RESEARCH ARTICLE

High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing [version 1; peer review: 1 approved with reservations, 1 not approved]

Devika Ganesamoorthy[1,2], Mengjia Yan[1], Valentine Murigneux[1], Chenxi Zhou[1,2], Minh Duc Cao[1], Tania P. S. Duarte[1], Lachlan J. M. Coin[1,2]

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, 4072, Australia
[2]Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, 3052, Australia

v1 **First published:** 02 Sep 2020, **9**:1084
https://doi.org/10.12688/f1000research.25693.1

Latest published: 02 Sep 2020, **9**:1084
https://doi.org/10.12688/f1000research.25693.1

Abstract

Background: Tandem repeats (TRs) are highly prone to variation in copy numbers due to their repetitive and unstable nature, which makes them a major source of genomic variation between individuals. However, population variation of TRs has not been widely explored due to the limitations of existing approaches, which are either low-throughput or restricted to a small subset of TRs. Here, we demonstrate a targeted sequencing approach combined with Nanopore sequencing to overcome these limitations.
Methods: We selected 142 TR targets and enriched these regions using Agilent SureSelect target enrichment approach with only 200 ng of input DNA. We barcoded the enriched products and sequenced on Oxford Nanopore MinION sequencer. We used VNTRTyper and Tandem-genotypes to genotype TRs from long-read sequencing data. Gold standard PCR sizing analysis was used to validate genotyping results from targeted sequencing data.
Results: We achieved an average of 3062-fold target enrichment on a panel of 142 TR loci, generating an average of 97X coverage per sample with 200 ng of input DNA per sample. We successfully genotyped an average of 75% targets and genotyping rate increased to 91% for the highest-coverage sample for targets with length less than 2 kb, and GC content greater than 25%. Alleles estimated from targeted long-read sequencing were concordant with gold standard PCR sizing analysis and highly correlated with alleles estimated from whole genome long-read sequencing.
Conclusions: We demonstrate a targeted long-read sequencing approach that enables simultaneous analysis of hundreds of TRs and accuracy is comparable to PCR sizing analysis. Our approach is feasible to scale for more targets and more samples facilitating large-scale analysis of TRs.

Open Peer Review

Reviewer Status ? ✗

	Invited Reviewers	
	1	**2**
version 1 02 Sep 2020	? report	✗ report

1. **Rick M. Tankard**, Murdoch University, Murdoch, Australia

2. **Mark T. W. Ebbert**, Mayo Clinic, Jacksonville, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Tandem repeats, targeted sequencing, long-read sequencing

Corresponding authors: Devika Ganesamoorthy (d.ganesamoorthy@imb.uq.edu.au), Lachlan J. M. Coin (lachlan.coin@unimelb.edu.au)

Author roles: Ganesamoorthy D: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Yan M**: Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Murigneux V**: Formal Analysis, Validation, Visualization, Writing – Review & Editing; **Zhou C**: Data Curation, Formal Analysis, Visualization, Writing – Review & Editing; **Cao MD**: Formal Analysis, Software, Writing – Review & Editing; **Duarte TPS**: Investigation, Validation, Writing – Review & Editing; **Coin LJM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Australian Government National Health and Medical Research Council Project Grant APP1052303 to Prof Lachlan J. M. Coin.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Ganesamoorthy D, Yan M, Murigneux V *et al.* **High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing [version 1; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2020, **9**:1084 https://doi.org/10.12688/f1000research.25693.1

First published: 02 Sep 2020, **9**:1084 https://doi.org/10.12688/f1000research.25693.1

Introduction

Repeated sequences occur in multiple copies throughout the genome; they make up almost half of the human genome[1]. Repeat sequences can be divided into two categories: interspersed repeats and tandem repeats (TRs). Interspersed repeats are scattered throughout the genome and are remnants of transposons[2]. TRs consists of repeat units that are located adjacent to each other (i.e. in tandem). There are almost 1 million TRs in the human genome, covering 10.6% of the entire genome[3]. TRs can be further divided into two types based on the length of the repeat unit; repeats with one to six base pair repeat units are classified as microsatellites or short tandem repeats (STRs) and those with more than six base pair repeat units are known as minisatellites[4].

TRs are prone to high rates of copy number variation and mutation due to the repetitive unstable nature, which makes them a major source of genomic variation between individuals. Variation in TRs may explain some of the phenotypic variation observed in complex diseases as it is poorly tagged by single nucleotide variation[5,6]. Recent studies have shown that 10% to 20% of coding and regulatory regions contain TRs and suggested that variations in TRs could have phenotypic effect[7]. Although TRs represent a highly variable fraction of the genome, analysis of TRs so far are limited to known pathogenic regions, mainly STRs due to the limitations in analysis techniques.

Traditionally, TR analysis has been carried out via restriction fragment length polymorphism (RFLP) analysis[8] or PCR amplification of the target loci followed by fragment length analysis[9]. These techniques are only applicable to a specific target region and not scalable to high-throughput analysis, which limits the possibility of genome-wide TR analysis. In the recent decade, significant progress has been made in utilising high-throughput short-read sequencing data for genotyping STRs[10]. Our group and others have also demonstrated targeted sequencing approaches using short-read sequencing for TR analysis[11,12]. Several computational tools have been developed to improve the accuracy of TR genotyping from short-read sequencing data with varying performance[13–19]. Yet, most of these tools have focused mainly on the analysis of STRs and analysis of longer TRs remains a hurdle for these approaches. We reported a new approach GtTR in Ganesamoorthy *et al.*[12], which utilizes short-read sequencing data to genotype longer TRs. GtTR reports absolute copy number of the TRs, but it does not report the exact genotype of two alleles due to the use of short-read sequencing data.

Sequencing reads that span the entire repeat region are informative to accurately genotype TRs[11], and therefore are ideal for genome-wide TR analysis. Long-read sequencing technologies have the potential to span all TRs in human genome, including long TRs. There have been few reports on the use of long-read sequencing for the analysis of specific TRs implicated in diseases[20–22]. Genotyping tools utilizing long-read sequencing data, such as Nanosatellite[21], RepeatHMM[23] and Tandem-genotypes[24] have been reported in the recent years with varying performance across different length of repeat units and repeat length. We reported VNTRTyper in Ganesamoorthy *et al.*[12] to genotype TRs from long-read whole genome sequencing data. Despite the availability of genotyping tools, long-read sequencing is not widely used for TR analysis, due to the high costs associated with whole genome long-read sequencing. Cost-effective long-read sequencing approaches will be an important and attractive option to genotype TRs in large-scale studies. However, there has been limited progression on targeted long-read sequencing of TRs.

We have previously demonstrated that targeted sequence capture of repetitive TR sequences are feasible using short-read sequencing technologies[12]. In this study, we demonstrate the targeted sequence capture of repetitive TRs using Oxford Nanopore long-read sequencing technologies. There have been previous reports on the use of targeted sequencing combined with long read sequencing technologies[25]; however, enrichment of repeat sequences requires optimization in probe design and probe hybridization approaches. We optimized the protocols and report successful enrichment of repetitive sequences followed by long-read sequencing. We demonstrate the accuracy of genotype estimates from targeted long-read sequencing by comparison with gold-standard PCR sizing analysis. In this study, we predominantly targeted longer TRs (i.e. minisatellites); however, our approach is applicable to all TRs. Our targeted long-read sequencing method presented here provides an accurate and cost-effective approach for large-scale analysis of TRs, which will be useful for researchers to explore the impact of TR variants on diseases and phenotypes.

Methhods

Samples for sequencing

DNA samples of CEPH/UTAH pedigree 1463 were purchased from Coriell Institute for Medical Research (USA). Seven family members from the pedigree used for sequencing analysis were NA12877, NA12878, NA12879, NA12881, NA12882, NA12889 and NA12890.

Selection of TRs and probe design

The selection of TRs and design of probes were described in Ganesamoorthy *et. al.* (2018)[12]. Briefly, 142 TRs were selected; they range from 112 to 25236 bp in length in the reference human genome (hg19) and the number of repeat units range from 2 to 2300 repeats. TRs used in this study were selected as part of another study to investigate association between TRs and Obesity and these targeted TRs are not disease associated. Agilent SureSelect DNA design (Agilent Technologies) was used to design target probes to capture the targeted regions (including 100-bp flanking regions) and regions flanking the TRs (at least 1000 bp).

Nanopore targeted sequencing of TRs

All seven family members from the CEPH pedigree 1463 were used for Nanopore targeted sequencing analysis (NA12877, NA12878, NA12879, NA12881, NA12882, NA12889 and NA1289). Target sequence capture for Nanopore sequencing was performed using Agilent SureSelect XT HS Target Enrichment System (Agilent Technologies) according to the manufacturer's instructions with slight modifications. Briefly, 200 ng of DNA was fragmented to 3 kb using Covaris Blue miniTUBE (Covaris). Greater than 90% of the targeted TRs are less than 3 kb and SureSelect capture protocol works effectively on fragments less than 4 kb in length; therefore, DNA products were sheared to 3 kb. Fragmented DNA was end-repaired, adapter-ligated and amplified prior to target capture. Extension time for pre-capture amplification was increased to 4 minutes to allow for the amplification of long fragments and 14 cycle amplification was used. Purified pre-capture PCR products were hybridized to the designed capture probes for 2 hours. Streptavidin beads (Thermo Fisher) were used to pull down the DNA fragments bound to the probes. Finally, captured DNA was amplified with long extension time (4 minutes) using Illumina Index adapters provided in the enrichment kit. Post capture PCR products were purified using 0.8X - 1X AMPure XP beads (Beckman Coulter).

Nanopore sequencing library preparation was performed using 1D Native barcoding genomic DNA (with EXP-NBD103 and SQK-LSK108) (Oxford Nanopore Technologies) protocol according to the manufacturer's instructions with minor modifications. Briefly, 100–200ng of post capture PCR products were end repaired and incubated at 20°C for 15 mins and 65°C for 15 mins. End repaired products were ligated with unique native barcodes. Purification steps after end repair and barcode ligation were avoided to minimize the loss of DNA. Barcoded samples were pooled in equimolar concentrations prior to adapter ligation. Adapter ligated samples were purified using 0.4X AMPure XP beads (Beckman Coulter). Samples were split into two sequencing groups: NA12877, NA12878, NA12879 and NA12890 (group 1); and NA12881, NA12882 and NA12889 (group 2). Sequencing was performed using a MinION sequencer (Oxford Nanopore Technologies) using R9.5 flow cell. Both groups were sequenced for 48 hours. Nanopore sequencing data were base called using Albacore (version 2.2.7) and reads were demultiplexed using Albacore (version 2.2.7) based on the barcode sequences.

Public data used in the study

Nanopore WGS data on CEPH Pedigree 1463 sample NA12878 were obtained from the Nanopore WGS consortium (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/nanopore-human-genome/rel_3_4.md)[26]. PacBio WGS data on NA12878 sample were downloaded from SRA with accession numbers SRX627421 and SRX638310[27]

VNTRTyper

Sequencing reads were mapped to hg19 reference genome using Minimap2 (version 2.13)[28]. For Nanopore sequencing '-ax map-ont' and for PacBio WGS '-ax map-pb' parameters were used. VNTRTyper, our in-house tool described by Ganesamoorthy *et al.*[12] was used to genotype TRs from long-read sequencing data. Briefly, VNTRTyper takes advantage of the long-read sequencing to identify the number of repeat units in the TR regions. Firstly, the tool identifies reads that span the repeat region and applies hidden Markov models (HMM) to align the repetitive portion of each read to the repeat unit. Then it estimates the multiplicity of the repeat units in a read using a profile HMM.

Recently, we further improved the accuracy of genotyping estimates by clustering the copy number counts from reads to identify the likely genotypes per target. We used Kmeans clustering and the number of clusters are fixed at two clusters for two alleles. A minimum threshold of two supporting reads per genotype was used to assign genotypes. Furthermore, for heterozygous alleles, both alleles should have at least 10% of reads supporting the allele, if not allele with less than 10% of reads was excluded during the analysis. The updated version of VNTRTyper can be accessed from GitHub Japsa release 1.9–3c and can be deployed using script name jsa.tr.longreads. Details of VNTRTyper analysis are previously reported by Ganesamoorthy *et al.*[12].

Tandem-genotypes

We also used another independent method, Tandem-genotypes, to estimate genotypes from long-read sequencing data. Tandem-genotypes was recently reported for analysis of TR genotypes from long-read sequencing data[24] and it can be utilised for both Nanopore and PacBio sequencing technologies.

Nanopore and PacBio sequencing data were mapped to the hg19 reference genome using LAST v959[29]. Calculation of repeat length per sequencing read was performed with Tandem-genotypes as reported by Mitsuhashi *et al.*[24]. Copy number changes in reads covering the repeat's forward and reverse strands were merged and the two alleles with the highest number of supporting reads for each VNTR were extracted. A minimum threshold of two supporting reads per genotype was used to assign genotypes.

PCR analysis of VNTRs

A total of 10 targeted VNTR regions which are less than 1 kb in repetitive sequence were validated by PCR sizing analysis in this study (PCR primer sequences provided in *Extended data*, Supplementary Table 1)[30]. These ten targets include various repeat unit length and repeat sequence combinations to assess the accuracy of the genotypes determined from sequencing data. The majority of these targets were tested in our previous study[12] and the results from the previous PCR analysis were used for these regions. PCRs were performed using HotStar Taq DNA Polymerase (Qiagen) and PCR conditions were optimized for each PCR target. PCR products were purified and subjected to capillary electrophoresis on an ABI3500xL Genetic Analyzer (Applied BioSystems). Fragment sizes were analyzed using GeneMapper 4.0 (Applied BioSystems). Alternatively, STRand or Osiris could be used for fragment size analysis. Capillary electrophoresis plots provided in *Extended data*, Supplementary Information PCR data[30].

Statistical analysis

Linear regression analysis was used to determine correlation between genotype estimates. All plots were generated using GraphPad Prism (version 7.00 for Windows; GraphPad Software, La Jolla California USA).

We investigated the effect of GC content, repeat length, repeat period and repeat copy number on target sequencing depth using a multivariate linear regression model. We used ggplot2 (version 3.2.0) to visualize the relationship between these factors and sequencing depth across all seven samples. Thresholds on GC and repeat length were chosen based on this visual analysis. Genotype rate was calculated as the proportion of sample, target pairs which had a predicted genotyped (based on VNTRtyper) amongst all targets which met the GC and repeat length thresholds.

Results

We demonstrate a targeted sequence capture approach combined with Nanopore long-read sequencing to genotype hundreds of TRs.

Targeted Capture Sequencing of Tandem repeats

We performed sequence enrichment of targeted TRs for 7 samples followed by long-read sequencing using Oxford Nanopore Technologies' MinION as described in the *Methods*. Figure 1a shows the read length distribution observed in targeted capture sequencing data. The median read length followed the expected read-length distribution, with the exception of an under-representation of repeats of length >3 kb (Figure 1a). The read length in this study was sufficient to analyse majority of targeted TRs of length less than 2 kb. Sequence coverage varied across targets and samples, on average 97X sequence coverage was achieved, with only 19 targets having less than 1X coverage (Figure 1b) and majority of the low coverage targets (16 of the 19 targets) have less than 25% GC content (*Extended data*, Supplementary Figure 1)[30].

Extended data, Supplementary Table 2[30] summarises the metrics for targeted sequencing on Nanopore sequencing technologies. Nanopore multiplexing (See *Methods*) group 1 samples had similar yield between samples; however, Nanopore multiplexing group 2 samples had varying yield per sample. Despite the differences in sequencing yield, we achieved an average of 3062-fold target enrichment and on target capture rate was approximately 50%.

Genotyping of TRs using targeted long-read sequencing

Genotype estimates from targeted long-read sequencing datasets were estimated using our tool VNTRTyper[12] with the improvements described in the *Methods*. We also applied Tandem-genotypes[24] to determine the genotypes of TRs from long-read sequencing data. We used a minimum of two reads as read threshold to determine the repeat number for each allele.

Prior to obtaining any sequence data, we generated PCR sizing results as a gold standard on 10 targets for comparison to sequencing analysis. These 10 targets were selected to include various repeat unit length and repeat sequence combinations to assess the accuracy of the genotypes determined from sequencing data. Of these 10 targets, two were excluded for comparison as all seven samples had insufficient number of spanning reads (minimum of two reads required for genotyping) to genotype these targets. Genotype estimates from VNTRtyper on these eight targets correlated well with PCR (Pearson correlation greater than 0.980 for all samples) (Table 1 and *Extended data*, Supplementary Figure 2)[30]. Genotype estimates by Tandem-genotypes also correlated well with PCR, with a correlation greater than 0.984 for all samples *Extended data*, (Supplementary Table 3 and Supplementary Figure 3)[30]; however, fewer targets had sufficient data to compare with PCR sizing results.

Genotype estimates from VNTRTyper and Tandem-genotypes for all 142 targets from Nanopore capture sequencing samples are provided in *Extended data*, Supplementary

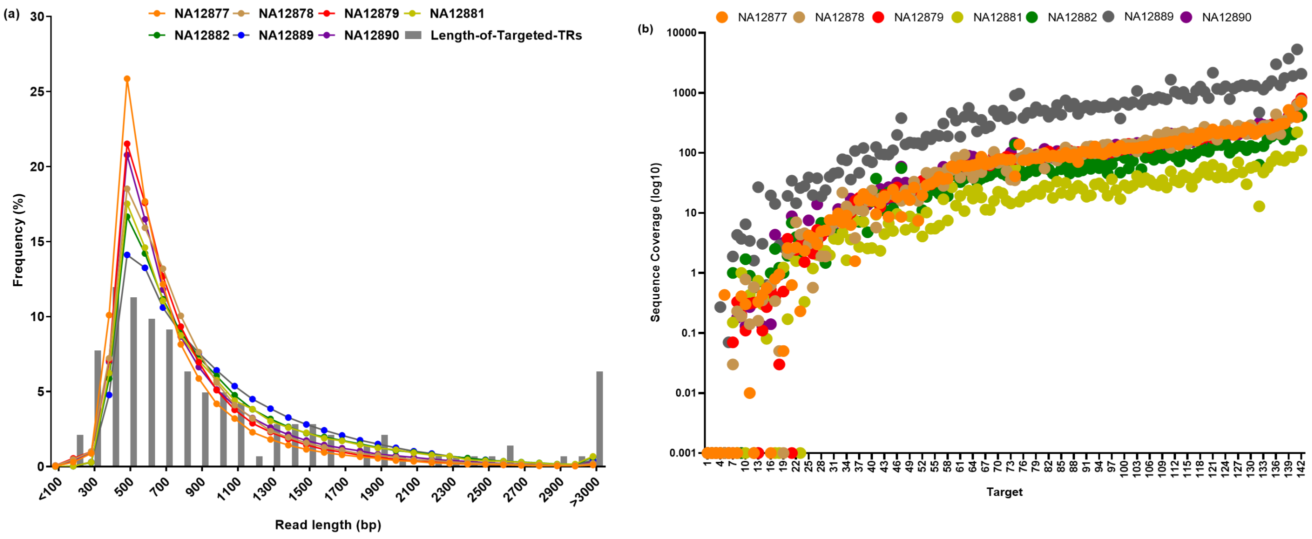

Figure 1. Read length and sequence coverage distribution. (**a**) Read length distribution of Nanopore targeted sequencing. Lines indicates the read length distribtuion for each sample and grey bars indicate the length distribution of targeted TRs and (**b**) Sequence coverage distribution of Nanopore targeted sequencing for all seven samples.

Table 1. Genotype estimates on Nanopore targeted capture sequencing using VNTRTyper.

Sample	Method	Genotype of Target*								Pearson correlation with PCR
		TR_8 (12.0)	TR_57 (15.6)	TR_86 (2.0)	TR_87 (9.0)	TR_93 (2.0)	TR_109 (15.3)	TR_112 (4.0)	TR_120 (2.2)	
NA12877	PCR	12.0/13.0	10.6/13.6	2.0/2.0	6.0/8.0	2.0/2.0	15.3/17.3	3.0/4.0	2.2/2.2	0.9988
	Nanopore	12.0/13.0	ND	2.0/2.0	6.2/8.4	2.0/2.0	15.3/17.3	3.0/4.1	2.2/3.2	
NA12878	PCR	12.0/12.0	10.6/12.6	2.0/2.0	6.0/9.0	2.0/2.0	15.3/17.3	3.0/4.0	2.2/2.2	0.9978
	Nanopore	11.0/12.1	ND	2.0/2.0	6.0/8.8	2.0/2.0	14.8 /17.1	3.0/4.0	2.2/3.2	
NA12879	PCR	12.0/13.0	10.6/10.6	2.0/2.0	8.0/9.0	2.0/2.0	15.3/17.3	3.0/3.0	2.2/2.2	0.9927
	Nanopore	12.4/12.4	10.6/10.6	2.0/2.0	6.0/8.6	2.0/2.0	14.8/16.9	3.0/4.0	2.2/3.2	
NA12881	PCR	12.0/12.0	10.6/10.6	2.0/2.0	8.0/9.0	2.0/2.0	15.3/15.3	3.0/4.0	2.2/2.2	0.9944
	Nanopore	12.0/12.0	10.6/10.6	2.0/2.0	8.0/9.0	ND	15.3/17.3	3.0/4.0	2.2/3.2	
NA12882	PCR	12.0/13.0	10.6/10.6	2.0/2.0	6.0/6.0	2.0/2.0	15.3/17.3	3.0/3.0	2.2/2.2	0.9919
	Nanopore	11.8/13.0	10.6/10.6	2.0/2.0	6.0/8.5	2.0/2.0	15.3/17.3	3.0/4.0	2.2/3.2	
NA12889	PCR	12.0/12.0	13.6/17.6	2.0/2.0	8.0/8.0	2.0/2.0	17.3/17.3	4.0/4.0	2.2/2.2	0.9800
	Nanopore	12.1/12.1	10.6/15.8	2.0/2.0	6.1/8.1	2.0/2.0	14.8/17.2	3.0/4.0	2.2/3.2	
NA12890	PCR	12.0/13.0	10.6/10.6	2.0/2.0	6.0/9.0	2.0/2.0	15.3/15.3	3.0/3.0	2.2/2.2	0.9936
	Nanopore	12.2/12.2	10.6/10.6	2.0/2.0	6.0/9.0	2.0/2.0	13.3/15.3	3.0/4.0	2.2/3.2	

*Repeat Number in reference hg19 is provided within brackets for each target

*Repeat numbers that do not agree with PCR results are highlighted in red.

ND – Sufficient data not available for genotype analysis

Spreadsheet Tables 1 and 2[30], respectively. Genotype estimates by VNTRtyper and Tandem-genotypes correlate well and the correlation values range from 0.904 to 0.994 for Nanopore targeted sequencing.

We were able to determine the genotype on average for 60% of the targets (range 48% to 75%) using VNTRTyper and 57% of the targets (range 41% to 74%) using Tandem-genotypes. Both VNTRTyper and Tandem-genotypes failed to genotype targets with low GC sequence content (<25% GC content) and targets which are greater than 2 kb in length, which accounts for approximately 22% of the targets (32 of the 142 targets). Targets with low GC sequence content (<25% GC content) did not have sufficient sequence coverage for analysis due to inefficient sequence enrichment in these regions (*Extended data*, Supplementary Figure 1)[30]. Targets which are greater than 2 kb in length did not have sufficient spanning reads for genotyping analysis (See *Methods* and Figure 1a).

It was evident that the GC content of the target and size (i.e. repeat length) affected the genotyping efficiency of our targeted capture sequencing approach. Therefore, we assessed the genotyping rate based on the size of the target and GC content of the target (Figure 2). For all 142 targets genotyping rate using VNTRTyper was only 59.8% (Figure 2a); however, the genotyping rate improved to 67% for 125 targets with a size threshold of 2 kb (Figure 2b) and 67.1% for 125 targets with 25% GC threshold (Figure 2c). Furthermore, genotyping rate improved to 75.2% for 110 targets with a combined 2 kb size threshold and 25% GC threshold (Figure 2d). Also, sample with high sequence coverage (NA12889) had the highest genotyping rate of 90.9% for 110 targets (*Extended data*, Supplementary Figure 4)[30]. Genotyping rate using Tandem-genotypes also improved to an average of 63.7% for 110 targets (range 43.6% to 85.5%) (*Extended data*, Supplementary Figure 5)[30].

Genotyping of Tandem repeats using long-read whole genome sequencing

To investigate the accuracy of genotype estimates of TRs from targeted sequence capture compared to WGS, we performed genotyping analysis on the targeted regions using VNTRTyper and Tandem-genotypes on whole genome long-read sequencing data. We downloaded whole genome long-read Nanopore and PacBio sequencing data on CEPH Pedigree 1463 NA12878 sample. We have previously reported genotyping estimates by VNTRTyper on PacBio NA12878 WGS data[12]. Here we use the genotype estimates by VNTRTyper on PacBio NA12878 WGS data to compare genotype estimates by Tandem-genotype and the results of targeted sequencing analysis.

We compared the accuracy of genotype estimates from WGS data with PCR sizing analysis. Genotype estimates by VNTRTyper

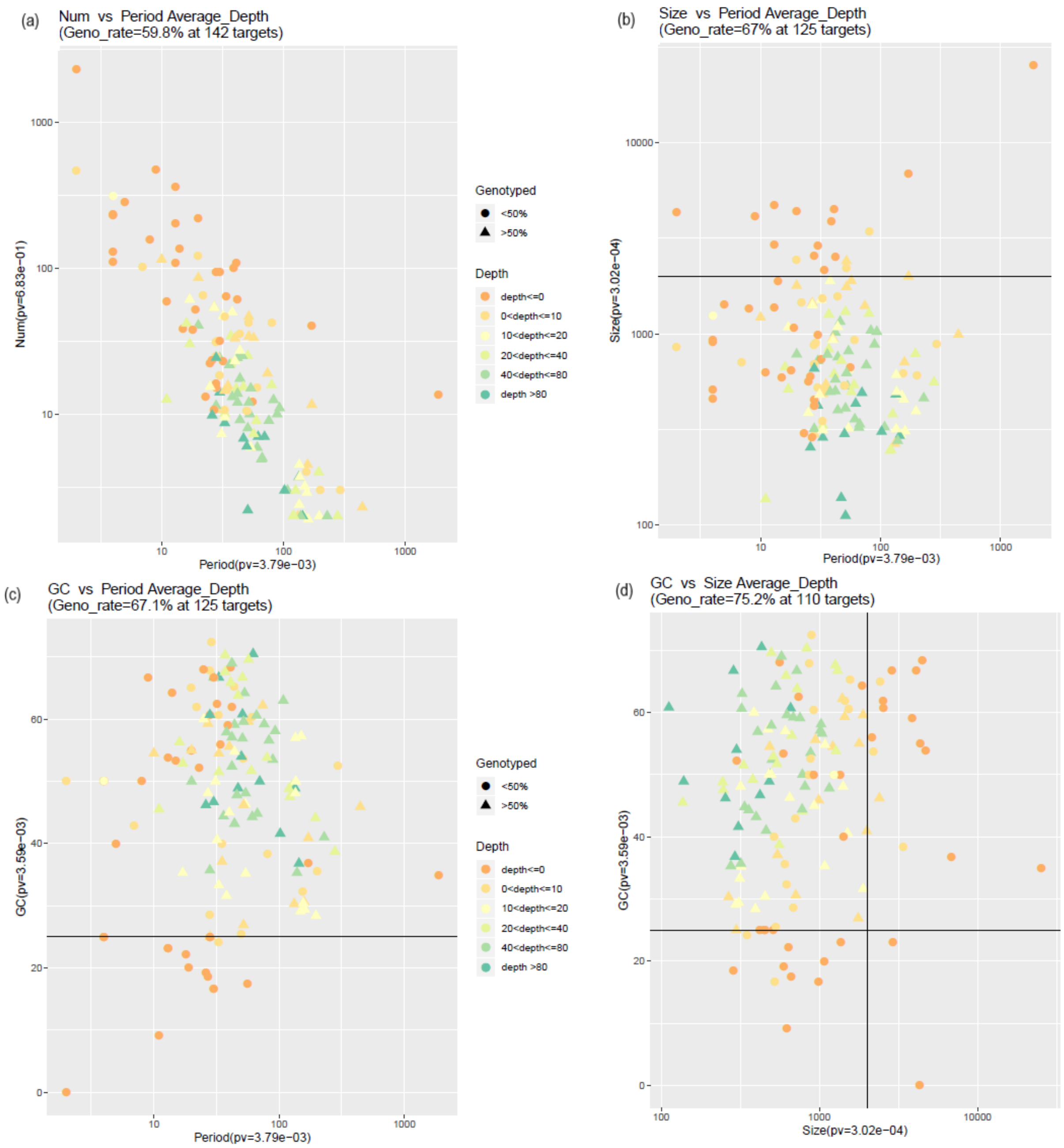

Figure 2. Assessment of genotyping rate using VNTRTyper based on the size and GC content of the target for all seven samples. Triangle indicates that greater than 50% of the samples had a genotyping estimate and circle indicates only less than 50% of the samples had a genotyping estimate for the given target. Colours indicate the depth, which is defined as the number of spanning reads detected for the target region. Black thick lines inside the plots indicate the 2-kb size threshold and 25% GC content size threshold. (**a**) Genotyping rate for all targets, shown as repeat number vs period (i.e. repeat unit). (**b**) Genotyping rate with 2-kb size threshold. (**c**) Genotyping rate with 25% GC threshold. (**d**) Genotyping rate with both 25% GC threshold and 2-kb size threshold.

and Tandem-genotypes on WGS data were compared with PCR sizing results on 10 targets (Table 2). VNTRTyper and Tandem-genotypes had comparable correlation with PCR sizing analysis for both Nanopore and PacBio WGS. Genotype estimates for all 142 targets from Nanopore and PacBio WGS data determined using VNTRTyper and Tandem-genotypes are provided in *Extended data*, Supplementary Spreadsheet Table 3[30].

We compared the genotype estimates between WGS data and targeted capture sequencing data (77 targets which had results for both WGS and targeted sequencing). Genotype estimates by VNTRTyper between WGS data and targeted capture sequencing data showed a correlation of 0.9782 (correlation on 154 alleles) (Figure 3a). Genotype estimates by Tandem-genotypes had lower correlation between WGS and targeted capture

Table 2. Genotype estimates on NA12878 WGS and Capture sequencing data using VNTRTyper and Tandem-genotypes.

Method^	Genotype of Target*										Pearson Correlation with PCR
	TR_8 (12.0)	TR_32 (38.4)	TR_57 (15.6)	TR_64 (18.3)	TR_86 (2.0)	TR_87 (9.0)	TR_93 (2.0)	TR_109 (15.3)	TR_112 (6.0)	TR_120 (2.2)	
PCR	12.0/12.0	9.4/10.4	10.6/12.6	17.3/18.3	2.0/2.0	6.0/9.0	2.0/2.0	15.3/17.3	3.0/4.0	2.2/2.2	
NPW_VT	12.0/12.0	8.2/10.7	10.6/12.6	17.3/18.6	2.0/2.0	6.0/9.0	2.0/2.0	15.3/17.3	3.0/4.0	2.2/3.2	0.9980
NPW_TG	11.0/12.0	4.4/9.4	10.6/12.6	18.3/19.3	2.0/2.0	6.0/9.0	2.0/2.0	15.3/17.3	3.0/4.0	2.2/2.2	0.9800
PBW_VT	12.0/12.0	11.0/12.4	10.6/12.6	18.3/19.3	2.0/2.0	*6.0/9.0*	2.0/2.0	15.3/17.3	3.0/4.0	2.2/2.2	0.9957
PBW_TG	12.0/12.0	ND	12.6/12.6	ND	2.0/2.0	ND	2.0/2.0	15.3/15.3	4.0/4.0	2.2/2.2	0.9900
NPC_VT	11.0/12.1	ND	ND	ND	2.0/2.0	6.0/8.8	2.0/2.0	14.8/17.1	3.0/4.0	2.2/3.2	0.9978
NPC_TG	11.0/12.0	ND	ND	ND	2.0/2.0	6.0/9.0	ND	15.3/17.3	3.0/4.0	2.2/2.2	0.9988

^NPW_VT – Nanopore WGS VNTRTyper; NPW_TG – Nanopore WGS Tandem-genotypes; PBW_VT – PacBio WGS VNTRTyper; PBW_TG – PacBio WGS Tandem-genotypes; NPC_VT – Nanopore Capture sequencing VNTRTyper; NPC_TG – Nanopore Capture sequencing Tandem-genotypes

*Repeat Number in reference hg19 is provided within brackets for each target

*Repeat numbers that do not agree with PCR results are highlighted in red.

ND – Sufficient data not available for genotype analysis.

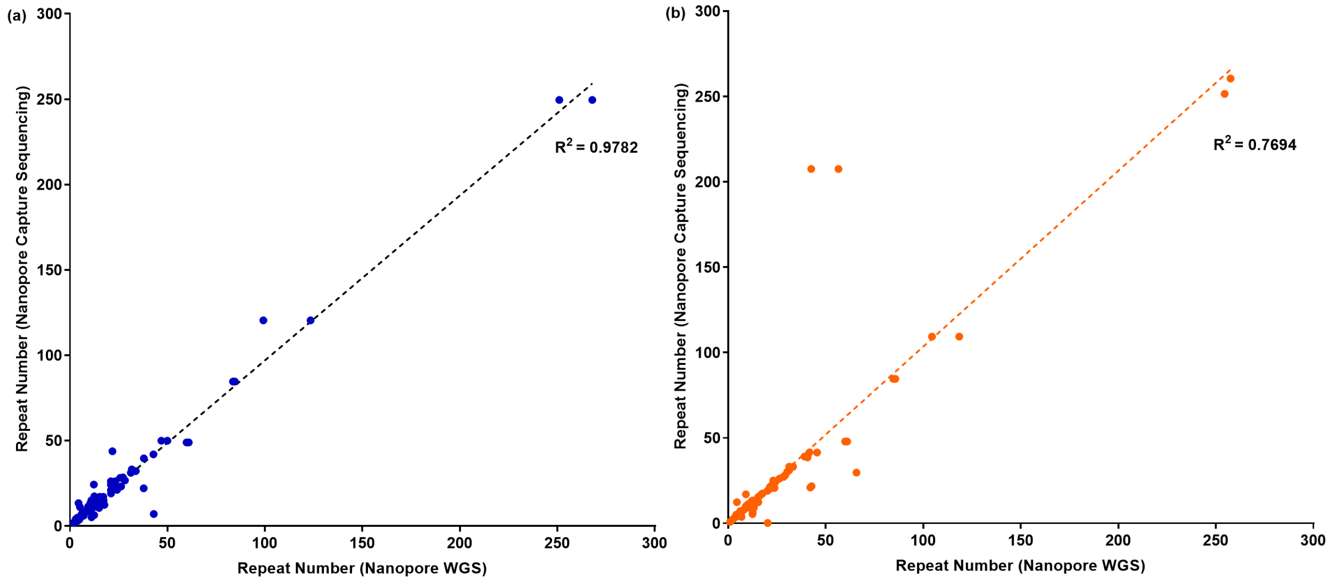

Figure 3. Correlation between whole genme sequencing and targeted sequencing genotype estimates. Using (**a**) VNTRTyper and (**b**) Tandem-genotypes for the NA12878 sample.

sequencing data of 0.7694 (correlation on 152 alleles – 76 targets, removing the two outliers improved correlation to 0.9084) (Figure 3b). On the subset of seven targets for which we had generated PCR sizing analysis, Nanopore WGS data correlated with 12/14 genotype estimates on Nanopore capture sequencing using VNTRTyper precisely compared to PCR sizing (Table 2 and Figure 4a). Genotype estimates using Tandem-genotypes on Nanopore WGS data correlated with 11/12 genotype estimates on Nanopore capture sequencing precisely compared to PCR sizing (Figure 4b).

Variation in Tandem repeats

To assess the extent of variation in repeat numbers between individuals, we compared the genotype estimates to the reported reference (hg19) repeat number. Genotype estimates determined by VNTRTyper on Nanopore capture sequencing on seven members of CEPH pedigree 1463 were used to assess the variation. We found that for a given sample, on average 51% (range 45–60%) of the targets have a genotype which is different to the reference, with more deletions (28%) than duplications (23%) (Figure 5).

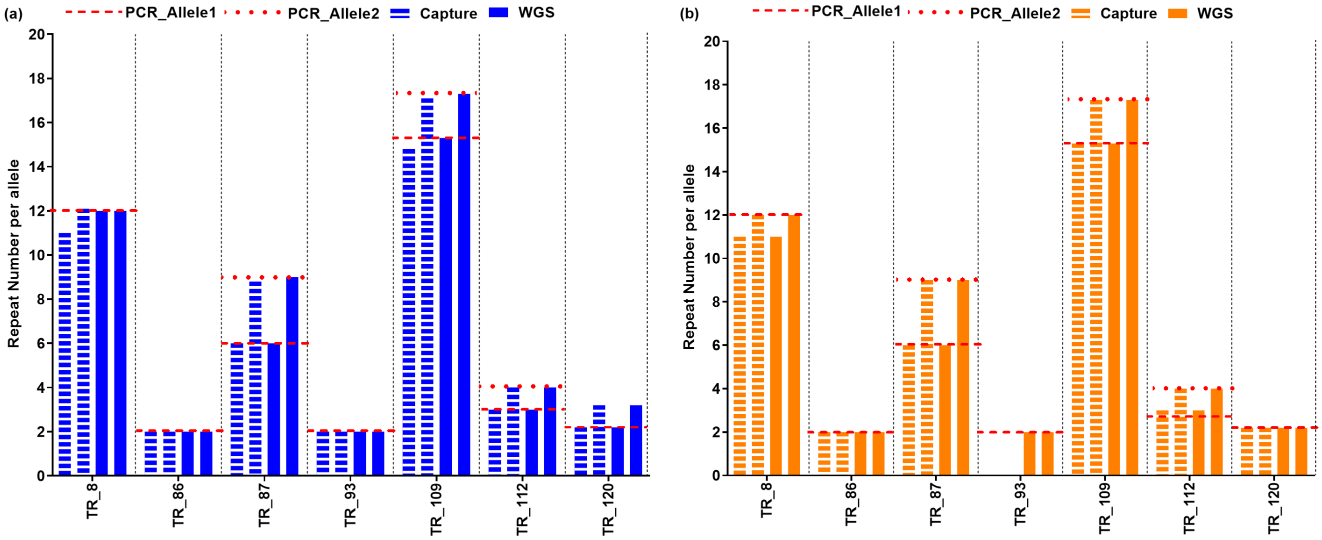

Figure 4. Comparison of genotype estimates between whole genme sequencing and target capture sequencing for NA12878 sample. Using (**a**) VNTRTyper and (**b**) Tandem-genotypes. Red line indicates PCR sizing results. Targets with no genotype estimates are shown as a gap for the corresponding column.

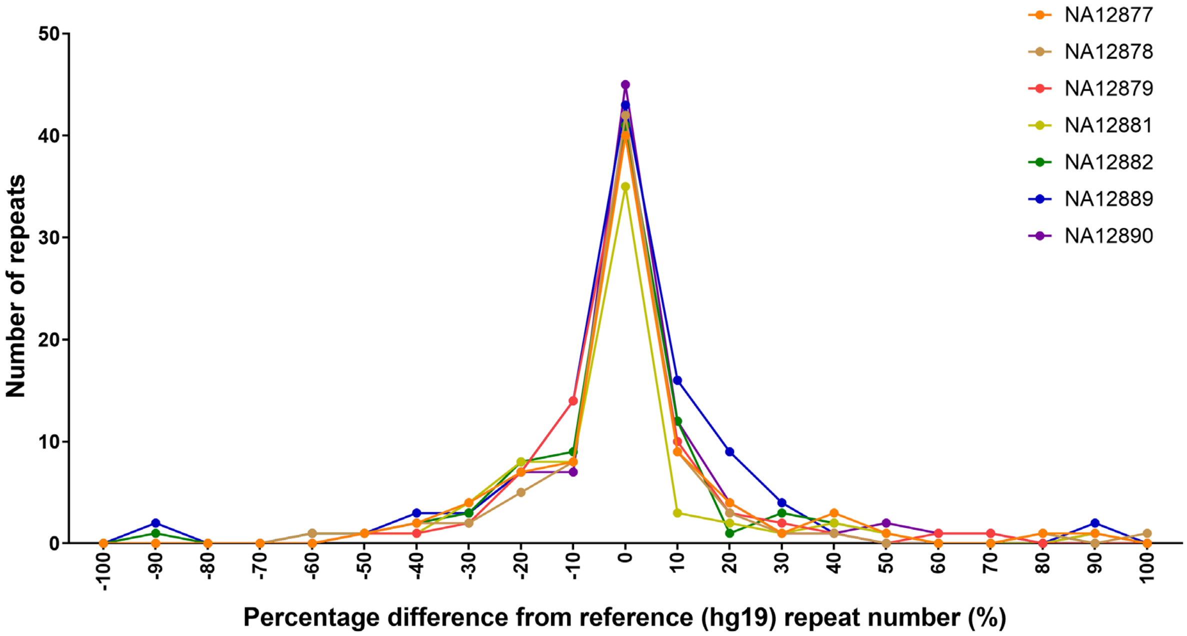

Figure 5. Percentage difference between reported repeat number in reference genome (hg19) and estimated repeat number based on genotype estimates using VNTRTyper on Nanopore targeted sequencing.

Discussion

In this study, we present a targeted sequencing approach combined with long-read sequencing technology to genotype TRs. To our knowledge, this is the first report on genotyping analysis of hundreds of TRs using targeted long-read sequencing approach. Sequencing reads that span the entire repeat region and flanking region are often useful in providing an accurate estimation of the repeat genotype. Long-read sequencing technologies have the ability to generate reads which can span the entire repeat region and flanking regions. However, whole genome long-read sequencing analysis is still expensive for large-scale population analysis; hence, we developed a targeted long-read sequencing approach for TR analysis.

We showed that 1) target enrichment of repetitive sequences followed by long-read sequencing is feasible and 2) genotype predictions on targeted TR sequencing are comparable to the accuracy of PCR sizing analysis of repeats. Overall, we achieved an average genotyping rate of 75% for 110 TR loci with repeat length less than 2 kb and GC content greater than 25%. Genotyping rate improved to 91% for the highest-coverage sample, indicating that more sequencing could improve genotyping rate.

Targets with low GC sequence content (<25% GC content) did not have sufficient sequence coverage with targeted sequencing. We have previously performed short-read target capture on these regions[12] and observed low sequence coverage in low GC targets. However, both Nanopore and PacBio WGS data do not have any bias in sequence coverage in low GC regions. Hence, the lack of sequence coverage in low GC region for targeted sequencing is likely due to the capture protocol. To overcome the issue of low capture efficiency for low GC regions, it is feasible to increase the number of probes in low GC regions during probe design. This will improve sequence enrichment in low GC targets. Furthermore, use of simulation tools[31,32] which can simulate sequencing data from probe sequences designed for capture sequencing can be used to assess the probe design efficiency prior to sequencing. This will allow to improve probe design in regions with low capture efficiency and subsequently improve coverage.

We also observed targets greater than 2 kb in length could not be genotyped due to the lack of spanning reads for genotyping analysis. This is primarily due to the limitation in sequence read length observed from the capture process. Streptavidin beads used during the capture process has limitations on the size of the fragments it can bind to, which limits the fragment length attainable with this capture protocol. Although there are longer TRs (greater than 2 kb) in the human genome, more than 99% of the TRs reported in human reference genome (hg38) are less than 2 kb in length[3]. Therefore, our protocol would still be able to successfully genotype most of the TRs in the human genome. TRs greater than 2 kb might need further optimized enrichment protocols.

Our target panel included eight (out of the 142 targets) STRs with longer expansions (>200 number of copies) and seven of these targets failed to genotype. However, three of these had low GC content and one was greater than 4 kb in repeat length. The longer expansions which failed to genotype also had low sequence coverage, however due to the low number of targets we could not conclusively identify the cause for failure for these targets.

We used VNTRTyper, an in-house genotyping tool described by Ganesamoorthy *et al*. (2018)[12] to determine the repeat number of TRs from long-read sequencing technologies. For comparison, we used Tandem-genotypes[24], recently reported genotyping tool for the detection of TR expansions from long-read sequencing. Both genotyping methods were comparable to PCR sizing analysis and genotyping estimates were comparable between the approaches. However, Tandem-genotypes genotyped fewer targets than VNTRTyper. The differences are likely due to the different algorithms used between the methods. Both VNTRTyper and Tandem-genotypes uses reads spanning the repeat region. However, for Tandem-genotypes the flanking length used for analysis is depended on the length of the repeat unit, with a maximum of 100 bp on both sides of the repeat unit. On the other hand, VNTRTyper uses a default 30 bp flanking length for analysis, but it is feasible to change the flanking length. Due to the longer flank length requirement, Tandem-genotypes could have possibly failed to genotype more targets compared to VNTRTyper.

Variations in TRs are a major source of genomic variation between individuals. TRs targeted in this study were initially selected due to the variation observed between case and control samples for obesity analysis[12] and these TRs are variable in the population. We show that approximately 50% of the targeted TRs differ from reported reference copy number. However, the major limitation in this analysis is that the sample size is small, and the individuals are related, which introduces a bias in the analysis. Nevertheless, these findings indicate the possibility of variation in TR copy number between individuals and further large-scale studies are required to ascertain the extent of variation.

We demonstrated that the accuracy of genotype estimates between WGS and targeted capture sequencing were comparable to the accuracy of PCR sizing analysis. However, targeted capture enrichment protocols used in this study have amplification steps, which can introduce errors in TR analysis. This could possibly explain the differences in genotype estimates observed between WGS and targeted capture sequencing for some targets.

An amplification free targeted analysis with long-read sequencing is an ideal option for accurate genotyping of TRs. Targeted cleavage with Cas9 enzyme followed by Nanopore sequencing[33] or PacBio sequencing[34] has been recently reported as alternative option for enrichment of regions of interest. This method does not have any amplifications and can be adapted for multiple targets in a single assay. However, currently the DNA input requirements are high and sequencing output are low, which currently restricts wide use of this technique for large-scale analysis.

Selective sequencing approaches utilising Nanopore real-time sequencing capabilities has been reported recently as an alternative approach to enrich regions of interest[35,36]. Selective sequencing works by mapping a section of the sequence read generated to the regions of interest and if the fragment matches to a region of interest, it will proceed with sequencing the fragment, if not the fragment is ejected from the pore. This approach will be a cost-effective approach to genotype TRs as it removes the need for specific sample preparation for target enrichment; however, the efficiency of this approach for TR analysis is yet to be determined.

The targeted long-read sequencing approach presented in this study is a cost-effective approach to analyse hundreds of TRs simultaneously. Long-read Nanopore WGS can cost approximately $4000 for 30X coverage of human genome and often with varying coverage across the genome. However, targeted long-read sequencing can be performed for a fraction of cost (less than $300 per sample depending on the multiplexing level) to enrich up to 25 Mb of genomic sequence of interest. The ability to analyse hundreds of TRs for a fraction of cost allows to explore TRs in large-scale studies.

In summary, we present a targeted approach combined with long-read sequencing to enable cost-effective and accurate approach to genotype TRs using long-read sequencing. Using this method, we have successfully demonstrated the feasibility of targeted capture sequencing of repetitive sequences and genotyping TRs using Nanopore long-read sequencing technology. Our targeted long-read sequencing approach would provide a cost-effective tool for large-scale population analysis of tandem repeats.

Data availability

Underlying data

NCBI BioProject: Capture Sequencing of Tandem Repeats. Accession number PRJNA422490, https://identifiers.org/ncbi/bioproject:PRJNA422490.

Extended data

Figshare: Supplementary Information for the "High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing" article. https://doi.org/10.6084/m9.figshare.12789278.v1[30].

This project contains the following extended data:

- Supplementary_Information,pdf, which contains the following:
  - Supplementary Figure 1. Sequence coverage distribution on targets vs GC%.
  - Supplementary Figure 2. Correlation of genotype estimates between PCR sizing and VNTRTyper.
  - Supplementary Figure 3. Correlation of genotype estimates between PCR sizing and Tandem-genotypes.
  - Supplementary Figure 4. VNTRTyper genotyping rate with 25% GC and 2Kb size threshold.
  - Supplementary Figure 5. Tandem-genotypes genotyping rate with 25% GC and 2Kb size threshold.
  - Supplementary Table 1. PCR primer sequences.
  - Supplementary Table 2. Targeted Sequencing metrics for Nanopore Capture sequencing of tandem repeats.
  - Supplementary Table 3. Genotype estimates on Nanopore targeted capture sequencing using Tandem-Genotypes.
- Supplementary_Spreadsheet_Table1.csv. (Genotype predictions on Nanopore Capture Sequencing data using VNTRtyper.)
- Supplementary_Spreadsheet_Table2.csv. (Genotype predictions on Nanopore Capture Sequencing data using Tandem-genotypes.)
- Supplementary_Spreadsheet_Table3.csv (Genotype predictions on NA12878 sample Nanopore WGS data and PacBio WGS data using VNTRTyper and Tandem-genotypes.)
- Supplementary_Information_PCR_data.pdf. (Capillary electrophoresis results of PCR sizing analysis.)

Acknowledgements

We thank Agilent Technologies for their technical assistance with targeted sequencing experiments.

References

1. Lander ES, Linton LM, Birren B, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature.* 2001; **409**(6822): 860–921.
PubMed Abstract | Publisher Full Text

2. Jurka J, Kapitonov VV, Kohany O, *et al.*: **Repetitive sequences in complex genomes: structure and evolution.** *Annu Rev Genomics Hum Genet.* 2007; **8**: 241–59.
PubMed Abstract | Publisher Full Text

3. Gelfand Y, Rodriguez A, Benson G: **TRDB--the Tandem Repeats Database.** *Nucleic Acids Res.* 2007; **35**(Database issue): D80–7.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Gemayel R, Cho J, Boeynaems S, *et al.*: **Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences.** *Genes (Basel).* 2012; **3**(3): 461–480.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Armour JA: **Tandemly repeated DNA: why should anyone care?** *Mutat Res.* 2006; **598**(1–2): 6–14.
PubMed Abstract | Publisher Full Text

6. Hannan AJ: **TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease.** *Discov Med.* 2010; **10**(53): 314–21.
PubMed Abstract

7. Gemayel R, Vinces MD, Legendre M, *et al.*: **Variable tandem repeats accelerate evolution of coding and regulatory sequences.** *Annu Rev Genet.* 2010; **44**: 445–477.
PubMed Abstract | Publisher Full Text

8. Bidwell JL, Bignon JD: **DNA-RFLP methods and interpretation scheme for HLA-DR and DQ typing.** *Eur J Immunogenet.* 1991; **18**(1–2): 5–22.
PubMed Abstract | Publisher Full Text

9. Tagliabracci A, Buscemi L, Sassaroli C, *et al.*: **Allele typing of short tandem repeats by capillary electrophoresis.** *Int J Legal Med.* 1999; **113**(1): 26–32.
PubMed Abstract | Publisher Full Text

10. Bahlo M, Bennett MF, Degorski P, *et al.*: **Recent advances in the detection of repeat expansions with short-read next-generation sequencing [version 1; peer review: 3 approved].** *F1000Res.* 2018; **7**: F1000 Faculty Rev-736.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Duitama J, Zablotskaya A, Gemayel R, *et al.*: **Large-scale analysis of tandem repeat variability in the human genome.** *Nucleic acids research.* 2014; **42**(9): 5728–5741.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Ganesamoorthy D, Cao MD, Duarte T, *et al.*: **GtTR: Bayesian estimation of absolute tandem repeat copy number using sequence capture and high throughput sequencing.** *BMC Bioinformatics.* 2018; **19**(1): 267.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Gymrek M, Golan D, Rosset S, *et al.*: **lobSTR: A short tandem repeat profiler for personal genomes.** *Genome Res.* 2012; **22**(6): 1154–62.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Highnam G, Franck C, Martin A, *et al.*: **Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles.** *Nucleic Acids Res.* 2013; **41**(1): e32.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Cao MD, Tasker E, Willadsen K, *et al.*: **Inferring short tandem repeat variation from paired-end short reads.** *Nucleic Acids Res.* 2014; **42**(3): e16.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Willems T, Zielinski D, Yuan J, *et al.*: **Genome-wide profiling of heritable and *de novo* STR variations.** *Nat Methods.* 2017; **14**(6): 590–592.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Dashnow H, Lek M, Phipson B, *et al.*: **STRetch: detecting and discovering pathogenic short tandem repeat expansions.** *Genome Biol.* 2018; **19**(1): 121.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Dolzhenko E, van Vugt JJF, Shaw RJ, *et al.*: **Detection of long repeat expansions from PCR-free whole-genome sequence data.** *Genome Res.* 2017; **27**(11): 1895–1903.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Mousavi N, Shleizer-Burko S, Yanicky R, *et al.*: **Profiling the genome-wide landscape of tandem repeat expansions.** *Nucleic Acids Res.* 2019; **47**(15): e90.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Schule B, McFarland KN, Lee K, *et al.*: **Parkinson's disease associated with pure *ATXN10* repeat expansion.** *NPJ Parkinsons Dis.* 2017; **3**: 27.
PubMed Abstract | Publisher Full Text | Free Full Text

21. De Roeck A, De Coster W, Bossaerts L, *et al.*: **Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION.** *Genome Biol.* 2019; **20**(1): 239.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Ebbert MTW, Farrugia SL, Sens JP, *et al.*: **Long-read sequencing across the *C9orf72* 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease.** *Mol Neurodegener.* 2018; **13**(1): 46.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Liu Q, Zhang P, Wang D, *et al.*: **Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing.** *Genome Med.* 2017; **9**(1): 65.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Mitsuhashi S, Frith MC, Mizuguchi T, *et al.*: **Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads.** *Genome Biol.* 2019; **20**(1): 58.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Karamitros T, Magiorkinis G: **Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure.** *Methods Mol Biol.* 2018; **1712**: 43–51.
PubMed Abstract | Publisher Full Text

26. Jain M, Koren S, Miga KH, *et al.*: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol.* 2018; **36**(4): 338–345.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Pendleton M, Sebra R, Pang AWC, *et al.*: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nat Methods.* 2015; **12**(8): 780–6.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Kielbasa SM, Wan R, Sato K, *et al.*: **Adaptive seeds tame genomic sequence comparison.** *Genome Res.* 2011; **21**(3): 487–93.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Ganesamoorthy D, Yan M, Murigneux V, *et al.*: **Supplementary Information for the "High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing" article.** *figshare* . Dataset. 2020.
http://www.doi.org/10.6084/m9.figshare.12789278.v1

31. Kim S, Jeong K, Bafna V: **Wessim: a whole-exome sequencing simulator based on *in silico* exome capture.** *Bioinformatics.* 2013; **29**(8): 1076–7.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Cao MD, Ganesamoorthy D, Zhou C, *et al.*: **Simulating the dynamics of targeted capture sequencing with CapSim.** *Bioinformatics.* 2018; **34**(5): 873–874.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Gilpatrick T, Lee I, Graham JE, *et al.*: **Targeted nanopore sequencing with Cas9-guided adapter ligation.** *Nat Biotechnol.* 2020; **38**(4): 433–438.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Hafford-Tear NJ, Tsai YC, Sadan AN, *et al.*: **CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated *TCF4* triplet repeat.** *Genet Med.* 2019; **21**(9): 2092–2102.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Payne A, Holmes N, Clarke T, *et al.*: **Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels.** *BioRxiv.* 2020; 2020.02.03.926956.
Publisher Full Text

36. Kovaka S, Fan Y, Ni B, *et al.*: **Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED.** *bioRxiv.* 2020; 2020.02.03.931923.
Reference Source

Open Peer Review

Current Peer Review Status: ? ✗

Version 1

Reviewer Report 21 September 2020

https://doi.org/10.5256/f1000research.28357.r70688

✗ **Mark T. W. Ebbert**
Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA

Ganesamoorthy *et al*. highlight an important issue and challenge in genotyping large tandem repeats (TR) that cannot be accurately genotyped using short-read sequencing data. This is an important issue because variations in TRs are known to cause many diseases, such as in repeat-expansion and repeat-retraction diseases. Important examples of repeat-expansion diseases include amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), Huntington's disease, Fuch's disease, muscular dystrophy, and many more. A prime example of a repeat-retraction disease is facioscapulohumeral muscular dystrophy (FSHD). Thus, being able to accurately genotype TRs is a critical issue for human health and disease research, as it will help understand disease etiology and how to diagnose and treat diseases.

While the questions being asked are important, there are unfortunately many limitations and issues with this particular study. My critique is honestly meant to be helpful to both the authors and the readers of this manuscript, and hope it will be received that way.

The major issues are as follows:

1. What may be the biggest issue is the methods used to answer the questions at hand. The authors set out to genotype 142 TRs of various sizes across seven individuals, ranging "from 112 to 25236 bp in length". This is a fantastic goal and I was excited to see the results. The first methodological issue is that the authors used amplification methods as part of the sequencing. I understand that the authors were seeking to maximize the number of targets they could include in their study based on costs, and to ensure deep coverage, but amplification simply will not work for long TRs (e.g., a TR that is 25236 nucleotides long), or those with high GC content; it should work fine for shorter and less GC-rich TRs, however. Surprisingly, the authors even acknowledge that the Agilent SureSelect protocol they employed "works effectively on fragments less than 4 kb in length", which clearly indicates that interrogating those ≥4kb will not work.

2. What surprised me most, however, was that the authors then intentionally sheared the amplicons to 3kb, making any attempt to sequence TRs >3kb a nonstarter—and realistically

making it difficult to sequence even TRs that approach 3kb, which the authors found in their results. Specifically, the authors found they were most successful sequencing and characterizing TRs <2kb.

This alone does not invalidate the utility for all of the authors' results, however; it simply limits their results to TRs <3kb (maybe <2kb). Thus, these results are still useful, but all downstream results and conclusions should be kept within these bounds.

Additional minor issues:

1. Early in the introduction, the authors make important points about repetitive sequences in the human genome. Most of the data they present, however, are extremely outdated, including data from the original human genome in 2001 and the number of TRs in the human genome from hg18 (2006). It's great to cite these early papers, but we now know that they are inaccurate because the reference genome has improved dramatically. Here are some specific points that need to be corrected in the introduction:
    1. Data from the original genome should not be stated as the current estimate for repeated sequences in the human genome, as was stated in the first sentence of the introduction. There are papers that are much more recent and that use a more updated reference genome.
    2. The authors make an important point about the number of TRs in the human genome, citing Gelfand *et al*. The data they present, however, is far outdated. The data from Gelfand *et al*. is from 2006 using the reference genome hg18. Reference genome hg18 is far too outdated to be used in a paper to be published in 2020. Most researchers are already moving past hg19 (to hg38). The authors need to update these statistics to represent data from hg38, which is actually to their advantage, as it will only further emphasize their point that TRs are plentiful in the human genome.
    3. Authors state that "...repeats with one to six basepair repeat units are classified as microsatellites or short tandem repeats (STRs) and those with more than six basepair repeat units are known as minisatellites", citing Gemayel *et al.* The paper by Gemayel *et al*., however, specifically states that microsatellites are between 1 and 10 nucleotides and that minisatellites are >10. After reading a bit more, I see that there is some discrepancy on the exact cutoff for each group, but the authors need to be clearer about this. It is certainly not accurate to state a cutoff of six, citing a paper that clearly states a different cutoff. Authors should clarify and cite additional papers on the matter—perhaps including a more recent publication, though that may not be required.

2. Development for Albacore ended over 1.5 years ago (last release was January 2019), and the authors used a version that is almost 2 years old (2.2.7 was released in October 2018). Software this old may not normally be an issue, but the technology and algorithms used for Nanopore sequencing have evolved so rapidly over the past two years that it may be important. The authors need to do redo analyses using a current version of Guppy (the current basecaller).

3. The authors use two different aligners (Minimap2 and LAST) for genotyping with VNTRTyper and Tandem-genotypes. It appears they did so because authors of Tandem-genotypes recommends LAST for their genotyper. I also assume that VNTRTyper works best with minimap2. It is reasonable to use different aligners if they have been validated for a given pipeline, but the authors should clarify in the methods why they used different aligners

rather than simply stating that they did.

4. Authors mention optimizing the PCR conditions for each TR, but I do not see specific details for these optimizations. If they are present in supplementary material, they should be clearly referenced in the manuscript.

5. Based on the methods, it is unclear whether the authors are only using reads that fully span the TRs to genotype (i.e., reads that contain adjacent sequence on both ends of the TR). It would not make sense to include any reads that do not fully span the TR. Authors should also make it clear how they determine whether reads span the TR.

6. Methods for how authors performed most correlations are unclear. For example, for correlations to PCR, did they perform PCR for each sample across all of the TRs tested? i.e., how many data points are even included?

7. Figures 3 & 4 misspelled 'genome' in the figure title.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, long-read sequencing, statistics, TRs, bioinformatics, computational biology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 17 September 2020

https://doi.org/10.5256/f1000research.28357.r70683

?

Rick M. Tankard

Mathematics and Statistics, Murdoch University, Murdoch, WA, Australia

This is an important work in sizing tandem repeats from long-read sequencing data in a cost-effective manner.

Some more STR analysis tools should be cited for "Several computational tools have been developed to improve the accuracy of TR genotyping from short-read sequencing data with varying performance", being exSTRa[3], ExpansionHunter Denovo[2], TREDPARSE[4] and TRhist[1]. The sentence “Yet, most of these tools have focused mainly on the analysis of STRs and analysis of longer TRs remains a hurdle for these approaches.”, it may give the impression that these tools cannot deal with repeat units longer than 6 bp (as defined for STRs in the Introduction), though both GangSTR and ExpansionHunter Denovo deal with up to 20 bp repeat units by default (adjustable with ExpansionHunter Denovo at run time, though there may not be evidence this will give good results).

Regarding reproducibility, the data for comparisons is available. It is not apparent that the software VNTRTyper is available for use, being labelled as an in-house tool. This will make it difficult for others to verify the performance of VNTRTyper on other data. The availability of VNTRTyper would allow other researchers to make use of this work on their own data.

In Tables 1 and 2, I would appreciate a Root Mean Square Error (RMSE) between the PCR genotypes and the genotypes of other methods to get a sense of the scale of errors. Similarly, for Figure 3, being careful this isn't the RMSE of the linear model.

Overall, I would recommend accepting the article with some small changes.

Minor:
Typo in some figure captions: ‘genme’

References

1. Doi K, Monjo T, Hoang PH, Yoshimura J, et al.: Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing.*Bioinformatics*. 2014; **30** (6): 815-22 PubMed Abstract | Publisher Full Text
2. Dolzhenko E, Bennett M, Richmond P, Trost B, et al.: ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biology*. 2020; **21** (1). Publisher Full Text
3. Tankard R, Bennett M, Degorski P, Delatycki M, et al.: Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *The American Journal of Human Genetics*. 2018; **103** (6): 858-873 Publisher Full Text
4. Tang H, Kirkness EF, Lippert C, Biggs WH, et al.: Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes.*Am J Hum Genet*. 2017; **101** (5): 700-715 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: Author of the repeat expansion detection method exSTRa [3] for Illumina Whole Genome Sequencing (WGS). This is not a direct competitor as this relied on short-read data instead of long-read data, and exSTRa focused on short tandem repeats (STRs) with much shorter repeat units than presented here.

Reviewer Expertise: bioinformatics, statistics, genomics, epigenetics, repeat expansions

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author/s:

Ganesamoorthy, D; Yan, M; Murigneux, V; Zhou, C; Cao, MD; Duarte, TPS; Coin, LJM

Title:

High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing

Date:

2020-09-02

Citation:

Ganesamoorthy, D., Yan, M., Murigneux, V., Zhou, C., Cao, M. D., Duarte, T. P. S. & Coin, L. J. M. (2020). High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing. F1000Research, 9, pp.1084-1084. https://doi.org/10.12688/f1000research.25693.1.

Persistent Link:

http://hdl.handle.net/11343/280242

File Description:

Published version

License:

CC BY