High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing [version 1; peer review: awaiting peer review]

Devika Ganesamoorthy¹², Mengjia Yan¹, Valentine Murigneux¹, Chenxi Zhou¹², Minh Duc Cao¹, Tania P. S. Duarte¹, Lachlan J. M. Coin¹²

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

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.
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.

Copyright: © 2020 Ganesamoorthy D et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

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: awaiting peer review] 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. 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. 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. 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.

TRs are prone to high rates of 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. 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. 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 or PCR amplification of the target loci followed by fragment length analysis. 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. Our group and others have also demonstrated targeted sequencing approaches using short-read sequencing for TR analysis. Several computational tools have been developed to improve the accuracy of TR genotyping from short-read sequencing data with varying performance. 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., 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, 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. Genotyping tools utilizing long-read sequencing data, such as Nanosatellite, RepeatHMM and Tandem-genotypes 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., 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. 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; 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.

Methods

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. Briefly, 142 TRs were selected; they range from 112 to 25236 bp in length in the human reference 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 NA12890). 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 MiniION 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/ref_3.4.md)26. PacBio WGS data on NA12878 sample were downloaded from SRA with accession numbers SRX627421 and SRX63831027.

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-ph’ 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 data29 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 v95930. Calculation of repeat length per sequencing read was performed with Tandem-genotypes as reported by Mitsuhashi et al.31. 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 study12 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 data30.

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 VNTRtypert) 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).

Extended data, Supplementary Table 2 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 VNTRtypert with the improvements described in the Methods. We also applied Tandem-genotypes 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 VNTRtypert 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). 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); however, fewer targets had sufficient data to compare with PCR sizing results.

Genotype estimates from VNTRtypert 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.](image) (a) Read length distribution of Nanopore targeted sequencing. Lines indicates the read length distribution 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.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Method</th>
<th>Genotype of Target*</th>
<th>Pearson correlation with PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TR_8 (12.0)</td>
<td>TR_57 (15.6)</td>
</tr>
<tr>
<td>NA12877</td>
<td>PCR</td>
<td>12.0/13.0</td>
<td>10.6/13.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>12.0/13.0</td>
<td>ND</td>
</tr>
<tr>
<td>NA12878</td>
<td>PCR</td>
<td>12.0/12.0</td>
<td>10.6/12.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>ND</td>
<td>11.0/12.1</td>
</tr>
<tr>
<td>NA12879</td>
<td>PCR</td>
<td>12.0/13.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>12.4/12.4</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td>NA12881</td>
<td>PCR</td>
<td>12.0/12.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>12.0/12.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td>NA12882</td>
<td>PCR</td>
<td>12.0/13.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>11.8/13.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td>NA12889</td>
<td>PCR</td>
<td>12.0/12.0</td>
<td>13.6/17.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>12.1/12.1</td>
<td>10.6/15.8</td>
</tr>
<tr>
<td>NA12890</td>
<td>PCR</td>
<td>12.0/13.0</td>
<td>10.6/10.6</td>
</tr>
<tr>
<td></td>
<td>Nanopore</td>
<td>12.2/12.2</td>
<td>10.6/10.6</td>
</tr>
</tbody>
</table>

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

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 insufficient sequence enrichment in these regions (Extended data, Supplementary Figure 1). 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). 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).

**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. 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
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.

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

---

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

<table>
<thead>
<tr>
<th>Method</th>
<th>TR_8 (12.0)</th>
<th>TR_32 (38.4)</th>
<th>TR_57 (15.6)</th>
<th>TR_64 (18.3)</th>
<th>TR_86 (2.0)</th>
<th>TR_87 (9.0)</th>
<th>TR_93 (2.0)</th>
<th>TR_109 (15.3)</th>
<th>TR_112 (6.0)</th>
<th>TR_120 (2.2)</th>
<th>Pearson Correlation with PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>12.0/12.0</td>
<td>9.4/10.4</td>
<td>10.6/12.6</td>
<td>17.3/18.3</td>
<td>2.0/2.0</td>
<td>6.0/9.0</td>
<td>2.0/2.0</td>
<td>15.3/17.3</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9980</td>
</tr>
<tr>
<td>NPW_VT</td>
<td>12.0/12.0</td>
<td>8.2/10.7</td>
<td>10.6/12.6</td>
<td>17.3/18.6</td>
<td>2.0/2.0</td>
<td>6.0/9.0</td>
<td>2.0/2.0</td>
<td>15.3/17.3</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9980</td>
</tr>
<tr>
<td>NPW_TG</td>
<td>11.0/12.0</td>
<td>4.4/9.4</td>
<td>10.6/12.6</td>
<td>18.3/19.3</td>
<td>2.0/2.0</td>
<td>6.0/9.0</td>
<td>2.0/2.0</td>
<td>15.3/17.3</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9957</td>
</tr>
<tr>
<td>PBW_VT</td>
<td>12.0/12.0</td>
<td>11.0/12.4</td>
<td>10.6/12.6</td>
<td>18.3/19.3</td>
<td>2.0/2.0</td>
<td>6.0/9.0</td>
<td>2.0/2.0</td>
<td>15.3/17.3</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9990</td>
</tr>
<tr>
<td>PBW_TG</td>
<td>12.0/12.0</td>
<td>ND</td>
<td>12.6/12.6</td>
<td>ND</td>
<td>2.0/2.0</td>
<td>ND</td>
<td>2.0/2.0</td>
<td>15.3/15.3</td>
<td>4.0/4.0</td>
<td>2.2/2.2</td>
<td>0.99</td>
</tr>
<tr>
<td>NPC_VT</td>
<td>11.0/12.1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>2.0/2.0</td>
<td>6.0/8.8</td>
<td>2.0/2.0</td>
<td>14.8/17.1</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9978</td>
</tr>
<tr>
<td>NPC_TG</td>
<td>11.0/12.0</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>2.0/2.0</td>
<td>6.0/9.0</td>
<td>ND</td>
<td>15.3/17.3</td>
<td>3.0/4.0</td>
<td>2.2/2.2</td>
<td>0.9988</td>
</tr>
</tbody>
</table>

^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 genome 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).
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 regions 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 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 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. 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) to determine the repeat number of TRs from long-read sequencing technologies. For comparison, we used Tandem-genotypes, 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 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 or PacBio sequencing 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. 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 analyze 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 analyze 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

Extended data

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.)

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

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

References

The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com