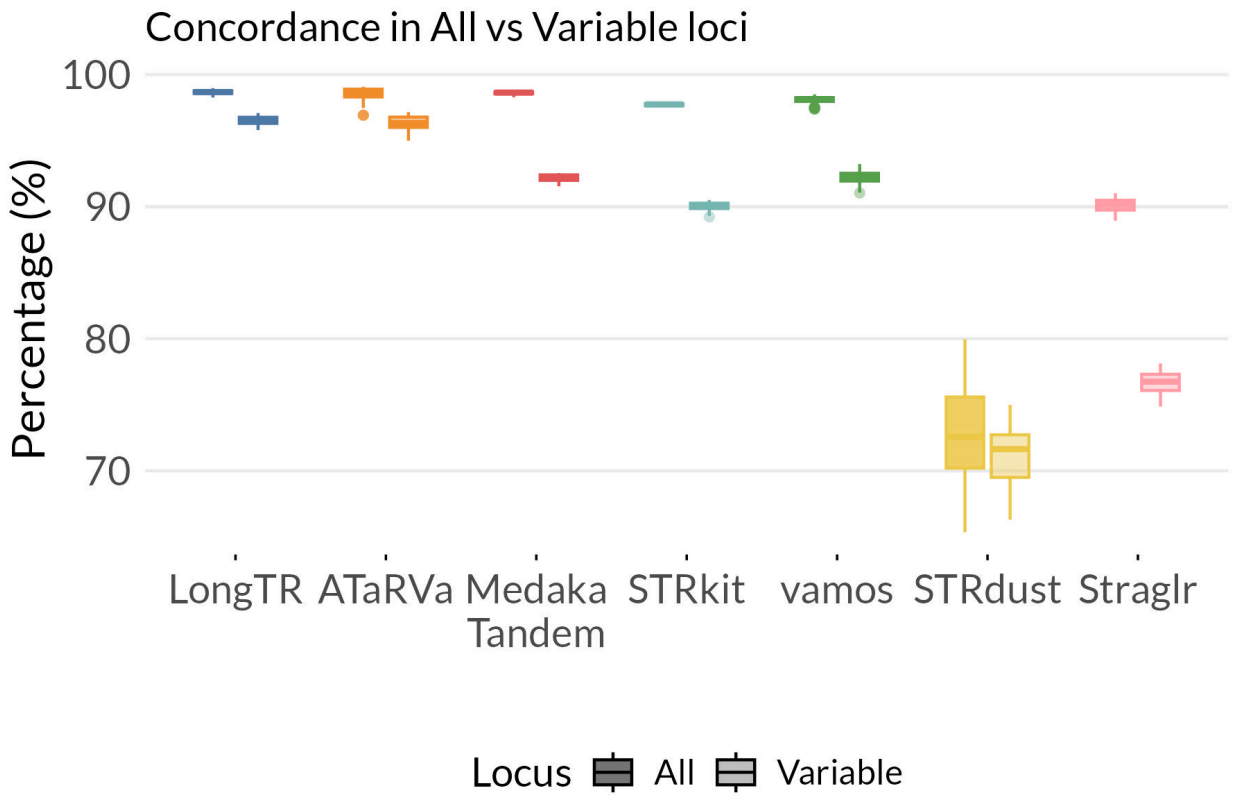

Fig. S2: **Assembly concordance of tools across all loci and variable loci in R10.4.1 samples.** Box plots display the distribution of concordance for each tool across samples sequenced with R10.4.1 pore chemistry. Fully opaque boxes represent concordance calculated across all loci, while transparent boxes represent concordance restricted to variable loci (excluding homozygous reference calls), allowing direct comparison of tool performance in the presence and absence of non-polymorphic loci.

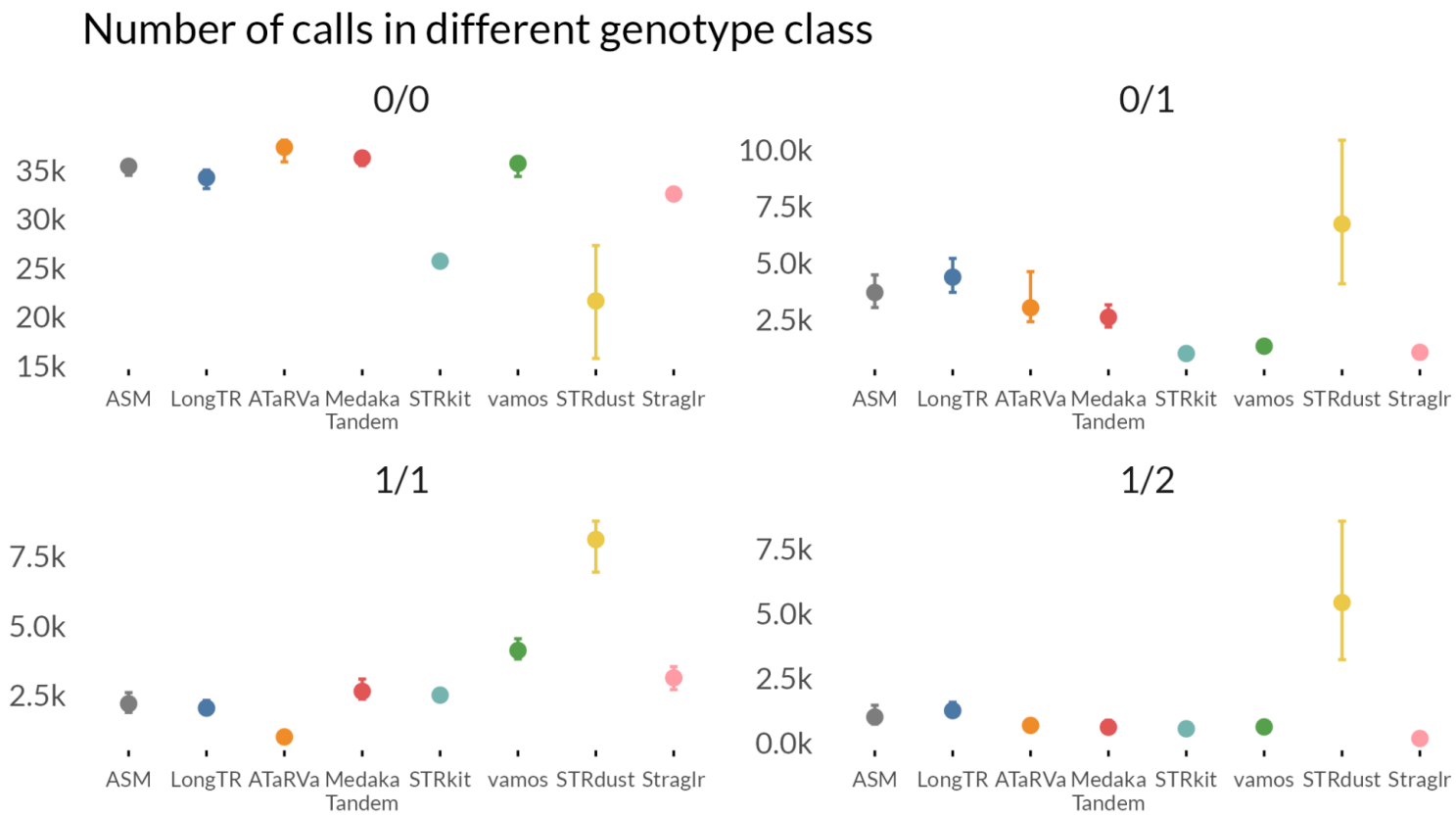

Fig. S3: **Number of loci called by each tool across genotype classes.** Points represent the mean number of loci called by each tool within each genotype class (0/0: homozygous reference, 0/1: heterozygous reference, 1/1: homozygous alternate, and 1/2: heterozygous alternate). Error bars indicate the range between the minimum and maximum number of calls observed across samples within each genotype class.

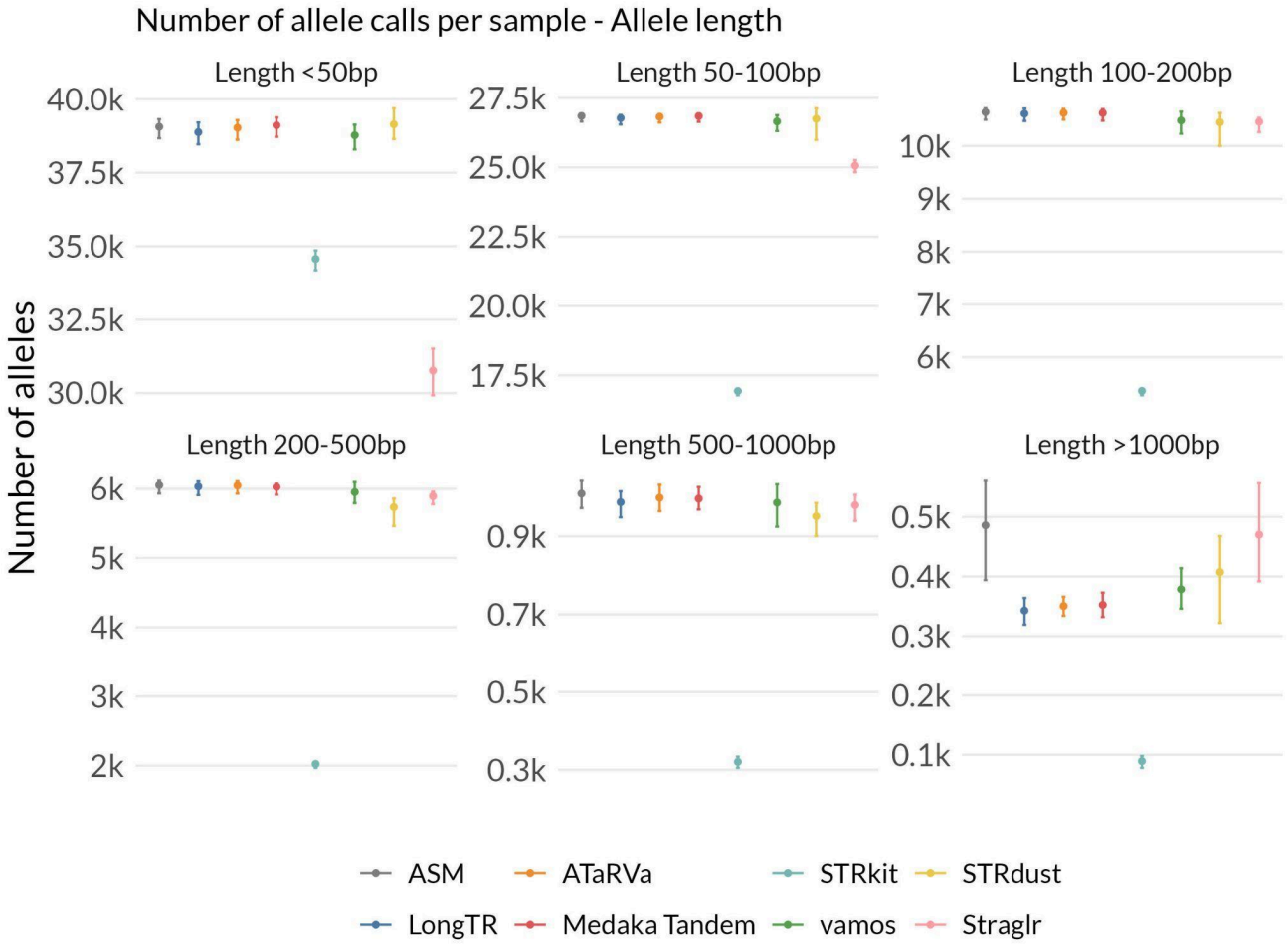

Fig. S4: **Mean number of alleles called per tool across allele length categories.** Points represent the mean number of alleles called by each tool across samples within allele length categories (<50 bp, 50–100 bp, 100–200 bp, 200–500 bp, 500–1000 bp, and >1000 bp). Error bars indicate the range between the minimum and maximum number of calls observed across samples within each category.

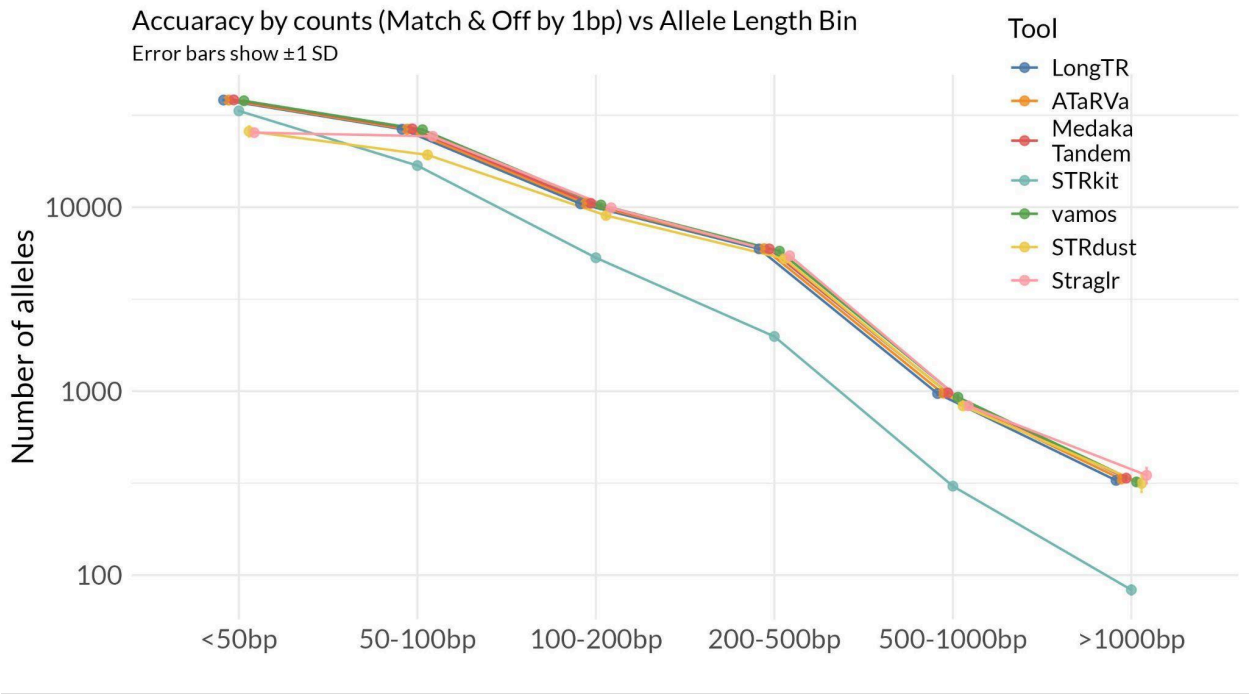

Fig. S5: **Mean number of accurate calls per allele length category across all samples.** Points represent the mean number of accurate calls made by each tool across all samples within allele length categories (<50 bp, 50–100 bp, 100–200 bp, 200–500 bp, 500–1000 bp, and >1000 bp). Y-axis indicates the number of alleles (log scale). Error bars indicate the range between the minimum and maximum number of accurate calls observed across samples within each category.

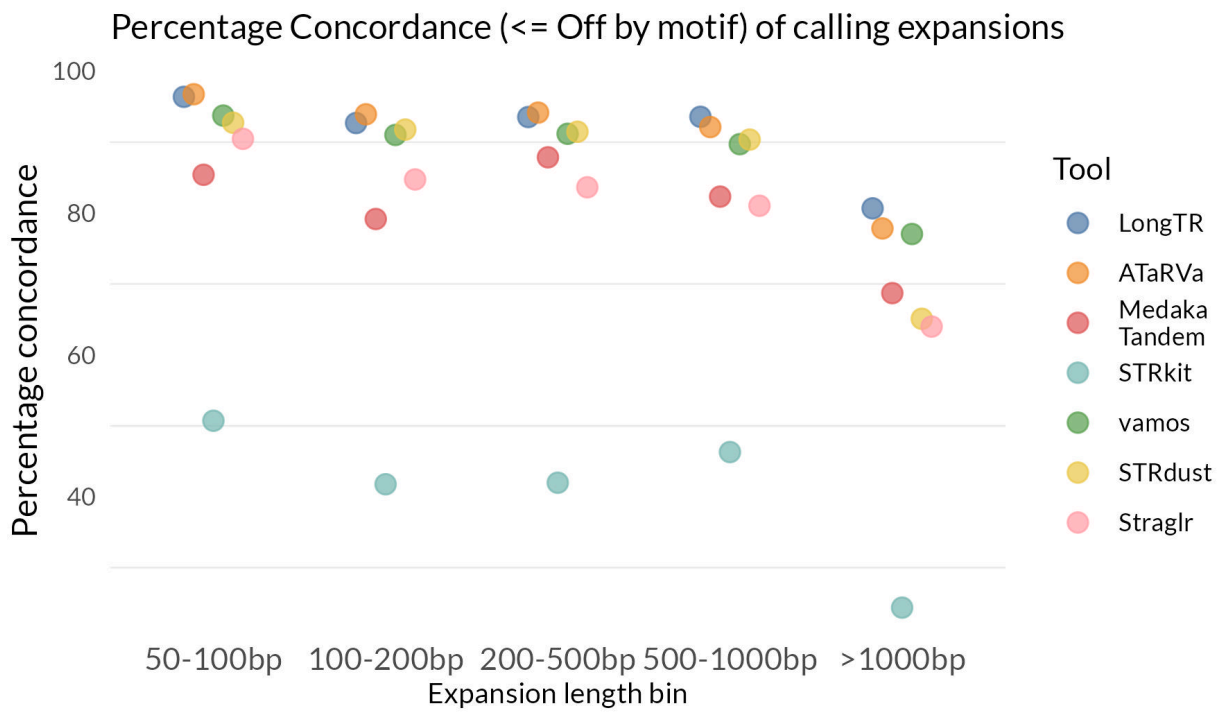

Fig. S6: **Assembly concordance of expansion calls per tool across R10.4.1 samples.** Points represent the percentage concordance of each tool in different expansion length bin, aggregated across all R10.4.1 samples.

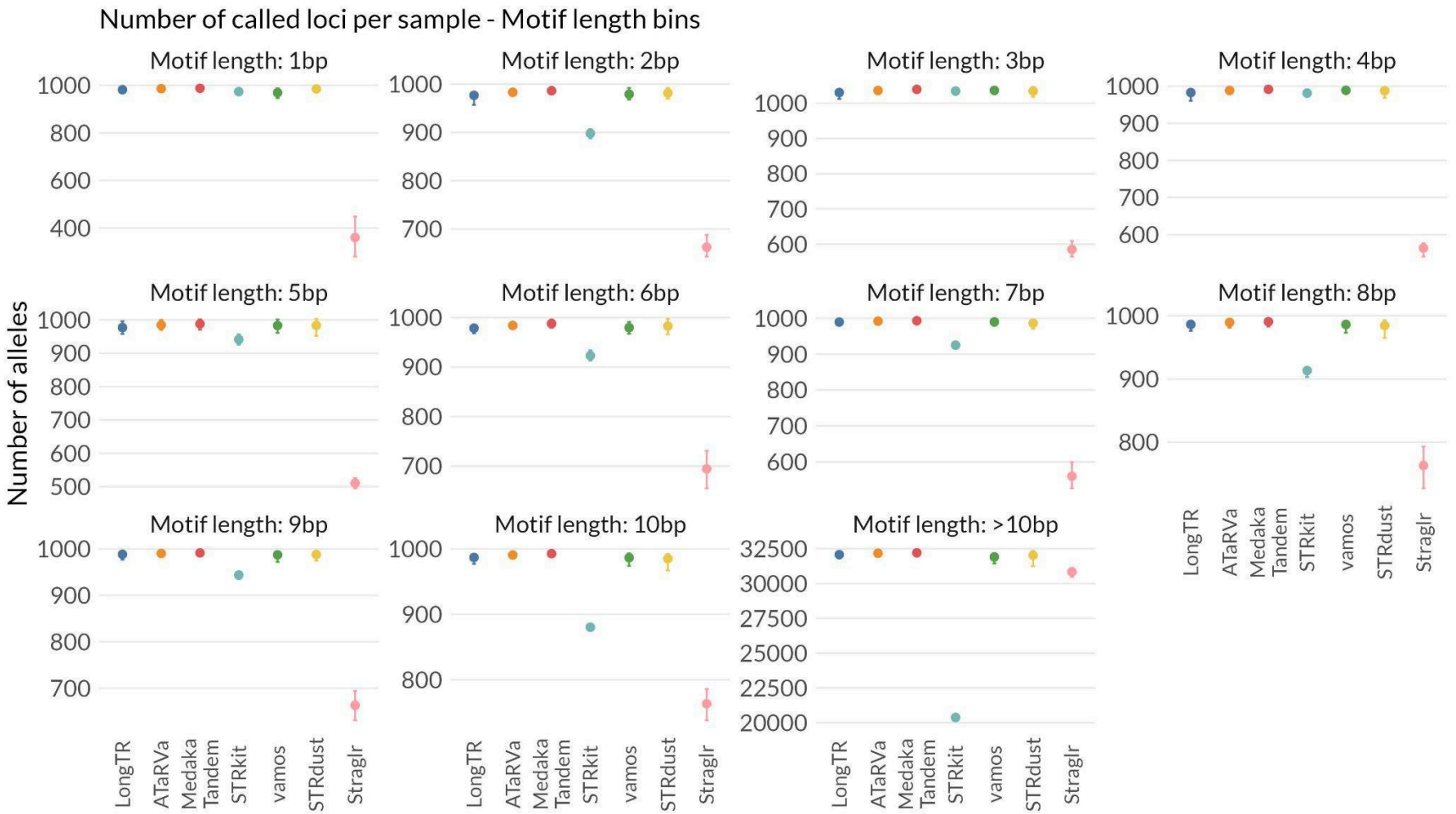

Fig. S7: **Average number of loci called per motif length category across all samples.** Points represent the mean number of loci called by each tool across all samples for motif length categories spanning 1–10 bp and >10 bp. Error bars indicate the range between the minimum and maximum number of calls observed across samples within each category.

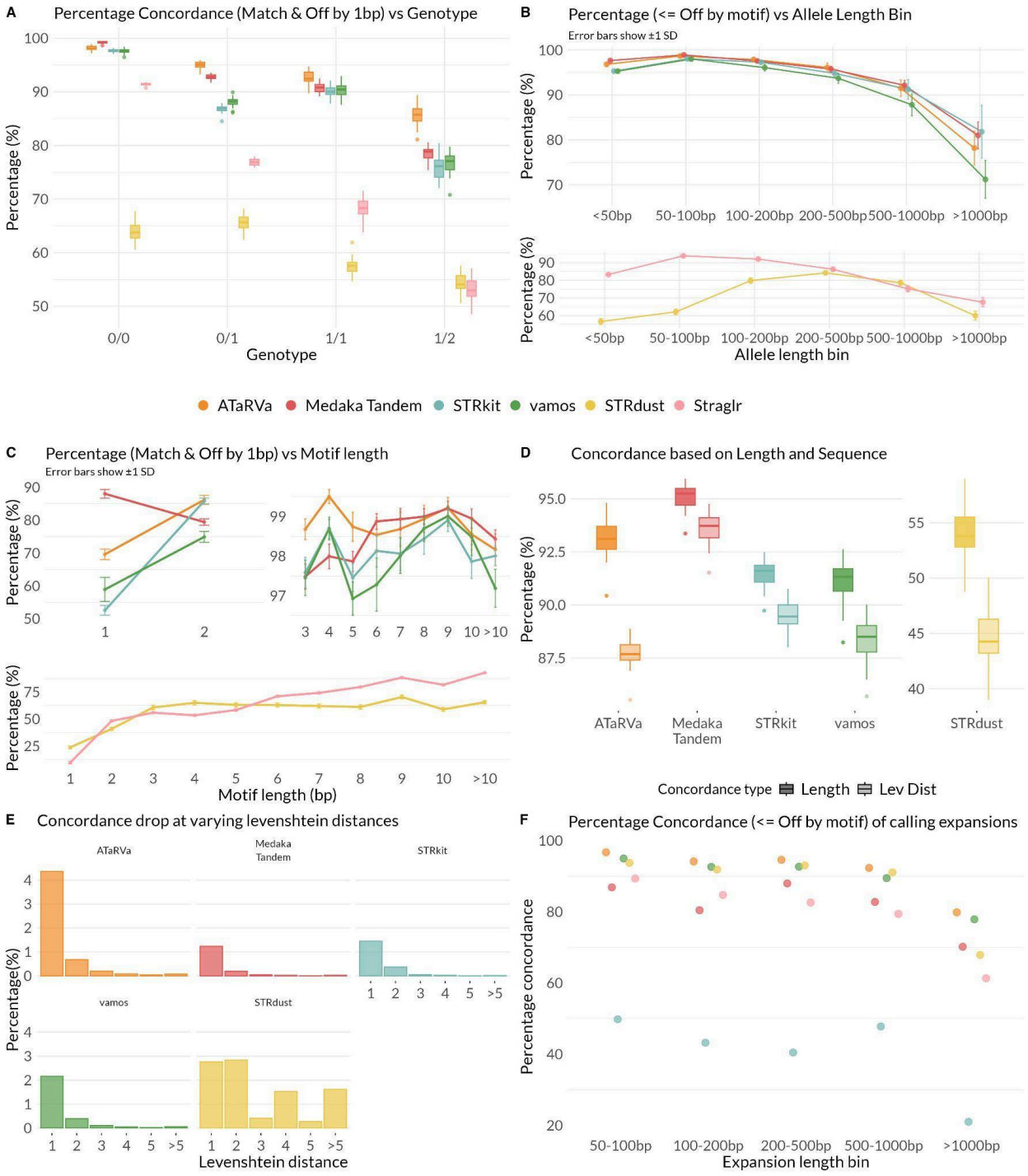

Fig. S8: **Assembly concordance of genotypers across different stratifications calculated as the percentage of allele calls with length within 1 bp for HPRC R9.4.1 samples. A.** Concordance of each tool stratified by genotype class: homozygous reference (0/0), heterozygous reference (0/1), homozygous alternate (1/1), and heterozygous alternate (1/2). Points represent mean concordance per genotype class across all samples. Points represent mean concordance, and error bars indicate ±1 standard deviation across samples. **B.**

Concordance of each tool across allele length bins (<50 bp, 50–100 bp, 100–200 bp, 200–500 bp, 500–1000 bp, and >1000 bp) in R9.4.1 samples. Points represent mean concordance, and error bars indicate ±1 standard deviation across samples. Tools are displayed in two separate panels for readability. **C.** Concordance of each tool across motif length categories (1–10 bp and >10 bp) in R9.4.1 samples. Points represent mean concordance, and error bars indicate ±1 standard deviation across samples. Tools are displayed across three panels for readability. The upper two panels show the five higher-performing tools (LongTR, ATaRVa, Medaka Tandem, STRkit, and vamos), with mononucleotide and dinucleotide repeats shown separately on a y-axis range of 50–95% and motif lengths ≥3 bp displayed on an expanded y-axis of 98–100%. The lower panel shows STRdust and Straglr on a broader scale to capture their wider concordance range across all motif lengths. **D.** Concordance calculated for each tool on two levels: length-based concordance, considering all calls with a matching allele length, and sequence-based concordance, calculated using Levenshtein distance for the subset of calls with a matching allele length. The difference between the two values reflects the proportion of length-concordant calls that differ at the sequence level from the assembly allele. **E.** Contribution of Levenshtein distance categories to sequence-level concordance drop. Bar plots showing the percentage of total calls at each Levenshtein distance value for each tool, restricted to calls with matching allele lengths. Each bar represents the proportion of calls at a given Levenshtein distance, illustrating the relative contribution of different magnitudes of sequence deviation to the overall sequence-level concordance drop observed for each tool. **F.** Concordance of expansion calls per tool across R9.4.1 samples. Points represent the percentage concordance of each tool in different expansion length bin, aggregated across all R9.4.1 sample.

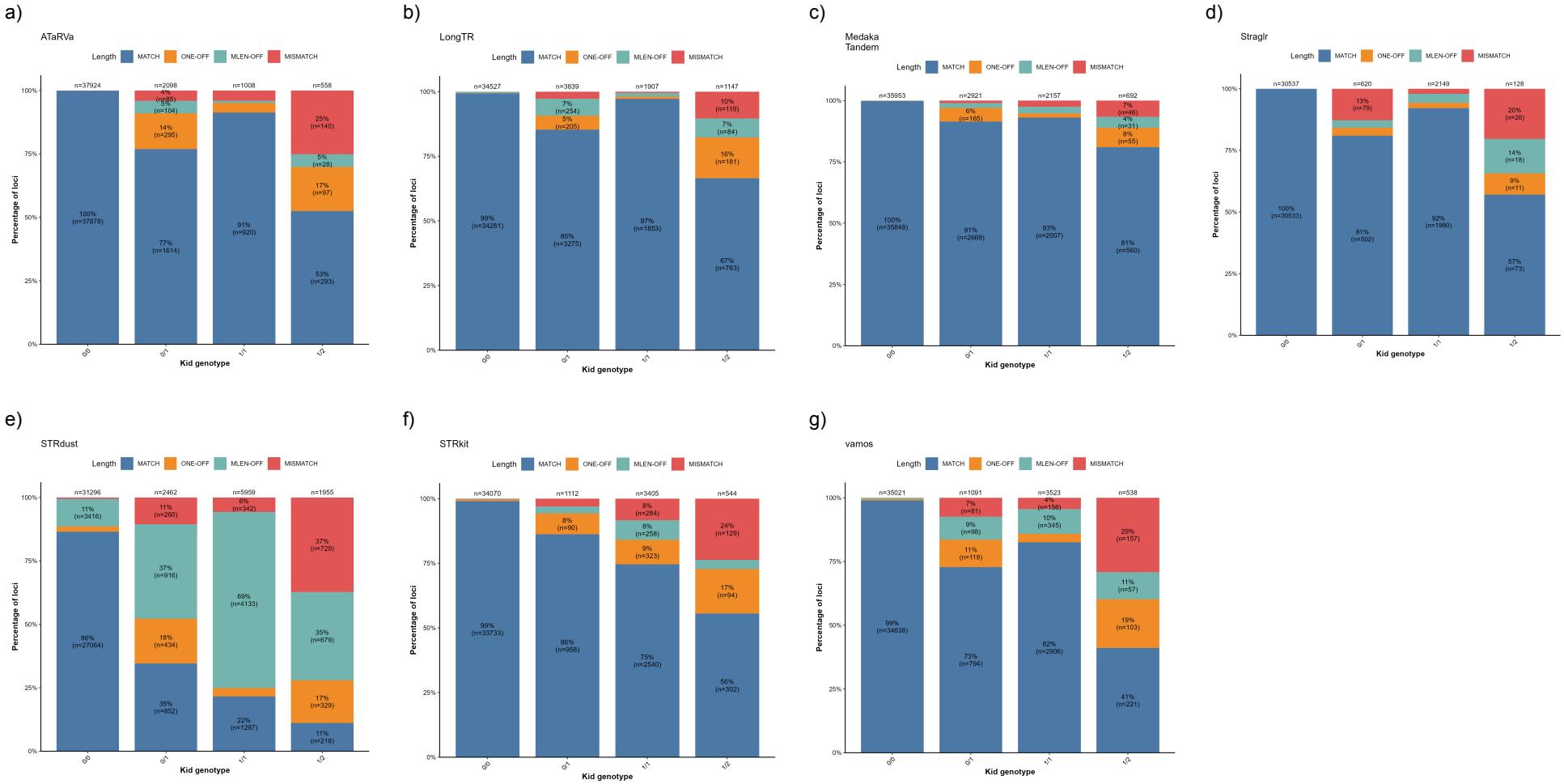

Fig. S9. Mendelian consistency of tandem repeat genotyping tools across genotype classes in the HG002 GIAB trio. Stacked bar charts showing the proportion of loci in each Mendelian consistency category for seven TR genotyping tools (ATaRVa, LongTR, Medaka Tandem, Straglr, STRdust, STRkit, and vamos), stratified by child genotype class (0/0, 0/1, 1/1, and 1/2) in the HG002 GIAB trio. Each bar represents all loci assigned to a given genotype class, with the total locus count shown above. Loci are classified into four categories: MATCH (child allele lengths exactly consistent with parental alleles); ONE-OFF (consistent within ±1 bp); MLEN-OFF (consistent within less than one full repeat motif unit); and MISMATCH (not consistent under any of the above criteria). Tools are arranged alphabetically.

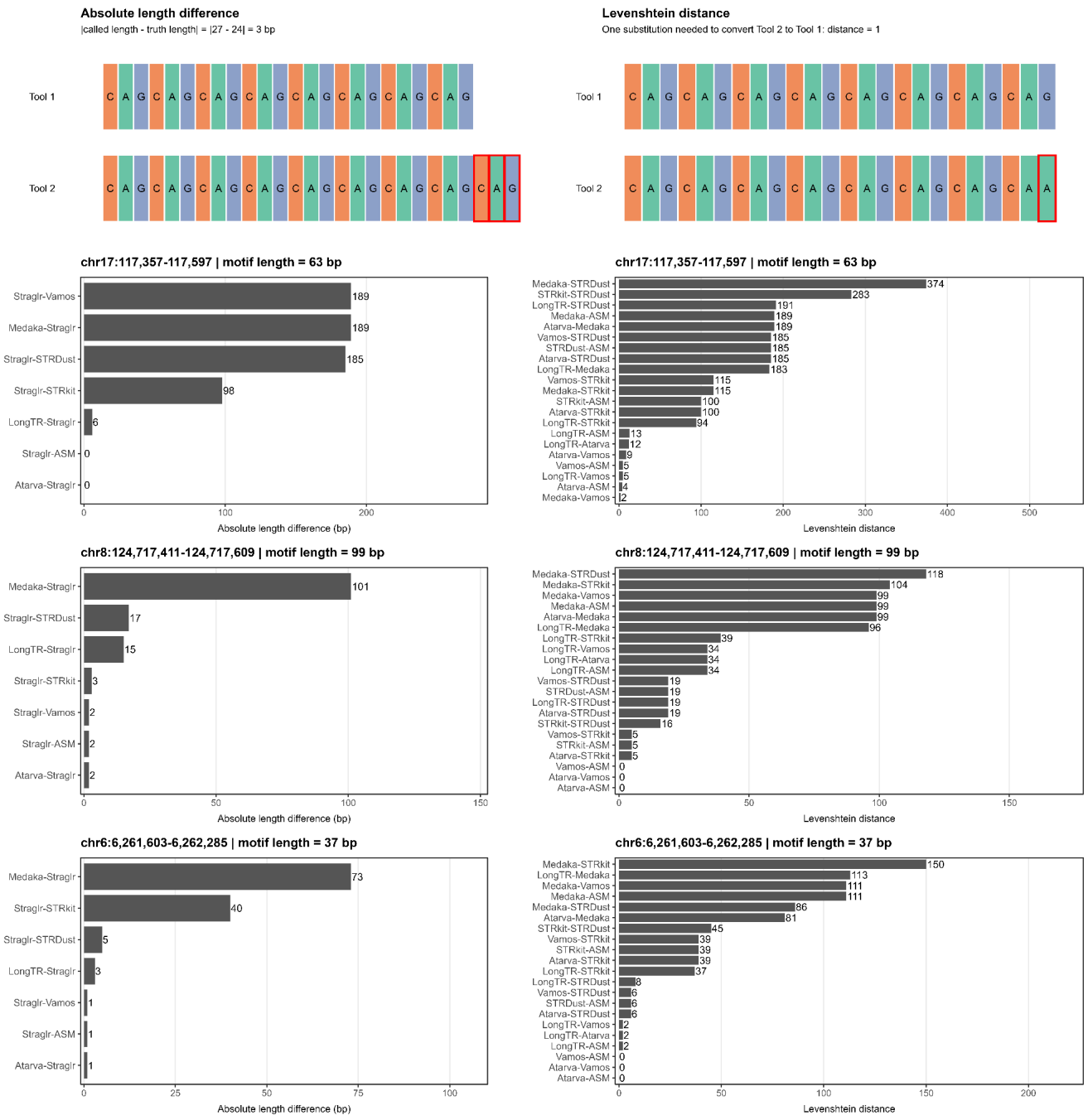

Fig. S10 | Illustrative examples of length and sequence disagreement across TR genotypers. Top, schematic examples showing how the absolute length difference and Levenshtein distance were calculated. For absolute length difference, the metric is defined as the absolute difference between the allele lengths reported by the two compared sequences. For Levenshtein distance, the metric is defined as the minimum number of single-character edits required to transform one sequence into the other. Bottom, three representative loci selected from shared catalog positions after excluding extreme outliers. Left panels show absolute length differences for comparisons involving Straglr only. Right panels show Levenshtein distances across all available tool pairs.

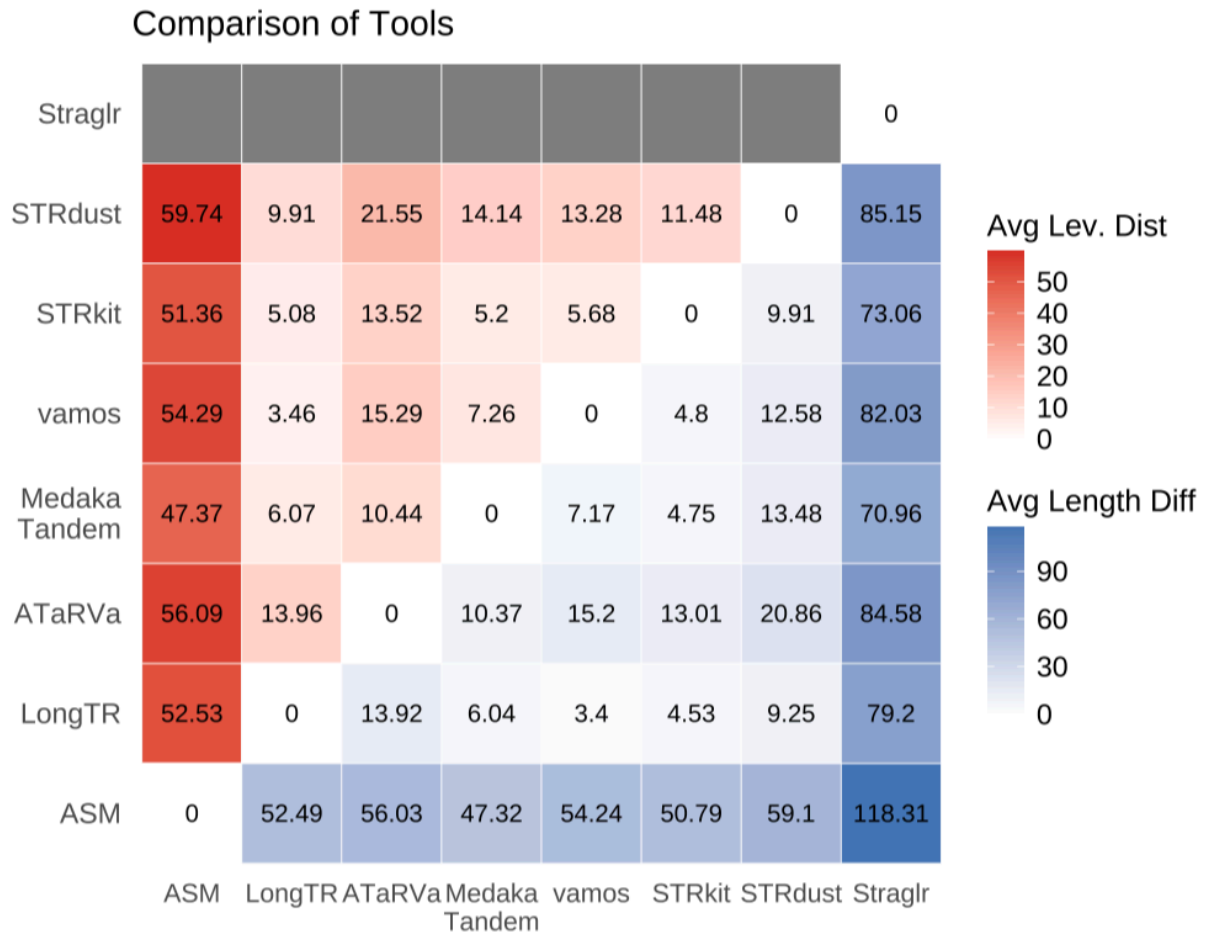

Fig. S11: **Pairwise comparison of genotyping tools for sample HG00290 at assembly discordant loci.** The heatmap displays pairwise comparisons between all tools, including genome assemblies as a reference genotyping method for sample HG00290, restricted to loci where at least five tools disagree with the assembly genotype. The upper triangle represents the mean Levenshtein distance between allele sequences, and the lower triangle represents the mean absolute length difference between alleles, averaged across all commonly called loci and samples. Straglr is excluded from sequence-level comparisons as it does not report allele sequences

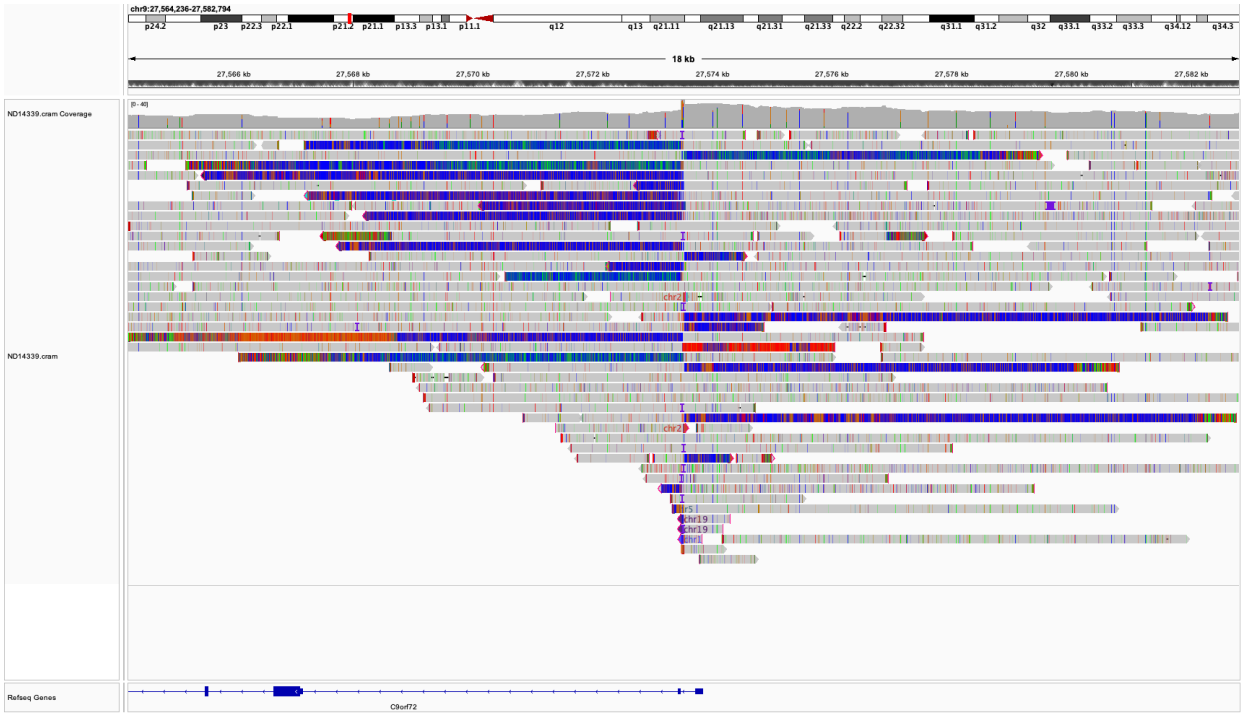

Fig. S12: The C9orf72 expansion in sample ND14339 was missed by all genotypers. While there was evidence of an expansion in the soft-clips, the aligner did not identify any spanning reads, and potentially spanning soft-clipped reads were low-quality.

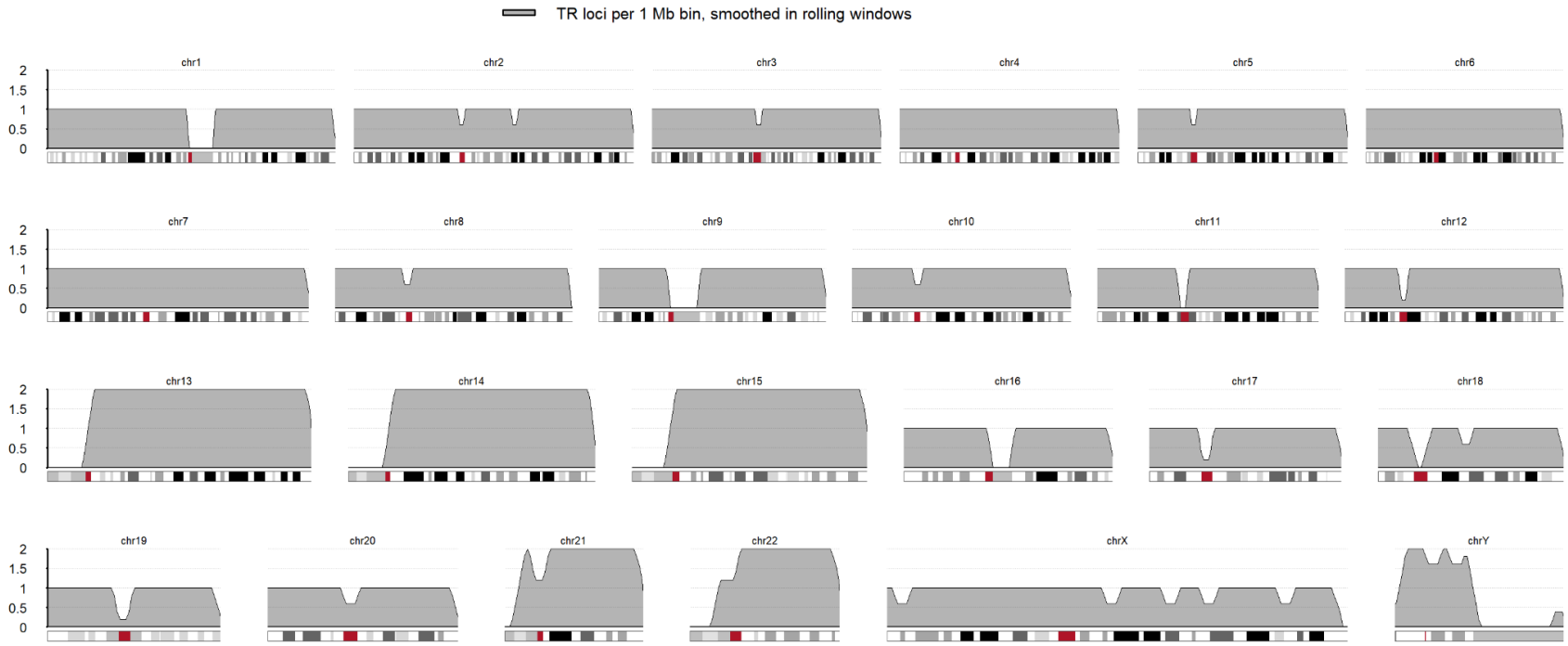

Fig. S13. Genome-wide distribution of the benchmark TR catalog used for genotyping. TR loci were counted from the benchmark catalog used for genotyping in 1 Mb genomic bins and smoothed using a rolling window. Gray profiles show the relative density of catalog loci along each chromosome. Cytogenetic ideograms are shown below; light bands indicate Giemsa-negative regions, dark bands indicate Giemsa-positive regions, and red indicates the centromere.