ATaRVa: Analysis of Tandem Repeat Variation from Long Read Sequencing data

Abishek Kumar Sivakumar[1], Sriram Sudarsanam[1], Anukrati Sharma[1], Akshay Kumar Avvaru[1,2,*], Divya Tej Sowpati[1,*]

1 - CSIR Centre for Cellular and Molecular Biology, Hyderabad, India - 500007
2 - Current Affiliation: Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

* - Corresponding authors:
Akshay Kumar Avvaru - avvaruakshay@gmail.com
Divya Tej Sowpati - tej@ccmb.res.in

Address for correspondence:
Divya Tej Sowpati
E509, East Wing 4th Floor
CSIR Centre for Cellular and Molecular Biology
Uppal Road, Habsiguda
Hyderabad - 500007
India
Ph: +91-40-2719-2862

Abstract

Long-read sequencing propelled comprehensive analysis of tandem repeats (TRs) in genomes. Current long-read TR genotypers are either inaccurate, platform-specific, or computationally inefficient. Here we present ATaRVa, a sequencing technology-agnostic genotyper that outperforms existing tools while running an order of magnitude faster. ATaRVa also supports short-read data, multi-threading, consensus sequence derivation, and motif decomposition, making it an invaluable tool for population scale TR analyses.

Availability

ATaRVa is implemented in Python and is freely available on PyPI. The source code is deposited to GitHub at https://github.com/SowpatiLab/ATaRVa under an MIT license.

Main

Tandem repeats (TRs) are contiguous repetitions of DNA motifs, either exactly or with minor variations. TRs are categorised based on the motif length: homopolymers (1nt motifs), short tandem repeats (STRs; 2-6nt), and variable number tandem repeat (VNTRs; >7nt)[1]. TRs are highly polymorphic and contribute to more than 70% of structural variants (SVs) longer than 50 bp[1,2]. TRs are distributed non-randomly in genomes, and their length variations impact gene expression and genome organization[3]. Expansions in TR regions are associated with over 50 monogenic disorders, including Huntington's disease, amyotrophic lateral sclerosis, and Fragile X syndrome[4]. Earlier studies primarily focused on analyzing length variations of TR loci, but recent research suggests that sequence composition is just as crucial in determining the pathogenicity of certain TRs[5,6].

Tandem repeats, particularly VNTRs, are tricky to genotype using short read sequencing (SRS) data, as these reads often lack sufficient non-repetitive flanks to achieve precise mapping to the reference[7]. Recent, long read sequencing (LRS) methods such as those offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) produce reads that are tens of kilobases long and span many TR loci in full. Leveraging this benefit, several long read TR genotypers have been developed[8–11]. However, these tools are limited in their speed, accuracy, efficiency, or in producing output that allows nuanced downstream analysis. One of the tools is also proprietary and is licensed to be used only with PacBio data.

Here, we present ATaRVa (**A**nalysis of **Ta**ndem **R**epeat **Va**riation), a technology-agnostic TR genotyper that accepts read alignments and a set of TR regions of interest, and outputs the TR genotypes (consensus sequence, length, and decomposed motifs). A detailed description of the ATaRVa algorithm is provided in the supplementary information. Briefly, ATaRVa creates a map of the repeat coordinates of interest from BED input, and processes the input BAM file read-wise, with the assumption that most long reads span info of multiple TR loci. After flank re-alignment and adjustment of read wise allele lengths, ATaRVa clusters the reads into haplotypes based on

nearby informative SNVs, or using a k-means approach where no SNV information is available. It derives consensus allele sequence using partial order alignment, and decomposes the TR allele into their motif level representations, and outputs this info in VCF format .

We evaluated the accuracy and performance of ATaRVa on three different whole-genome sequencing (WGS) datasets of the widely catalogued Genome-In-A-Bottle[12] HG002 sample at 30x coverage: High Fidelity reads from Pacific Biosciences (PacBio HiFi), Duplex data from Oxford Nanopore Technologies (ONT Duplex), and normal data from ONT (ONT Simplex). The analysis was performed on 2,927,916 human repeat loci from Project Adotto[1]. We also compared the performance of ATaRVa to other long read base genotypers: LongTR[10], Straglr[8], TREAT[11], and TRGT[9] (PacBio only). All tools showed a mean genotyping rate of ~99% except for Straglr, which genotyped 92.6% regions on average (Table S1).

The accuracy was evaluated by comparing the allele lengths derived from the high quality phased assembly of the HG002 sample from the HPRC consortium[13]. For this, we selected a subset of 2,774,873 loci exhibiting exactly two alleles in the reference HG002 assembly. On this subset, all tools showed a genotyping rate close to 100%, except for Straglr (Table S1). On the PacBio HiFi data, ATaRVa achieved an exact match concordance rate of 91.2%, compared to the next best tools, TRGT (88%) and LongTR (83%; Fig 1a left). TREAT and Straglr showed an exact match rate of 76.1% and 66.8% respectively. When allowing for a ±1 bp tolerance, the concordance of ATaRVa improves to 96.9%, followed by other tools: LongTR (95.8%), TRGT (92.5%), TREAT (82.9%), Straglr (77.1%).

ATaRVa also surpassed other tools on datasets generated by Oxford Nanopore (ONT). For ONT Duplex and ONT Simplex data, ATaRVa achieved a concordance rate of 95.7% and 96.4% respectively, allowing for ±1 bp tolerance (Fig 1a, middle and right). In comparison, LongTR, the next best tool, showed a concordance of 93.2% on both datasets. TREAT performed substantially better on the Duplex data (88.4%) compared to Simplex (77.7%), while Straglr showed the lowest concordance of 77.6%. The concordance results of all the tools on the three HG002 datasets are summarized in Supplementary Table S1. Of particular note, ATaRVa showed a significantly high exact concordance rate compared to the other tools.

We sought to understand the effect of the TR allele length on accuracy of genotyping. For this, we analyzed the accuracy of the tools by categorizing the TR loci into three allele length groups: <100bp, 100-500bp, >500bp. In addition to allowing for “off-by-one” errors, we also calculated loci where the difference between the reported and actual allele lengths was within 50% of the TR motif size, as these lengths would round off to the correct number of repeat motifs. For the <100bp and 100-500bp categories, ATaRVa performed better than all the other tools (Fig S2). For the >500bp loci, on the PacBio HiFi dataset, LongTR performed the best (92%), followed by ATaRVa (89.9%). Surprisingly, the accuracy of TRGT on long alleles dropped to 47.9%. This was due to the use of non-default parameter ‘--flank-len 10’ to be consistent with other tools. With the default flank length of 250bp, the accuracy of TRGT on long alleles improves to 88.1%, albeit at a substantial cost to runtime (8.2 loci/s vs 3.3 loci/s). On the ONT datasets for this length category,

ATaRVa was on par with LongTR (Duplex: 91% vs 91.2%, Simplex: 91.5% for both tools). Though we accounted for length differences of up to 50% of the repeat motif size, the lengths reported by ATaRVa for most (>93%) of the long alleles was within ±10 bp (Fig 1b, Fig S3). These results show that ATaRVa accurately genotypes TR loci from long read data, even at long allele lengths.

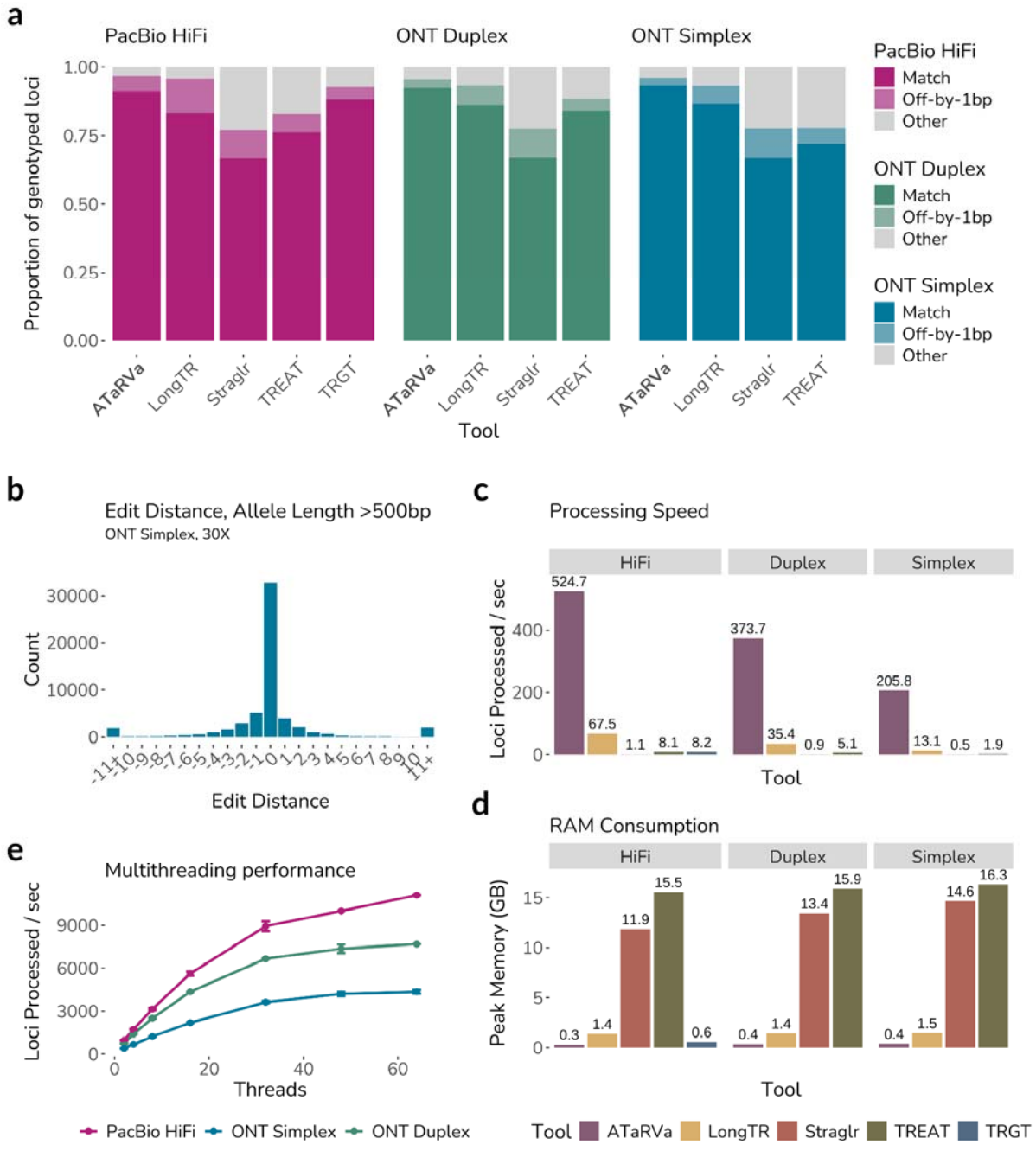


Figure 1: Performance of ATaRVa on HG002 datasets. **a)** Accuracy metrics of ATaRVa and other TR genotypers on 30x WGS PacBio HiFi (left), ONT Duplex (middle) and ONT Simplex (right) datasets. Analysis performed on 2.77M loci where exactly two assembly alleles were recovered. Match: exact match of both allele lengths; Off-by-1bp: the length reported by tool is

±1bp compared to the assembly allele; Other: length difference is >1bp. **b)** Difference between the assembly allele length and that reported by ATaRVa on TR loci longer than 500bp, on the ONT Simplex 30x WGS dataset. A value of 0 indicates exact match, values <0 indicate underestimation and >0 indicate overestimation. Differences >=11nt are clubbed into a single bin. **c)** Processing speed of the tools tested, reported as the number of loci processed per second. Analysis was done on a subset of 100k loci from the 2.9M Adotto TR catalog. The times represented are an average of 5 independent runs. **d)** Peak memory consumed by various tools on the 100k loci subset, reported in GB, as measured by the unix time command. The values reported are an average of 5 independent runs. **e)** Multithreading performance of ATaRVa, reported as the number of loci processed per second. The number of CPU cores tested was 2, 4, 8, 16, 32, 48, and 64. Points represent arithmetic mean of 5 independent measures, and the error bars indicate standard deviation.

On a regular desktop computer, ATaRVa could genotype the entire 2.9M Adotto catalog on a 30x PacBio HiFi data in under 15 minutes using 8 CPU cores. For performance comparison, we chose to evaluate the time metrics in single-threaded mode, as there was inconsistent support for multithreading across tools. Due to the prohibitively slow nature of some tools, we evaluated the time taken by various tools on a subset of 100k loci. On the 30x PacBio HiFi data, ATaRVa attained a processing speed of ~525 loci/sec, >7 fold faster than LongTR, which operated at a speed of ~70 loci/sec (Fig 1c). The difference was particularly stark on the ONT data. On the 30x Simplex data, ATaRVa was >15x faster than LongTR (~205 loci/sec vs ~13 loci/sec), and over two orders of magnitude faster than other tools.

We next evaluated the memory consumed by various tools. For the PacBio dataset, ATaRVa needed the least peak RAM (0.3GB), closely followed by TRGT (0.6GB) and LongTR (1.4GB; Fig 1d). ATaRVa outperformed all the other tools in terms of memory requirement across the datasets. On an average, ATaRVa used 4-5x less RAM than LongTR, and >30x less RAM than Straglr and TREAT. TREAT required the most amount of RAM across all the datasets. We further tested the multithreading capability of ATaRVa. We observe that the speed gains are negligible beyond 32 cores (Fig 1e). However, we expect this result is subject to other bottlenecks such as I/O throughput.

To ensure that our algorithm is not over optimized for the HG002 sample, which has become the de facto standard in the field for benchmarking, we evaluated the performance of ATaRVa on the PacBio HiFi data of four other samples from the HPRC: HG00438, HG00673, HG02630, and HG00621[14]. We compared the accuracy to that of LongTR and TRGT, the two next best performing tools. Similar to what was observed for HG002, ATaRVa consistently surpassed LongTR and TRGT in accurately genotyping the Adotto repeat catalog (Fig 2a).

To further check the accuracy, we assessed Mendelian concordance for repeat lengths using the Ashkenazi Jewish familial trio (HG002, HG003, and HG004) data, which was previously used by the authors of TRGT for their analysis[15]. The Mendelian concordance rate for all loci was 98.57% for exact matches and 99.57% when allowing for a ±1 bp tolerance between a parent and child (Table S2). This rate is marginally better than LongTR (99.5%) and TRGT (99.25%). We repeated this analysis after removing loci which were homozygous for the reference allele length in all the

family members. On this "non-ref" subset, the Mendelian concordance dropped to 91.32% for exact matches, 97.38% when allowing for a ±1 bp tolerance, and 99.35% when allowing for ±1 motif differences (Fig 2b). In comparison, with ±1 bp tolerance LongTR and TRGT showed a rate of 98% (99.43% allowing for 1 motif difference) and 95.54% (98.2% with ±1 motif difference) respectively.

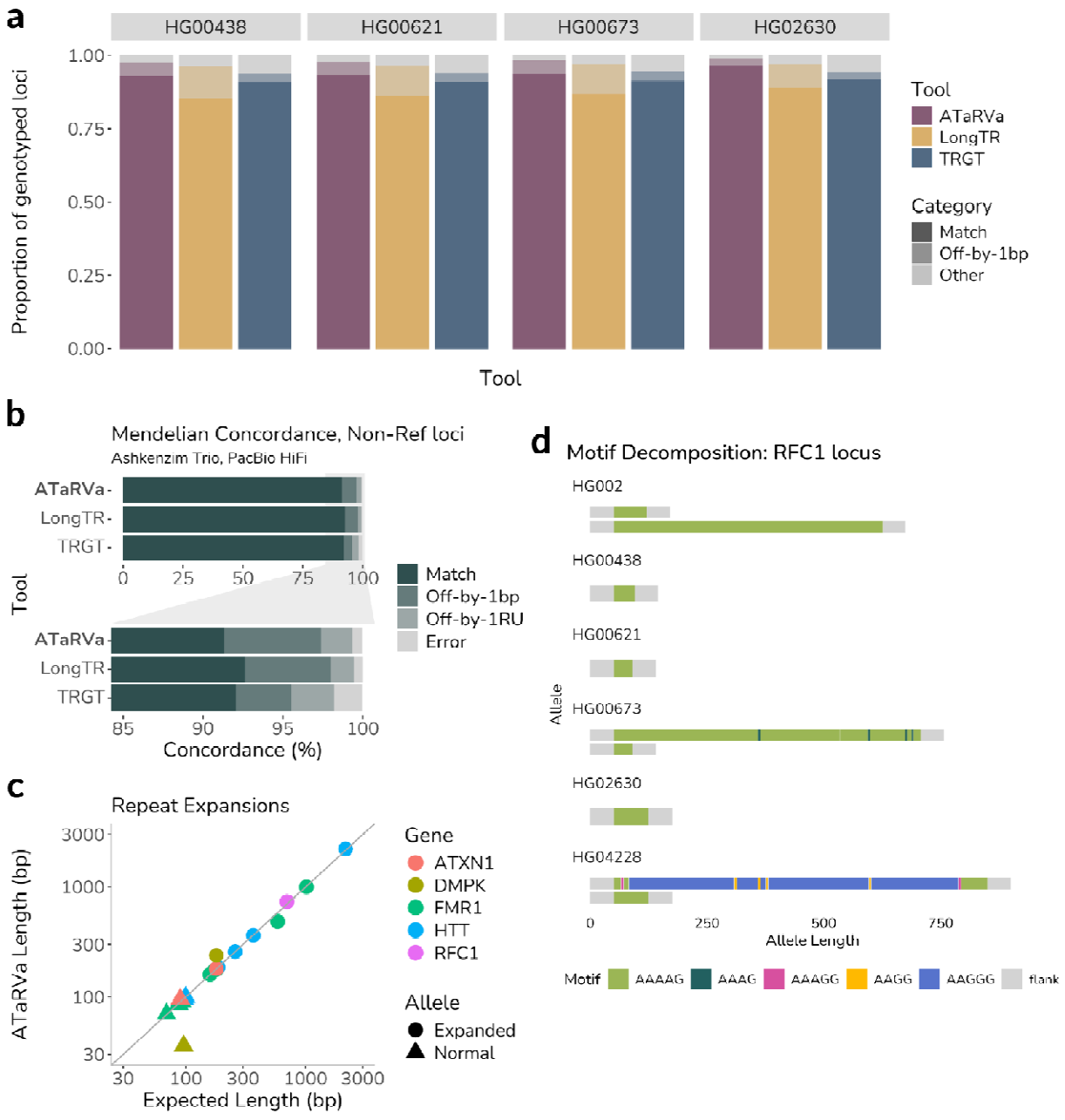


Figure 2: a) Accuracy metrics of ATaRVa, LongTR, and TRGT on 4 different PacBio HiFi WGS datasets. Match: exact match of both allele lengths; Off-by-1bp: the length reported by tool is ±1bp compared to the assembly allele; Other: length difference is >1bp. **b)** Mendelian concordance rates of ATaRVa, LongTR, and TRGT, calculated on the PacBio HiFi data of the

Ashkenzi Trio samples (HG002-HG004). Top: analysis on all the loci which are successfully genotyped in all the samples. Bottom: analysis limited to a subset where at least one non-reference allele is present in any of the three samples. Match: exact match of allele lengths in parents and son; Off-by-1bp: the length difference between the parent and son alleles is ±1bp; Off-by-1RU: the length difference is less than 1 repeat unit (based on the reference TR motif length); Other: larger errors. **c)** Performance of ATaRVa on picking known pathogenic expansions. Data of 12 samples compiled from different sources is depicted here. The X-axis indicates the length of the alleles reported by the source, and the Y-axis represents the length reported by ATaRVa. The median diagonal indicates perfect match. **d)** Motif decomposition analysis of the RFC1 locus across samples. Motifs mentioned in the plot are derived from the decomposed VCF output of ATaRVa. Heterozygous alleles are represented by two bars, whereas homozygous alleles are drawn as a single taller bar.

To assess ATaRVa's performance in identifying pathogenic repeat expansions, we used data from 12 samples with known pathogenic expansions. These data were collected from multiple sources[9,16,17] (see Data availability) and have expansions in 5 different genes (ATXN1, DMPK, FMR1, HTT, RFC1). ATaRVa could successfully and accurately identify the expansions in all the tested samples with default parameters, without needing any optimization (Fig 2c). This includes a >2Kb CAG motif expansion in the sample NA14044 (Table S3). Recent evidence suggests that the sequence variation of expansions can impact the phenotypic and pathogenic consequences[18,19]. Hence, merely reporting the allele length may not be sufficient to delineate the subtle yet important TR variations. ATaRVa supports motif level decomposition of the reported alleles with negligible impact on runtime (~2% slower). We showcase the utility of this feature by analyzing the RFC1 locus of various samples. Concomitant with previous reports, we observe expansion of the non-canonical AAGGG motif in the sample HG04228 as opposed to the AAAAG motif seen in the other samples (Fig 2d).

We then asked if ATaRVa can genotype STRs from short-read sequencing data. For this, we used the 1.7M HipSTR loci[20,21] and the Illumina data of HG002[22] (see Data Availability). The assembly alleles were extracted for these loci following the same approach that was used for the 2.9M loci. When ATaRVa detects short read data (average read length of <350nt), it processes the data locus wise, akin to other genotypers. We compared the performance of ATaRVa with another short read genotyper, GangSTR[23]. Unsurprisingly, ATaRVa genotyped fewer loci (1.52M) compared to GangSTR (1.62M), as ATaRVa only uses reads that fully enclose at least one STR locus. However, ATaRVa achieved a significantly higher accuracy across all the motif sizes, including homopolymers (Fig S4). ATaRVa finished processing a 30X Illumina BAM file in 6 minutes using 32 CPU cores.

In summary, ATaRVa is an efficient tandem repeat genotyper that is technology-agnostic, and runs an order of magnitude faster while utilizing 4X less RAM than the current best performing long read genotyper. It achieves this performance without sacrificing accuracy. In fact, ATaRVa was significantly more accurate than other tools in most scenarios, and is close behind the best tool in cases such as genotyping very long alleles using PacBio HiFi data. The accuracy, speed, versatility, and features such as motif decomposition make ATaRVa an ideal choice for TR genotyping, particularly for fast-growing population scale datasets.

Methods

Overview of ATaRVa algorithm

ATaRVa begins by creating a map of the alignment file based on the coordinates provided in the input region file and iterates only through the corresponding chromosomal chunks. It then iterates through the aligned reads and stores essential information from the reads that meet the minimum mapping quality cutoff and enclose at least one TR region. The stored info includes read and TR region boundaries with flanks, number of enclosed TRs, InDel positions, and SNVs with their base qualities. When the start of the current read exceeds the end coordinate of a TR region—indicating that all supporting reads for that region have been collected—ATaRVa initiates the genotyping process. This involves local realignment and clustering of the reads. For each of these reads, ATaRVa performs local realignment of the flanking regions using the striped Smith-Waterman algorithm[24]. Realignment is performed between the inserted sequence in the flanking regions and the reference repeat region to determine whether the insertion includes repeat motifs matching the reference. If alignment is confirmed, the insertion along with the intermediate flanking sequence is incorporated into the repeat region. After updating the repeat region boundaries, the allele length is calculated. Each locus is then classified as either 'Homozygous' or 'Ambiguous', based on the supporting distribution of allele lengths. A locus is tagged as Homozygous if 75% of supporting reads correspond to a single allele length, and all other lengths are supported by only one read each. These loci are directly written to the VCF after generating a consensus sequence using partial order alignment (abPOA)[25], bypassing the clustering step. Loci classified as Ambiguous are subjected to clustering. If SNVs occur in 20–80% of the reads at a particular position (heterozygous SNVs) — clustering is performed based on these variants. If no such SNVs are found, k-means clustering is applied to the allele lengths[26]. For each cluster, a consensus sequence is generated using abPOA, and the details of the genotyped locus are then written to a VCF file. Simultaneously, all information from reads that fall completely outside the span of the current read and belong to already genotyped loci is cleared, ensuring efficient memory and runtime performance.

Assembly concordance

The concordance was measured using a 2.9 million TR loci catalog derived from Project Adotto[1]. The Adotto catalog of 1.7M loci contains TR annotations grouped together based on distance. Each "parent" locus may harbor several TRs, each with their own motif size and length. Additionally, a parent entry also has a flank of up to 25nt. We parsed this catalog to derive a "flattened" set of 2.9M loci, where each entry has its own motif size, and the flank coordinates are removed.

The ONT datasets of HG002 were downloaded from HPRC (Duplex) or ONT open data repository hosted on AWS (Simplex). The HG002 HiFi data and four additional HiFi datasets (HG00438, HG00673, HG02630, and HG00621) were downloaded from the HPRC[14]. Their corresponding diploid assemblies were also obtained from the HPRC and aligned to the GRCh38 reference genome using minimap2 v2.28[27] command '`minimap2 -ax asm5 -t 4 GRCh38.fasta.gz`

`$sample.fa.gz'`. The allele lengths for each of the 2.9M loci were then calculated using a custom script from the maternal and paternal assembly-to-assembly alignment files. Only loci supported by a single read or contig in both the maternal and paternal BAM files were retained, while all other loci were discarded. The common loci between the assembly-derived loci and each tool's genotyped loci were extracted, compared, and categorized based on their match type (match, match with ±1 bp difference, or mismatch). Finally, the overall accuracy percentage was calculated.

Mendelian Concordance

Whole-genome sequencing (WGS) data of the Ashkenazi trio (HG002, HG003, and HG004), generated using PacBio HiFi sequencing as part of the TRGT analysis, were downloaded from the NCBI PRJNA1028149 and genotyped using ATaRVa, LongTR, and TRGT. Initially, common genotyped loci across the trio were extracted for each tool. The alleles at each locus were then categorized as either Mendelian or non-Mendelian, based on whether they were present in the parental alleles. Mendelian alleles were further classified into match, ±1 bp difference, or ±1 motif difference categories. Finally, the Mendelian concordance percentage was calculated as the ratio of Mendelian alleles to the total number of alleles (2 × total common loci).

Analysis of samples with known pathogenic expansions

A total of 12 pathogenic samples with repeat expansions in *HTT*, *FMR1*, *RFC1*, *DMPK*, and *ATXN1* were obtained from three different platforms (See Data Availability). Three HPRC samples with expansions in *FMR1* (HG00438, HG04184) and *RFC1* (HG04228) were included. Seven samples with expansions in *HTT* (NA13505, NA13509, NA20253, NA14044) and *FMR1* (NA13664, NA06896, NA07537) were sourced from PacBio targeted sequencing data. Two additional samples with expansions in *DMPK* (NA23265) and *ATXN1* (NA13537) were obtained from NCBI (PRJNA786382).

Comparison with other tools

ATaRVa

ATaRVa v0.1.0 was run using non-default parameters --karyotype XY to consider chromosome X & Y as haploid chromosomes. An example command for HG002 sample is

```
`atarva -fi Homo_sapiens_assembly38.fasta --bams hg002.bam -bed ATaRVa_compatible_adotto_TRs_v1.2.bed.gz -o hg002 -p 32 --karyotype XY`
```

LongTR

LongTR v1.2.0 was run using non-default parameters --min-reads 10, --min-mapq 5, --max-tr-len 50000, --min-mean-qual 10, --indel-flank-len 10 and --haploid-chrs chrX,chrY to genotype repeats with at least 10 overlapping reads, reads with mapping quality of at least 5, genotype repeats with length up to 50,000 bp, reads with mean quality of at least 10, considering InDels up to 10bp around the repeats and to consider chromosome X & Y as haploid chromosomes. The command used was

```
`LongTR --bam-samps hg002 --bam-libs lib1 --min-mapq 5 --max-tr-len 50000 --min-reads 10 --min-mean-qual 10 --silent --haploid-chrs chrX,chrY --indel-flank-len 10 --bams hg002.bam --fasta Homo_sapiens_assembly38.fasta --tr-vcf hg002_tr_calls.vcf.gz --regions LongTR_compatible_adotto_TRs_v1.2.bed`
```

TRGT

TRGT v1.5.1 was run using non-default parameters --max-depth 100, --flank-len 10, and -k XY, to genotype repeats utilizing maximum of 100 reads, considering InDels up to 10bp around the repeats and to consider chromosome X & Y as haploid chromosomes. The command used was

```
`trgt genotype -g Homo_sapiens_assembly38.fasta -r hg002.bam -b TRGT_compatible_adotto_TRs_v1.2.bed -o TRGT_hg002 -t 32 --flank-len 10 -k XY --disable-bam-output --max-depth 100`
```

TREAT

TREAT v1.0.0 was run using non-default parameters -minCov 10 to genotype repeats with at least 10 overlapping reads. The command was

```
`python3 TREAT.py reads -b TREAT_compatible_adotto_TRs_v1.2.bed -i hg002.bam -r Homo_sapiens_assembly38.fasta -t 32 -minCov 10 -o TREAT_output`
```

Straglr

Straglr v1.5.1 was run using non-default parameters --genotype_in_size, --min_support 10, --max_str_len 1000 and --sex m to report genotypes in terms of allele sizes instead of copy numbers, to genotype repeats with at least 10 overlapping reads, genotype repeats with motif length up to 1000 bp and to specify the sample sex as male. The command used was

```
`python straglr.py hg002.bam Homo_sapiens_assembly38.fasta straglr_hg002 --genotype_in_size --min_support 10 --loci Straglr_compatible_adotto_TRs_v1.2.bed --max_str_len 1000 --sex m --nprocs 32`
```

GangSTR

GangSTR v2.5.0 was run using default parameters with the command

```
`GangSTR-2.5.0 --bam hg002_srs.bam --ref Homo_sapiens_assembly38.fasta --regions Gangstr_compatible_1.7M_TR_region.bed --out gangstr_hg002 --include-ggl`
```

All processes were run on the same server running Ubuntu 22.04.4 LTS, with an Intel(R) Xeon(R) Gold 6242R CPU @ 3.10GHz with 503 GB RAM and 2.4TB SAS storage. Time and memory consumption were measured using the UNIX command 'time -v' and the elapsed (wall clock) time and maximum resident set size were reported. The average of 5 independent runs was reported

for each tool on each dataset, except in the case of Straglr, which was a single measurement for each dataset. Time and RAM consumption was measured on a subset of the 100k loci from the 2.9M catalog, whereas the accuracy was measured across the entire catalog.

Motif decomposition

ATaRVa analyses the TR structure by decomposing them into constituent units using bitwise comparisons and pattern frequency analysis, based on our previously developed algorithm Ribbit for identification of complex TRs[28]. ATaRVa begins with canonical motif decomposition using the motif size provided in the BED file. In case of an interspersed or flanking region, motif length estimation is performed via a Shift-and-Match approach. Each nucleotide in the ALT allele sequence is encoded into bitarrays and are aligned with their shifted versions. This function evaluates multiple candidate shift values (including motif length ±1 and values 1–6) and selects the one that maximises matched bits. Motif frequency analysis identifies potential frequent motifs and a non overlapping version of Knuth-Morris-Pratt (KMP) algorithm identifies the appropriate motif with minor interruptions. These detected motifs are collapsed into compressed repeat notation while performing recursive motif decomposition of leftover sequence. This strategy enables ATaRVa to annotate complex repeat structures. Due to the highly redundant and imperfect nature of large TR motifs, currently motif decomposition is performed only on motif sizes <=10bp.

Data availability

HG002 assembly : https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.0.1.fasta.gz
HG002 PacBio HiFi reads : https://human-pangenomics.s3.amazonaws.com/submissions/80d00e88-7a92-46d8-88c7-48f1486e11ed--HG002_PACBIO_REVIO/m84039_230117_233243_s1.hifi_reads.default.bam
HG002 illumina reads: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.GRCh38.60x.1.bam
Ashkenazi trio(HG002, HG003, HG004) PacBio HiFi reads : https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1028149
HPRC assemblies and PacBio HiFi reads : https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC/
PacBio HiFi samples with repeat expansion: https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/
Oxford Nanopore targeted sequencing reads with repeat expansion: https://www.science.org/doi/10.1126/sciadv.abm5386 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA786382
Project Adotto Tandem-Repeat Regions and Annotations: https://zenodo.org/records/8387564
HipSTR tandem repeat regions:

https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/GRCh37.hipstr_reference.bed.gz
Oxford Nanopore Duplex data for HG002:
https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDADE--UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/
Oxford Nanopore Simplex data for HG002:
https://epi2me.nanoporetech.com/gm24385_ncm23_preview/

Author contributions

AKA and DTS conceived and designed the study. AKA contributed to initial code development. AKS wrote the code for ATaRVa. Motif decomposition code was developed by AS. Analysis was performed by AKS and DTS with help from SS and AS. AKS, SS and AS wrote the initial draft. DTS acquired funding and wrote the final draft. All authors read and approved the manuscript.

Funding
This work was supported by the Department of Biotechnology grants BT/PR40264/BTIS/137/44/2022 and BT/PR40270/BTIS/137/65/2023.

Conflict of interest statement
The authors declare no competing interests.

References

1. English, A. C. *et al.* Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat. Biotechnol.* **43**, 431–442 (2025).
2. Reis, A. L. M. *et al.* The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**, 602–610 (2023).
3. Liao, X. *et al.* Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* **6**, 954 (2023).
4. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
5. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A. & Friedman, J. M. Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat.*

Rev. Genet. **25**, 476–499 (2024).

6. Danzi, M. C. *et al.* Detailed tandem repeat allele profiling in 1,027 long-read genomes reveals genome-wide patterns of pathogenicity. Preprint at https://doi.org/10.1101/2025.01.06.631535 (2025).
7. Javadzadeh, S. *et al.* Analysis of targeted and whole genome sequencing of PacBio HiFi reads for a comprehensive genotyping of gene-proximal and phenotype-associated Variable Number Tandem Repeats. *PLoS Comput. Biol.* **21**, e1012885 (2025).
8. Chiu, R., Rajan-Babu, I.-S., Friedman, J. M. & Birol, I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.* **22**, 224 (2021).
9. Dolzhenko, E. *et al.* Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.* **42**, 1606–1614 (2024).
10. Ziaei Jam, H. *et al.* LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol.* **25**, 176 (2024).
11. Tesi, N. *et al.* Characterizing tandem repeat complexities across long-read sequencing platforms with TREAT and *otter*. *Genome Res.* **34**, 1942–1953 (2024).
12. Olson, N. D. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).
13. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
14. HPRC assemblies and PacBio HiFi data. https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC/.
15. PacBio HiFi data for Askenazim Trio. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1028149.
16. Stevanovski, I. *et al.* Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci. Adv.* **8**, eabm5386 (2022).

17. PacBio targeted sequence data for 7 Repeat Expansion Disorders samples. https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/.

18. Chintalaphani, S. R., Pineda, S. S., Deveson, I. W. & Kumar, K. R. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.* **9**, 98 (2021).

19. Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).

20. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).

21. HipSTR hg38 repeat regions. https://github.com/HipSTR-Tool/HipSTR-references/tree/master/human.

22. HG002 Illumina data. https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/.

23. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90–e90 (2019).

24. Farrar, M. Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**, 156–161 (2007).

25. Gao, Y. *et al.* abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics* **37**, 2209–2211 (2021).

26. KMeans — scikit-learn 1.6.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

28. Avvaru, A. K., Sharma, A. & Sowpati, D. T. Ribbit: Accurate identification and annotation of complex tandem repeat sequences in genomes. Preprint at https://doi.org/10.1101/2025.02.06.636828 (2025).

Figure 1

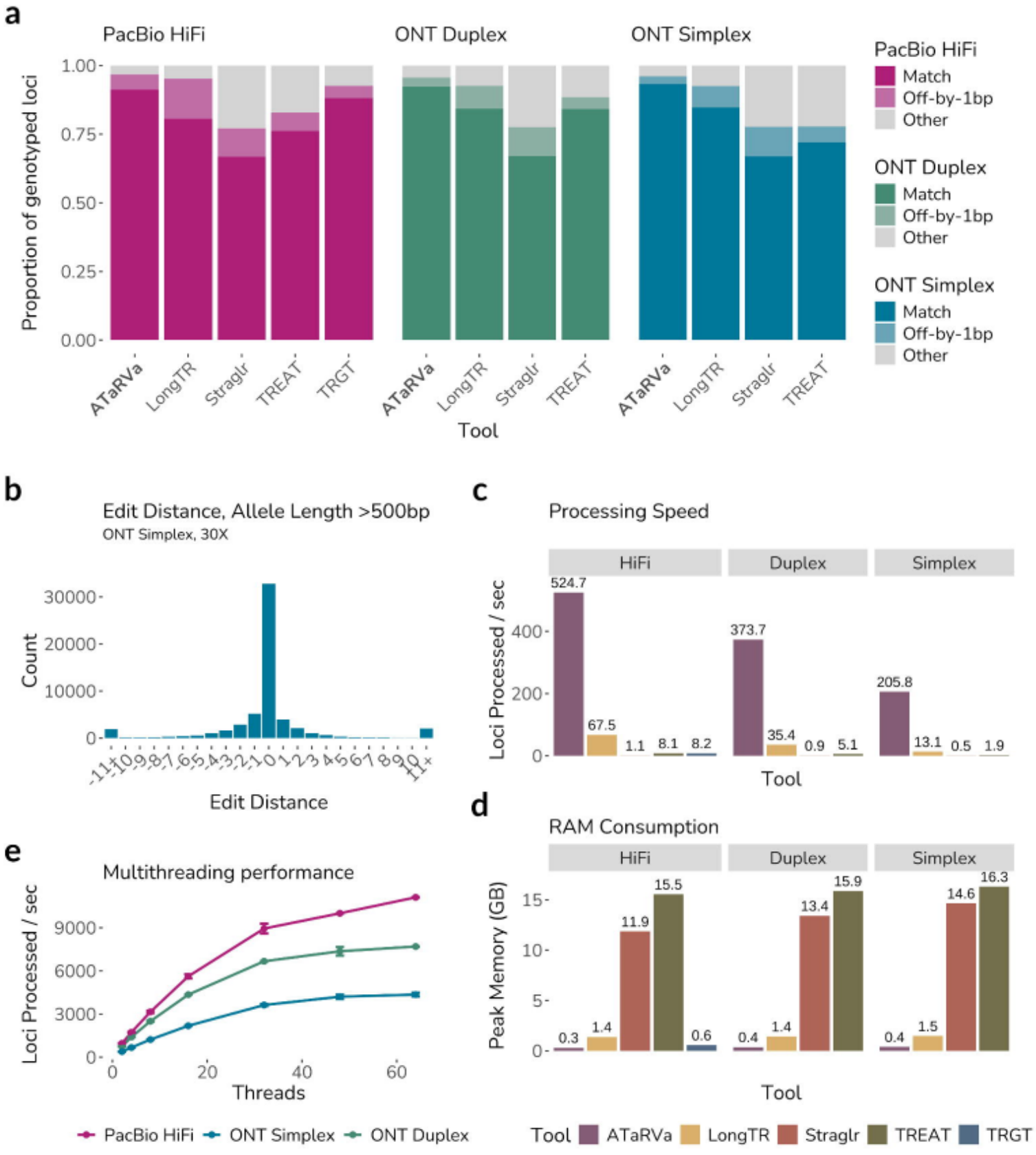
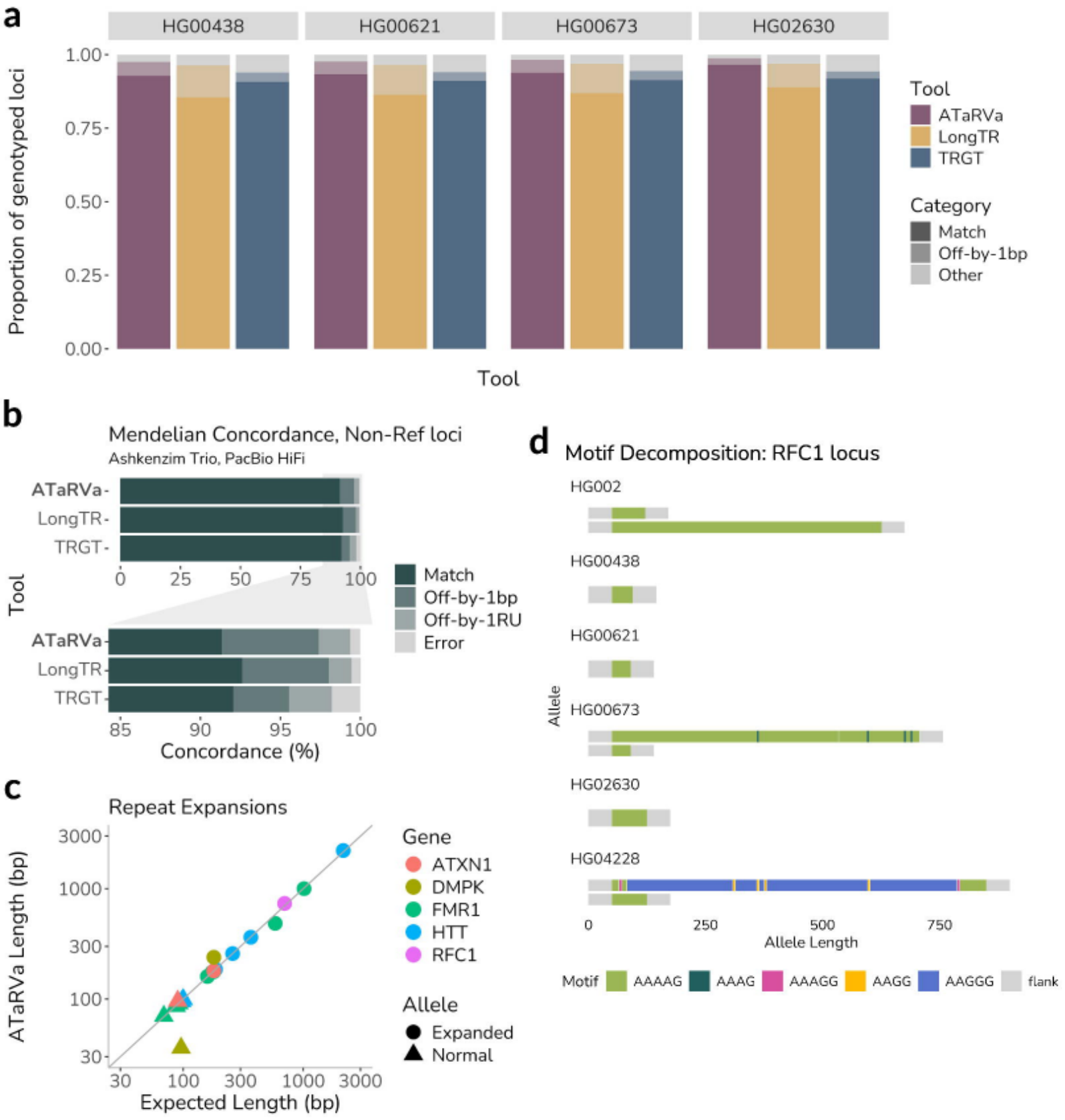

a
PacBio HiFi
ONT Duplex
ONT Simplex
Proportion of genotyped loci
1.00
0.75
0.50
0.25
0.00
ATaRVa
LongTR
Straglr
TREAT
TRGT
Tool
PacBio HiFi
Match
Off-by-1bp
Other
ONT Duplex
Match
Off-by-1bp
Other
ONT Simplex
Match
Off-by-1bp
Other
b
Edit Distance, Allele Length >500bp
ONT Simplex, 30X
Count
30000
20000
10000
0
Edit Distance
c
Processing Speed
HiFi
Duplex
Simplex
Loci Processed / sec
524.7
67.5
1.1
8.1
8.2
373.7
35.4
0.9
5.1
205.8
13.1
0.5
1.9
Tool
d
RAM Consumption
HiFi
Duplex
Simplex
Peak Memory (GB)
0.3
1.4
11.9
15.5
0.6
0.4
1.4
13.4
15.9
0.4
1.5
14.6
16.3
Tool
e
Multithreading performance
Loci Processed / sec
9000
6000
3000
0
0
20
40
60
Threads
PacBio HiFi
ONT Simplex
ONT Duplex
Tool
ATaRVa
LongTR
Straglr
TREAT
TRGT

Figure 2


a
HG00438
HG00621
HG00673
HG02630
Proportion of genotyped loci
1.00
0.75
0.50
0.25
0.00
Tool
Tool
ATaRVa
LongTR
TRGT
Category
Match
Off-by-1bp
Other
b
Mendelian Concordance, Non-Ref loci
Ashkenzim Trio, PacBio HiFi
ATaRVa
LongTR
TRGT
0
25
50
75
100
Tool
Match
Off-by-1bp
Off-by-1RU
Error
85
90
95
100
Concordance (%)
c
Repeat Expansions
3000
1000
300
100
30
ATaRVa Length (bp)
30
100
300
1000
3000
Expected Length (bp)
Gene
ATXN1
DMPK
FMR1
HTT
RFC1
Allele
Expanded
Normal
d
Motif Decomposition: RFC1 locus
HG002
HG00438
HG00621
HG00673
HG02630
HG04228
Allele
0
250
500
750
Allele Length
Motif
AAAAG
AAAG
AAAGG
AAGG
AAGGG
flank