

METHOD

Open Access



Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads

Satomi Mitsuhashi^{1*†}, Martin C. Frith^{2,3,4*†}, Takeshi Mizuguchi¹, Satoko Miyatake¹, Tomoko Toyota⁵, Hiroaki Adachi⁵, Yoko Oma⁶, Yoshihiro Kino⁷, Hiroaki Mitsuhashi⁸ and Naomichi Matsumoto¹

Abstract

Tandemly repeated DNA is highly mutable and causes at least 31 diseases, but it is hard to detect pathogenic repeat expansions genome-wide. Here, we report robust detection of human repeat expansions from careful alignments of long but error-prone (PacBio and nanopore) reads to a reference genome. Our method is robust to systematic sequencing errors, inexact repeats with fuzzy boundaries, and low sequencing coverage. By comparing to healthy controls, we prioritize pathogenic expansions within the top 10 out of 700,000 tandem repeats in whole genome sequencing data. This may help to elucidate the many genetic diseases whose causes remain unknown.

Keywords: Tandem repeat, Repeat diseases, Long-read sequencing, Nanopore, PacBio

Background

A tandem repeat is a region where multiple adjacent copies of sequence reside in the genomic DNA. These regions are highly variable among individuals due to replication error during cell division. They are a source of phenotypic variability in disease and health. More than 30 human diseases are caused by copy number alterations in tandem repeats [1].

The range of pathogenic copy number change relative to the reference varies from a few copies to several thousand, and the length of repeating unit varies from, e.g., three (triplet repeat disease) to several thousand (macro-satellite repeat). As might be expected from such diverse underlying genetic causes, disease mechanisms are also variable. Well-known examples of triplet-repeat expansion diseases in protein-coding regions are polyglutamine diseases (e.g., spinal and bulbar muscular atrophy, Huntington's disease) [2, 3]. Triplet repeat expansion of CAG or CAA codons, which encode

glutamine, leads to toxic protein aggregation and neuronal cell death. Another example of triplet-repeat disease is caused by CUG repeat expansion in the 3'UTR of the transcript from the *DMPK* gene, producing a toxic gain-of-function transcript which sequesters splicing factor proteins and causes aberrant splicing, resulting in multiple symptoms [4]. Not only gain-of-function mutations, but also loss-of-function repeat change in the promoter region due to transcriptional silencing has been reported (e.g., fragile X syndrome) [5]. In addition to short tandem repeat diseases, repeat copy number aberration in human disease is also reported in a macro-satellite repeat (D4Z4). Shortening of the D4Z4 repeat causes aberrant expression of the flanking gene *DUX4*, which has a toxic effect in muscle cells [6]. The thresholds of pathogenic repeat expansion in coding regions are usually less than 100 copies and sometimes even a few copy differences can cause disease (e.g., oculopharyngeal muscular dystrophy) [7]. In contrast, some disease-causing tandem repeat expansions in introns or UTRs can be very long (e.g., *DMPK*) [4]. Moreover, some repeats are interrupted by different sequences (e.g., *DMPK*, *ATXN10*, *SAMD12*) [8–10], making it difficult to analyze the precise repeat structure.

It has been roughly a decade since the introduction of high throughput short read sequencers to clinical genetics.

* Correspondence: satomits@yokohama-cu.ac.jp; mcfirth@edu.ku-tokyo.ac.jp

†Satomi Mitsuhashi and Martin C. Frith contributed equally to this work.

¹Department of Human Genetics, Yokohama City University Graduate School of Medicine, Fukuura 3-9, Kanazawa-ku, Yokohama 236-0004, Japan

²Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26 Aomi, Koto-ku, Tokyo 135-0064, Japan

Full list of author information is available at the end of the article



There have been numerous successful identifications of small nucleotide changes, especially in coding regions, mainly thanks to targeted sequencing (e.g., whole exome sequencing). However, the diagnostic rate remains 30% (depending on the diagnostic platform used) [11], leaving a large population of Mendelian diseases unsolved. The remaining patients may have mutations in “non-coding regions,” or mutations in coding regions which were overlooked due to the limitations of short read sequencing [12]. One candidate is tandem repeat regions, which are difficult to analyze genome-wide by conventional techniques. Identification of disease-causing tandem repeat changes is usually realized by classical genetic technologies (i.e., linkage analysis, Southern blot) and targeted repeat region analysis in a large number of families.

The recent advancement of long-read sequencing technologies may provide a good solution, because long enough reads can encompass whole expanded repeats, and can be analyzed using the flanking unique sequences. Long-read sequencers (PacBio and nanopore) have begun to appear in clinical genetics very recently [13]. As of 2018, these technologies are continually improving in accuracy and throughput. In the clinical laboratory, however, these technologies are still problematic due to both sequencing cost efficiency and the computing burden for large data. It would be preferable, and practical, if low coverage data ($\sim 10\times$) can be used to detect alteration of tandem repeats.

We are aware of two previous methods for determining tandem repeat copy number from long DNA reads: PacmonSTR and RepeatHMM [14, 15]. These methods align the reads to a reference genome, then get the reads that cover a tandem repeat region of the reference, and perform sophisticated probability-based comparisons of these reads to the sequence of the repeating unit. In this study, however, we find that these methods do not always succeed with current long-read data.

We have recently advocated a method (using LAST software) for aligning DNA reads to a genome allowing for rearrangements and duplications [16]. This method has two key features. First: it determines the rates of insertion, deletion, and each kind of substitution in the data and uses these rates to determine the most probable alignments [17]. Second: it finds the most probable division of each read into (one or more) parts together with the most probable alignment of each part. This method found diverse types of rearrangement, the most common of which was tandem multiplication (e.g., hep-tuplication), often of tandem repeat regions [16].

Here, we detect tandem repeat copy number changes by aligning long DNA reads to a reference genome with LAST, and analyzing these alignments in a crude-but-effective way (Fig. 1). Our analysis is based on LAST's division of reads into non-overlapping parts, with a mismatch probability (i.e.,

alignment ambiguity) for each part, and its treatment of “simple” sequence (see the “Methods” section). We point out several practical difficulties with analyzing tandem repeat sequences, which motivate our crude analysis method. Our approach is capable of analyzing tandem repeats genome-wide even with relatively low coverage sequencing data. We believe that this tool will be useful for identifying disease-causing mutations in tandem repeat regions which have been overlooked by short read sequencing in human disease.

Results

Nanopore sequencing of tandem repeat containing plasmids

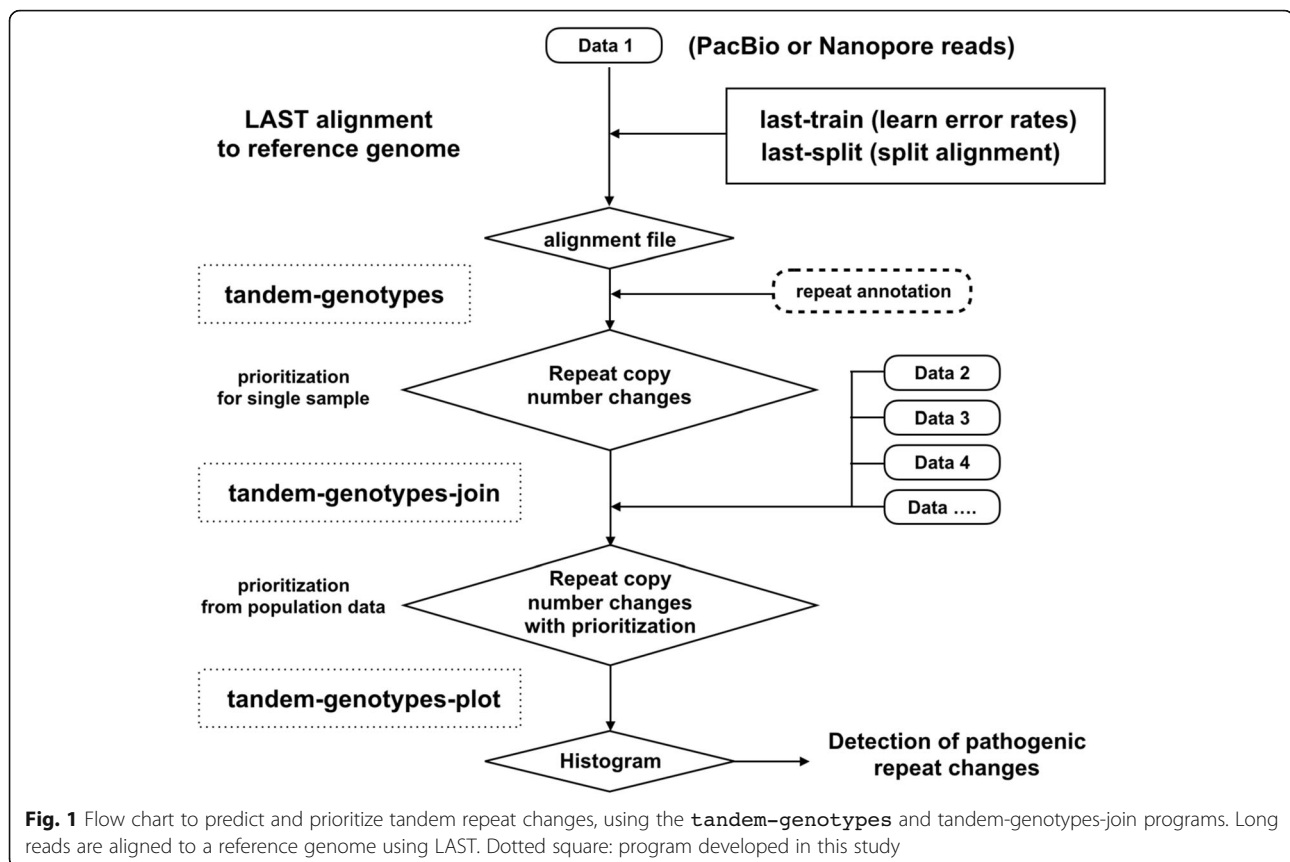
To test our method, we analyzed artificial DNA plasmids containing four different kinds of tandem repeat (CAG, CAA, GGGGCC, and iCCTG) that are known to cause human diseases. We used plasmids with various copy numbers of the repeat unit (e.g., 15 and 109 for CAA; see Fig. 2). These plasmids were subjected to Oxford Nanopore Technologies' (ONT) MinION sequencing. The MinION reads were aligned to plasmid reference sequences with copy number 6 (CAG), 15 (CAA), 3 (GGGGCC), and 15 (CCTG). `tandem-genotypes` predicted the copy number change in each read (Fig. 1): these predictions roughly agree with the actual copy number changes (Fig. 2, Additional file 1: Figure S1). There is a minority of unexpectedly low copy number predictions (Fig. 2d, i; black arrow), especially for the longer plasmids: manual inspection of alignment dotplots (not shown) suggests that these are correct, and the copy numbers in these plasmids may not be completely stable.

In one case, pBS-(CAG)₃₀, `tandem-genotypes` failed with almost no predictions. pBS-(CAG)₃₀ was linearized by cutting it very near to the repeat region (10-bp upstream, see the “Methods” section), so there is only 10 bp of non-repeat sequence upstream of the repeat, which is too short for step 2 of `tandem-genotypes` (see the “Methods” section). Thus, we cut the same plasmid with a different enzyme, DraIII, far from the repeat. As expected, the `tandem-genotypes` prediction agrees with the actual copy number change (Additional file 1: Figure S2a, red arrow).

The GGGGCC repeats have bimodal copy number predictions, where the two modes correspond to reads from each DNA strand (Fig. 2h, i, Additional file 1: Figure S2b). A plausible explanation of this would be if sequencing errors are not independent of the sequence. For example, deletion errors might be more common in GGCCCC repeats than in GGGGCC repeats.

Analyzing chimeric human/plasmid nanopore reads

We performed further tests on semi-artificial data. We obtained human nanopore reads (from “rel3” [18]) that



cover 10 disease-associated repeat regions and replaced the repeat region in each read with the repeat region of a plasmid nanopore read. We used plasmid repeat regions with disease-causing and healthy repeat copy numbers in 1:1 ratio (Table 1). These chimeric reads were aligned to a reference human genome, and copy-number changes were predicted. For each repeat, the predictions have clear bimodal distributions close to the expected copy numbers (Fig. 3, Additional file 1: see the “Results” section, Figure S3).

We also analyzed this semi-artificial data using RepeatHMM and PacmonSTR. RepeatHMM did not clearly detect any of the projected repeat expansions (Additional file 1: Figure S4). PacmonSTR produced no output.

PacBio sequencing datasets of patients with SCA10

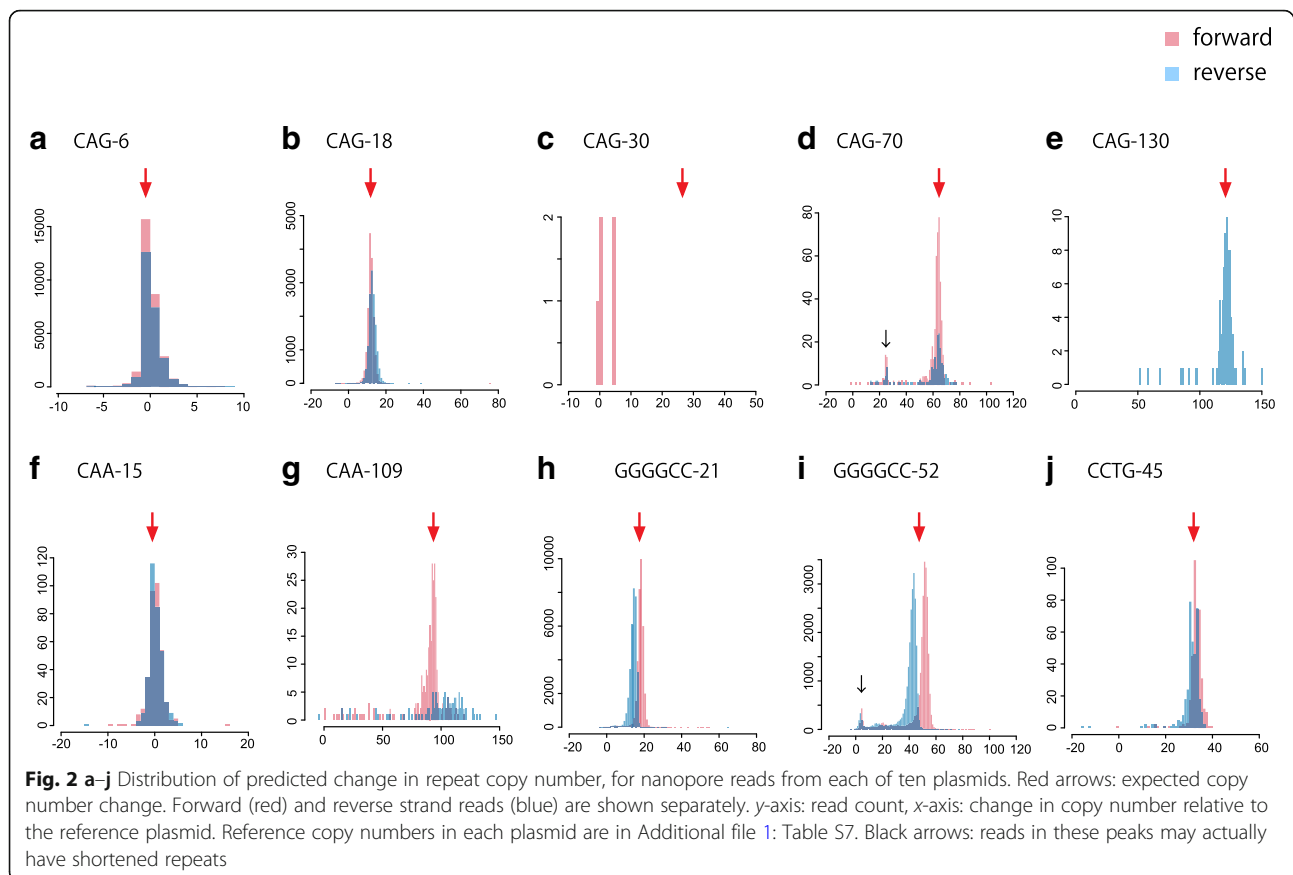
Next, we examined four PacBio sequencing datasets of cloned PCR amplification products from the SCA10 disease locus (spinocerebellar ataxia 10, MIM 603516). SCA10 is caused by ATTCT expansion in the intron of *ATXN10*. These datasets are from three unrelated patients: A, B, and C [9]. Patient C has two datasets, C-1 and C-2, which are different clones sequenced with different PacBio chemistries. According to McFarland et al., subjects A, B, and C have 4.5 kb, 3.9 kb, and 2.7 kb

repeat expansions (since PCR product sizes are 6.5 kb, 5.9 kb, and 4.7 kb and they contain 2 kb flanking sequences), respectively [9]. `tandem-genotypes` predicted datasets A, B, C-1, and C-2 have average expansions of 913, 841, 484, and 486 copies relative to the reference (14 copies), hence repeat lengths 4.6 kb, 4.3 kb, 2.5 kb, and 2.5 kb, respectively (Fig. 4a). Thus, subject B is predicted to have 0.4 kb larger repeat size than the PCR product; however, from McFarland et al. Fig. 1b [9], the purified clone they sequenced by PacBio had > 6 kb insertion, making the actual repeat size > 4 kb, closer to our prediction.

The same datasets were also analyzed with RepeatHMM and PacmonSTR. We first ran RepeatHMM with straightforward parameters: for subject A, it found a similar peak to us but also an unexpected peak around zero, and it did not find the expected peaks for the other three datasets (Fig. 4b). We then consulted the RepeatHMM authors, who suggested non-obvious parameters that improved the C-2 result, but there was still a peak around zero (Fig. 4c). PacmonSTR did produce some output in this case, but none for the SCA10 locus.

Expanded inexact repeats in NA12878

Repeat annotations (i.e., RepeatMasker from UCSC) include non-exact tandem repeats. Non-exact or interrupted



tandem repeats sometimes cause human disease [8]. We detected inexact repeat expansions in the NA12878 human genome, by applying `tandem-genotypes` to PacBio and MinION datasets.

In an intron of *PCDH15*, RepeatMasker annotates a “TATAT” tandem repeat (chr10:54421448-54421530), though the actual sequence is not exactly TATAT (tataaaat aaactttat atatttagca ttgattttt attatgtat attataaat gaatatagt tatattataa ta). `tandem-genotypes` found two peaks for this repeat, indicating a heterozygous ~300 bp insertion (Fig. 5a). PCR amplification of this region from NA12878 DNA showed two different products. One had the same length as the reference sequence (*PCDH15*-intron-repeat-S) (Fig. 5a). Sanger sequencing of the other longer PCR product (*PCDH15*-intron-repeat-L) revealed a ~332 bp insertion. Surprisingly, this insertion was not a tandem expansion, but rather an AluYb8 SINE (according to RepeatMasker).

We also examined an intergenic GT tandem repeat (chr8:48173947-48174212). `tandem-genotypes` found one peak indicating an insertion of ~1000 bp (Fig. 5b). PCR of this region showed a single product, estimated to contain a ~1000 bp insertion. Sanger sequencing revealed that this expansion includes not only GT but also some unknown sequence. The expanded sequence is

present in the chimpanzee genome (Fig. 5b), indicating that this is actually a deletion in the human reference genome (which may have occurred by recombination between GT tandem repeats).

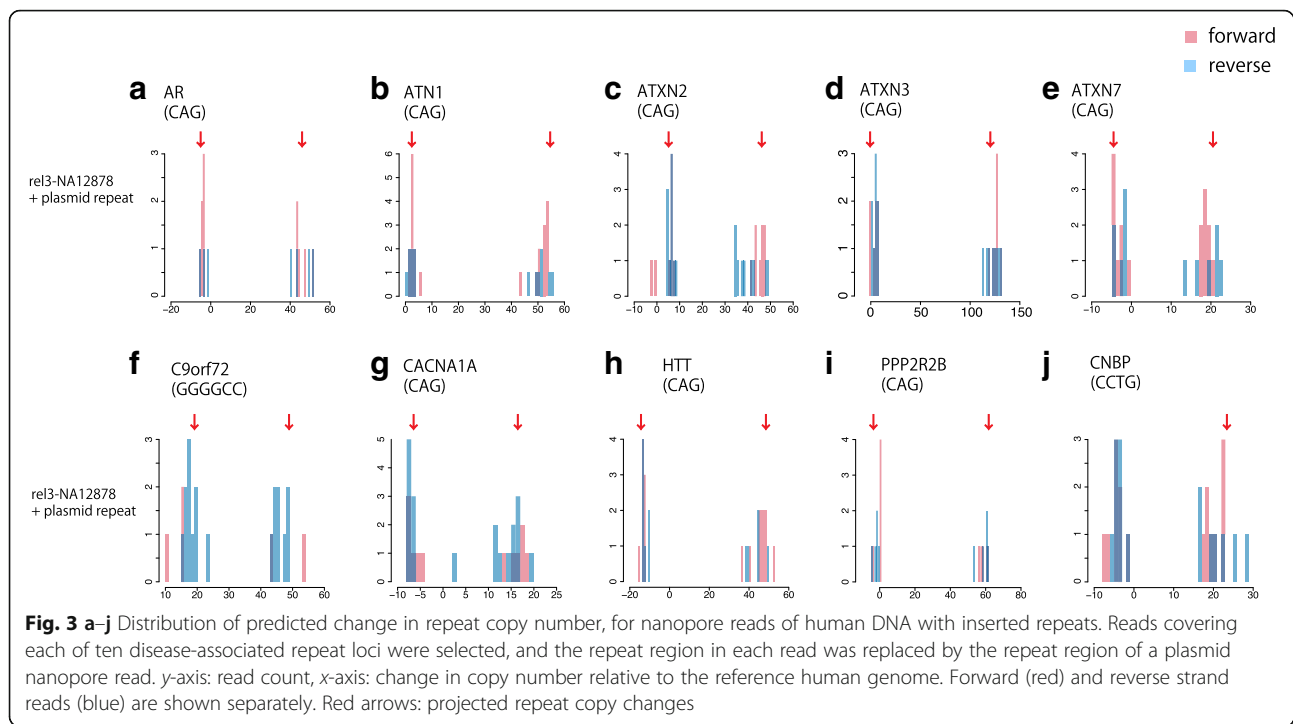
These two examples indicate that `tandem-genotypes` can also find complex and interrupted expansions (or non-deletions) of tandem repeats. We also tested three structural variation (SV) finding methods: NGMLR-sniffles, nanoSV [19, 20], and PBSV (<https://github.com/PacificBiosciences/pbsv>). Only PBSV detected both expansions (Additional file 1: Table S1).

PacBio sequencing of a patient with BAFME

We further analyzed PacBio whole genome sequencing of a patient with a phenotype of benign adult familial myoclonic epilepsy (BAFME). In another large number of BAFME patients in Japan [10], the cause was attributed to large expansions of intronic TTTCA and TTTT A repeats in *SAMD12*. We wished to know whether our patient has such an expansion in *SAMD12*. We sequenced this patient’s genomic DNA using a PacBio Sequel sequencer. `tandem-genotypes` detected ~5 kb insertion in three reads at the BAFME locus, where the coverage is 6× (Fig. 6a). We also applied RepeatHMM, PacmonSTR, NGMLR-sniffles, PBSV, and nanoSV to

Table 1 Chimera reads

chr	Start	End	Repeat	Gene	Region	Repeat number in reference	Inserted repeat 1	Inserted repeat 2	Expected change 1	Expected change 2	Read number 1	Read number 2	Tandem-genotypes detected	Total mapped reads
chr3	63,912,686	63,912,715	CAG	ATXN7	Coding	10	6	30	-4	20	24	24	28	30
chr3	129,172,577	129,172,656	CCTG	CNBP	Intron	21	Raw data	45	NA	24	18	17	31	33
chr4	3,074,877	3,074,933	CAG	HIT	Coding	21	6	70	-15	49	21	20	31	38
chr5	146,878,729	146,878,758	CAG	PPP2R2B	Intron	10	6	70	-4	60	11	10	21	21
chr9	27,573,529	27,573,546	GGGGCC	C9orf72	Intron	3	21	52	18	49	14	14	23	26
chr12	6,936,729	6,936,773	CAG	ATN1	Coding	15	18	70	3	55	18	20	36	38
chr12	111,598,951	111,599,019	CAG	ATXN2	Coding	23	30	70	7	47	20	20	34	36
chr14	92,071,011	92,071,034	CAG	ATXN3	Coding	10	6	130	-4	120	15	15	30	30
chr19	13,207,859	13,207,897	CAG	CACNA1A	Coding	13	6	30	-7	17	18	17	31	34
chrX	67,545,318	67,545,383	CAG	AR	Coding	23	18	70	-5	47	21	20	18	22



this dataset, but they failed to predict any expansions in the BAFME locus (Additional file 1: Table S1). In this study, we used PBSV version 2.0.1 (see Additional file 1). In a parallel study conducted while this paper was in peer-review, surprisingly, an older version (PBSV v0.1.0) found this BAFME expansion as an “insertion” [21]. These structural-variant-finding tools do not indicate which variants are repeat expansions, or of which repeats, and lack our critical prioritization/ranking functionality.

Some difficulties with tandem repeat analysis

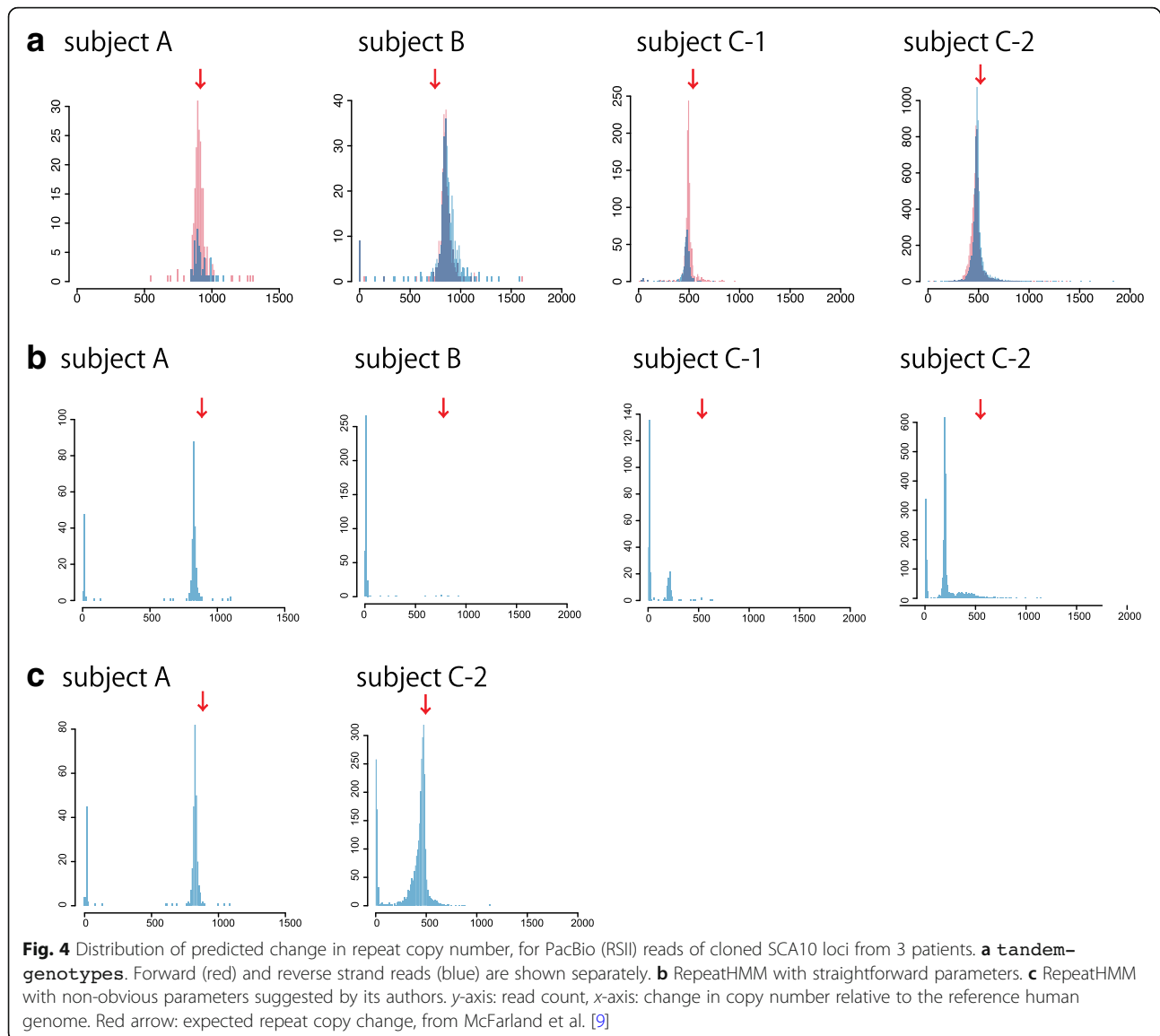
The three expanded BAFME reads do not align with the repetitive region as would be expected for a straightforward repeat expansion (Fig. 6b). Read 2, from the forward genome strand, does not align to the repetitive region at all, because its expanded region consists mostly of TCCC repeats whereas the forward strand of the reference genome has TAAA repeats (Additional file 1: Figure S5). Reads 1 and 6, from the reverse genome strand, align to the repeat at only one side of the expansion. The expanded regions of these two reads start with TTTTA repeats, which match the reverse strand of the reference, but mostly consist of TTTC repeats. Since the expanded region of read 2 does not match the reverse complement of read 1 or 6, we infer that systematic sequencing error has occurred on at least one strand. It is plausible that short-period tandem repeats suffer a nasty kind of sequencing error: if a systematic error occurs for one repeat unit, the same error will tend

to occur for all the other units, producing a different repeat (which may align elsewhere in the genome: the main reason for step 2 in `tandem-genotypes`). Systematic TAAA to TCCC and TTTTA to TTTC conversions occur in some other reads at other TTTTA repeat loci (Additional file 1: Figure S6).

Another kind of difficulty is illustrated by our chimeric human/plasmid reads for *ATXN7* (Fig. 6c). Here, the reference sequence adjacent to the annotated repeat is similar to the sequence within the repeat. Depending on the exact sequences and alignment parameters, the expanded region of a read may align outside the repeat annotation (Fig. 6c top) or appear as alignment gaps some distance beyond the repeat (Fig. 6c bottom). `tandem-genotypes` handles such cases, up to a point, by examining the alignments out to ad hoc distances beyond the annotated repeat.

Specificity of repeat expansion predictions

`tandem-genotypes` can handle custom-made repeat annotation files in BED-like format. We made an annotation file with 31 repeat expansion disease loci, including BAFME, and analyzed these 31 repeats with our BAFME data. No large pathological expansions other than BAFME were predicted (Additional file 1: Figure S7). We also analyzed these 31 repeats with each of the nanopore and PacBio datasets for NA12878: no obvious pathological expansions were predicted and peaks are around zero in most cases Additional file 1: Figure S8, S9). These results suggest that our method does not



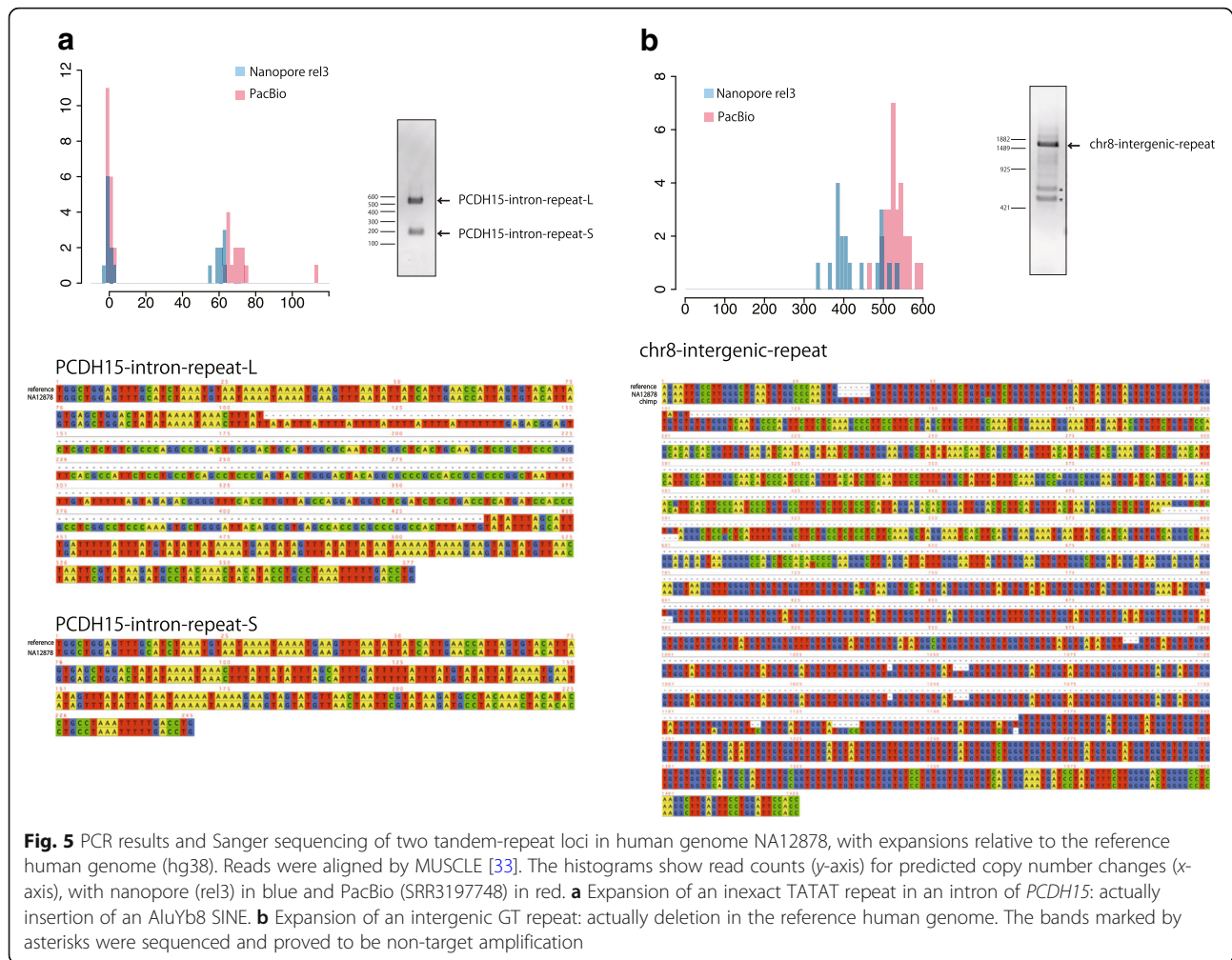
spuriously predict pathological repeat expansions, although there may be some difficulties detecting small disease-causing expansions (e.g., +2 alanine expansion in *PABPN1* causes disease) due to deviations toward copy number increase in PacBio sequences. We believe this will be solved when sequencing quality improves.

Prioritization of copy number changes: needles in a haystack
 Since genome-wide sequencing covers ~1 million highly variable tandem repeats, it is necessary to predict which repeat alterations are likely to be important or pathological. Our prioritization method ranked the BAFME repeat expansion 4th out of 0.7 million tandem repeat regions in *rmsk.txt* (Fig. 7a). When prioritization was done without any control datasets, it was ranked 13th, so using controls greatly improved prioritization (Fig. 7c).

Repeat expansions in protein-coding regions can cause disease with less than 100 extra copies. So these expansions may be difficult to prioritize. To test this, we combined **tandem-genotypes** output for the whole genome (NA12878 rel3) with outputs for the plasmid-rel3 chimeric reads with coding-region expansions (*ATNI*, *HTT*, *ATXN2*, *ATXN3*, *CACNA1A*, *ATXN7*, and *AR*). All 7 chimeric expansions were ranked in the top 10 out of 0.7 million repeat regions (Fig. 7b). Again, controls greatly improved prioritization (Fig. 7c).

Comparing tandem repeat copy number change distribution of MinION, PromethION, and PacBio RSII sequencing data

We next examined the genome-wide repeat copy number changes in the NA12878 human genome sequenced



by both PacBio RSII (SRR3197748) and Oxford Nanopore Technology's MinION (rel3). There was marked discordance between MinION and PacBio when the repeat unit size was one or two (Fig. 8a, Additional file 1: Figure S10). This is probably because MinION tends to make small deletion errors and PacBio small insertion errors, which are hard to distinguish from copy number changes of these tiny repeat units. Note that a repeat unit size of one means homopolymers (such as AAAA AAAAAA). The triplet-repeat distribution showed a slight difference between MinION and PacBio (Fig. 8a). However, where the repeat unit is longer than 3, MinION and PacBio had similar distributions of copy number changes (Additional file 1: Figure S10), so both sequencing platforms work on these tandem repeats. In locus-by-locus comparisons, the repeat copy number changes in nanopore and PacBio reads roughly agree (Additional file 1: Figures S8 versus S9, Figure S11). Locus-by-locus scatter plots show a clear correlation for repeat unit length > 1 (Additional file 1: Figure S12). We also verified that the GGGGCC strand bias, which we

observed in the plasmids, is also seen in the rel3 dataset. The distribution of GGGGCC copy number change showed a slight difference between forward and reverse strands (Additional file 1: Figure S2b).

We examined the distribution of copy number change for all kinds of di- and tri-nucleotide repeat. There are four possible types of di-nucleotide repeat (AG, AT, CG, and AC) and ten kinds of triplet repeat (TAA, GTC, AAC, GAT, CTT, CTG, GTA, CGG, CCT, and CAC). MinION sequences (rel3) showed obvious systematic strand biases for AC, AG, CCT, and CTT (Fig. 8b, d, Additional file 1: Figures S13–S15). Note that we do not expect any strand bias for CG and AT because they are palindromic.

The MinION data we tested (rel3) was published in 2017. Nanopore basecallers and chemistries have been improved recently. We tested a recent nanopore MinION dataset analyzed by MinKNOW1.11.5, which uses the Albacore 2.0 basecaller. The strand bias for AC, AG, CCT, and CTT was greatly improved (Additional file 1: Figure S16). We also tested a recently published human

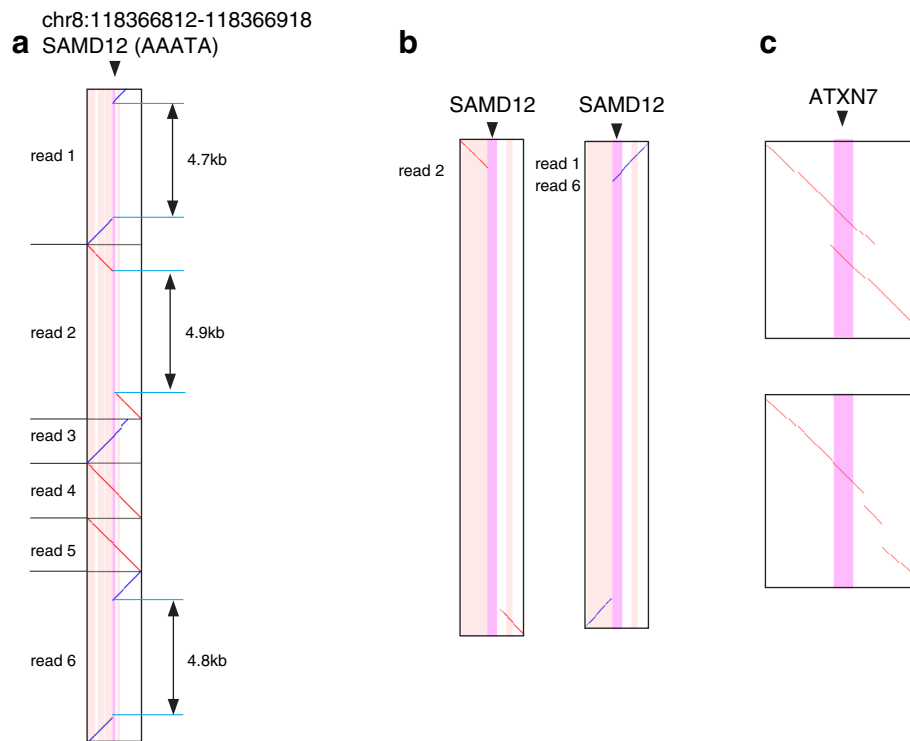


Fig. 6 Alignments of DNA reads (vertical) to the reference human genome (horizontal). Diagonal lines indicate alignments, of the same strands (red) and opposite strands (blue). The vertical stripes indicate repeat annotations in the reference genome: tandem repeats (purple) and transposable elements (pink). **a** Six reads from a BAFME patient that cover the disease-causing *SAMD12* AAAAT repeat locus. **b** Close-ups of three reads with ~ 5 kb expansions. **c** Two examples of chimeric human reads (rel3) with expanded CAG repeats at the *ATXN7* disease locus

genome dataset sequenced by ONT's new high throughput sequencer PromethION [22]. We also found that strand biases for AC, AG, CTT, and CCT are greatly improved in PromethION reads (Fig. 8b, d, Additional file 1: Figures S17, S18).

Computation time and repeat masking

The slowest computational step was aligning the reads to the genome (*lastal*). For some datasets, we made it much faster by “masking” repeats (both interspersed and tandem) with *WindowMasker*. Here, masking means that the repeats (indicated by lowercase letters) are excluded from the similarity-search steps of the alignment algorithm, but are included when making the final alignments: the hope is to find the same alignments faster.

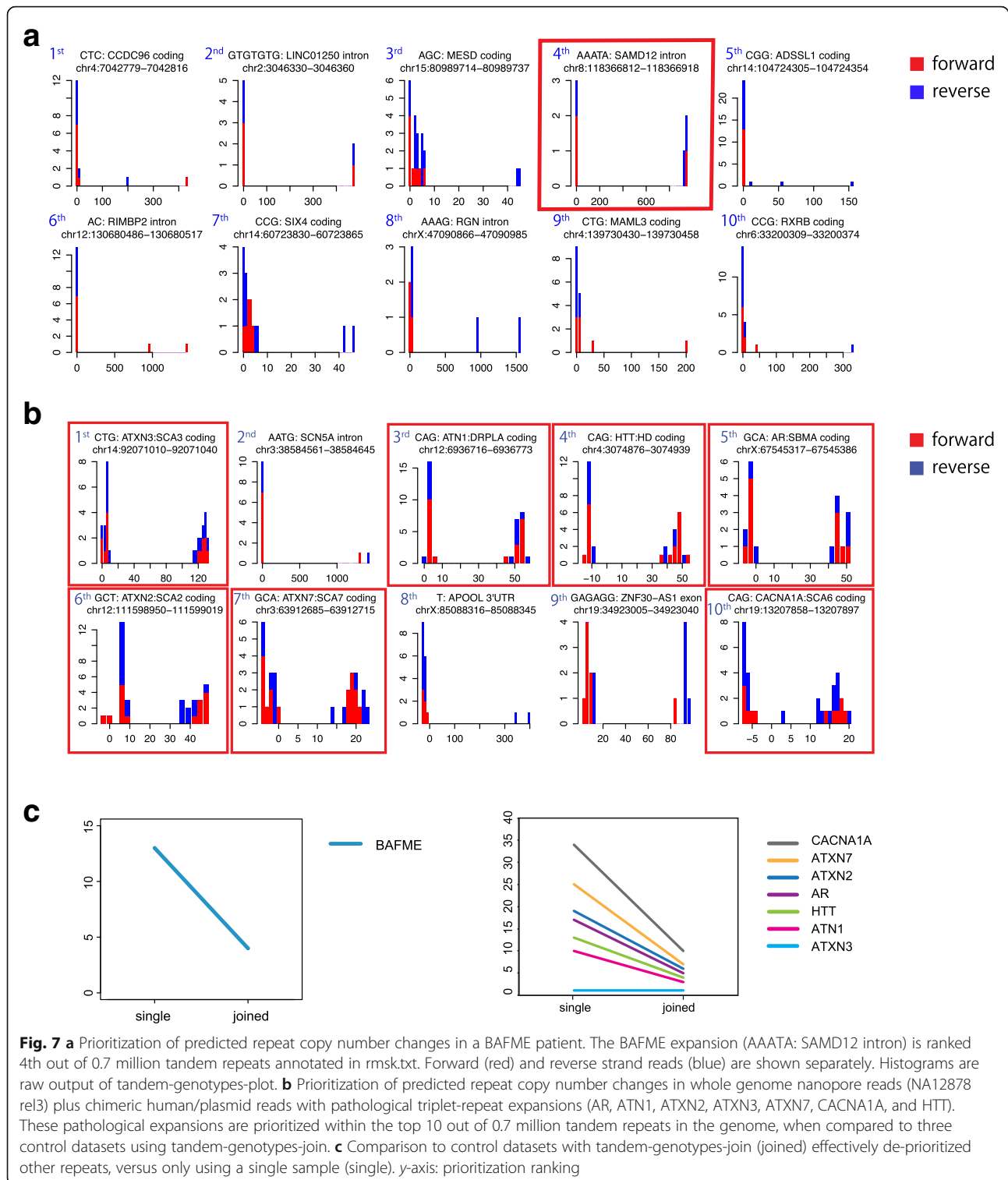
In practice, this masking is often harmless, but sometimes harmful. It did not prevent us from detecting the BAFME expansion, or expansions at ten other disease loci in chimeric human/plasmid reads. On the other hand, it prevented detection of the *SCA10* expansions (result not shown). This is because one flank of the *SCA10* tandem repeat consists of transposable elements and is almost completely masked. Note this dataset has somewhat short reads (Table 2): the problem would be

solved by longer reads that extend beyond the masked region.

When we do not mask, the total run time for our analysis is competitive with those of RepeatHMM and PacmonSTR (Additional file 1: Table S2). (Note the *last-train* run time does not increase much for larger datasets, because it uses a fixed-size sample of the data.) When we do mask, the computation is much faster (Additional file 1: Table S3), and usually, the results do not change significantly (Additional file 1: Figure S19). The computing time for whole human genome PacBio and nanopore datasets (with masking) is practical, compared to other aligners such as NGMLR (Additional file 1: Table S4) (i.e., 37 h for NGMLR and 13 h for LAST with the BAFME data).

Discussion

We have presented several lines of evidence that we can robustly detect pathological expansions of tandem repeats. We successfully detected them in constructed plasmids, semi-artificial plasmid/human sequences, and real human sequences from PCR-amplified *SCA10* locus, PacBio No-Amp sequence of Huntington's disease locus [23] (Additional file 1: Figure S20), and PacBio whole genome data from a BAFME patient. We also did *not* detect



unexpected (false-positive) large known-pathological expansions in three whole genome datasets: PacBio reads from a BAFME patient, and PacBio and nanopore reads from NA12878. Importantly, we can also rank copy number changes by priority, such that pathological expansions

are ranked near the top out of ~ 0.7 million tandem repeats in the genome.

Our method is not specific to tandem expansions, but detects any kind of expansion of a tandem repeat. For example, we detected an expansion due to the insertion

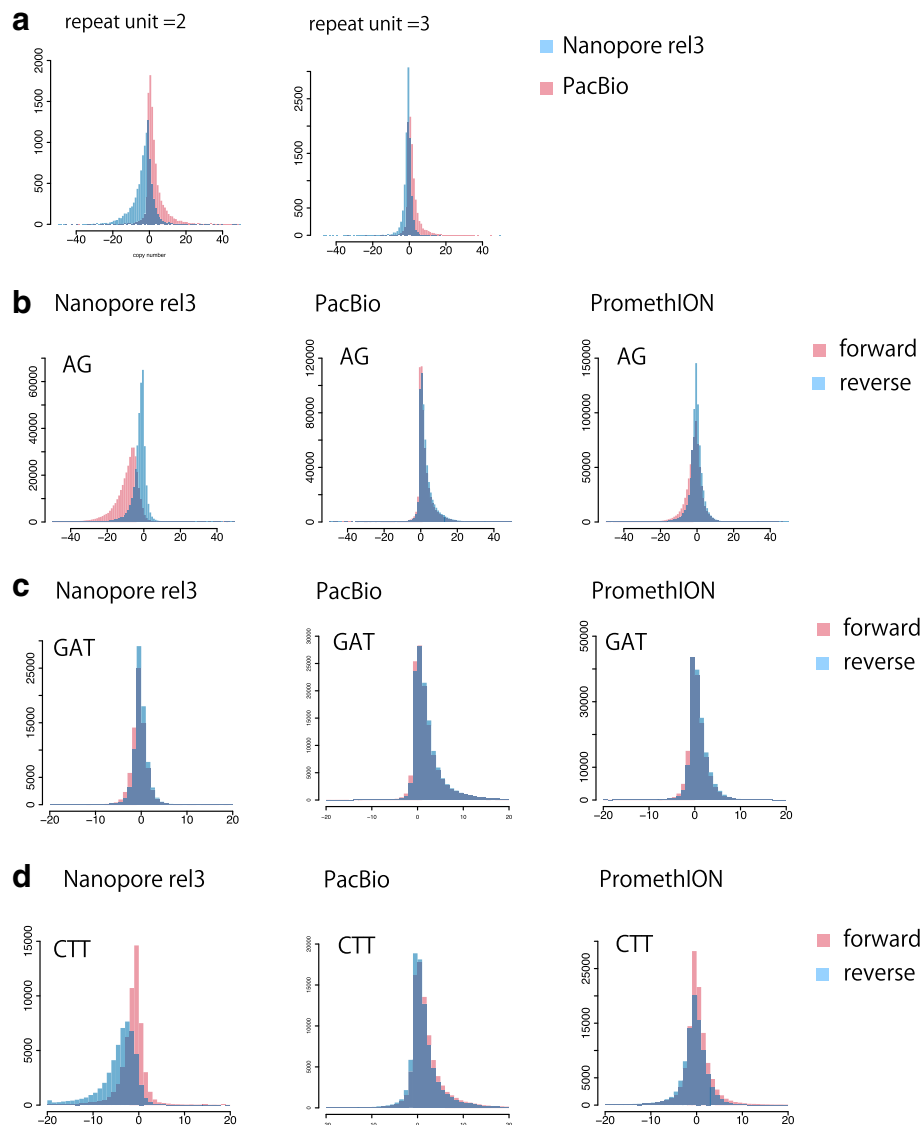


Fig. 8 a Genome-wide distribution of predicted change in repeat number, for nanopore MinION (rel3) and PacBio (SRR3197748) reads from the same human (NA12878). Nanopore tends to have negative and PacBio positive predicted changes, especially for short repeat units. Read number = 10,000 (randomly sampled). Nanopore reads are shown in blue and PacBio in red. **b–d** Genome-wide distribution of predicted change in repeat copy number for nanopore MinION (rel3) and PacBio (SRR3197748) reads from the same human (NA12878), and nanopore PromethION (ERR2585112-5) from a different individual (NA19240). **b** Distributions for AG di-nucleotide repeats. **c** Distributions for GAT. **d** Distributions for CTT. CTT shows the most prominent strand bias in nanopore rel3 reads among all types of triplet repeat (all types are in Additional file 1: Figures S9, S10, S13–S16). PromethION shows less strand bias compared to rel3. y-axis: read count, x-axis: change in copy number relative to the reference human genome

of an Alu SINE within a tandem repeat (Fig. 5a). We also detected an expansion that contained a non-repeat sequence, which turned out to be a deletion in the reference genome (Fig. 5b). Such non-tandem expansions may impact genomic function and health, so we believe it is useful to detect them too during first-round genome-wide screening.

If a repeat expansion is actually the ancestral state, with the reference genome having a contraction or deletion (e.g., Fig. 5b), then it is plausible that the expansion

is less likely to be pathological. Thus, our prioritization of copy number changes likely benefits from comparing the changes to ape genomes (see the “Methods” section). An ancestral reference human genome would be ideal [16]. A similar idea is to de-prioritize expansions commonly present in healthy humans (Fig. 7a, b): this will become more powerful as tandem repeat data accumulates.

We have also pointed out some interesting difficulties with analyzing tandem repeat sequences. Some DNA reads do not align with the repetitive region of the

Table 2 Dataset information

Name	Type	Sequencer	Individual ID	Accession number	Bases	Reads	Average length	Median
rel3	Human genome	MinION	NA12878	*	91,240,120,433	14,183,584	6432.8	4617.0
SRR3197748	Human genome	RSII	NA12878	SRR3197748	176,931,641,323	40,820,704	4334.4	3053.0
ERR2585112-5	Human genome	PromethION	NA19240	ERR2585112-5	185,056,318,464	13,134,890	14,088.9	12,551.0
BAFME	Human genome	Sequel	Private	Private	44,835,599,221	6,174,384	7261.6	6324.0
CAA-15	Plasmid	MinION	–	DRR140497	3,312,623	726	4562.8	4689.5
CAA-109	Plasmid	MinION	–	DRR140498	2,861,064	597	4792.4	4962.0
GGGGCC-52	Plasmid	MinION	–	DRR140499	228,298,911	50,884	4486.7	4439.0
GGGGCC-21	Plasmid	MinION	–	DRR140500	305,112,100	71,016	4296.4	4600.0
CAG-6	Plasmid	MinION	–	DRR140501	170,853,052	59,808	2856.7	2935.0
CAG-18	Plasmid	MinION	–	DRR140502	177,739,795	31,644	5616.9	5887.0
CAG-30	Plasmid	MinION	–	DRR140503	420,957,360	142,000	2964.5	2999.0
CAG-30-Drall	Plasmid	MinION	–	DRR140504	104,545,063	33,551	3116.0	3026.0
CAG-70	Plasmid	MinION	–	DRR140505	64,609,201	18,232	3543.7	3091.5
CAG-130	Plasmid	MinION	–	DRR140506	10,215,540	2135	4784.8	5002.0
CCTG-45	Plasmid	MinION	–	DRR140507	3,675,433	875	4200.5	4605.0
SCA10-subjectA	Plasmid	RSII	–	SRR2080459	92,385,186	81,623	1131.9	642.0
SCA10-subjectB	Plasmid	RSII	–	SRR2081063	303,000,762	81,740	3706.9	2605.0
SCA10-subjectC-	Plasmid	RSII	–	SRR2082412	157,578,422	81,741	1927.8	893.0
SCA10-subjectC-	Plasmid	RSII	–	SRR2082428	611,539,771	163,476	3740.9	1834.0
AR-chimera	rel3-plasmid	MinION	–	–	869,853	41	21,215.9	13,224.0
HTT-chimera	rel3-plasmid	MinION	–	–	518,235	40	12,955.9	7462.5
CACNA1A-chimera	rel3-plasmid	MinION	–	–	387,962	35	11,084.6	9411.0
C9orf72-chimera	rel3-plasmid	MinION	–	–	330,265	26	12,702.5	9054.5
ATXN2-chimera	rel3-plasmid	MinION	–	–	627,855	40	15,696.4	8155.0
ATXN3-chimera	rel3-plasmid	MinION	–	–	465,481	30	15,516.0	14,225.0
ATXN7-chimera	rel3-plasmid	MinION	–	–	795,136	48	16,565.3	9921.5
ATN1-chimera	rel3-plasmid	MinION	–	–	374,363	38	9851.7	8863.0
PPP2R2B-chimera	rel3-plasmid	MinION	–	–	334,450	21	15,926.2	8899.0
CNBP-chimera	rel3-plasmid	MinION	–	–	361,861	34	10,643.0	9349.0

*Downloaded from <https://github.com/nanopore-wgs-consortium/NA12878>

reference genome (e.g., Fig. 6b), and systematic sequencing errors may turn a tandem repeat into a different tandem repeat. The analysis becomes harder when the reference sequence next to an annotated tandem repeat resembles the sequence in the repeat (e.g., Fig. 6c). Some (inexact) tandem repeats do not have unambiguous boundaries, and different annotations (e.g., RepeatMasker versus Tandem Repeats Finder [24]) sometimes disagree on the boundaries. In some cases, there may be no unambiguous distinction between the expansion of a tandem repeat and sequence insertion near the repeat. A user may consider adjusting the default treatment of this (see Additional file 1).

Systematic sequencing errors can have different effects on the two strands of tandemly repeated DNA, causing the predicted copy number changes to have a bimodal

distribution (e.g., Fig. 2i). So it is important to indicate which predictions come from which strands, in order to not misinterpret this as two alleles. We report length and strand biases of several long-read sequencers for every possible type of di- and tri-nucleotide repeat: these biases are prominent for specific repeats (e.g., CTT and CCT in older MinION data), and the worst biases are greatly improved in more recent sequencing systems.

If sequencing accuracy continues to improve, tandem repeat analysis will obviously benefit. The alignment will automatically become faster, due to a lower tolerance of gaps and substitutions. Copy number will be predictable more accurately and with lower coverage.

While this manuscript was in peer-review, tandem-genotypes identified a novel repeat expansion from both nanopore and PacBio whole genome sequencing of

multiple families with neuronal intranuclear inclusion disease [25]. Thus, our method can find the unknown cause of a genetic disease.

Conclusion

This study demonstrates a practical and robust way to identify changes in tandem repeats that may have biologically impactful consequences. Although there are still limitations due to the developing sequencing technologies and cost to immediately apply this approach in clinical sequencing, we clearly show that there is hope that long-read sequencing is useful to identify overlooked changes in the genome and may give an answer to the large numbers of patients with genetic diseases whose causes and mechanisms have remained unsolved for many years.

Methods

tandem-genotypes method

`tandem-genotypes` has two required inputs: annotations of tandem repeats in a reference genome and alignments of DNA reads to the same genome.

The annotations supply a start and end coordinate for each repeat, and the length u of its repeating unit. The repeat length need not be an integer multiple of the unit length. We define two ad hoc distances: “far” $f = \max[100, u]$ and “near” $n = \max[60, u]$. (Actually, we truncate f at the edge of the sequence: where we speak of f , we really mean $\min[f, \text{distance to the edge of the reference sequence}]$.)

`Last-split` finds a division of each DNA read into (one or more) parts and an alignment of each part to the genome. It gives each alignment a “mismatch probability,” which is high if that part of the read aligns almost equally well to other loci [16]. We regard one read’s alignments as ordered by their 5’ to 3’ positions in the read.

For each DNA read, `tandem-genotypes` performs these steps:

1. Discard alignments with mismatch probability $> 10^{-6}$.
2. Discard alignments of mostly lowercase sequence. This removes alignments that consist almost entirely of simple sequence (such as `atatatata`), which are less reliable. Simple sequence is detected and lowercased by `lastdb` and `lastal`, using `tantan` [26]. An alignment is discarded if it lacks any segment with score \geq `lastal`’s score threshold, when “gentle masking” is applied [27].
3. Join consecutive alignments that are colinear on the same strand of the same chromosome and separated by $\leq 10^6$ bp.
4. Find all alignments that overlap a given repeat.

If there is one such alignment:

- Require that it extends beyond both sides of the repeat by at least f , else give up (i.e., do not use this DNA read for this repeat).
- Extract all “gaps” from the alignment. Here, one “gap” may have d unaligned reference bases and i unaligned query bases, flanked by aligned bases. For each gap:
 - If it does not overlap the repeat and $i \leq u/2$: ignore it.
 - If it is wholly $\geq f$ away from the repeat: ignore it.
 - If it is partly $\geq f$ away from the repeat: give up.
 - If it is wholly $> n$ away from the repeat: ignore it.
 - Define r to be the number of unaligned reference bases in the gap that are in the repeat.
 - Define the gap’s net deletion size: $D = \min(d-i, r)$.
 - Assume this gap contributes an integer (or zero) copy number change. Find this contribution by rounding D to the nearest multiple of u (breaking ties by rounding toward zero).

If there is more than one such alignment:

- Require that they are consecutive in the read and on the same strand.
- Define the “left” alignment to be the first one if the strand is “+,” else the last one. Define the “right” alignment in the opposite way.
- Require that the left alignment extends leftwards of the repeat and leftwards of the other alignments, by at least f . Require that the other alignments do not extend leftwards of the repeat by f or more.
- Likewise for the right alignment.
- Define the insertion size as the number of query bases, minus the number of reference bases (which could be negative), between the end of the left alignment and the start of the right alignment.
- Find the nearest multiple of u to this insertion size (as above).

Prioritization of copy number changes

The repeats are ranked by priority score. Each repeat has multiple predictions of copy number change, one per DNA read. If the average number of predictions (for repeats with at least 1 prediction) is ≥ 3 , ignore the most extreme expansion and contraction per repeat. (Our long-read datasets all have coverage > 3 . This coverage criterion is relevant for using an assembled genome instead of long reads, e.g., chimp versus human.) For each

repeat, take the most extreme remaining change and calculate:

$$(\text{Length increase in bases})/(\text{reference repeat length} + 30).$$

(The +30 provides robustness against dubious repeat annotations, e.g., UCSC RepeatMasker includes triplet repeats of length 6.) This score is multiplied by an ad hoc value per gene annotation, currently 50 for coding, 20 for UTR, 15 for the promoter, 15 for an exon of non-protein-coding RNA, and 5 for intron. This is multiplied by 2 for coding annotations where the repeating unit codes polyglutamine or polyalanine in any reading frame (out of 6). The repeats are ranked by an absolute value of this priority score. The results are robust to small changes of the gene annotation scores (Additional file 1: Tables S5, S6).

Multi-dataset prioritization

Suppose we find tandem repeat changes in several individuals with a disease and several individuals without this disease. We wish to prioritize disease-associated changes. For each repeat, calculate d = the cubic mean ($\sqrt[3]{\text{avg.}[x^3]}$) of the diseased individuals' priority scores, and h = the cubic mean of the non-diseased individuals' priority scores. (We use cubic mean in order to emphasize large values: for example, if just 1 of 3 non-diseased individuals has a large expansion, we still wish to strongly de-prioritize this expansion.) If $d < 0$, negate d and h . Finally, the joint priority score is $\max(d - \max(h, 0), 0)$.

Public and patient human genome data and alignment to reference genome

Human whole genome nanopore (rel3) and PacBio (SRR3197748) sequence data from the same individual (NA12878) were downloaded from (<https://github.com/nanopore-wgs-consortium/NA12878>) and from the SRA database, respectively [18]. Another 60× coverage human whole genome nanopore dataset from a different individual (NA19240) using PromethION was downloaded from <https://www.ebi.ac.uk/ena/data/view/PRJEB26791>.

Genomic DNA from a human patient with BAFME phenotype was sequenced by PacBio Sequel according to the manufacturer's protocol. Briefly, genomic DNA was sheared by g-TUBE (Covaris, MA, USA), then size selection was done by BluePipin (Sage science, MA, USA) according to the standard method. These datasets were aligned to the human genome (GRCh38) with LAST version 936:

```
WindowMasker -mk_counts -in genome.fa >
genome.wmstat
```

```
WindowMasker -ustat genome.wmstat -outfmt fasta
-in genome.fa > genome-wm.fa
lastdb -P8 -uNEAR -R11 -c GRCh38 genome-wm.fa
last-train -P8 GRCh38 reads.fasta > train.out
lastal -P8 -p train.out GRCh38 reads.fasta | last-
split > alns.maf
```

Generating tandem repeat containing plasmids

Plasmids containing various numbers of CAG, GGGG CC, and CAA used for this study were generated as described elsewhere [28–30] and are available upon request (Additional file 1: Table S7). Sequence data of the plasmids were deposited in (DRA007012, Table 2).

A plasmid with CCTG repeats was generated as follows. Briefly, exon 1 with flanking 225 bp of intron 1 and exon 2 with flanking 1051 bp of intron 1 (including interrupted CCTG₁₂ repeats) of the human *CNBP* gene were amplified by PCR using human genome DNA (cat.# G304A, Promega, WI, USA) and then cloned into pBluescriptII-KS(-) using In-Fusion cloning kit (Clontech Takara, Shiga, Japan). The resulting construct was digested with Sall and XhoI and then ligated with T4 DNA ligase to delete the Sall site in the multi-cloning site of the pBluescriptII-KS(-). A new Sall enzyme site was generated by site-directed mutagenesis after 13 bp of the interrupted CCTG₁₂ repeats, to obtain a pBS-CNBP-Sall vector. Oligo DNAs containing CCTG₁₅ repeats and flanking sequence, 5'-TCGA (CCTG)₁₅C-3' and 5'-TCGAC (CAGG)₁₅-3', were phosphorylated by T4 polynucleotide kinase, annealed, and then ligated using T4 DNA ligase (Takara, Shiga, Japan). The resulting ligated oligo DNA was digested with Sall and XhoI to remove undesired DNA fragments. Three tandemly ligated oligos were cloned into the Sall site of the pBS-CNBP-Sall vector. This plasmid has CCTG₄₅ repeats interrupted by CTCGA in every 15 CCTG repeats, named interrupted CCTG₄₅ (iCCTG₄₅).

Nanopore sequencing and alignment of tandem repeat containing plasmids

These repeat containing plasmids were cut (linearized) with restriction enzymes NheI, EcoRI-HF, BamHI-HF, or DraIII (NEB, MA, USA) (Additional file 1: Table S7) and then treated with Klenow Fragment DNA Polymerase (Takara, Shiga, Japan) at 37 °C for 30 min. The whole DNA fragments were purified using AmPureXT beads (Agilent Technologies, CA, USA), then subjected to nanopore sequencing. Library preparation was performed using a 1D native barcoding genomic DNA kit (EXP-NBD103 and SQK-LSK108) and then subjected to MinION (Oxford Nanopore Technologies) sequencing using one FLA-MIN106 (R9.4.1) flow cell according to the manufacturer's protocol. Basecalling and fastq conversion were performed with MinKNOW ver1.11.5.

De-barcoding was done using EPIME software (Oxford Nanopore Technologies).

Obtained fastq files were transformed to fasta files using `seqkit fq2fa` option (<http://bioinf.shenwei.me/seqkit>). fasta files were aligned to plasmid references like this:

```
lastdb -P8 -uNEAR -R01 plasmid-ref plasmid.fasta
last-train -P8 plasmid-ref reads.fasta > train.out
lastal -P8 -p train.out plasmid-ref reads.fasta |
last-split > alns.maf
```

SCA10 data

SCA10 sequences from three patients with spinocerebellar ataxia 10 (MIM 603516) [9] were downloaded from SRA (subjectA SRR2080459, subjectB SRR2081063, subjectC-1 SRR2082412, subjectC-2 SRR2082428); then fasta files were generated by this command:

```
fastq-dump --fasta --table SEQUENCE --split-spot
--skip-technical -I --gzip
```

Reads were aligned to the human genome (GRCh38) like this:

```
lastdb -P8 -uNEAR -R01 GRCh38 reference.fasta
last-train -P8 GRCh38 reads.fasta > train.out
lastal -P8 -p train.out GRCh38 reads.fasta | last-
split > alns.maf
```

Comparison to RepeatHMM and PacmonSTR

The SCA10, BAFME, and chimeric reads were analyzed as follows:

PacmonSTR:

```
blasr fasta reference.fa -m 5 --out blasr.out.m5
makeBinnedAnchors.py blasr.out.m5
simpleRepeat.txt 100
pacMonStr_v1.py blasr.out.m5 binned_anchors 10
8 current_directory_path
```

RepeatHMM:

```
repeatHMM.py FASTQinput --fastq fastq
--Patternfile hg38.predefined.pa --repeatName gene
--hgfile reference.fa
repeatHMM.py BAMinput --Onebamfile bam
--Patternfile hg38.predefined.pa --repeatName gene
--hgfile reference.fa
```

For SCA10 subjectA and subjectC-2, after consulting the RepeatHMM authors, we re-ran it with these options added: “-CompRep ALTIT50/C50ICIT/C --SplitAndReAlign 1.”

Running tandem-genotypes

The copy number changes of tandem repeats were determined by `tandem-genotypes v1.1.0` with some different options.

To check all tandem repeats in the human genome in NA12878 data (rel3 and SRR3197748);

```
tandem-genotypes -u 1 -g refFlat.txt rmsk.txt
alns.maf
```

To check disease-related tandem repeats in chimeric reads, BAFME and SCA10 data;

```
tandem-genotypes hg38-disease-tr.txt alns.maf
```

To check the plasmid sequences;

```
tandem-genotypes plasmid-repeat.bed alns.maf
```

Tandem repeat (rmsk.txt) and gene (refFlat.txt) annotations were obtained from the UCSC genome database (<http://genome.ucsc.edu/>) [31]. We made the file hg38-disease-tr.txt, with 31 disease-associated tandem repeats, based on Tang et al. [1].

Tandem repeat changes in a chimpanzee, relative to the reference human genome, were found like this:

```
tandem-genotypes -g refFlat.txt rmsk.txt hg38-
panTro5-1.maf
```

These alignments (from <https://github.com/mcfrith/last-genome-alignments>) [16] are of an assembled chimp genome, *not* long reads: our methods work in this case too.

Multi-dataset prioritization was done with commands of this form:

```
tandem-genotypes-join data 1: data 2
Data 1: tandem-genotypes output to be prioritized
Data 2: tandem-genotypes output to be de-prioritized
```

For the BAFME patient, two humans and one chimpanzee (NA12878 PacBio SRR3197748, NA19240 PromethION ERR2585112-5, and panTro5) were used for de-prioritizing possibly benign expansions. For rel3 data with coding-region expansions, NA19240 (PromethION ERR2585112-5), the BAFME patient (PacBio), and panTro5 were used for de-prioritization.

Chimeric reads of plasmid-derived repeats and human-derived flanks

Human nanopore reads covering repeat expansion disease loci were extracted from whole genome sequence data (rel3). The nanopore sequences flanking the repeat

were excised using an in-house script. Randomly selected expanded and non-expanded repeat sequences were excised from plasmid nanopore sequences, using `maf-cut`. Then, these expanded and non-expanded repeats were inserted between the flanks, to generate chimeric reads. The combinations of repeat copy and number of `rel3` reads are shown in Table 1, imitating the diploid genome. The chimeric reads were aligned to GRCh38 as mentioned above with `WindowMasker` [32].

Dotplot pictures were made with `last-dotplot` using the following command: `last-dotplot --max-gap2 = 0,inf --rmsk1 rmsk.txt aln.maf file.png`

PCR amplification of inexact repeats in `rel3` and PacBio data

Two inexact tandem repeats were tested by PCR and Sanger sequencing. Primers are described in Additional file 1: Table S8. PCR amplification was done using KAPA HiFi HS DNA polymerase (Kapa Biosystems, Basel, Switzerland). DNA for NA12878 was obtained from the Coriell Institute (<https://coriell.org>). PCR products were cloned into pCR-Blunt vector (Thermo Fisher Scientific, MA, USA) and subjected to Sanger sequencing.

Additional file

Additional file 1: Supplemental methods, results, figures, and tables. (PDF 9545 kb)

Acknowledgements

MCF thanks Masahiro Onoguchi for pointing out that strand bias must be discriminated from heterozygosity. SM thanks Qian Liu for suggesting parameters for SCA10 data analyses using RepeatHMM. YK thanks Jun-ichi Satoh and Mika Takitani for plasmid construction. Sanger sequencing was performed by the Support Center for Medical Research and Education, Tokai University.

Funding

This work was supported by AMED under grant numbers, JP18ek0109280, JP18dm0107090 and JP18ek0109301; JSPS KAKENHI Grant Numbers, JP17H01539, JP17K16132 and JP16K09683; Takeda Science Foundation; Kawano Masanori Memorial Public Interest Incorporated Foundation for Promotion of Pediatrics; and Dementia Drug Resource Development Center Project (S1511016).

Availability of data and materials

`tandem-genotypes` is available at <https://github.com/mcfrith/tandem-genotypes> under the open source license: GPL-3.0-or-later. The version (1.1.0) used in this study is archived with DOI: <https://doi.org/10.5281/zenodo.2565742>. The datasets generated and analyzed during the current study are available in the DDBJ repository (DRA007012). Human whole genome nanopore (`rel3`) and PacBio (SRR3197748) sequence data from the same individual (NA12878) were downloaded from <https://github.com/nanopore-wgs-consortium/NA12878> and from the SRA database, respectively [18]. Another human whole genome nanopore dataset from a different individual (NA19240) was downloaded from <https://www.ebi.ac.uk/ena/data/view/PRJEB26791> under accession numbers ERR2585112, ERR2585113, ERR2585114 and ERR2585115.

Authors' contributions

SM and MCF contributed to the conception of the work, acquisition, analysis, and interpretation of the data. TM, SM, TT, HA, YO, YK, HM, and NM

contributed to obtaining and making the materials and acquisition and analysis of the data. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The Institutional Review Board of Yokohama City University of Medicine approved the experimental protocols (IRB approval number: A180800011). Written informed consent was obtained from the patient, in accordance with Japanese regulatory requirements. Experimental methods comply with the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

¹Department of Human Genetics, Yokohama City University Graduate School of Medicine, Fukuura 3-9, Kanazawa-ku, Yokohama 236-0004, Japan. ²Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26 Aomi, Koto-ku, Tokyo 135-0064, Japan. ³Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Chiba, Japan. ⁴Computational Bio Big-Data Open Innovation Laboratory (CBBDO-ILL), AIST, Shinjuku-ku, Tokyo, Japan. ⁵Department of Neurology, University of Occupational and Environmental Health School of Medicine, Kitakyushu, Fukuoka, Japan. ⁶Department of Liberal Arts, Faculty of Medicine, Saitama Medical University, Iruma, Saitama, Japan. ⁷Department of Bioinformatics and Molecular Neuropathology, Meiji Pharmaceutical University, Kiyose, Tokyo, Japan. ⁸Department of Applied Biochemistry, School of Engineering, Tokai University, Hiratsuka, Kanagawa, Japan.

Received: 14 August 2018 Accepted: 1 March 2019

Published online: 19 March 2019

References

- Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, Ramakrishnan S, Lavrenko V, Kakaradov B, Hou C, et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am J Hum Genet.* 2017;101:700–15.
- La Spada AR, Roling DB, Harding AE, Warner CL, Spiegel R, Hausmanowa-Petrusewicz I, Yee WC, Fischbeck KH. Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. *Nat Genet.* 1992;2:301–4.
- MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell.* 1993;72:971–83.
- Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell.* 1992;68:799–808.
- Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, Richards RI. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p (CCG)_n. *Science.* 1991;252:1711–4.
- Lemmers RJ, van der Vliet PJ, Klooster R, Sacconi S, Camano P, Dauwerse JG, Snider L, Straasheijm KR, van Ommen GJ, Padberg GW, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science.* 2010;329:1650–3.
- Brais B, Bouchard JP, Xie YG, Rochefort DL, Chretien N, Tome FM, Lafreniere RG, Rommens JM, Uyama E, Nohira O, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet.* 1998;18:164–7.
- Musova Z, Mazanec R, Krepelova A, Ehler E, Vales J, Jaklova R, Prochazka T, Koukal P, Marikova T, Kraus J, et al. Highly unstable sequence interruptions

- of the CTG repeat in the myotonic dystrophy gene. *Am J Med Genet A*. 2009;149A:1365–74.
9. McFarland KN, Liu J, Landrian I, Godiska R, Shanker S, Yu F, Farmerie WG, Ashizawa T. SMRT sequencing of long tandem nucleotide repeats in SCA10 reveals unique insight of repeat expansion structure. *PLoS One*. 2015;10:e0135906.
 10. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet*. 2018;50:581–90.
 11. Nishikawa A, Mitsuhashi S, Miyata N, Nishino I. Targeted massively parallel sequencing and histological assessment of skeletal muscles for the molecular diagnosis of inherited muscle disorders. *J Med Genet*. 2017;54:104–10.
 12. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9:eaal5209.
 13. Ameer A, Kloosterman WP, Hestand MS. Single-molecule sequencing: towards clinical applications. *Trends Biotechnol*. 2018. <https://doi.org/10.1016/j.tibtech.2018.07.013>.
 14. Ummat A, Bashir A. Resolving complex tandem repeats with long reads. *Bioinformatics*. 2014;30:3491–8.
 15. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med*. 2017;9:65.
 16. Frith MC, Khan S. A survey of localized sequence rearrangements in human DNA. *Nucleic Acids Res*. 2018;46:1661–73.
 17. Hamada M, Ono Y, Asai K, Frith MC. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*. 2017;33:926–8.
 18. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36:338–45.
 19. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018;15:461–8.
 20. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 2017;8:1326.
 21. Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J Hum Genet*. 2018. <https://doi.org/10.1038/s10038-018-0551-7>.
 22. De Coster W, De Roeck A, De Pooter T, D'Hert S, De Rijk P, Strazisar M, Sleegers K. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *BioRxiv*. 2018. <https://doi.org/10.1101/434118>.
 23. Höjjer I, Tsai YC, Clark TA, Kotturi P, Dahl N, Stattin EL, Bondeson ML, Feuk L, Gyllenstein U, Ameer A. Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. *Hum Mutat*. 2018;39:1262–72.
 24. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27:573–80.
 25. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, Kohno Y, et al. Long-read sequencing identifies GGC repeat expansion in human-specific NOTCH2NLC associated with neuronal intranuclear inclusion disease. *bioRxiv*:515635. <https://doi.org/10.1101/515635>.
 26. Frith MC. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res*. 2011;39:e23.
 27. Frith MC. Gentle masking of low-complexity sequences improves homology search. *PLoS One*. 2011;6:e28819.
 28. Oma Y, Kino Y, Sasagawa N, Ishiura S. Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. *J Biol Chem*. 2004;279:21217–22.
 29. Kino Y, Washizu C, Kurosawa M, Oma Y, Hattori N, Ishiura S, Nukina N. Nuclear localization of MBNL1: splicing-mediated autoregulation and repression of repeat-derived aberrant proteins. *Hum Mol Genet*. 2015;24:740–56.
 30. Oma Y, Kino Y, Toriumi K, Sasagawa N, Ishiura S. Interactions between homopolymeric amino acids (HPAAs). *Protein Sci*. 2007;16:2195–204.
 31. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733–45.
 32. Morgulis A, Gertz EM, Schaffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*. 2006;22:134–41.
 33. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

