Characterising tandem repeat complexities across long-read sequencing platforms with TREAT

Niccolo' Tesi,[1,2,3*] Alex Salazar,[1*] Yaran Zhang,[1] Sven van der Lee,[1,2] Marc Hulsman,[1,2,3] Lydian Knoop,[1] Sanduni Wijesekera,[1] Jana Krizova,[1] Anne-Fleur Schneider,[1] Maartje Pennings,[5] Kristel Sleegers,[4] Erik-Jan Kamsteeg,[5] Marcel Reinders,[3] & Henne Holstege [1,2,3]

* Authors contributed equally

[1] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[4] Complex Genetics of Alzheimer's Disease Group, Antwerp Center for Molecular Neurology, VIB, Antwerp B-2610, Belgium

[5] Department of Genome Diagnostics, Radboud University Medical Center, Nijmegen, The Netherlands

Corresponding author

Dr. Henne Holstege

Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics

Amsterdam UMC

De Boelelaan 1117HV, Amsterdam, The Netherlands

h.holstege@amsterdamumc.nl

Keyworks

LRS Special Issue; Tandem repeats; Long-read sequencing; PacBio; Nanopore; Genotyping; Cross-platform compatibility;

Abstract

Tandem repeats (TR) play important roles in genomic variation and disease risk in humans. Long-read sequencing allows for the characterisation of TRs, however, the underlying bioinformatics perspective remains challenging.

We evaluated potential biases when genotyping >864k TRs using diverse Oxford Nanopore Technology (ONT) and PacBio long-read sequencing technologies. We showed that, in rare cases, long-read sequencing suffers from coverage drops in TRs, such as the disease-associated TRs in *ABCA7* and *RFC1* genes. Such coverage drops can lead to TR mis-genotyping, hampering accurate assessments of TR alleles and highlighting the need for bioinformatic tools to characterise TRs across different technologies and data-types.

For this reason, we have developed *otter* and TREAT: *otter* is a fast targeted local assembler, cross-compatible across different sequencing platforms. It is integrated in TREAT, an end-to-end workflow for TR characterisation, visualisation and analysis across multiple genomes. Together, these tools enabled accurate characterisation of >864k TRs in long-read sequencing data from ONT and PacBio technologies, with error rates ranging 0.6-1.1% and with limited computational resources.

This performance extends across diverse genomes: applied to clinically relevant TRs, TREAT significantly detected diseased individuals with extreme expansions (p=4.3x10-7 and p=1.4x10-5, TR expansions in RFC1 gene). Importantly, in a case-control setting, we significantly replicated previously reported TRs-associations with Alzheimer's Disease, including those near or within *APOC1* (p=2.63x10-9), *SPI1* (p=6.5x10-3) and *ABCA7* (p=0.04) genes.

Our tools overcome common limitations regarding cross-sequencing platform compatibility and allow end-to-end analysis and comparisons of tandem repeats in human genomes, with broad applications in research and clinical fields.

1. Introduction

Roughly 30% of the human genome consists of tandem repeats (TR) characterised by one or more repeat motifs that are defined by their consecutive repetition.[1] This repetitive pattern often leads to DNA instability, facilitating not only expansions and contractions of the repeating motif sequence, but also allelic diversity within sequence.[2,3] Several definitions of TRs have been introduced based on the motif length and size variability, including microsatellites, minisatellites, and macrosatellites. Microsatellites (or short tandem repeats, STR) are the most abundant TRs in the human genome and are characterised by a repetitive motif of less than 6 base pairs (bp), and tend to cluster in non-coding regions of the genome.[4] Minisatellites are characterised by a repetitive motif with a size between 7 and 100 bp, and they are highly enriched in the telomeric regions of the genome.[5] Macrosatellites are characterised by large tandem repeat units (>100 bp and up to several kilobases), and are enriched in the telomeric and centromeric portions of the genome.[6]

TRs can disrupt gene-expression regulation and contribute to over 40 neurological disorders.[1,7,8] Pathogenic TR expansions, surpassing critical lengths, are linked to conditions like spinocerebellar ataxias, Huntington's disease, Fragile-X syndrome, Amyotrophic lateral sclerosis (ALS), and Myotonic Dystrophy.[7,8] For instance, Fragile-X syndrome results from a GGC repeat expansion in the *FMR1* gene, with affected individuals having up to 4,000 copies compared to less than 50 in healthy individuals.[9] Similarly, ALS is associated with an intronic hexa-nucleotide repeat expansion (GCCCCG) in the *C9orf72* gene, exceeding a critical length of more than 200 copies.[10] Beyond causing diseases, TRs, such as the intronic VNTR in the *ABCA7* gene, are identified as risk factors for polygenic human diseases, like a 4.5-fold increased risk of Alzheimer's Disease (AD) when the TR exceeds 5720 base pairs.[11]

Traditionally, the evaluation of TRs lengths and sequences has been challenging. Conventional methods, such as repeat-primed polymerase chain reaction (PCR) and Southern blot assays, are time-consuming and limited in detecting TRs exceeding PCR-based boundaries. Short-read

sequencing approaches offer an alternative, but their limited read lengths often fail to span repetitive regions effectively. Despite heuristic methods and statistical modelling,[12–17] accurately assessing clinically relevant TRs remains difficult. The advent of long-read sequencing, particularly with PacBio's High Fidelity (HIFI) and Oxford Nanopore Technology's duplex technology (10-20kb on average, >99% accuracy),[18,19] has significantly improved TR evaluation by providing long and accurate fragments.

Characterising TRs with long-read sequencing technology currently has two main limitations. First, there is a lack of comprehensive studies that have investigated potential biases when sequencing TRs. For example, DNA methylation has been previously shown to impact base-calling accuracy in long-read sequencing data.[20–22] Similarly, the formation of secondary structures due to TRs could impact enzyme efficiency (e.g. polymerase or nanopores),[23] potentially reducing read-quality and sequencing throughput in current long-read sequencing technologies. Furthermore, some technologies require the alignment of noisy reads to generate high quality consensus sequences, which might be more difficult in case of repetitive regions. These problems may impact genotyping accuracy and lead to incorrect assessments of allele-sequences, including disease-associated TRs in patients.

Second, there is a lack of cross-compatible tools to characterise TRs across different (long-read) sequencing technologies and data-types.[24–26] This is critically important given the growing long-read sequencing initiatives aiming to comprehensively assess TRs in extensive genome datasets,[27] spanning both population-wide and clinical contexts. For example, some existing tools are constrained by predefined TR databases, hindering the identification of new TR features such as novel motif sequences;[28] other tools are technology and data-type-dependent,[24] or do not produce generalizable multi-sample outputs.[25,26]

In this study, we first evaluated potential biases when genotyping TRs with long-read sequencing data. We do this by evaluating sporadic coverage drops in >864k genome-wide TRs using the well-

characterised HG002 genome, based on long-read sequencing data from ONT (duplex), HIFI and non-HIFI data from PacBio's Revio and Sequel II instruments. Based on our findings, we developed TREAT (Tandem REpeat Annotation Toolkit), a unified workflow for characterising TRs across multiple genomes, cross-compatible across long-read technologies and data-types (e.g. read-alignments and de novo assemblies). TREAT employs a novel generic targeted local assembler, *otter*, that can adapt to different sequencing chemistries to accurately characterise TRs. To illustrate the performances of TREAT, we: (i) evaluated genotyping-accuracy across previously curated-TRs in long-read sequencing data from ONT and PacBio at various coverage levels; (ii) showcase clinical applicability in patients as well as healthy controls; and (iii) characterise TRs in case-control setting.

2. Results

2.1 Tandem repeats may be sensitive to coverage dropouts in long-read sequencing

We first curated a genome-wide set of 864,424 tandem repeats (TRs) based on available repeat annotations in the CHM13 reference genome (see Methods). The average TR-length in this curated set was 93 bp, with average GC-content of 32%. Next, we aligned long-read sequencing data of HG002 from Oxford Nanopore Technology (ONT) Duplex chemistry, HIFI and non-HIFI data from PacBio's Revio and Sequel II instruments to the CHM13 reference genome. These three datasets harbored genome-wide coverages at ~38x, with median read-lengths ranging from 16-22 Kbp.

A systematic analysis of TR-coverage profiles revealed a set of 454 TRs with unexpected coverage drops in at least one of the three sequencing technologies. The mean ratio was 1.01, 1.01, and 1.05, respectively, for the three technologies, indicating generally no unexpected coverage-drops of TRs (Figure S1A). However, 478 unique TRs had ratios below 0.25 (i.e. four-fold lower than expected average coverage), of which 454 (95%) are present in the HG002 T2T reference assembly (Figure S1B-H). Indeed, we observed higher read-support for these TRs when integrating non-HIFI data from Revio and Sequel II, indicating that although reads were generated for these TRs, most were flagged as low quality during HIFI-data generation (Figure S1I). Importantly, the majority of these TRs (294/454) overlap gene annotations, potentially leading to mis-genotyping that may impact biological interpretation.

Among this set is an intronic TR in *ABCA7*, previously identified as a risk factor for Alzheimer's Disease (AD). Coverage in this TR is significantly reduced in Revio and Sequel II data (Figure S1J). Notably, all three technologies highlight a potential mis-assembly in HG002 as they do not support the expanded version of the *ABCA7* TR (Figure S1J). Manual inspection revealed that the expanded TR in the HG002 assembly harbored TE-like sequences not supported by any of the long-read datasets, suggesting either variation specific to the sequenced HG002 cell-line, or a mis-assembly

116 in the HG002 reference. We find similar sequencing challenges in other coding TRs, including
117 complete coverage dropouts in expanded allele-sequences, such as an intronic TR in *NLRP9*, or
118 high length-variability present in only one of the three technologies, such as a TR overlapping the
119 non-long coding RNA, *PCCA-DT* (Figure S1J).
120
121 We further investigated challenges when sequencing the intronic TR in *ABCA7* by comparing TR-
122 lengths from Southern Blotting assay (see Supplementary Methods) and whole-genome long-read
123 sequencing data generated on the Sequel II instrument from nine Dutch individuals (see Methods,
124 Figure 1 and Figure S2). The genome-wide coverage ranged from 11.5-34x, with median read-
125 lengths of 14-19 Kbp. We found similar coverage drops in the *ABCA7* TR in all nine genomes, as
126 the local coverage of spanning (region of interest is fully contained in the reads) and non-spanning
127 (region of interest is partially contained) reads ranged from 1-7x. We found low-concordance
128 between lengths in HIFI data and the observed lengths from Southern Blotting assay, suggesting
129 challenges in accurately characterising this TR with HIFI technology.
130
131 Although rare, we cannot exclude larger impacts of challenges when sequencing TRs in other
132 genomes, especially in patients with expanded TRs and variable sequencing throughput. Our results
133 highlight the benefit of unified integration of long-read data from multiple technologies and data-types
134 for the comparison and characterisation of TRs.

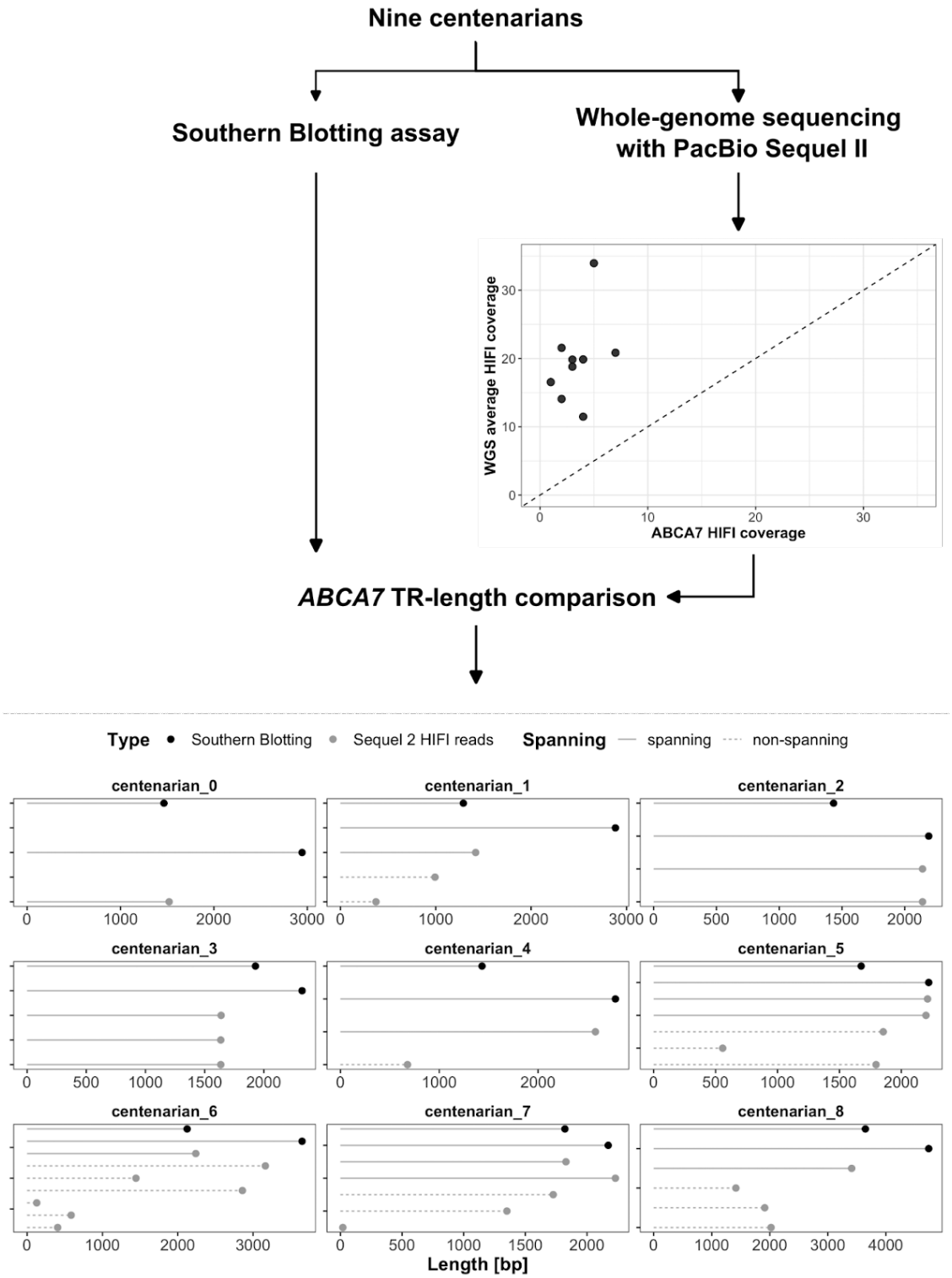


Figure 1: Allele and coverage dropouts in the TR intronic of ABCA7 gene. We performed PacBio long-read sequencing as well as experimental validation (by Southern Blotting) for 9 centenarian genomes. Based on HIFI data, we identify variable coverage drops in the ABCA7 TR compared to the relative average genome coverage. When looking at the individual HIFI reads aligning to the region, we observed that the majority of HIFI reads (56%) did not span the TR-alleles.

2.2 Cross-compatible workflow for characterising tandem repeats with *otter* and TREAT

Given our observations above, we present *otter* and TREAT, two bioinformatic tools that enable tandem repeat (TR) characterisation across different long-read sequencing technologies and data-types (see Figure 2). *otter* is a stand-alone generic targeted local assembler for long-read sequencing data, which automatically adapts to sequencing error-rates and coverage levels per target region. We integrated *otter* in TREAT, an end-to-end unified workflow to enable high-resolution characterisation, visualisation, and analysis of TRs, including length, motif composition, and zygosity. TREAT requires sequencing data aligned to the reference genome (.bam files), the coordinates of the regions of interest (chromosome, start and end positions encoded in a .bed file), and the reference genome used (.fasta file). There are two main analysis strategies: (1) the assembly-based analysis, which uses *otter* to genotype specific TR. TR genotyping is then followed by TR content characterization (see Methods) based on Tandem Repeat Finder;[29] (2) the read-based analysis, which extracts all individual reads spanning a given region of interest using Samtools,[30] estimate and TR content, and performs haplotype calling using a clustering framework (see Methods). TREAT outputs a multi-sample VCF (Variant Call Format) file reporting phased size, copy number and the relative motif characterization of each TR in each sample. Furthermore, TREAT implements an analysis module to identify extreme variations in TR sizes across a set of samples, and a plot module to display TR alleles across samples.

TREAT is written in Python and R (for plots). The source code is freely available at https://github.com/holstegelab/treat along with example datasets, documentation, and a Conda configuration file to ease the installation.

Otter is a tool written in C++: source code, example usage and documentation are freely available at https://github.com/holstegelab/otter. *Otter* can be used as part of the TREAT analysis workflow, or as a stand-alone application.

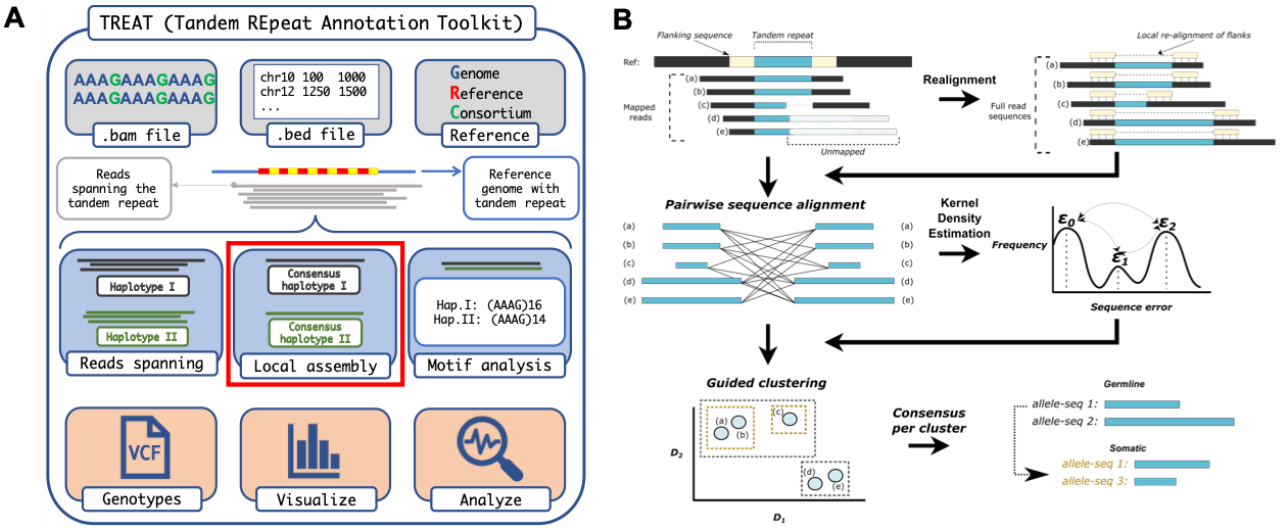


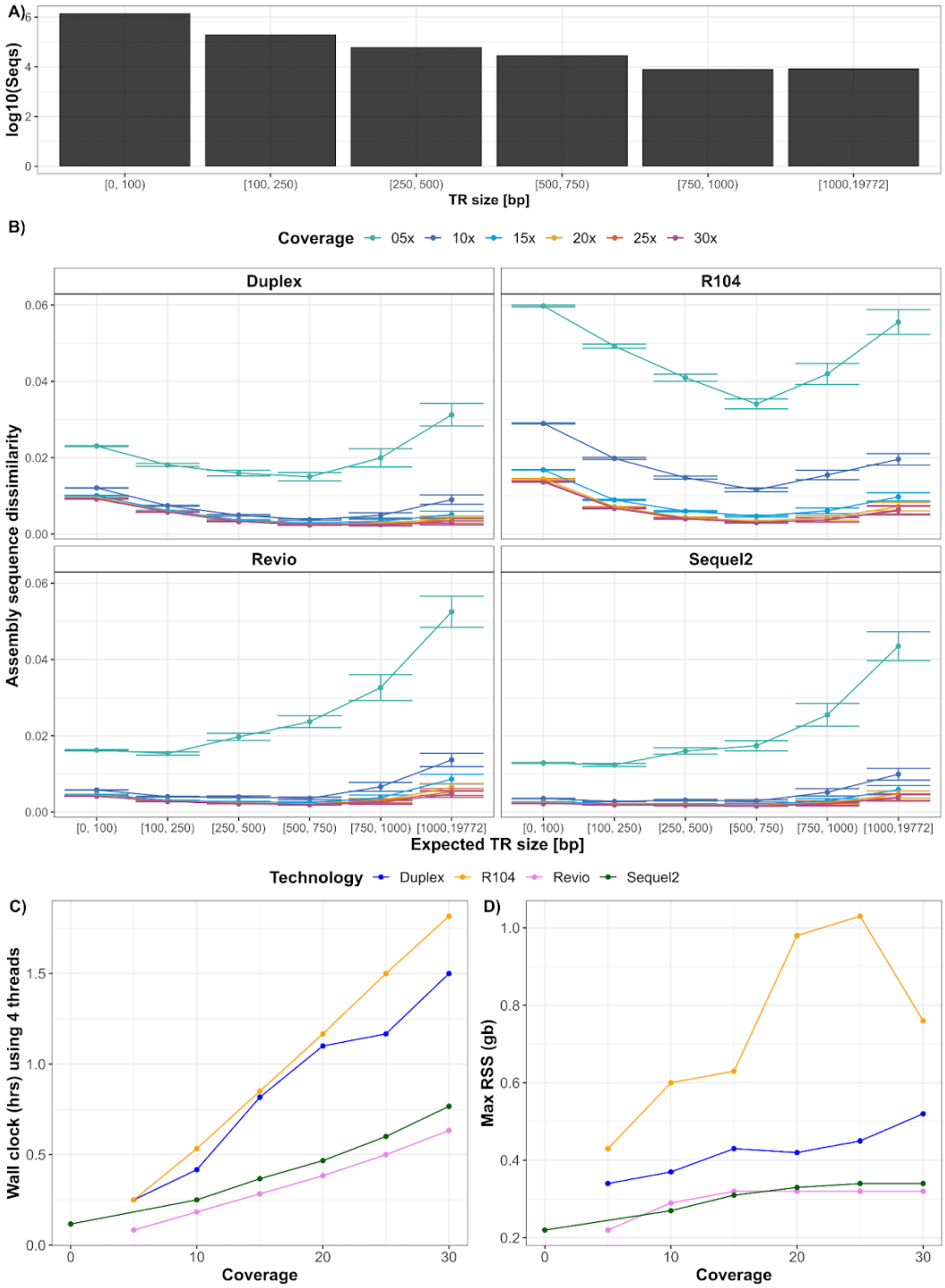
Figure 2: Schematic workflow of TREAT and otter. ***A.*** *Shows TREAT workflow, highlighting the required inputs, the main features, and the main outputs of the tool. The red box highlights the assembly-based analysis based on otter.* ***B.*** *Summarises the main algorithmic steps of otter, a novel targeted local assembler for long-read sequencing data.*

2.3 Otter accurately assembles tandem repeat sequences

To benchmark *otter*, we use the same set of 864,424 TR (corresponding to 1,728,848 allele sequences) used in our previous systematic analysis (Figure 2A). We identified both the maternal and paternal allele-sequences in HG002 for 863,343/864,424 TRs (99.9%), and identified a single parental allele-sequence for 839/864,424 TRs (0.09%). We were not able to identify any allele-sequences for 105 TRs (0.01%); 50 TRs (0.01%) were not fully spannable due to soft/hard-clipping in the whole-genome alignments; and 137 TRs (0.02%) had more than two parental alignments in the same expected chromosome due to alternative alignments.

We then used *otter* to perform targeted local-assembly of this curated set of TRs, and compared it to the expected allele-sequences using CHM13-aligned long-read sequencing data of HG002 generated on both ONT and PacBio sequencing instruments. Figure 2B shows the average assembly-sequence dissimilarity from *otter* as a function of TR-size and coverage across different sequencing technologies: ONT with duplex and R10.4 chemistries, and HIFI data from PacBio's Revio and Sequel II sequencing instruments. Across all three technologies, increasing coverage

189 yielded more accurate TR-sequence assemblies. At 15x coverage, the average sequence
190 dissimilarities were 0.923% and 1.52% for ONT's Duplex and R10.4 chemistries, respectively, and
191 0.442% and 0.262% for PacBios's Revio and Sequel II, with slight improvements at higher
192 coverages. Overall, we found that *otter's* assemblies were most accurate when using Sequel II HIFI
193 data, followed by Revio, Duplex, and R10.4 (Figure 2B). A closer analysis revealed that Sequel II
194 HIFI data had significantly lower read-sequence variability relative to Duplex, R10.4, and Revio
195 (Wilcoxon rank-sum test, p-value < 2.22e-16), based on the pairwise-sequence alignments of the
196 reads used by *otter* to assemble each allele-sequence (see Supplementary Results). This is possibly
197 due to a combination of increased sequencing errors and somatic variation in Revio data relative to
198 Sequel II, which ultimately yielded suboptimal assemblies relative to the expected allele-sequences
199 from the HG002 T2T assembly.
200
201 Importantly, for PacBio read-sets, the above benchmark required a maximum runtime of up-to 47
202 minutes across all coverage-levels for all 864,424 TRs (four threads), using no more than 0.37 Gb
203 of memory (Figure 2C). We observed higher runtime (≤ 1.82 hrs) and memory usage (≤ 1.03 GB) in
204 ONT data, likely due to additional resource requirements for processing larger-read lengths and
205 higher error-rates (Figure 2D).
206 It's important to note that *otter* only provides assembled sequences. Hence, downstream
207 characterisation and visualisation of TRs is performed via TREAT.



208

209 *Figure 3:* ***A.*** *Size distribution of the curated set of 864,170 unique TRs (1,727,475 allele-sequences in total),*

210 *with an average length of 93 bp.* ***B.*** *Assembly-sequence error-rates as a function of the TR-size and*

211 *sequencing technologies.* ***C.*** *Run time of otter as a function of the coverage and sequencing technologies.* ***D.***

212 *Memory consumption of otter as a function of the coverage and different sequencing technologies.*

213

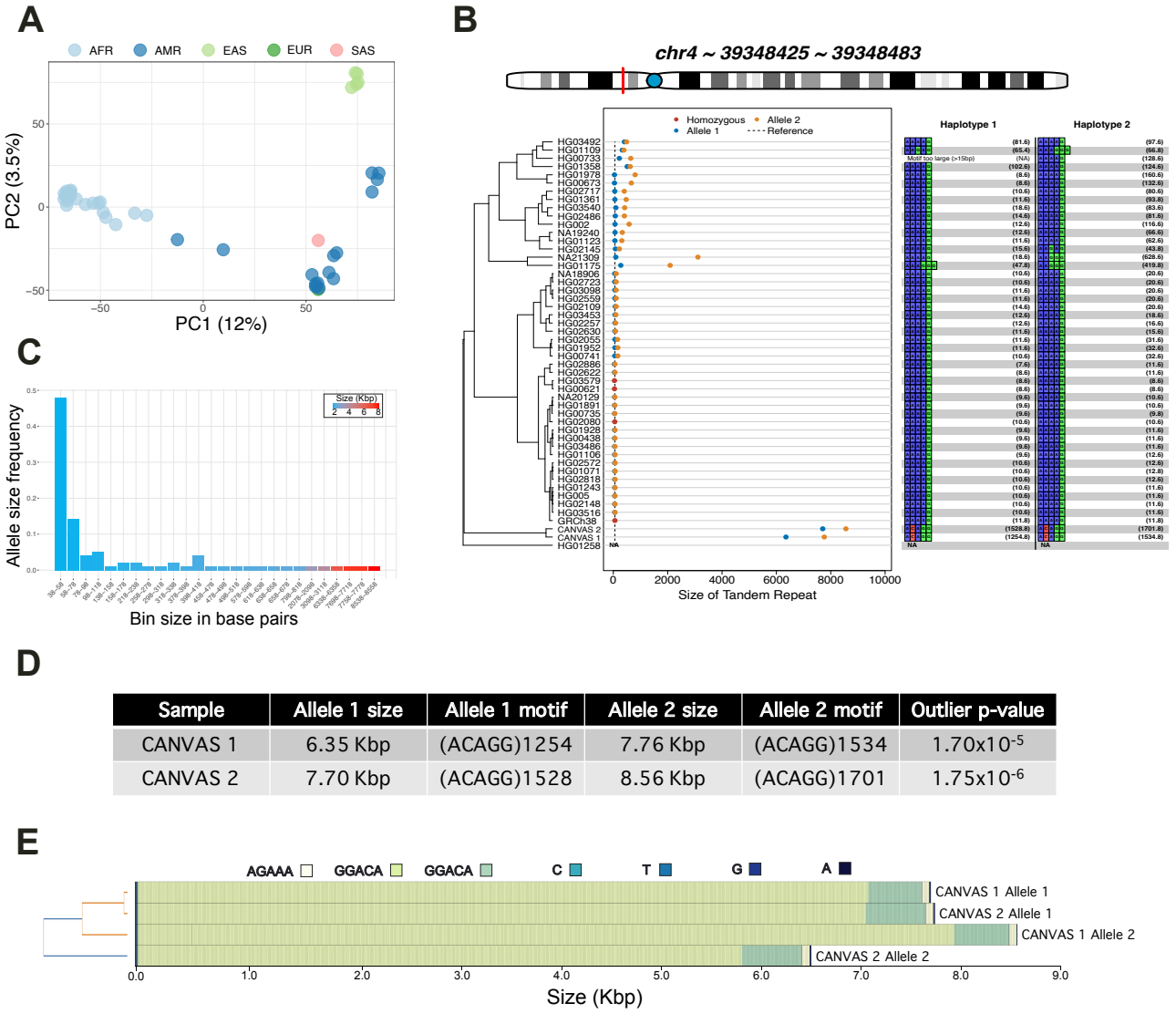
2.4 TREAT enables tandem repeat characterisation in patients and genomes with diverse ancestries

We evaluated TREAT's unified workflow for characterising TRs across genomes from diverse ancestries. We first characterised PacBio's curated set of 161,037 TRs in GRCh38 using HIFI data from 47 genomes from the Human Pangenome Consortium (HPRC). We did this by comparing individual read support (TREAT read-based analysis) to each of the unique allele-sequences assembled by *otter* (TREAT assembly-based analysis), across all genomes. We found nearly perfect concordance between the assembly and reads when correlating the size of the TRs across the shorter, longer, and joint allele sizes (Spearman correlation=0.99, p<2x10-16 in all cases), indicating robust utility to characterise TRs across different ancestries, including genome with relatively higher levels of heterozygosity.

We then used TREAT to perform a TR-based population stratification analysis on all 47 HPRC genomes. We extracted the top 20% most variable TRs (N=32,208, based on the coefficient of variation, see Methods), and performed a principal component analysis (PCA, Figure 4A) on the joint allele sizes (ie, the sum of the maternal and paternal alleles). We found that PC1 explained 12% of the total variance and genetically represented the African-American axis, while PC2 explained 3.5% of variance and corresponded to the American-Asian axis (Figure 4A). The explained variance was similar to that of a PCA including 40/47 matching samples and 30,544 random common (minor allele frequency >10%) SNPs (PC1: 14%, PC2: 4%, Figure S3).

Finally, we used TREAT's analysis module to detect and score extreme TR expansions or contractions in 35 clinically relevant TRs (Table S1) in 47 genomes from the HPRC, as well as two Dutch CANVAS patients. The two CANVAS patients were previously characterised to harbor expansions in the intronic TR in *RFC1*, and long-read whole-genome sequencing data were generated with PacBio's Sequel II instrument. In total, we identified 33 instances where the TR length in certain samples were significantly different from the distribution of tandem repeat lengths across

all 49 genomes. The most significant deviations were observed for the TR in the *RFC1* gene (p=4.3x10-7 and p=1.4x10-5, respectively, Figure 4B-E), the *ZIC2* gene (p=4.3x10-10 and p=6.4x10-03, Figure S4), the JPH3 gene (p=5.3x10-10, Figure S5), and the *CNBP* gene (1.33x10-9, Figure S6). While the repeat expansion in *RFC1* was homozygous in both patients, the repeat expansions in *ZIC2, JPH3*, and *CNBP* were all heterozygous, and thus likely not pathogenic.



Sample	Allele 1 size	Allele 1 motif	Allele 2 size	Allele 2 motif	Outlier p-value
CANVAS 1	6.35 Kbp	(ACAGG)1254	7.76 Kbp	(ACAGG)1534	1.70×10^{-5}
CANVAS 2	7.70 Kbp	(ACAGG)1528	8.56 Kbp	(ACAGG)1701	1.75×10^{-6}

Figure 4: TREAT visualisation and analysis modules. ***A.*** *The PCA of the ancestry-based analysis based on the 20% most variable TRs across 47 genomes.* ***B.*** *The main TR in the RFC1 gene. Y-axis: individuals, X-axis: TR size (in bp). Blue dots: smaller allele, orange dots: larger allele, red dots; homozygous genotypes. Dashed line: the allele in the reference genome GRCh38. The right side of the plot reports, for each sample and each allele, the motif and relative number of copies.* ***C.*** *The distribution of allele sizes for the TR in RFC1 gene.* ***D.*** *The output of the analysis module for the TR in the RFC1 gene. The TR length of the two CANVAS*

patients were identified as significant outliers compared to the length-distribution of 47 samples from the HPRC. E. Motif representation in CANVAS patients, as produced with MotifScope.[31]

2.5 TREAT can correct mis-characterisations of tandem repeats in clinical settings

We used TREAT to integrate HIFI and non-HIFI data in order to better characterise the Alzheimer's Disease-associated intronic TR in *ABCA7* in the nine centenarian genomes for which experimental validation by Southern Blotting was available. Then, we correlated the TR-alleles from Southern Blotting with those estimated by TREAT. Using HIFI-only data, we found a Pearson correlation of 0.58. The inclusion of non-HIFI data increased read-support to an average coverage of 22x. As a result, the Pearson correlation increased to 0.99 when including non-HIFI-only data and when merging both HIFI and non-HIFI data (Figure 5A). These results showcase TREAT's utility to integrate long-read data with different error-rates, and they confirm the value of including non-HIFI reads for accurate characterisations of complex TRs.

Separately, we characterised an intronic TR in the *RFC1* gene across HIFI long-read sequencing data from multiple tissues from a Dutch cognitively healthy centenarian parent-child duo. This TR has been previously associated with CANVAS, where the most common allele consists of an (AAAAG)11 motif, with a total size of ~55bp. In CANVAS patients, the TR can range from 2-10 Kbp in total length (Figure 4E).

We found that the intronic *RFC1* TR was expanded in this Dutch family (Figure 5B). More specifically, long-read data from blood of the parent yielded heterozygous allele-sequences: one similar to the wild-type allele-sequence with a total length of 244 bp with the (AAAAG)50 motif; the other allele was expanded with a total length of 2.49 Kbp composed primarily of the (AAGGG)490 motif (Figure). Although long-read data from brain of the same centenarian yielded similar heterozygous alleles with the same motif sequences, the expanded allele-sequence was 2.67 Kbp in total length. This

281 suggests somatic expansion in brain tissue by 180 bp (36 additional motif-copies), relative to blood
282 tissue.
283
284 Interestingly, long-read data from the child of the centenarian yielded a homozygous allele-sequence
285 of 63 bp with the (AAAAG)12 motif (Figure 5B). This was unexpected as at least one of the two
286 allele-sequences from the parent should be inherited in the child. A closer analysis of HiFi long-read-
287 pileup overwhelmingly supported this genotype. However, similar to our previous observations when
288 characterising TRs in HG002, we observed an abnormal drop in the relative coverage of spanning
289 and non-spanning both the centenarian and child, but increased coverage when including non-HIFI
290 reads (see Supplementary Results). Merging HIFI and non-HIFI data of the child enabled TREAT to
291 correctly assemble the expanded allele-sequence at 2.65 Kbp in total length with (AAGGG)>374
292 motif due two supporting non-HIFI reads. Indeed, penta-repeat primed PCR (RP-PCR, see
293 Supplementary Methods: RP-PCR assay and Figure S7) of whole-blood DNA confirmed that both
294 individuals harbored repeat expansions separately composed of the (AAAAG) and (AAGGG) motifs
295 (Figure S6) Hence, the HIFI sequencing data failed to capture this expanded allele-sequence, but
296 was recoverable when including the non-HIFI data.

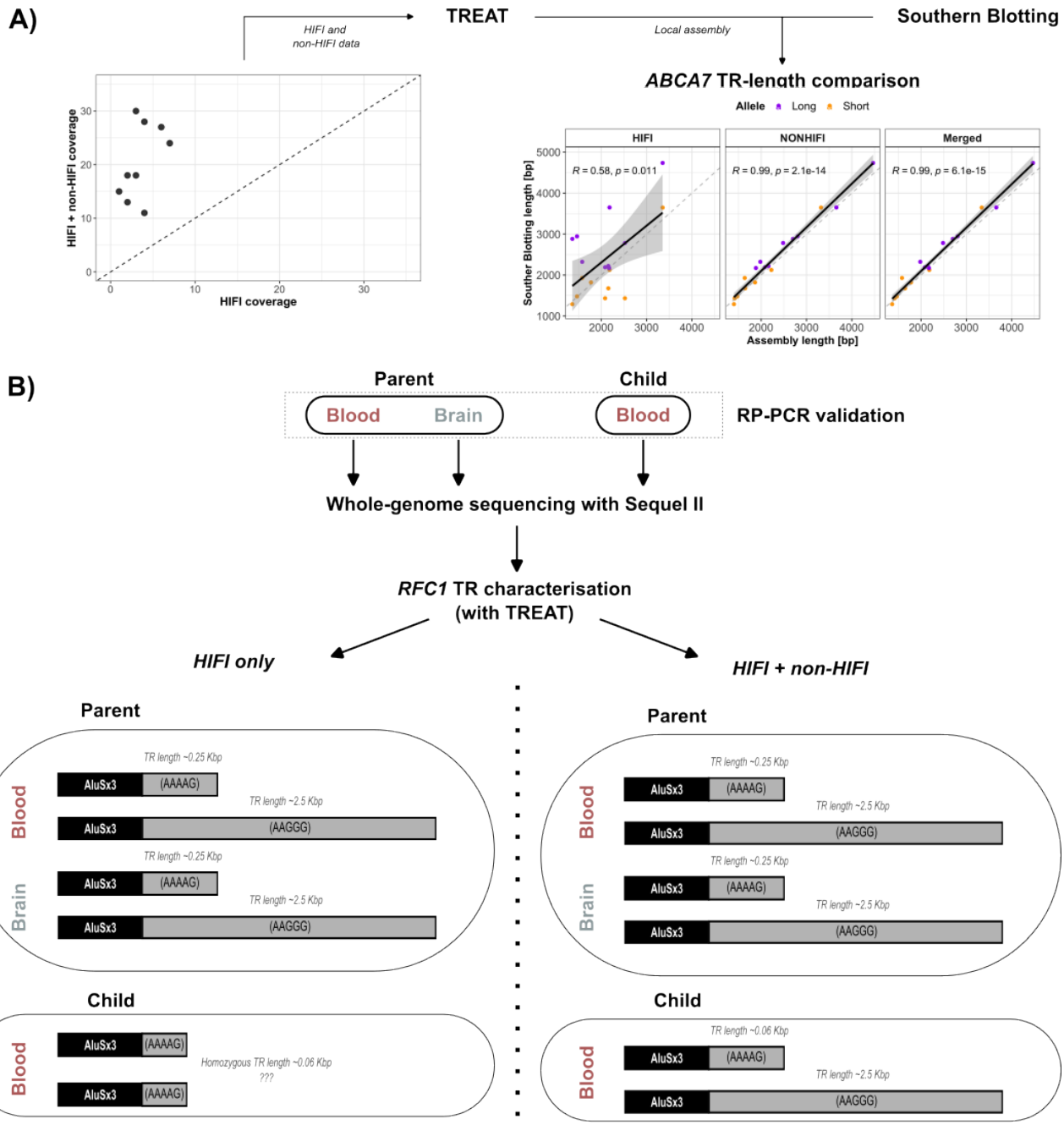
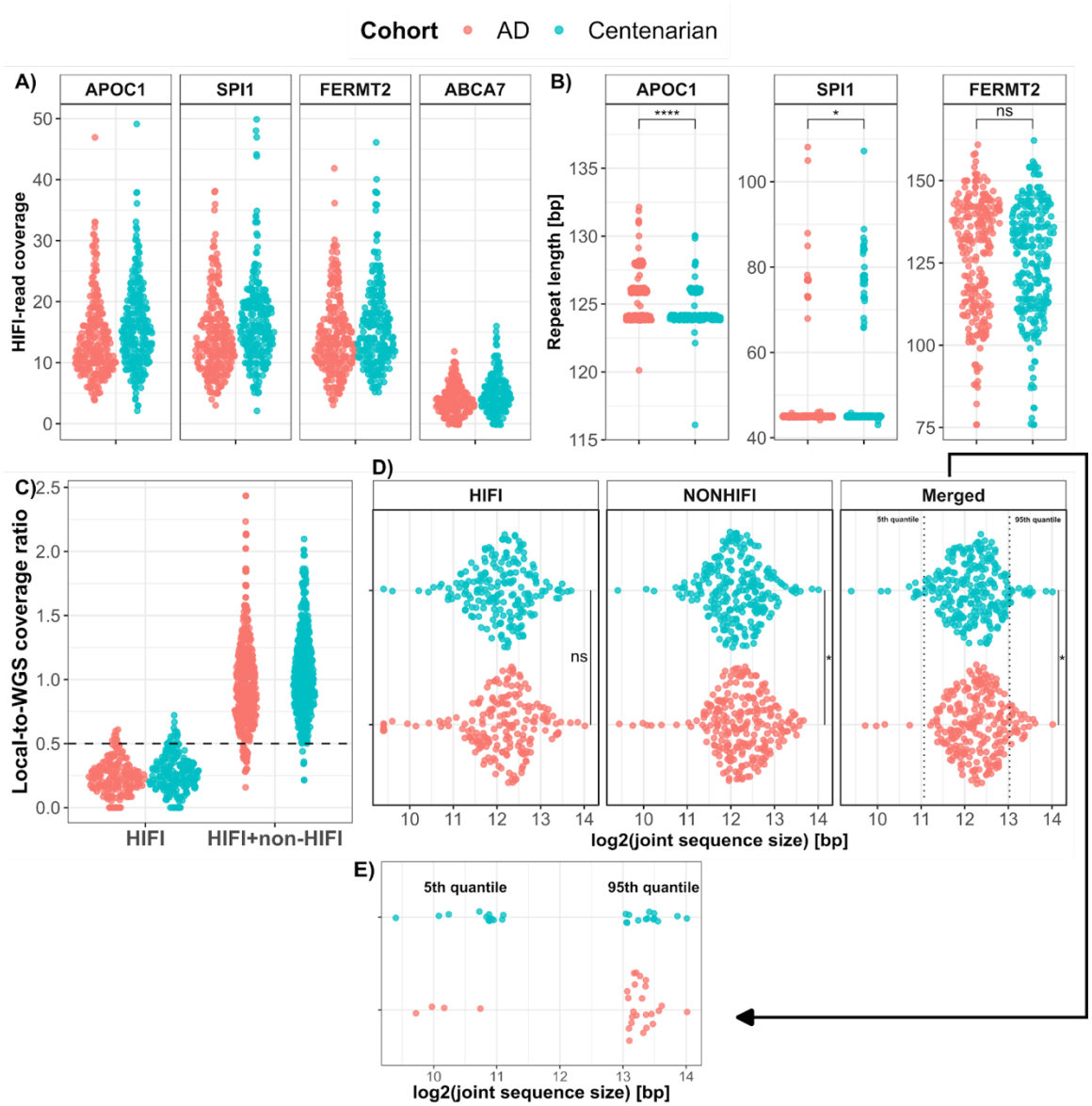


Figure 5: Recovery of allele dropout in ABCA7 and RFC1 region. ***A.*** *Comparison between experimentally validated alleles and TREAT's alleles for the TR in ABCA7 gene, when using HIFI data, non-HIFI data, and their combination. Correlations are based on 9 genomes for which both experimental validation and PacBio long-read sequencing was available. Adding non-HIFI data allowed the recovery of the majority of TR-alleles.* ***B.*** *Comparison of TREAT's allele for the TR in RFC1 gene when using HIFI and non-HIFI data, for the centenarian duo composed of blood-DNA and brain-DNA of the parent, and blood-DNA of the child. When adding non-HIFI data, we could recover the expanded allele in the child, missing in HIFI data alone.*

306 ## 2.6 Comparing tandem repeats across multiple genomes in a
307 ## case-control setting

308 To showcase the applicability of TREAT in a case-control setting, we replicated the association of
309 four TRs that were previously shown to associate with Alzheimer's Disease (AD) risk (Table 1 and
310 Table S2). We did so by using a set of 250 AD patients (mean age = 67.9±9.8, 70% females) and
311 N=248 cognitively healthy centenarians (mean age = 101.2±2.5, 70% females) that were sequenced
312 with PacBio Sequel II instrument (see Methods and Figure S8). Across all 498 genomes, we
313 observed a median coverage (HIFI data) of 14, 15, 14, and 4, respectively for the TRs in *APOC1*,
314 *SPI1*, *FERMT2*, and *ABCA7* (Figure 6A). The combined allele size (ie, the sum of the maternal and
315 paternal alleles) of the TR nearby *APOC1* (chr19:44921096-44921134) was significantly expanded
316 in AD patients compared to cognitively healthy centenarians (beta=0.38, p=2.63x10-9, Figure 6B,
317 Table 1). In contrast, the short allele of the TR within *SPI1* gene was significantly contracted in AD
318 patients compared to cognitively healthy centenarians (beta=-0.03, p=6.5x10-3, Figure 6B, Table 1).
319 The direction of effect of these TRs was in line with the original studies. We could not replicate the
320 association of the TR within *FERMT2* (beta=0.01, p=0.27, short allele) (Figure 6B, Table 1). Notably,
321 we found that expansion and contractions for these TRs were mainly driven by homopolymers,
322 especially in *SPI1* and *FERMT2*.
323
324 For the intronic TR in *ABCA7*, we found significant expansions in AD cases after integrating non-
325 HIFI data (beta=8.63x10-5, p=0.04, joint allele size, Figure 6C-D). We note that 22 samples were
326 omitted due to reduced coverage levels even after integrating HIFI and non-HIFI data. In addition,
327 we identified TR size boundaries in the centenarian controls corresponding to the 5th and 95th
328 percentiles of the joint TR allele sizes (2.2 Kbp and 8.4 Kbp, respectively). The number of
329 centenarians with a TR size lower than the 5th percentile was 3-folds higher than that of AD cases
330 (1-tailed Fisher's exact test p=0.023, OR=3.2, Figure 6E), and the number of AD cases with a TR
331 size larger than the 95th percentile was 2-folds higher than that of centenarians (1-tailed Fisher's
332 exact test p=0.04, OR=2.0, Figure 6E). Given the difficulties in correctly assessing the allele

333 sequences of this TR, we cannot exclude that additional samples suffer from allelic dropouts,
334 especially for the larger expanded allele-sequences.



335

336 *Figure 6: Replication of the association with AD of TRs in APOC1, SPI1, FERMT2 and ABCA7.* ***A.*** *The*
337 *coverage distribution of the four TRs in AD patients and cognitively healthy centenarians.* ***B.*** *The TR size*
338 *difference between AD patients and cognitively healthy centenarians in APOC1, SPI1 and FERMT2. For the*
339 *associations, we used logistic regression models using the TR size as predictor for AD case-control status.* ***C.***
340 *HIFI and non-HIFI coverage distribution of the TR intronic of ABCA7 gene.* ***D.*** *Comparison of the joint allele*
341 *size of ABCA7 TR between AD cases and cognitively healthy centenarians, respectively using HIFI data, non-*
342 *HIFI data, and the merged dataset of HIFI and non-HIFI.* ***E.*** *Number of AD cases and cognitively healthy*
343 *centenarians in the lower 5th quantile and upper 95th quantile. Quantiles were defined based on the distribution*

344 *of the joint TR-allele size in the centenarians. We tested the differential enrichment of AD and centenarians in*
345 *each quantile with Fisher's exact tests.*
346
347 *Table 1: Replication of TR previously associated with AD*

	TRs previously associated with AD			
Region	*chr19:44921096-44921134*	*chr11:47775208-47775243*	*chr19:1049436-1050066*	*chr14:52832909-52832938*
Gene	*APOC1*	*SPI1*	*ABCA7*	*FERMT2*
Best model	*Joint alleles*	*Short alleles*	*Joint alleles*	*Short alleles*
Beta (OR)	*0.38 (1.46)*	*-0.03 (0.97)*	*8.63x10-5 (1.01)*	*0.01 (1.01)*
P-value	*2.6x10-9*	*6.5x10-3*	*0.041*	*0.27*
Original study	*38014121*	*37745545*	*29589097*	*37745545*
Original OR	*NA*	*-0.01 (0.99)*	*4.5*	*0.01 (1.01)*
Original model	*Longer allele*	*Joint alleles*	*Individuals with alleles >5720 bp*	*Joint alleles*
Original method	*Logistic regression*	*Mixed linear models*	*Fisher's exact*	*Mixed linear models*
Original p-value	*4.3x10-10*	*NA*	*0.008*	*NA*
Original samples	*1489 AD vs. 1492 controls*	*6328 AD vs. 6580 controls*	*275 AD vs. 177 controls*	*6328 AD vs. 6580 controls*
Data type	*Short read sequencing*	*Short read sequencing*	*Southern blot*	*Short read sequencing*

348 *Region: genomic coordinates of the TR with respect to GRCh38; Gene: the closest gene as reported in the*
349 *original publications; Best model: model that yielded the most significant association, in our comparison: Short*
350 *allele, Long allele or Joint alleles size; Beta (OR): effect size and relative Odds Ratio with respect to AD: an*
351 *increased TR size leads to increased AD risk for positive estimates; P-value: p-value of association. We used*
352 *logistic regression models using TR size (short allele, long allele and combined allele size) as predictor for AD*
353 *case-control status, using 250 AD patients (cases) and 248 cognitively healthy centenarians (controls); Original*
354 *study: the Pubmed ID of the original study; Original OR: the odds ratio as reported in the original study; Original*
355 *model: model used for association in the original study; Original method: method used for association in the*
356 *original study; Original p-value: the p-value reported in the original study; Original samples: the number of AD*
357 *cases and controls used in the original study; Data type: the data on which the association were identified.*

3. Discussion

Characterising tandem repeats using long-read sequencing data can be challenging. We showed that, in rare cases, state-of-the-art long-read sequencing technologies suffer from coverage drops in TRs, such as the disease-associated intronic TRs in *ABCA7* and *RFC1* genes. Such coverage drops can lead to TR mis-genotyping, hampering accurate assessments of TRs. These observations highlighted the need for bioinformatic tools to characterise TRs across different technologies and data-types. For this reason, we have developed *otter*, a fast generic targeted local assembler for long-read sequencing data. It is integrated in TREAT, a unified workflow to genotype, analyse and visualise TRs. Together, these tools enabled accurate characterisation of >864k TRs from long-read sequencing data from ONT (Duplex and R10.4 chemistries) and PacBio (Revio and Sequel II), with limited computational resources. This performance extends across diverse genomes, enabling the successful identification of disease-associated TR expansions in CANVAS patients. Importantly, we replicated previously reported TR-associations with Alzheimer's Disease, including those near or within *APOC1*, *SPI1*, and *ABCA7* genes, showcasing the utility of *otter* and TREAT in case-control settings.

The set of TRs with coverage dropouts found in this study (which were concentrated in Revio and Sequel II data) may be consequences of alignment artefacts and spurious quality differences during whole-genome sequencing. However, we found that these TRs are not only present in the HG002 T2T reference assembly, but were consistently supported in ONT Duplex and R10.4 data. Furthermore, whole-genome sequencing of nine centenarian genomes in Sequel II and experimental validation with Southern Blotting revealed consistent coverage dropouts in the Alzheimer's Disease-associated intronic *ABCA7* TR across all genomes. Although coverage dropouts in TRs were rare, our analysis relied on HG002, a highly homozygous genome sequenced at high coverage (38x). Hence, TR coverage drop-outs may be more prevalent in other genomes with more expanded TR sequences and sequenced at lower coverages.

385 Our findings further highlight the complexity of analysing TRs, showcasing the importance of
386 bioinformatic tools that are cross-compatible across different technologies and data-types. For
387 example, all long-read sequencing datasets (Revio, Sequel II, and Duplex) did not support the
388 presence of transposable element (TE)-like sequences in the intronic TR in *ABCA7* in the HG002
389 T2T reference genome, suggesting a misassembly or variation specific to the sequenced cell culture.
390 The ability to compare different sequencing technologies (e.g. PacBio and ONT) and different data-
391 types (e.g. long-read alignments and de novo assemblies) are critical to accurately assess and
392 replicate TR findings in on-going long-read sequencing projects. Although other TR-based tools have
393 been recently developed, they are exclusive to a single long-read sequencing technology and data-
394 type (e.g. TRGT for only HIFI long-read data), expect pre-defined motif-compositions (e.g. TRGT,
395 vamos),[24,28] or do not enable multi-sample calling to facilitate comparisons across multiple genomes
396 (e.g. Stragl, uTR).[25,26] Hence, we consider *otter* and TREAT's cross-compatibility and multi-sample
397 capabilities a major strength.
398
399 Our benchmark of *otter* showed that PacBio's Revio and Sequel II data generated more accurate
400 TR-sequence assemblies than ONT's Duplex and R10.4 chemistries. In particular, we found
401 differences in assembly-accuracy for short and long TR-lengths in ONT and PacBio data. *Otter*
402 currently is unable to separate highly similar allele-sequences that differ by a few nucleotides in the
403 presence of high sequencing error-rates. This led to relatively more erroneous assemblies for short
404 TRs in ONT data relative to PacBio data. However, *otter* generated more accurate assemblies for
405 longer TRs (e.g. 500bp) than shorter TRs in ONT data. This is in contrast to PacBio data, where we
406 observed relatively higher error-rates at larger TRs, often due to discrepancies between the expected
407 parental allele-sequences found in HG002 and read-support from Revio and Sequel II. Importantly,
408 *otter* currently is unable to assemble TR-sequences that are not spanned by individual long-reads,
409 and hence will fail in TRs sequences longer than current typical read-lengths (e.g. >15-20 Kbp). In
410 such cases, ONT's utra-long (100Kbp+) chemistry may facilitate TR-characterisation of highly-
411 expanded TRs. Furthermore, *otter* is prone to reference bias as it relies on pre-mapped long-read
412 data, although it tries to minimise this limitation by performing local-realignment if it detects possible

413 mis-alignments due to high-sequence divergence between a read and the reference. Finally, *otter* is
414 not currently designed as a TR discovery tool, and requires a set of regions to be specifically
415 investigated. As *otter* is integrated in TREAT, these limitations extend to TREAT's unified workflow.
416
417 Despite these limitations, we were able to replicate previously reported TRs associated with
418 Alzheimer's Diseases (AD) by comparing a cohort of AD patients and cognitively healthy
419 centenarians. We acknowledge that these TRs were previously identified using different
420 experimental methods (eg, short-read sequencing, southern blotting), and analyses strategies
421 (logistic regressions, linear mixed models, fisher's exact test).[32–34] While this heterogeneity hampers
422 the direct comparison of the effect size estimates, we highlight that all associations we observed
423 were in the same direction as the original studies. In particular, the TR intronic of ABCA7 was shown
424 to carry an odds ratio for AD of 4.5 when one allele was expanded >5.7 Kbp.[32] Similarly, we observed
425 that individuals carrying larger allele-sequences were significantly associated with AD. However, the
426 effect was stronger in the 5th quantile as cognitively healthy centenarians have a shorter joint-allele
427 size, rather than AD cases having a more expanded TR-sizes in the 95th quantile. While we cannot
428 exclude that we have missed some genotypes due to allele dropouts, the centenarians we used may
429 be enriched with the protective haplotypes in this TR, as previously shown for other AD-associated
430 genetic variants.[35]
431
432 In summary, *otter* and TREAT are flexible and accurate bioinformatics tools that overcome common
433 limitations regarding cross-sequencing platform compatibility, and enable end-to-end analysis and
434 comparisons of tandem repeats in human genomes, with broad applications in research, clinical and
435 diagnostics fields.
436

4. Methods

4.1 Systematic analysis of allele dropouts in tandem repeats

We downloaded the CHM13 reference genome (version 2.0, https://github.com/marbl/CHM13), as well as the following repeat annotations from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/gbdb/hs1/): *SimpleRepeat, RepeatMasker, CenSat*, and *SegDups*. We extracted all *Satellite* and *Simple_repeat* annotations from RepeatMasker, concatenated them with annotations from *SimpleRepeat*, and merged all annotations within <= 100 bp. We excluded annotations that were within 10 Kbp of centromeric sequences (*Censat*) and segmental duplications (*SegDups*), and removed TRs from both sex chromosomes. This curated dataset counts 864,424 TRs genome-wide. We extracted the corresponding parental and maternal allele-sequences in HG002 for these TRs by aligning the HG002 T2T assembly (version 0.7) to CHM13, and traversing the CIGAR string in each contig-alignment to extract the corresponding allele-sequences of each overlapping TR.

We used the publicly available, unaligned data of HG002 genome, generated with ONT (Duplex) as well as PacBio (Revio) technologies.[36] In addition, we generated long-read sequencing data of HG002 with PacBio's Sequel II instrument, using the same wet-lab and data processing protocol previously described (see Supplementary Methods).[37] All data was aligned to GRCh38 (build GRCh38.p14) and the CHM13 reference genomes using minimap2 for ONT data and pbmm2 for PacBio data (see Methods: Genomes included for testing). Then, for each TR and each genome, we calculated the ratio between local TR coverage and average global coverage.

4.2 TREAT

Two main analysis strategies are implemented: the assembly-based analysis, and the read-based analysis. The assembly-based analysis is more efficient, and thus more applicable for the analysis of large datasets of tandem repeats. However, the read-based approach may be preferred when information from all individual reads is needed, for example for performing a multiple sequence

464 alignment, or when studying somatic instability of TRs. In both cases, TR content is characterised
465 with Tandem Repeat Finder (TRF).[29] When multiple motif annotations for the same sequence are
466 found by TRF, a consensus representation of the repeat content is generated. This is done by
467 combining motif annotations to maximise the fraction of sequence annotated by motifs.
468 TR alleles estimation in the assembly-based analysis assumes that different assembled sequences
469 reflect the different alleles. In the reads-based analysis, alleles are estimated from the TR sizes of
470 the reads spanning the TRs using an iterative clustering framework. A minimum number of reads
471 supporting each allele needs to be satisfied (by default, this equals 2). Then, an iterative k-means
472 clustering (*k*=2) framework is used to identify the allele sizes, while controlling for a maximum
473 deviation of the TR sizes within each allele based on the median absolute deviation (by default, 10%
474 of median allele size). This framework recognizes homozygous calls and its iterative nature makes
475 it robust to outliers.
476
477 TREAT's analysis module consists of an outlier-detection framework that scores extreme variations
478 in TR allele sizes across a set of samples. Outliers are detected using a distance-based method that
479 uses Mahalanobis distance to identify observations (ie. TR allele sizes) that deviate from the mean
480 of the data. Briefly, a distance matrix is calculated, and a threshold value *t* corresponding to the
481 number of standard deviations from the mean (by default *t*=3) is chosen such that any data point
482 with a distance greater than the threshold is considered an outlier. A p-value for each outlier is finally
483 calculated by comparing each data point's distance to a chi-squared distribution.
484

485 4.3 *Otter*: a stand-alone, fast, local assembler

486 Because TREAT read-based analysis of large datasets of TR can be computationally costly, we
487 developed *otter*, a generic stand-alone method for generating fast local assemblies of a given region
488 or genotyping whole-genome de novo assemblies. We implemented *otter* in the TREAT assembly-
489 based analysis. Briefly, given a region of interest, *otter* uses the htslib library to identify spanning
490 reads (region of interest is fully contained in the reads) and non-spanning reads (only partially
491 contained) in a given BAM file, and extracts the corresponding subsequence per read based on their

492 alignment (top-left of Figure 1B). If a reference genome is provided, it will perform local read-
493 realignments on non-spanning reads if it detects a clipping-signal, which can indicate suboptimal
494 mappings to due highly divergent sequences (top-right of Figure 1B). This is done by aligning (using
495 WFA2-lib alignment library)[38] the flanking sequences of a region derived from the reference genome
496 onto each read, which are then be used to recalibrate the corresponding subsequence of the region
497 of interest. Recalibrated non-spanning are reclassified as 'spanning' if both flanking sequences are
498 successfully aligned with a minimum length and sequence similarity. In the context of TRs, this
499 realignment procedure often correctly recalibrates the alignments of TRs with major length and/or
500 motif-composition differences relative to a reference genome.
501
502 *Otter* will then attempt to identify a candidate set of unique allele-sequences by attempting to cluster
503 spanning-reads via pairwise-sequence alignment (middle of Figure 1B). To deal with high somatic
504 variation and sequencing errors, *otter* will first estimate local base-line error-rates per region. More
505 specifically, we apply a gaussian-kernel density estimator across a one-dimensional distribution of
506 all spanning pairwise-sequence distances in a region. For single homozygous allele-sequences, this
507 distribution will me unimodal with centered at 0, representing the baseline error-rate of the
508 sequences. When two or more unique allele-sequences are present, the distribution will be multi-
509 modal such, that ε_0, ε_1, ε_2, …, ε_m are global maximas, and ε_1, ε_2, …, ε_m represent sequence errors
510 between reads from different allele-sequences. We thus identify global maximas and minimas
511 through a simple peak-finding procedure by sampling n points (by default n = 100) through this
512 distribution. *otter* uses this information to perform a hierarchical clustering of the reads and stopping
513 when the distance exceeds the densest peak among ε_1, ε_2, …, ε_m, ultimately partitioning reads into
514 initial clusters representing candidate allele-sequences.
515
516 An additional curation step is subsequently applied on the initial clusters to determine if there is
517 sufficient read support (automatically adapted to local coverage, see bottom of Figure 1B). If no
518 maximum number of alleles, α, is enforced, and *otter* will output all clusters, which can be >1 if
519 heterozygous or somatic variation exists. Otherwise, clusters below the adapted coverage threshold

520 are merged to the nearest cluster using a single-linkage approach. If the total number of remaining
521 clusters exceeds α, the hierarchical clustering procedure is continued until α final clusters remain.
522
523 A consensus sequence is then generated for each final cluster. To avoid the high computational and
524 memory costs of a multiple sequence alignment, we use a sparsity-based consensus algorithm
525 inspired from Ye and Ma 2016 to rapidly derive a consensus sequence per cluster.[39] Briefly, we find
526 a representative sequence in each cluster based on the smallest sum of pairwise-distances, and
527 construct a directed-acyclic graph for each representative such that each node corresponds to a
528 position-specific kmer in the sequence. We then align all reads in a cluster to their representative
529 sequence, and update the graph by either incrementing the weight of nodes if their position-specific
530 kmers are identical, or by inserting a new node due to mismatches. The consensus sequence is
531 generated by traversing the graph and retaining the highest-weighted path as previously described.[39]
532
533 Finally, non-spanning reads are assigned to at-most one cluster to help accurately infer allele-
534 frequencies. To minimise runtime, we recycle the existing pairwise alignments and infer cluster
535 membership by identifying the minimum edit-distance between non-spanning reads and reads in a
536 final cluster. A future implementation can make further use of the non-spanning reads in both the
537 assembly and consensus step.
538

539 4.4 Genomes included for testing

540 *HPRC*: Publicly available Pacbio long-read HiFi data of 47 individuals from the Human Pangenome
541 Reference Consortium (HPRC) were downloaded (see Data Accession).[40] For the well characterised
542 HG002 genome,[36] we also downloaded data generated with Oxford Nanopore (ONT, Duplex and
543 R10.4 chemistries) and PacBio Revio technologies.[40] ONT data was aligned to the reference
544 genomes (GRCh38 and CHM13) using minimap2 (2.21-r1071, specifying -x map-ont);[41] PacBio data
545 was aligned using pbmm2 (1.9.0, specifying –preset CCS and –preset SUBREADS respectively for
546 HIFI and non-HIFI data).[18]
547

548 *100-plus Study cohort and Alzheimer Dementia Cohort*: For the replication of TRs previously
549 associated with Alzheimer's Disease (AD), we used HIFI sequencing data from the blood DNA of
550 N=250 patients with AD from the Amsterdam Dementia Cohort (ADC),[42] and N=248 cognitively
551 healthy centenarians from the 100-plus Study cohort.[43] For one cognitively healthy centenarian we
552 sequenced a duo, including next to the blood-derived DNA from the centenarian, the brain-derived
553 DNA from the centenarian and blood-derived DNA from a (72 years old) child of the centenarian.
554 Data analysis and processing was conducted as previously described (see Supplementary
555 Methods).[37]
556
557 *CANVAS patients*: We used the HIFI data of 2 patients diagnosed with CANVAS (Cerebellar ataxia
558 with neuropathy and vestibular areflexia syndrome), caused by a TR expansion in RFC1 gene.[44]
559

560 4.5 Evaluating TREAT and *otter* performances

561 *Benchmark analysis*: we benchmarked the performances of TREAT assembly-based analysis
562 (based on *otter*) by interrogating 864,644 genome-wide TRs in the well-characterised HG002
563 genome from HPRC.[36] We compared *otter* performances with the HG002 telomere-to-telomere
564 assembly.[36] This assembly consists of fully phased maternal (3.06 Gb) and paternal (2.96 Gb)
565 haplotypes from a trio-based experiment. To benchmark *otter*, we extracted the corresponding
566 maternal and paternal allele-sequences for each TR in the composite set (see Supplementary
567 Methods).
568 We tested the flexibility of *otter* using sequencing data from different technologies (PacBio Sequel
569 II, PacBio Revio, ONT Duplex and R10.4), and at different levels of sequencing depth (by
570 subsampling data at 5-fold, 10-fold, 15-fold, 20-fold, 25-fold, and 30-fold). For each sequencing
571 technology and each coverage level, we calculated the assembly error as the normalised edit
572 distance of the allele-sequences assembled by *otter* and the expected parental allele-sequences
573 from the HG002 assembly. Additionally, for each allele-sequence assembled by *otter*, we calculated
574 the standard-errors of all pairwise edit distances between reads used to assemble the sequences.
575

576 *TREAT applications for population ancestry, outlier analysis and case-control analysis:* we compared
577 the performances of TREAT assembly- and read-based analyses by correlating the estimated TR
578 allele sizes with each other. For this analysis, we used the curated PacBio genome-wide catalogue
579 of TRs (version 0.3.0) including 161,037 regions in total, and all 47 genomes from the HPRC.
580 Spearman correlations were calculated for the shorter (in size), longer, and joint alleles (ie, the sum
581 of the shorter and longer alleles).
582 We used TRs for population stratification analyses: using the same set of 161,037 TRs, we selected
583 the top 20% most variable TRs based on the coefficient of variation (ratio of standard deviation to
584 the mean TR joint allele size). Then we applied Principal Component Analysis (PCA) based on the
585 joint allele sizes. For 40/47 matching samples with SNPs data from the 1000Genome project,[45] we
586 also performed PCA based on 30,544 randomly sampled common (minor allele frequency >10%)
587 SNPs.
588 To evaluate clinical applicability, we applied the TREAT analysis module on the combined dataset
589 of 47 HPRC genomes plus the two CANVAS patients. For this analysis, we focused on 35 clinically
590 relevant TRs from literature (Table S1) that were previously established as causative of several
591 neurological diseases (PMID: 20953213, 32060097, 29589097).[7,8,32]
592 Finally, we evaluated a case-control application by replicating the association of 4 TR that were
593 previously associated with Alzheimer's Disease (AD).[32–34] We did this using the long-read
594 sequencing data of N=250 AD genomes and N=248 cognitively healthy centenarians that were
595 sequenced with PacBio Sequel II instrument.[37] We extracted TR allele sequences for all individuals
596 using TREAT, and then compared the distribution of the short, long, and joint-alleles between AD
597 cases and centenarians using logistic regression models.

598 ## Data access

599 Human Pangenome Consortium data is publicly available and can be downloaded from
600 https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0?tab=readme-ov-file.
601 Long-read sequencing data generated with PacBio Sequel II for the 2 CANVAS patients as well as
602 250 AD patients and 250 cognitively healthy centenarians is available upon reasonable request to
603 the corresponding author.
604

605 ## Consent statement

606 The Medical Ethics Committee of the Amsterdam UMC approved all studies. All participants and/or
607 their legal representatives provided written informed consent for participation in clinical and genetic
608 studies.
609

610 ## Funding

611 Part of the work in this manuscript was carried out on the Cartesius supercomputer, which is
612 embedded in the Dutch national e-infrastructure with the support of SURF Cooperative. Computing
613 hours were granted to H. H. by the Dutch Research Council (100plus: project# vuh15226, 15318,
614 17232, and 2020.030; Role of VNTRs in AD; project# 2022.31, Alzheimers Genetics Hub project#
615 2022.38). N.T is appointed at ABOARD and H.H, M.R, S.L are recipients of ABOARD, a public-
616 private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector
617 Life Sciences & Health (PPP-allowance; #LSHM20106).[46,47] More than 30 partners, including de
618 Hersenstichting (Dutch Brain Foundation) participate in ABOARD (https://www.alzheimer-
619 nederland.nl/onderzoek/projecten/aboard). This work is supported by a VIDI grant from the Dutch
620 Scientific Counsel (#NWO 09150172010083) and a public-private partnership with TU Delft and
621 PacBIo, receiving funding from ZonMW and Health~Holland, Topsector Life Sciences & Health
622 (PPP-allowance), and by Alzheimer Nederland WE.03-2018-07. S.L. is recipient of ZonMW funding
623 (#733050512). H.H. was supported by the Hans und Ilse Breuer Stiftung (2020), Dioraphte 16020404

(2014) and the HorstingStuit Foundation (2018). Acquisition of the PacBio Sequel II long read sequencing machine was supported by the ADORE Foundation (2022).

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**, 286–298 (2018).
2. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**, 9272–9277 (2008).
3. Pearson, C. E., Edamura, K. N. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**, 729–742 (2005).
4. Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**, R13 (2003).
5. Linthorst, J. *et al.* Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. *Transl Psychiatry* **10**, 369 (2020).
6. Dumbovic, G., Forcales, S.-V. & Perucho, M. Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics* **12**, 515–526 (2017).
7. McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* **11**, 786–799 (2010).
8. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J Biol Chem* **295**, 4134–4170 (2020).
9. Yu, S. *et al.* Fragile X Genotype Characterized by an Unstable Region of DNA. *Science* **252**, 1179–1181 (1991).
10. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
11. De Roeck, A., Van Broeckhoven, C. & Sleegers, K. The role of ABCA7 in Alzheimer's disease: evidence from genomics, transcriptomics and methylomics. *Acta Neuropathol* **138**, 201–220 (2019).
12. Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
13. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).

14. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B. & Halldórsson, B. V. popSTR: population-scale detection of STR variants. *Bioinformatics* **33**, 4041–4048 (2017).
15. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).
16. Gelfand, Y., Hernandez, Y., Loving, J. & Benson, G. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Res* **42**, 8884–8894 (2014).
17. Eslami Rasekh, M., Hernández, Y., Drinan, S. D., Fuxman Bass, J. I. & Benson, G. Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic Acids Research* **49**, 4308–4324 (2021).
18. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155–1162 (2019).
19. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* **19**, 823–826 (2022).
20. Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays in Biochemistry* **63**, 639–648 (2019).
21. Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol* **22**, 295 (2021).
22. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).
23. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
24. Dolzhenko, E. *et al.* Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* (2024) doi:10.1038/s41587-023-02057-3.
25. 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).
26. Masutani, B., Kawahara, R. & Morishita, S. Decomposing mosaic tandem repeats accurately from long reads. *Bioinformatics* **39**, btad185 (2023).
27. Gustafson, J. A. *et al. Nanopore Sequencing of 1000 Genomes Project Samples to Build a Comprehensive Catalog of Human Genetic Variation*. http://medrxiv.org/lookup/doi/10.1101/2024.03.05.24303792 (2024) doi:10.1101/2024.03.05.24303792.
28. Ren, J., Gu, B. & Chaisson, M. J. P. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol* **24**, 175 (2023).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
30. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

31. Zhang, Y. *et al. MotifScope: A Multi-Sample Motif Discovery and Visualization Tool for Tandem Repeats*. http://biorxiv.org/lookup/doi/10.1101/2024.03.06.583591 (2024) doi:10.1101/2024.03.06.583591.
32. On Behalf of the BELNEU Consortium *et al.* An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathologica* **135**, 827–837 (2018).
33. Guo, M. H., Lee, W.-P., Vardarajan, B., Schellenberg, G. D. & Phillips-Cremins, J. Polygenic burden of short tandem repeat expansions promote risk for Alzheimer's disease. *medRxiv* 2023.11.16.23298623 (2023) doi:10.1101/2023.11.16.23298623.
34. Wang, H. *et al.* Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. *medRxiv* 2023.09.13.23295505 (2023) doi:10.1101/2023.09.13.23295505.
35. Tesi, N. *et al. Cognitively Healthy Centenarians Are Genetically Protected against Alzheimer's Disease Specifically in Immune and Endo-Lysosomal Systems*. http://medrxiv.org/lookup/doi/10.1101/2023.05.16.23290049 (2023) doi:10.1101/2023.05.16.23290049.
36. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
37. Salazar, A. *et al. An AluYb8 Retrotransposon Characterises a Risk Haplotype of TMEM106B Associated in Neurodegeneration*. http://medrxiv.org/lookup/doi/10.1101/2023.07.16.23292721 (2023) doi:10.1101/2023.07.16.23292721.
38. Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2021).
39. Ye, C. & Ma, Z. S. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ* **4**, e2016 (2016).
40. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
42. van der Flier, W. M. & Scheltens, P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. *Journal of Alzheimer's Disease* **62**, 1091–1111 (2018).
43. Holstege, H. *et al.* The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *European Journal of Epidemiology* (2018) doi:10.1007/s10654-018-0451-3.
44. van de Pol, M. *et al.* Detection of the ACAGG Repeat Motif in RFC1 in Two Dutch Ataxia Families. *Mov Disord* **38**, 1555–1556 (2023).
45. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
46. Dreves, M. A. E. *et al.* Rationale and design of the ABOARD project (A Personalized Medicine Approach for Alzheimer's Disease). *Alzheimers Dement (N Y)* **9**, e12401 (2023).

723 47. Van Der Flier, W. M., De Vugt, M. E., Smets, E. M. A., Blom, M. & Teunissen, C. E. Towards a future where Alzheimer's
724 disease pathology is stopped before the onset of dementia. *Nat Aging* **3**, 494–505 (2023).
725