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

Devika Ganesamoorthy[1#], Mengjia Yan[1], Valentine Murigneux[1], Chenxi Zhou[1], Minh Duc Cao[1], Tania PS Duarte[1] and Lachlan JM Coin[1#].

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

[#] Correspondence to d.ganesamoorthy@imb.uq.edu.au or l.coin@imb.uq.edu.au

Devika Ganesamoorthy	d.ganesamoorthy@imb.uq.edu.au
Mengjia Yan	mengjia.y2@gmail.com
Valentine Murigneux	v.murigneux@imb.uq.edu.au
Chenxi Zhou	chenxi.zhou@imb.uq.edu.au
Minh Duc Cao	minhduc.cao@gmail.com
Tania PS Duarte	t.duarte@imb.uq.edu.au
Lachlan JM Coin	l.coin@imb.uq.edu.au

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, TRs are not been explored in the context of genomic variation, due to the limitations of existing tools, which are either low-throughput or restricted to a small subset of TRs. Here, we report a targeted sequencing approach combined with long-read sequencing technologies to overcome these limitations. Our approach described here presents an accurate and cost-effective method to explore TRs at the population level.

Results:

We report targeted enrichment of 142 TR loci using sequence capture combined with Nanopore and PacBio sequencing. We achieved 50-60% on-target sequence reads and an average of 100X sequence coverage per sample with Nanopore sequencing using 200ng of input DNA. Genotyping estimates from targeted long-read sequencing were equivalent to the accuracy of gold standard PCR sizing analysis with correlation greater than 0.98 for both Nanopore and PacBio sequencing. Genotyping estimates from targeted long-read sequencing were determined using two different methods (VNTRTyper and Tandem-genotypes) and results were comparable. Furthermore, genotyping estimates from targeted long-read sequencing were highly correlated with genotyping estimates 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

BACKGROUND

Repeated sequences occur in multiple copies throughout the genome, and they make up almost half of the human genome [1]. Repeat sequences can be divided into two categories, namely 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 less than six base pair repeat units are classified as microsatellites or short tandem repeats (STRs) and those with more than 6 base pair repeat units are known as minisatellites [4]. TRs that have variable copy numbers are known as variable number tandem repeats (VNTRs).

VNTRs 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 VNTRs 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

69 (RFLP) analysis [8] or PCR amplification of the target loci followed by fragment length
70 analysis [9]. These techniques are only applicable to a specific target region and not scalable
71 to high-throughput analysis, which limits the possibility of genome-wide TR analysis. In the
72 recent decade, significant progress has been made in utilising high throughput short-read
73 sequencing data for genotyping STRs [10]. Our group and others have also demonstrated
74 targeted sequencing approaches using short-read sequencing for TR analysis [11, 12]. Several
75 computational tools have been developed to improve the accuracy of TR genotyping from
76 short-read sequencing data with varying performance [13-17]. Yet, most of these tools have
77 focused mainly on the analysis of STRs and analysis of longer TRs remains a hurdle for these
78 approaches. We reported GtTR in Ganesamoorthy *et. al.* (2018) [12], which utilizes short-read
79 sequencing data to genotype longer TRs. GtTR reports absolute copy number of the TRs, but
80 it does not report the exact genotype of two alleles due to the use of short-read sequencing data.
81
82 Sequencing reads that span the entire repeat region are informative to accurately genotype TRs
83 [11], and therefore are ideal for genome-wide TR analysis. Long-read sequencing technologies
84 have the potential to span all TRs in human genome, including long TRs. There have been few
85 reports on the use of long-read sequencing for the analysis of specific TRs implicated in
86 diseases [18-20]. Genotyping tools utilizing long-read sequencing data, such as
87 Nanosatellite[19], RepeatHMM [21] and Tandem-genotypes [22] have been reported in the
88 recent years with varying performance across different length of repeat units and repeat length.
89 We reported VNTRTyper in Ganesamoorthy *et. al.* (2018) [12] to genotype TRs from long-read
90 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.

In this study, we demonstrate a targeted sequence capture approach combined with long-read sequencing for large-scale analysis of TRs. We used a hybridization based target enrichment system to enrich regions of interest followed with long-read sequencing [23]. 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 both Oxford Nanopore and PacBio long-read sequencing technologies. We demonstrate the accuracy of genotype estimates from targeted long-read sequencing by comparison with gold-standard PCR sizing analysis. Our targeted long-read sequencing approach presented in this study provides an accurate and cost effective approach for large-scale analysis of TRs.

RESULTS

We demonstrate a targeted sequence capture approach combined with long-read sequencing to genotype hundreds of TRs. We evaluated this approach using both Nanopore and PacBio long-read sequencing technologies.

Targeted Capture Sequencing of Tandem repeats

We performed sequence enrichment of targeted TRs followed by Nanopore or PacBio long-read sequencing as described in Methods. Table 1 summarises the metrics for targeted

113 sequencing on both Nanopore and PacBio sequencing technologies. Sequencing yield per
114 sample was lower for Nanopore compared to PacBio due to multiplexing of samples. Nanopore
115 multiplexing (See Methods) group 1 samples had similar yield between samples, however,
116 Nanopore multiplexing group 2 samples had varying yield per sample. Despite the differences
117 in sequencing yield, on target capture rate was approximately 50% for Nanopore sequencing,
118 however only 8% of reads were on target for PacBio sequence capture. Hence, the fold
119 enrichment of targets were lower for PacBio sequence capture compared to Nanopore capture.
120 It is worth noting that, the sequencing capture protocols used were different between Nanopore
121 and PacBio sequencing as described in Methods.
122
123 Figure 1a shows the read length distribution observed in targeted capture sequencing data using
124 Nanopore and PacBio sequencing technologies. The median read length was higher for PacBio
125 sequence capture protocol compared to Nanopore capture sequencing protocol. However, it is
126 worth noting that lengths of targeted TR range from 112bp to 25236bp, with a median length
127 of 667bp (Additional File 1: Figure 1). The read length in this study was sufficient to analyse
128 majority of targeted TRs. Figure 1b shows the sequence coverage (*i.e.* average per base
129 coverage in the region) distribution in targeted TR regions for Nanopore and PacBio
130 sequencing. On average 80X and 100X sequence coverage was achieved for PacBio and
131 Nanopore sequencing respectively. Furthermore, 8% and 14% of targets had less than 1X
132 coverage in PacBio and Nanopore capture sequencing respectively and majority of these targets
133 had less than 25% GC content (Additional File 1: Figure 2).
134

135 *Genotyping of Tandem repeats using targeted long-read sequencing*

136 Genotype estimates from targeted long-read sequencing datasets were estimated using our tool
137 VNTRTyper [12] with the improvements described in Methods. We also applied Tandem-
138 genotypes [22] for genotyping TRs from long-read sequencing data. We used a minimum of
139 two reads as read threshold to determine the repeat number for each alleles.

140

141 Genotype estimates by VNTRTyper were compared to the PCR sizing results on 8 targets which
142 had sufficient coverage (Table 2). Genotype estimates from both Nanopore and PacBio targeted
143 sequencing correlated well with PCR results and the correlation values were greater than 0.980
144 for all samples. Genotype estimates by Tandem-genotypes were also compared to the PCR
145 sizing results (Additional File 1: Table 1) and correlation values were greater than 0.984 for all
146 samples, however fewer targets had sufficient data to compare with PCR sizing results.

147

148 The three samples (NA12877, NA12878 and NA12879) which were sequenced using both
149 Nanopore and PacBio sequencing technologies were used to assess the performance of different
150 sequencing technologies. Both VNTRTyper and Tandem-genotypes correlated well with PCR
151 sizing results with correlation values greater than 0.984 for both sequencing technologies
152 (Table 2 and Additional File 1: Table 1). However, Nanopore targeted sequencing had
153 marginally better correlation than PacBio targeted sequencing (Additional File 1: Figure 3 and
154 4).

155

156 We were able to determine the genotype on 60% of the targets (range 48% to 75%) with
157 Nanopore targeted sequencing and on 54% of the targets (range 54% to 55%) with PacBio
158 targeted sequencing using VNTRTyper. On the other hand, Tandem-genotypes was only able
159 to genotype 51% of the targets (range 34% to 70%) with Nanopore sequencing and 45% of the
160 targets (range 44% to 46%) with PacBio sequencing. Both genotyping approaches failed to
161 genotype targets with low GC sequence content (< 25% GC content) and targets which are
162 greater than 2Kb in length, which accounts for 20% of the targets. Targets with low GC
163 sequence content (< 25% GC content) didn't have sufficient sequence coverage for analysis
164 due to inefficient sequence enrichment in these regions (Additional File 1: Figure 2). Targets
165 which are greater than 2Kb in length didn't have sufficient spanning reads for genotyping
166 analysis (See Methods and Additional File 1: Figure 1).
167
168 Tandem-genotypes had lower genotyping rate compared to VNTRTyper and this is likely due
169 to the difference in algorithms. Targets which were only genotyped by VNTRTyper, had less
170 than 5 reads supporting the alleles, which indicates that the coverage in these targets were not
171 sufficient for Tandem-genotypes analysis. Despite these differences, genotype estimates by
172 VNTRtyper and Tandem-genotypes correlate well and the correlation values range from 0.950
173 to 0.995 for Nanopore (except 0.904 for sample NA12881, which had low sequence coverage)
174 and 0.982 to 0.999 for PacBio targeted sequencing
175

176 Genotype estimates from VNTRTyper and Tandem-genotypes for all 142 targets from
177 Nanopore capture sequencing samples and PacBio capture sequencing samples are provided in
178 Additional File 2: Table 1 and Table 2 respectively.
179
180 *Genotyping of Tandem repeats using long-read whole genome sequencing*
181 To investigate the accuracy of genotype estimates of TRs from targeted sequence capture
182 compared to WGS, we performed genotyping analysis on targeted regions using VNTRTyper
183 and Tandem-genotypes on whole genome long-read sequencing data. We downloaded whole
184 genome long-read Nanopore and PacBio sequencing data on CEPH Pedigree 1463 NA12878
185 sample. We have previously reported genotyping estimates by VNTRTyper on PacBio
186 NA12878 WGS data [12]. Here we use the genotype estimates by VNTRTyper on PacBio
187 NA12878 WGS data to compare genotype estimates by Tandem-genotype and the results of
188 PacBio targeted sequencing analysis.
189
190 We compared the accuracy of genotype estimates from WGS data with PCR sizing analysis.
191 Genotype estimates by VNTRTyper and Tandem-genotypes on WGS data were compared with
192 PCR sizing results on 10 targets (Table 3). VNTRTyper and Tandem-genotypes had comparable
193 correlation with PCR sizing analysis for both Nanopore and PacBio WGS.
194
195 We also compared genotype estimates from targeted capture sequencing data and WGS data
196 with PCR sizing analysis on 7 targets which had sufficient coverage (Figure 2 and Table 3).
197 Nanopore WGS data correlated precisely with 13/14 genotype estimates on Nanopore capture

198 sequencing and 12/14 genotype estimates correlated accurately between PacBio WGS and
199 capture sequencing using VNTRTyper. Genotype estimates using Tandem-genotypes on
200 Nanopore WGS data correlated precisely with 9/10 genotype estimates on Nanopore capture
201 sequencing and 10/12 genotype estimates correlated accurately between PacBio WGS and
202 capture sequencing. Genotype estimates for all 142 targets from Nanopore and PacBio WGS
203 data determined using VNTRTyper and Tandem-genotypes are provided in Additional File 2:
204 Table 3.
205
206 We compared the genotype estimates on all targets between WGS data and targeted capture
207 sequencing data for both Nanopore and PacBio sequencing. Genotype estimates by
208 VNTRTyper correlated well between WGS data and targeted capture sequencing data with a
209 correlation of 0.9782 and 0.9378 for Nanopore and PacBio sequencing respectively (Figure 3).
210 Genotype estimates by Tandem-genotypes correlated well for PacBio WGS data and targeted
211 capture sequencing with a correlation value of 0.9954, however had low correlation (r^2=0.8618)
212 for Nanopore WGS and targeted capture sequencing data (Additional File 1: Figure 5).
213
214 *Variation in Tandem repeats*
215 To assess the extent of variation in repeat numbers between individuals, we compared the
216 genotype estimates to the reported reference (hg19) repeat number. Genotype estimates
217 determined by VNTRTyper on Nanopore capture sequencing on 7 members of CEPH pedigree
218 1463 were used to assess the variation. On average, 49% (range 40% - 55%) of the targets were
219 similar to the reported reference genotype and 51% (range 45% - 60%) of the targets differed

220 from the reference (Figure 4). Furthermore, more deletions (28%) than duplications (23%)
221 were observed in repeat number.
222
223 **DISCUSSION**
224 In this study, we presented a targeted sequencing approach combined with long-read
225 sequencing technology to genotype TRs. To our knowledge, this is the first report on
226 genotyping analysis of hundreds of TRs using targeted long-read sequencing approach.
227 Sequencing reads that span the entire repeat region and flanking region are often useful in
228 providing an accurate estimation of the repeat genotype. Long-read sequencing technologies
229 have the ability to generate reads which can span the entire repeat region and flanking regions.
230 However, whole genome long-read sequencing analysis is still expensive for large-scale
231 population analysis; hence, we developed a targeted long-read sequencing approach for TR
232 analysis.
233
234 We showed that 1) target enrichment of repetitive sequences followed by long-read sequencing
235 is feasible, 2) both Nanopore and PacBio long-read sequencing technologies can be utilised to
236 genotype TRs and 3) genotype predictions on targeted TR sequencing are comparable to the
237 accuracy of PCR sizing analysis of repeats. We modified target enrichment protocol between
238 Nanopore and PacBio sequencing technologies to minimize sequencing cost and to achieve
239 required input DNA requirements (See Methods). Hence, there were differences in sequencing
240 output between Nanopore and PacBio targeted sequencing experiments. On target capture was
241 less than 10% for PacBio sequencing, but greater than 50% for Nanopore sequencing. However,

242 median read length was greater than 1.1Kb for PacBio but less than 770bp for Nanopore
243 sequencing. These differences are likely due to the difference in probe hybridization conditions
244 and reagents between Nanopore and PacBio capture experiments. Despite these biases,
245 estimated genotypes from both Nanopore and PacBio capture sequencing were highly
246 concordant.
247
248 On average, 60% of the targets could be genotyped by targeted long-read sequencing analysis.
249 However, this improved to 75% with increased sequence output, indicating more sequencing
250 could improve genotyping rate. We also observed targets greater than 2Kb in length could not
251 be genotyped due to the lack of spanning reads for genotyping analysis. This is primarily due
252 to the limitation in sequence read length observed from the capture process. Streptavidin beads
253 used during the capture process has limitations on the size of the fragments it can bind to, which
254 limits the fragment length attainable with the capture protocol. Furthermore, we noticed targets
255 with low GC sequence content (< 25% GC content) didn't have sufficient sequence coverage
256 in both Nanopore and PacBio sequencing experiments. We have previously performed short-
257 read target capture on these regions [12] and did not observe any significant bias in sequence
258 coverage due to GC sequence content. Furthermore, both Nanopore and PacBio WGS data
259 didn't have any bias in sequence coverage in low GC regions. Hence, the lack of sequence
260 coverage in low GC region for targeted sequencing is likely due to the capture protocol used
261 in this study.
262

263 We used VNTRTyper, an in-house genotyping tool described in Ganesamoorthy et. al. (2018)
264 [12] to determine the repeat number of TRs from long-read sequencing technologies. For
265 comparison, we used Tandem-genotypes [22], recently reported genotyping tool for the
266 detection of TR expansions from long-read sequencing. Both genotyping methods were
267 comparable to PCR sizing analysis and genotyping estimates were comparable between the
268 approaches. However, Tandem-genotypes genotyped less targets than VNTRTyper. The
269 differences is likely due to the different algorithms used between the methods. Both
270 VNTRTyper and Tandem-genotypes uses reads spanning the repeat region. However, for
271 Tandem-genotypes the flanking length used for analysis is depended on the length of the repeat
272 unit, with a maximum of 100bp on both sides of the repeat unit. On the other hand, VNTRTyper
273 uses a default 30bp flanking length for analysis, but it is feasible to change the flanking length.
274 Due to the longer flank length requirement, Tandem-genotypes could have possibly failed to
275 genotype more targets compared to VNTRTyper.
276
277 Variations in TRs are a major source of genomic variation between individuals. TRs targeted
278 in this study were initially selected due to the variation observed between case and control
279 samples for obesity analysis [12] and these TRs are variable in the population. We show that
280 approximately 50% of the targeted TRs differ from reported reference copy number. However,
281 the major limitation in this analysis is that the sample size is small and the individuals are
282 related, which introduces a bias in the analysis. Nevertheless, these findings indicate the
283 possibility of variation in TR copy number between individuals and further large-scale studies
284 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 TRs. Targeted cleavage with Cas9 enzyme followed by Nanopore sequencing [24] or PacBio sequencing [25] 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.

CONCLUSIONS

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 both Nanopore and PacBio long-read sequencing technologies. Our targeted long-read sequencing approach would provide a cost-effective tool for large-scale population analysis of tandem repeats.

307

308 **METHODS**

309 *Samples for sequencing:*

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

313

314 *Selection of Tandem Repeats and Probe design*

315 The selection of TRs and design of probes were described in Ganesamoorthy *et. al*. (2018) [12].
316 Briefly, 142 TRs were selected; they range from 112bp to 25236bp in length in the reference
317 human genome (hg19) and the number of repeat units range from 2 to 2300 repeats. Agilent
318 SureSelect DNA design was used to design target probes to capture the targeted regions
319 (including 100bp flanking regions) and regions flanking the TRs (at least 1000bp).

320

321 *Nanopore targeted sequencing of TRs:*

322 All 7 family members from the CEPH pedigree 1463 were used for Nanopore targeted
323 sequencing analysis (NA12877, NA12878, NA12879, NA12881, NA12882, NA12889 and
324 NA1289). Target sequence capture for Nanopore sequencing was performed using Agilent
325 SureSelect XT HS Target Enrichment System according to the manufacturer's instructions with
326 slight modifications. Briefly, 200ng of DNA was fragmented to 3Kb using Covaris Blue
327 miniTUBE. Greater than 90% of the targeted TRs are less than 3Kb and SureSelect capture
328 protocol works effectively on fragments less than 4Kb in length; therefore, DNA products were

329 sheared to 3Kb. Fragmented DNA was end repaired, adapter ligated and amplified prior to
330 target capture. Extension time for pre-capture amplification was increased to 4 minutes to allow
331 for the amplification of long fragments and 14 cycle amplification was used. Purified pre-
332 capture PCR products were hybridized to the designed capture probes for 2 hours. Streptavidin
333 beads were used to pull down the DNA fragments bound to the probes. Finally, captured DNA
334 was amplified with long extension time (4 minutes) using Illumina Index adapters provided in
335 the enrichment kit. Post capture PCR products were purified using 0.8X - 1X Agencourt
336 AMPure XP beads.
337
338 Nanopore sequencing library preparation was performed using 1D Native barcoding genomic
339 DNA (with EXP-NBD103 and SQK-LSK108) protocol according to the manufacturer's
340 instructions with minor modifications. Briefly, 100ng – 200ng of post capture PCR products
341 were end repaired and incubated at 20℃ for 15mins and 65℃ for 15mins. End repaired products
342 were ligated with unique native barcodes. Purification steps after end repair and barcode
343 ligation were avoided to minimize the loss of DNA. Barcoded samples were pooled in
344 equimolar concentrations prior to adapter ligation. Adapter ligated samples were purified using
345 0.4X Agencourt AMPure XP beads. Samples were split into 2 sequencing groups; NA12877,
346 NA12878, NA12879 and NA12890 – group 1; NA12881, NA12882 and NA12889 – group 2.
347 Sequencing was performed using Oxford Nanopore Technologies MinION sequencer using
348 R9.5 flow cell. Both groups were sequenced for 48 hours. Nanopore sequencing data were base
349 called using Albacore (version 2.2.7) and reads were demultiplexed using Albacore (version
350 2.2.7) based on the barcode sequences.

351

352 *PacBio targeted sequencing of TRs:*

353 Three family members from the CEPH pedigree 1463 were selected for PacBio targeted
354 sequencing analysis (NA12877, NA12878 and NA12879). Target capture process for PacBio
355 sequencing was similar to Nanopore target capture process as described above. However, to
356 facilitate high input requirement for PacBio sequencing a high input target capture protocol
357 was used. Agilent SureSelect XT Target Enrichment System was used according to the
358 manufacturer's instructions with slight modifications. Briefly, 3000ng of DNA was used as
359 input and fragmented to 3Kb using Covaris Blue miniTUBE. Fragmented DNA was end
360 repaired, adapter ligated and amplified prior to target capture. Purified pre-capture PCR
361 products were hybridized to the designed capture probes for 24 hours. Streptavidin beads were
362 used to pull down the DNA fragments bound to the probes. Captured DNA was amplified using
363 InPE 1.0 short and InPE 2.0 short primers provided by Agilent Technologies. Amplified
364 products (approximately 1000ng) were used as input for PacBio library preparation.

365

366 PacBio sequencing library preparation was performed using PacBio® Template Prep Kit
367 protocol according to the manufacturer's instructions. Briefly, amplified DNA fragments were
368 end repaired, blunt adapter ligated and SMRTbell templates were prepared. Purified SMRTbell
369 libraries were sequenced on PacBio RS sequencing system using one SMRTcell per sample.
370 Sequencing data were processed using SMRTportal and sub reads were extracted for
371 downstream analysis.

372

373 *Public data used in the study:*

374 Nanopore WGS data on CEPH Pedigree 1463 sample NA12878 were obtained from Nanopore
375 WGS consortium [26]. PacBio WGS data on NA12878 sample were downloaded from SRA
376 with accession numbers SRX627421 and SRX638310 [27]

377

378 *VNTRTyper:*

379 Both Nanopore and PacBio sequencing data were mapped to hg19 reference genome using
380 Minimap2 (version 2.13) [28]. For Nanopore sequencing reads '-ax map-ont' and for PacBio
381 sequencing reads '-ax map-pb' parameters were used. VNTRTyper, our in-house tool described
382 in Ganesamoorthy *et. al.* (2018) [12] was used to genotype TRs from long-read sequencing
383 data. Briefly, VNTRTyper takes advantage of the long-read sequencing to identify the number
384 of repeat units in the TR regions. Firstly, the tool identifies reads that span the repeat region
385 and applies Hidden Markov Models (HMM) to align the repetitive portion of each read to the
386 repeat unit. Then it estimates the multiplicity of the repeat units in a read using a profile HMM.
387 Recently, we further improved the accuracy of genotyping estimates by clustering the copy
388 number counts from reads to identify the likely genotypes per target. A minimum threshold of
389 two supporting reads per genotype was used to assign genotypes. Furthermore, for
390 heterozygous alleles, both alleles should have at least 10% of reads supporting the allele, if not
391 allele with less than 10% of reads was excluded during the analysis.

392

393 The updated version of VNTRTyper can be accessed from https://github.com/mdcao/japsa
394 release 1.9-3c and can be deployed using script name jsa.tr.longreads. Details of VNTRTyper

analysis are previously reported in Ganesamoorthy *et. al.* (2018) [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 [22] 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 in Mitsuhashi et. al. (2019) [22]. 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:

Ten targeted VNTR regions which are less than 1Kb in repetitive sequence were validated by PCR sizing analysis in this study. These ten targets include various repeat unit length and repeat sequence combinations to assess the accuracy of the genotypes determined from sequencing data. Majority of these targets were tested in our previous study [12] and the results from the previous PCR anlaysis were used for these regions. PCRs were performed using HotStar Taq DNA Polymerase (Qiagen) and PCR conditions were optimized for each PCR target. PCR

417 products were purified and subjected to capillary electrophoresis on an ABI3500xL Genetic
418 Analyzer (Applied BioSystems). Fragment sizes were analyzed using GeneMapper 4.0
419 (Applied BioSystems).
420
421 *Statistical analysis:*
422 Linear regression analysis was used to determine correlation between genotype estimates. All
423 plots were generated using GraphPad Prism version 7.00 for Windows, GraphPad Software,
424 La Jolla California USA, www.graphpad.com
425
426 **LIST OF ABBREVIATIONS**
427 TR – Tandem Repeats
428 STR _Short Tandem Repeats
429 VNTR – Variable Number of Tandem Repeats
430 PCR – Polymerase Chain Reaction
431 HMM - Hidden Markov Model
432 WGS – Whole Genome Sequencing
433
434 **DECLARATIONS**
435
436 **Ethics approval and consent to participate:** Not applicable
437
438 **Consent for publication:** Not applicable

Availability of data and materials: The sequencing datasets generated during the study are available in NCBI SRA repository, project number PRJNA422490.

Competing interests: The authors declare that they have no competing interests.

Funding: This study is supported by the funding from Australian National Health and Medical Research Council (APP1052303).

Authors' contributions: DG and LC conceived the study. DG, MY and TD performed the experiments. DG, MY, VM, CZ, MDC and LC performed the sequencing and statistical analysis. DG and MY wrote the paper with input from the other authors. All authors read and approved the final manuscript.

Acknowledgements: Not Applicable.

REFERENCES:

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Jurka J, Kapitonov VV, Kohany O, Jurka MV: **Repetitive sequences in complex**

461 **genomes: structure and evolution.** *Annu Rev Genomics Hum Genet* 2007, **8:**241-259.

462 3. Gelfand Y, Rodriguez A, Benson G: **TRDB--the Tandem Repeats Database.** *Nucleic*
463 *Acids Res* 2007, **35:**D80-87.

464 4. Gemayel R, Cho J, Boeynaems S, Verstrepen KJ: **Beyond junk-variable tandem**
465 **repeats as facilitators of rapid evolution of regulatory and coding sequences.** *Genes*
466 2012, **3:**461-480.

467 5. Armour JA: **Tandemly repeated DNA: why should anyone care?** *Mutat Res* 2006,
468 **598:**6-14.

469 6. Hannan AJ: **TRPing up the genome: Tandem repeat polymorphisms as dynamic**
470 **sources of genetic variability in health and disease.** *Discov Med* 2010, **10:**314-321.

471 7. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ: **Variable tandem repeats**
472 **accelerate evolution of coding and regulatory sequences.** *Annual review of genetics*
473 2010, **44:**445-477.

474 8. Bidwell JL, Bignon JD: **DNA-RFLP methods and interpretation scheme for HLA-**
475 **DR and DQ typing.** *Eur J Immunogenet* 1991, **18:**5-22.

476 9. Tagliabracci A, Buscemi L, Sassaroli C, Paoli M, Rodriguez D: **Allele typing of short**
477 **tandem repeats by capillary electrophoresis.** *Int J Legal Med* 1999, **113:**26-32.

478 10. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ: **Recent**
479 **advances in the detection of repeat expansions with short-read next-generation**
480 **sequencing.** *F1000Res* 2018, **7**.

481 11. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, Verstrepen
482 KJ, Froyen G: **Large-scale analysis of tandem repeat variability in the human**

483 **genome.** *Nucleic acids research* 2014, **42:**5728-5741.

484 12. Ganesamoorthy D, Cao MD, Duarte T, Chen W, Coin L: **GtTR: Bayesian estimation**
485 **of absolute tandem repeat copy number using sequence capture and high**
486 **throughput sequencing.** *BMC Bioinformatics* 2018, **19:**267.

487 13. Gymrek M, Golan D, Rosset S, Erlich Y: **lobSTR: A short tandem repeat profiler for**
488 **personal genomes.** *Genome Res* 2012, **22:**1154-1162.

489 14. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D: **Accurate**
490 **human microsatellite genotypes from high-throughput resequencing data using**
491 **informed error profiles.** *Nucleic Acids Res* 2013, **41:**e32.

492 15. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S,
493 Balasubramanian S, Boden M: **Inferring short tandem repeat variation from paired-**
494 **end short reads.** *Nucleic Acids Res* 2014, **42:**e16.

495 16. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y: **Genome-wide**
496 **profiling of heritable and de novo STR variations.** *Nat Methods* 2017, **14:**590-592.

497 17. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, Davis M, Lamont
498 P, Clayton JS, Laing NG, et al: **STRetch: detecting and discovering pathogenic short**
499 **tandem repeat expansions.** *Genome Biol* 2018, **19:**121.

500 18. Schule B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, Liu M, Byrne C, Gopi
501 R, Huang N, et al: **Parkinson's disease associated with pure ATXN10 repeat**
502 **expansion.** *NPJ Parkinsons Dis* 2017, **3:**27.

503 19. De Roeck A, De Coster W, Bossaerts L, Cacace R, De Pooter T, Van Dongen J, D'Hert
504 S, De Rijk P, Strazisar M, Van Broeckhoven C, Sleegers K: **Accurate characterization**

505 **of expanded tandem repeat length and sequence through whole genome long-read**
506 **sequencing on PromethION.** *bioRxiv* 2018**:**439026.
507 20. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, Prudencio M,
508 McLaughlin IJ, Bowman B, Seetin M, DeJesus-Hernandez M, et al: **Long-read**
509 **sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for**
510 **clinical use and genetic discovery efforts in human disease.** *Mol Neurodegener* 2018,
511 **13:**46.
512 21. Liu Q, Zhang P, Wang D, Gu W, Wang K: **Interrogating the "unsequenceable"**
513 **genomic trinucleotide repeat disorders by long-read sequencing.** *Genome Med*
514 2017, **9:**65.
515 22. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, Oma Y, Kino
516 Y, Mitsuhashi H, Matsumoto N: **Tandem-genotypes: robust detection of tandem**
517 **repeat expansions from long DNA reads.** *Genome Biol* 2019, **20:**58.
518 23. Karamitros T, Magiorkinis G: **Multiplexed Targeted Sequencing for Oxford**
519 **Nanopore MinION: A Detailed Library Preparation Procedure.** *Methods Mol Biol*
520 2018, **1712:**43-51.
521 24. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Sedlazeck FJ, Timp
522 W: **Targeted Nanopore Sequencing with Cas9 for studies of methylation,**
523 **structural variants and mutations.** *bioRxiv* 2019**:**604173.
524 25. Hafford-Tear NJ, Tsai YC, Sadan AN, Sanchez-Pintado B, Zarouchlioti C, Maher GJ,
525 Liskova P, Tuft SJ, Hardcastle AJ, Clark TA, Davidson AE: **CRISPR/Cas9-targeted**
526 **enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-**

527 **associated TCF4 triplet repeat.** *Genet Med* 2019.

528 26. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD,
529 Dilthey AT, Fiddes IT, et al: **Nanopore sequencing and assembly of a human genome**
530 **with ultra-long reads.** *Nat Biotechnol* 2018, **36:**338-345.

531 27. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman
532 W, Anantharaman T, Hastie A, et al: **Assembly and diploid architecture of an**
533 **individual human genome via single-molecule technologies.** *Nat Methods* 2015,
534 **12:**780-786.

535 28. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
536 **34:**3094-3100.

537 29. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC: **Adaptive seeds tame genomic**
538 **sequence comparison.** *Genome Res* 2011, **21:**487-493.

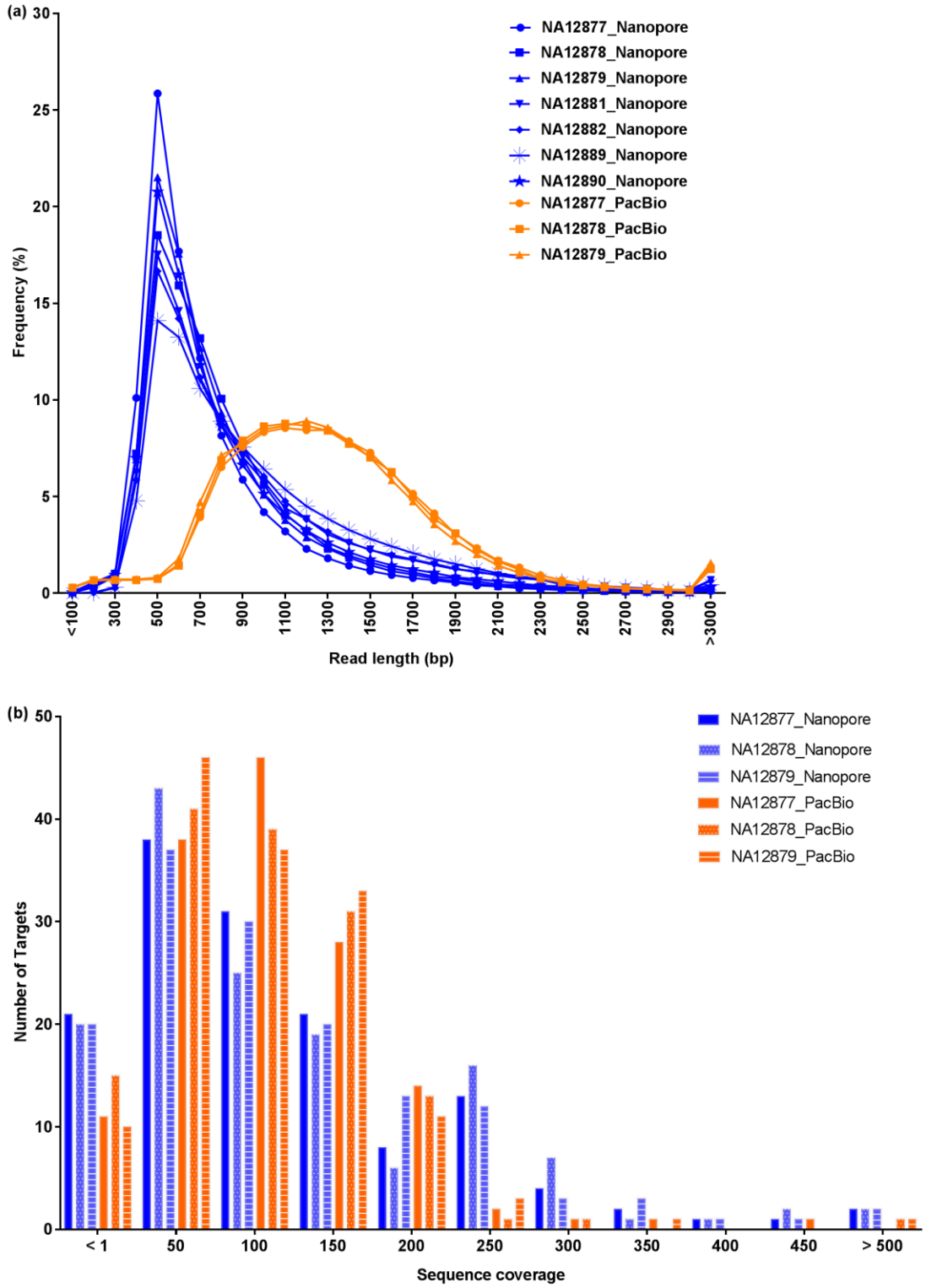
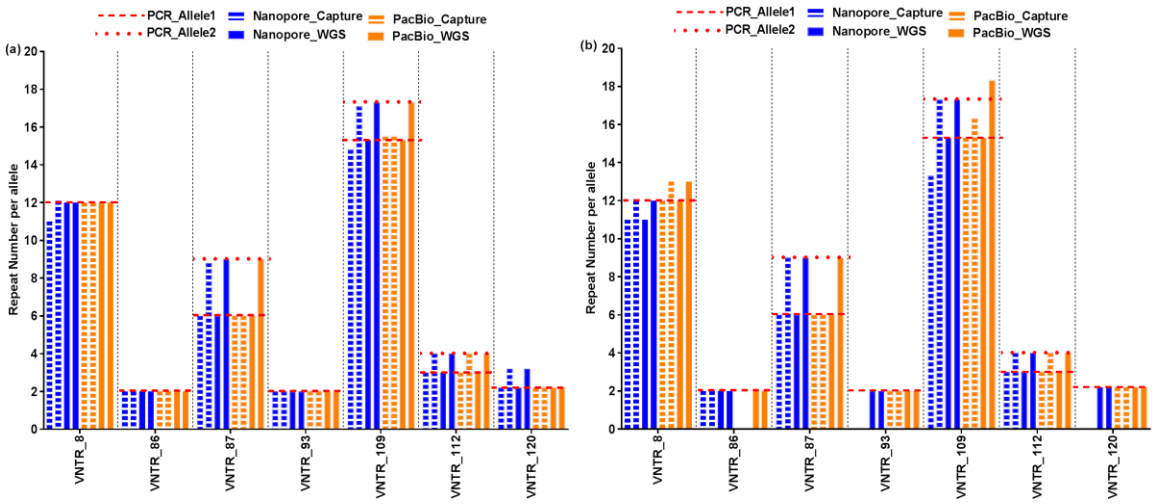


Figure 1: Targeted sequence capture (a) read length and (b) sequence coverage distribution for Nanopore and PacBio sequencing.

543



544

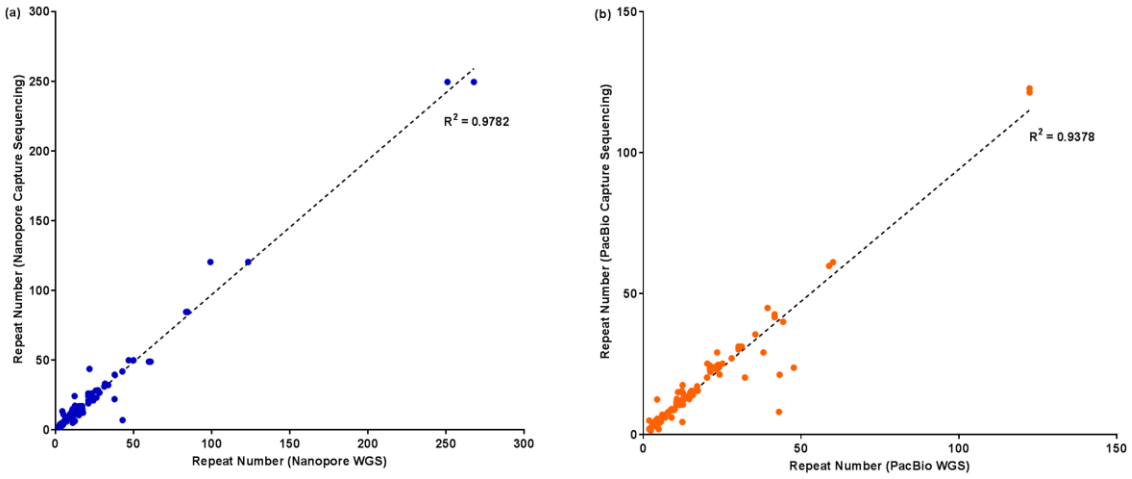
545 **Figure 2:** Comparison of genotype estimates between WGS and Target capture sequencing for

546 NA12878 sample using (a) VNTRTyper and (b) Tandem-genotypes. Red line indicates PCR

547 sizing results. Targets with no genotype estimates are shown as a gap for the corresponding

548 column.

549



550

551 **Figure 3:** Correlation of VNTRTyper genotype estimates between WGS and Targeted

552 sequencing on (a) Nanopore and (b) PacBio sequencing data for NA12878 sample.

553

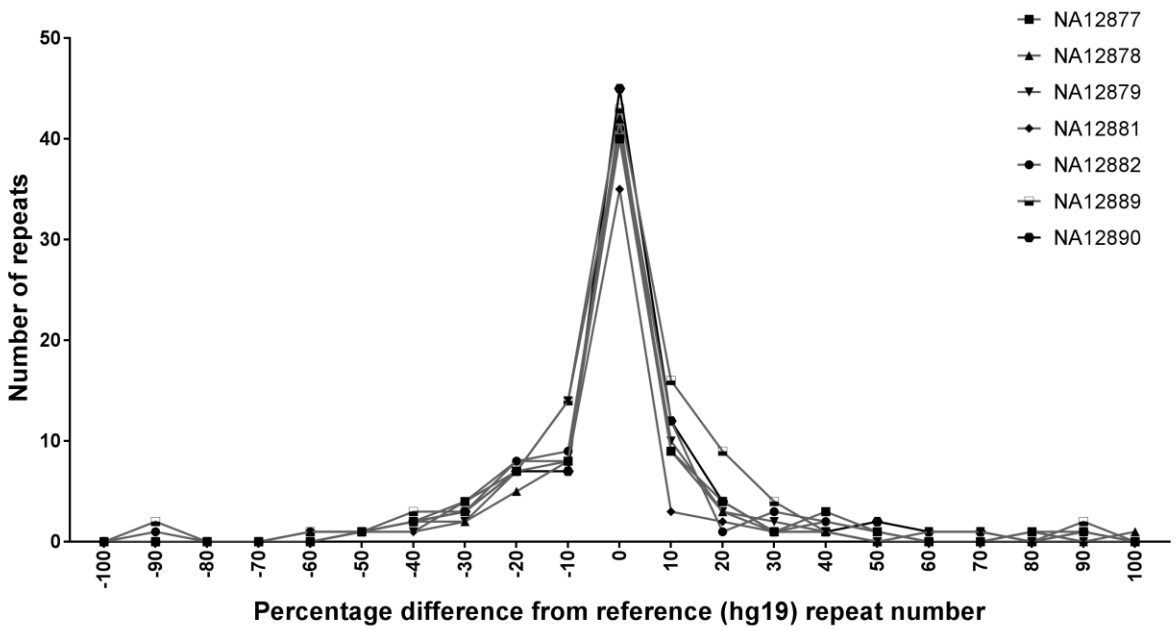


Figure 4: Percentage difference between reported repeat number in reference genome (hg19) and estimated repeat number based on genotype estimates using VNTRTyper on Nanopore targeted sequencing.

Table 1: Targeted Sequencing metrics for Nanopore and PacBio Capture sequencing of tandem repeats

Sample	Number of reads	Number of Aligned bases	On-bait bases (%)^	Off-bait bases (%)^	Fold enrichment^	Median Read length (bp)	Max Read length (bp)
NA12877_Nanopore*	207976	63384429	59%	41%	3497	568	11207
NA12878_Nanopore*	189050	70210003	62%	38%	3566	646	6781
NA12879_Nanopore*	209341	72340656	56%	44%	3284	614	24657
NA12890_Nanopore*	206077	79329086	52%	48%	2955	633	8040
NA12881_Nanopore#	44363	21288912	48%	52%	2739	702	5720
NA12882_Nanopore#	106259	52366827	49%	51%	2823	718	7383
NA12889_Nanopore#	1021376	539165551	45%	55%	2573	776	10127
NA12877_PacBio	742698	740195633	8%	92%	325	1225	44272
NA12878_PacBio	651117	649109211	8%	92%	362	1206	38765
NA12879_PacBio	725437	691795401	7%	93%	321	1192	40661

^ Output from Picard CollectHS Metrics (Version 2.18.29) (On-bait bases - Number of bases that are uniquely mapped to the targeted regions of the genome; Off-bait bases - Number of bases that are uniquely mapped away from the targeted regions of the genome; Fold enrichment - The fold by which the targeted region has been amplified above genomic background)

*Nanopore Multiplexing Group – 1; #Nanopore Multiplexing Group – 2

Table 2: Genotype estimates on Nanopore and PacBio targeted capture sequencing using VNTRTyper

Sample	Method	Genotype of Target*								Correlation with PCR
		VNTR_8	VNTR_57	VNTR_86	VNTR_87	VNTR_93	VNTR_109	VNTR_112	VNTR_120	
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	
	Nanopore_Capture	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	0.9988
	PacBio_Capture	12.0/12.0	10.6/13.6	2.0/2.0	3.0/7.4	2.0/2.0	15.3/15.3	4.0/4.0	2.2/3.2	0.9845
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	
	Nanopore_Capture	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	0.9978
	PacBio_Capture	12.0/12.0	10.6/12.6	2.0/2.0	6.0/6.0	2.0/2.0	15.5/15.5	3.0/4.0	2.2/2.2	0.9880
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	
	Nanopore_Capture	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	0.9927
	PacBio_Capture	ND	ND	2.0/2.0	8.0/8.0	2.0/2.0	17.3/17.3	3.0/3.0	2.2/2.2	0.9944
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	
	Nanopore_Capture	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	0.9944
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	
	Nanopore_Capture	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	0.9919
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	
	Nanopore_Capture	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	0.9800
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	
	Nanopore_Capture	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	0.9936

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

ND – Sufficient data not available for genotype analysis

Table 3: Genotype estimates on NA12878 WGS and Capture sequencing data using VNTRTyper and Tandem-genotypes

Method^	Genotype of Target*										Corr. with PCR
	VNTR_8	VNTR_32	VNTR_57	VNTR_64	VNTR_86	VNTR_87	VNTR_93	VNTR_109	VNTR_112	VNTR_120	
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	5.4/9.4	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	0.9861
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	13.3/17.3	3.0/4.0	ND	0.9942
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	12.0/13.0	9.4/10.4	12.6/12.6	18.3/18.3	2.0/2.0	6.0/9.0	2.0/2.0	15.3/18.3	3.0/4.0	2.2/2.2	0.9966
PBC_VT	12.0/12.0	ND	10.6/12.6	ND	2.0/2.0	6.0/6.0	2.0/2.0	15.5/15.5	3.0/4.0	2.2/2.2	0.9880
PBC_TG	12.0/13.0	ND	10.6/12.6	ND	ND	6.0/6.0	2.0/2.0	15.3/16.3	3.0/4.0	2.2/2.2	0.9865

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

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

ND – Sufficient data not available for genotype analysis