1 **Nanopore-based single molecule sequencing of the D4Z4 array**

2 **responsible for facioscapulohumeral muscular dystrophy**

3

4 Satomi Mitsuhashi[1,2], So Nakagawa[1,3], Mahoko Ueda[1,3], Tadashi Imanishi[1],

5 Hiroaki Mitsuhashi[4]

6

7 1. Biomedical Informatics Laboratory, Department of Molecular Life Science,

8   Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan

9 2. Department of Human Genetics, Yokohama City University

10   Graduate School of Medicine, Yokohama, Kanagawa 236-0004, Japan

11 3. Micro/Nano Technology Center, Tokai University, Hiratsuka, Kanagawa,

12   Japan

13 4. Department of Applied Biochemistry, School of Engineering, Tokai

14   University, Hiratsuka, Kanagawa, Japan

15

16 **Corresponding author**

17 Satomi Mitsuhashi

18 Department of Human Genetics Yokohama City University Graduate School of

19 Medicine, Yokohama, Kanagawa 236-0004, Japan

20 Tel. +81-45-787-2606

21 Fax. +81-45-786-5219

22 E-mail: satomits@yokohama-cu.ac.jp

23

24

1

Abstract

Subtelomeric macrosatellite repeats are difficult to sequence using conventional sequencing methods owing to the high similarity among repeat units and high GC content. Sequencing these repetitive regions is challenging, even with recent improvements in sequencing technologies. Among these repeats, a haplotype of the telomeric sequence and shortening of the D4Z4 array on human chromosome 4q35 causes one of the most prevalent forms of muscular dystrophy with autosomal-dominant inheritance, facioscapulohumeral muscular dystrophy (FSHD). Here, we applied a nanopore-based ultra-long read sequencer to sequence a BAC clone containing 13 D4Z4 repeats and flanking regions. We successfully obtained the whole D4Z4 repeat sequence, including the pathogenic gene *DUX4* in the last D4Z4 repeat. The estimated sequence accuracy of the total repeat region was 99.7% based on a comparison with the reference sequence. Errors were typically observed between purine or between pyrimidine bases. Further, we analyzed the D4Z4 sequence from publicly available ultra-long whole human genome sequencing data obtained by nanopore sequencing. This technology may become a new standard for the molecular diagnosis of FSHD in the future and has the potential to widen our understanding of subtelomeric regions.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is one of the most prevalent adult-onset muscular dystrophies. The genomes of most patients with FSHD have a common feature, i.e., a contracted subtelomeric macrosatellite

2

49 repeat array called D4Z4 on chromosome 4q35. The D4Z4 array consists of a

50 highly similar 3.3-kb single repeat unit. Normally, the D4Z4 array is highly

51 methylated and forms heterochromatin. Patients with FSHD have less than 11

52 D4Z4 repeats (1-3). In Japan, the majority of patients with FSHD have less than

53 7 repeats (4). Shortening of the D4Z4 array causes the de-repression of the

54 flanking genes as well as *DUX4*, located in the last D4Z4 repeat. The ectopic

55 expression of *DUX4* is toxic in muscle tissues and is thought to be a causal

56 factor for FSHD (5-9). In addition to the repeat number, the haplotype of the last

57 D4Z4 repeat is important for the development of FSHD (1, 2). The telomeric

58 flanking region of D4Z4 contains the 3′ UTR of *DUX4* and is called the pLAM

59 region. The presence of a polyadenylation signal in this region allows *DUX4*

60 expression and disease manifestation (10). In contrast, individuals without

61 polyadenylation signals do not manifest the disease (2).

62   Molecular diagnosis of FSHD is commonly made by Southern blotting

63 of genomic DNA after restriction enzyme digestion to measure the D4Z4 array

64 length and estimate the number of repeats. Haplotype analysis requires a

65 different probe (1). Sequencing of this D4Z4 array using Sanger sequencing or

66 short-read sequencers (up to 600 bp for Illumina and IonTorrent) is technically

67 difficult owing to the high similarity and the high GC content of the repeats. The

68 Oxford Nanopore Technologies MinION (Oxford, UK) is a single-molecule

69 sequencer that can produce long reads exceeding 100 kbp (11). Therefore,

70 MinION sequencing may enable the determination of pathogenicity by

71 sequencing the complete D4Z4 array.

72

Results

Nanopore-based D4Z4 sequencing using a BAC clone

The D4Z4 array on 4q35 has *Eco*RI sites in its flanking region. We took advantage of this restriction enzyme to excise the full-length D4Z4 repeats with flanking sequences, for a total of 49,877 bp. Both sides of the *EcoRI*-digested DNA fragment had unique sequences that are not found in the D4Z4 repeats (4,823 bp on centromeric side and 865 bp on the telomeric side). RP11-242C23 contained multiple *Eco*RI sites. pBACe3.6 vector-derived DNA was digested, yielding fragments of less than 10 kb (Figure 1a). We were able to easily separate the D4Z4-containing DNA fragment (49877 bp) from vector-derived DNA by agarose gel electrophoresis and gel extraction (Figure 1b). We extracted the D4Z4 array-containing DNA and subjected it to MinION 1D sequencing (Oxford Nanopore Technologies, Oxford, UK). Base-calling was initially performed using MinKNOW ver. 1.5.12 and fastq conversion was performed using poretools to obtain 20,761 reads (12). Base-calling was not possible for 87,410 reads using real-time MinKNOW basecaller probably due to out of computer memory; we used Albacore (v.1.1.0) to obtain the fastq sequences in these cases. A total of 128,171 reads were obtained, with an average read length of 7,577 bp (Supplemental Table 1).

We mapped the reads to the reference BAC clone sequence (GenBank accession number CT476828.7) using LAST (13). Visualization of mapped reads using IGV showed coverage of the whole D4Z4 array (Figure 2). The longest read mapped to the D4Z4 repeat was 29,060 bp. The consensus sequence had an accuracy of 99.72%. We also used BWA-MEM for mapping

4

97   and found that the consensus sequence had a lower accuracy (99.18%). Thus,

98   we used LAST for subsequent analyses.

99   The haplotype of the telomeric flanking region of the final D4Z4 repeat

100   known as pLAM is important for disease manifestation. There are two equally

101   common haplotypes, A and B. Haplotype A has an added polyadenylation signal

102   at the 3′ UTR of the *DUX4* gene(10). This polyadenylation signal allows the

103   ectopic expression of *DUX4*, which is toxic in muscle cells of patients with

104   FSHD with the contracted D4Z4 array (14). Haplotype B lacks homologous

105   sequence to pLAM. Individuals with haplotype B do not manifest the disease,

106   despite having the contracted D4Z4 allele. Thus, it is important to identify the

107   pLAM sequence for the molecular diagnosis of FSHD. Using MinION, we

108   successfully sequenced the whole pLAM region with an accuracy of 100%

109   (Figure 3). In total, 135 bases were different from the reference genome

110   sequence among the whole D4Z4 array sequence of 49,877 bp (0.27%)

111   (Supplemental Figure 1a). Among 135 bases, 115 (85.2%) substitutions were

112   between purines or between pyrimidines (Supplemental Figure 1b). Most of

113   these errors were repeatedly detected at the same position in the repeats

114   (indicated by asterisks in Supplemental Figure 1a). Interestingly, 16 out of 18

115   recurrent errors were seen in the CCXGG sequence at the X position.

116   We also compared the nanopore-sequenced *DUX4* open-reading

117   frame (ORF) to the reference and the Sanger sequencing results for the

118   subcloned *DUX4* ORF. The accuracy of the *DUX4* ORF sequence was 99.9%

119   (Supplemental Figure 2) and there were 3 errors. These errors were also

120   located in the X position in the CCXGG sequence.

5

121

D4Z4 detection using nanopore-based whole human genome sequencing

We tested whether we can identify the D4Z4 array from whole genome

sequencing data obtained from the MinION sequencer. We used the publicly

available human reference standard genome NA12878 with R9.4 chemistry

(11). This project contains two sets of data. The rel3 dataset had approximately

26× coverage with an N50 length of 10.6 kb. Rel4 had 5× coverage of ultra-long

reads with an N50 of 99.7 kb, indicating that rel4 contained reads that possibly

cover the whole D4Z4 region. We performed a blastn similarity search against

rel4 reads using the pLAM region as a query and obtained 18 hits (e-value =

0.0). These reads were aligned to the D4Z4 repeat reference sequence with the

pLAM region. The consensus sequence identity to the last D4Z4 repeat and the

pLAM region was 96.3% (Figure 4). Among 18 reads, 2 had homologous

sequences to both centromeric and telomeric flanking regions of D4Z4 (Figure

2, Supplemental Table 2). These two reads are expected to cover the whole

D4Z4 array, which usually includes more than 16 D4Z4 repeats. Although, we

could not determine the exact number of D4Z4 repeats owing to the high rate of

deletion errors, this range of read fragment is capable of detecting contracted

D4Z4 array which is seen in the most of the FSHD patients.

Discussion

Sequencing a highly repetitive subtelomeric region is extremely

challenging. There is variation in the number of repeats among individuals and

6

145   sometimes within individuals, i.e., somatic mosaicism. It has been reported that

146   subtelomeric regions form heterochromatin, functioning as an insulator or

147   repressor of near-by genes or preventing telomere shortening (15, 16). It is

148   important to determine the relationship between phenotypic differences and

149   either sequence or structural variation in repeats not only to decipher the

150   pathomechanisms of the disease, but also to obtain a deeper understanding of

151   human genomes. Here, we applied a nanopore-based sequencer to investigate

152   the subtelomeric repeat array associated with FSHD for the first time. In the

153   near future, it will be feasible to search for these sequence-difficult regions to

154   find causal relationships between the human genome and genetic diseases;

155   even given the prevailing use of high-throughput sequencing of coding regions,

156   the genetic causes of many diseases remain unsolved.

157   The disease locus of FSHD was identified at 4q35 more than 20 years

158   ago; however, the mechanism underlying the disease has been a mystery for

159   years and the causative genes have not been identified until recently, when

160   accumulating evidence has shown that the misexpression of *DUX4* is

161   associated with the disease. Further, it is still unclear whether there is any

162   sequence polymorphism in the *DUX4* gene or flanking regions, as it is difficult to

163   sequence the gene or the *DUX4* transcript, which is expressed at the very low

164   levels even in the muscle tissues of patients (14, 17). Since therapeutic

165   approaches including nucleic acid drugs targeting *DUX4* mRNA are being

166   studied (18, 19), it may be useful to determine the exact *DUX4* sequence of

167   patients for the development of effective therapies as well as an integrative

168   diagnostic method.

7

169   Currently, the number of D4Z4 repeats is usually determined by

170   Southern blotting using a probe that hybridizes to the centromeric flanking

171   sequence, p13E-11 (1). If the patient has a deletion at this probe site, it is not

172   possible to detect the D4Z4 repeat by Southern blotting. The Southern blotting

173   technique is complicated and time-consuming. Alternative methods have been

174   investigated, but are not widely used (4, 20).

175   Morioka et al. sequenced D4Z4 using the PacBio sequencer (21) and

176   analyzed random fragments from the BAC clone. The advantage of the

177   nanopore sequencer over the PacBio sequencer is the ultra-long read capability

178   (11). It has the potential to obtain reads of more than 100 kbp, the approximate

179   mean size of D4Z4 in healthy individuals. Currently, we could only obtained two

180   reads that potentially cover all D4Z4 repeats from human genome data with 5×

181   coverage using 14 flow-cells (11). Our estimate of the number of D4Z4 repeats

182   was more than 16, the normal size observed in healthy individuals. As the

183   NA12878 standard DNA originated from healthy individuals without FSHD, this

184   repeat number is reasonable. As FSHD patients have D4Z4 number less than

185   11, we think this ultra-long read sequencing may be usable to detect the

186   disease-causing contracted D4Z4 array. If the data output for the MinION

187   sequencer improves, it will be possible to obtain sequence data with better

188   resolution. This approach is potentially applicable to subtelomeric regions of

189   other chromosomes or even to centromere sequences.

190   In nanopore-based sequencing, changes in electric current are

191   detected as nucleotides pass through the pore. We observed that the errors

192   tend to occur between purines or between pyrimidines, probably because they

193   have similar chemical structures (11). In addition, we also observed that

194   substitution errors tend to occur at the same nucleotide position across repeats

195   (Supplemental Figure 1, asterisk). This may reflect the fact that the nanopore

196   detects combinations of nucleotides and the specific combination CCXGG was

197   prone to be misread. We anticipate further improvements of the base-calling

198   algorithm, which will make MinION more beneficial for medical applications.

199   Sequencing technologies are continuously developed. During the

200   preparation of this manuscript, the new chemistry R9.5 with the new flow-cell

201   FLO-MIN107 was released. Considering the rapid improvements in this

202   technique, it may not be very long before this sequencing technology is used for

203   D4Z4 repeat analyses for patients with FSHD.

204

205   **Conclusions**

206   Using MinION with a R9.4 flow-cell and 1D sequencing chemistry, we

207   successfully sequenced the complete *EcoRI*-digested D4Z4 array from a BAC

208   clone that contained the D4Z4 repeat region of human chromosome 4. Our

209   deep sequencing results had an accuracy of 99.8% for the whole D4Z4 array

210   and flanking region. This includes the pLAM region, with an accuracy of 100%,

211   and the whole ORF of the pathogenic gene *DUX4,* with the accuracy of 99.9%,

212   which are important regions for determining the pathogenesis. This short report

213   may provide a basis for the future use of nanopore sequencing to deepen our

214   understanding of highly heterogenous subtelomeric regions that may contribute

215   to human disease.

216

217 **Materials and Methods**

218 **BAC clone**

219 The RP11-242C23 human BAC clone was obtained from BAC PAC Resources

220 Center (https://bacpaacresources.org). This BAC clone was sequenced and

221 deposited at GenBank under accession number CT476828.7 by the Wellcome

222 Trust Sanger Institute. It contained 13 3306-bp D4Z4 repeats.

223

224 **Preparation of D4Z4 repeats from the BAC clone**

225 RP11-242C23 was digested using EcoRI and treated with Klenow Fragment

226 DNA Polymerase (Takara, Shiga, Japan) at 37°C for 20 min. DNA was

227 subjected to electrophoresis on a 0.5% agarose gel. Bands larger than the 10-

228 kb marker (GeneRuler 1kb DNA Ladder; Thermo Fisher Scientific, Waltham,

229 MA, USA) were excised using a razor under ultraviolet light. The DNA

230 fragments larger than 1 kb were subjected to phenol-chloroform DNA

231 preparation. Agarose gels were soaked in phenol and incubated for 30 min at -

232 80°C. Then, the aqueous phase was collected and phenol-chloroform DNA

233 preparation was performed. The EcoRI-digested whole D4Z4 repeat was

234 enriched in the DNA sample.

235

236 **MinION 1D sequencing**

237 Library preparation was performed using a SQK-LSK108 Sequencing Kit R9.4

238 version (Oxford Nanopore Technologies, Oxford, UK) using 500 ng of DNA.

239 MinION sequencing was performed using one FLO-MIN106 (R9.4) flow cell with

240 the MinION MK1b sequencer (Oxford Nanopore Technologies). Base-calling

241   and fastq conversion were performed with MinKNOW ver. 1.5.12 followed by

242   poretools or Albacore.

243

Sequence alignment by LAST and BWA-MEM

245   Sequence reads were aligned to the EcoRI-digested D4Z4 repeat reference

246   (Figure 1b, Supplemental material) using the LAST aligner (13). Sequences

247   were also mapped to the reference genome hg19 using BWA-mem with default

248   settings (22). Consensus sequences were obtained and sequence identity was

249   calculated using UGENE (23). Mapped reads were visualized using IGV

250   software (24).

251

Subcloning of the last D4Z4 repeat

253   An *Escherichia coli* transformant with the RP11-242C23 human BAC clone was

254   cultured in LB medium containing 12.5 µg/ml chloramphenicol at 37°C The

255   human BAC clone DNA was purified using the QIAGEN Plasmid Midi Kit

256   (Hilden, Germany) according to the "User-Developed Protocol (QP01)." Briefly,

257   bacterial lysate from a 100-ml scale culture was passed through a QIAGEN-tip

258   100 column. The BAC clone DNA was eluted with buffer QF prewarmed to 65°C

259   and concentrated by isopropanol precipitation.

260

261   To obtain the DNA clone containing the last D4Z4 repeat with the pLAM region,

262   50 ng of the purified BAC clone was used as a template for PCR with the

263   forward primer 5′-cgcgtccgtccgtgaaattcc-3′ and the reverse primer 5′-

264   caggggatattgtgacatatctctgcac-3′. PCR was performed with PrimeSTAR GXL

265 DNA Polymerase (Takara) with the following cycling conditions: 98°C for 2 min

266 and 30 cycles of 98°C for 10 s, 60°C for 15 s, and 68°C for 30 min. PCR

267 products were gel-purified and cloned into a pCR blunt vector (ThermoFisher

268 Scientific) with the Mighty Mix DNA Ligation Kit (Takara). The sequence of the

269 resulting plasmid was confirmed by Sanger sequencing with M13 forward and

270 M13 reverse primers.

271

272

273 **D4Z4 sequence analysis using the ultra-long human whole genome**

274 **sequence**

275 The human whole genome sequenced by MinION sequencers was downloaded

276 (https://github.com/nanopore-wgs-consortium/NA12878) (11). A Blastn search

277 was performed against the ultra-long read dataset, rel4, using the pLAM

278 sequence as a query. A total of 18 sequence reads containing pLAM hits were

279 extracted (e-value = 0.0). These extracted reads were mapped to the D4Z4

280 reference sequence using LAST and consensus sequences for the last D4Z4

281 and pLAM regions were obtained as described above.

282

283 **References**

284

285 1. Wijmenga C, Hewitt JE, Sandkuijl LA, Clark LN, Wright TJ, Dauwerse HG,

286   et al. Chromosome 4q DNA rearrangements associated with

287   facioscapulohumeral muscular dystrophy. Nat Genet. 1992;2(1):26-30.

288 2. Lemmers RJ, de Kievit P, Sandkuijl L, Padberg GW, van Ommen GJ,

289   Frants RR, et al. Facioscapulohumeral muscular dystrophy is uniquely

290   associated with one of the two variants of the 4q subtelomere. Nat Genet.

291   2002;32(2):235-6.

292 3. van Deutekom JC, Wijmenga C, van Tienhoven EA, Gruter AM, Hewitt JE,

293   Padberg GW, et al. FSHD associated DNA rearrangements are due to

294   deletions of integral copies of a 3.2 kb tandemly repeated unit. Hum Mol

295   Genet. 1993;2(12):2037-42.

296 4. Goto K, Nishino I, Hayashi YK. Rapid and accurate diagnosis of

297   facioscapulohumeral muscular dystrophy. Neuromuscul Disord.

298   2006;16(4):256-61.

299 5. Mitsuhashi H, Mitsuhashi S, Lynn-Jones T, Kawahara G, Kunkel LM.

300   Expression of DUX4 in zebrafish development recapitulates

301   facioscapulohumeral muscular dystrophy. Hum Mol Genet. 2013;22(3):568-

302   77.

303 6. Kowaljow V, Marcowycz A, Ansseau E, Conde CB, Sauvage S, Matteotti C,

304   et al. The DUX4 gene at the FSHD1A locus encodes a pro-apoptotic

305   protein. Neuromuscul Disord. 2007;17(8):611-23.

306 7. Snider L, Asawachaicharn A, Tyler AE, Geng LN, Petek LM, Maves L, et al.

307   RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4

308   units: new candidates for the pathophysiology of facioscapulohumeral

309   dystrophy. Hum Mol Genet. 2009;18(13):2414-30.

310 8. Wallace LM, Garwick SE, Mei W, Belayew A, Coppee F, Ladner KJ, et al.

311   DUX4, a candidate gene for facioscapulohumeral muscular dystrophy,

312   causes p53-dependent myopathy in vivo. Ann Neurol. 2011;69(3):540-52.

313 9. Wuebbles RD, Long SW, Hanel ML, Jones PL. Testing the effects of FSHD

314   candidate gene expression in vertebrate muscle development. Int J Clin

315   Exp Pathol. 2010;3(4):386-400.

316 10. Lemmers RJ, van der Vliet PJ, Klooster R, Sacconi S, Camano P,

317   Dauwerse JG, et al. A unifying genetic model for facioscapulohumeral

318   muscular dystrophy. Science. 2010;329(5999):1650-3.

319 11. Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, et al. Nanopore

320   sequencing and assembly of a human genome with ultra-long reads.

321   BioRxiv. 2017. **doi:** https://doi.org/10.1101/128835

322 12. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore

323   sequence data. Bioinformatics. 2014;30(23):3399-401.

324 13. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame

325   genomic sequence comparison. Genome Res. 2011;21(3):487-93.

326 14. Snider L, Geng LN, Lemmers RJ, Kyba M, Ware CB, Nelson AM, et al.

327   Facioscapulohumeral dystrophy: incomplete suppression of a

328   retrotransposed gene. PLoS Genet. 2010;6(10):e1001181.

329 15. Ottaviani A, Rival-Gervier S, Boussouar A, Foerster AM, Rondier D,

330   Sacconi S, et al. The D4Z4 macrosatellite repeat acts as a CTCF and A-

331   type lamins-dependent insulator in facio-scapulo-humeral dystrophy. PLoS

332   Genet. 2009;5(2):e1000394.

333   16. Stadler G, Rahimov F, King OD, Chen JC, Robin JD, Wagner KR, et al.

334   Telomere position effect regulates DUX4 in human facioscapulohumeral

335   muscular dystrophy. Nat Struct Mol Biol. 2013;20(6):671-8.

336   17. Jones TI, Chen JC, Rahimov F, Homma S, Arashiro P, Beermann ML, et al.

337   Facioscapulohumeral muscular dystrophy family studies of DUX4

338   expression: evidence for disease modifiers and a quantitative model of

339   pathogenesis. Hum Mol Genet. 2012;21(20):4419-30.

340   18. Ansseau E, Vanderplanck C, Wauters A, Harper SQ, Coppee F, Belayew A.

341   Antisense oligonucleotides used to target the DUX4 mRNA as therapeutic

342   approaches in faciosscapulohumeral muscular dystrophy (FSHD). Genes.

343   2017;8(3).

344   19. Wallace LM, Liu J, Domire JS, Garwick-Coppens SE, Guckes SM, Mendell

345   JR, et al. RNA interference inhibits DUX4-induced muscle toxicity in vivo:

346   implications for a targeted FSHD therapy. Mol Ther. 2012;20(7):1417-23.

347   20. Vasale J, Boyar F, Jocson M, Sulcova V, Chan P, Liaquat K, et al. Molecular

348   combing compared to Southern blot for measuring D4Z4 contractions in

349   FSHD. Neuromuscul Disord. 2015;25(12):945-51.

350   21. Morioka MS, Kitazume M, Osaki K, Wood J, Tanaka Y. Filling in the gap of

351   human chromosome 4: single molecule real time sequencing of

352   macrosatellite repeats in the facioscapulohumeral muscular dystrophy

353   locus. PLoS One. 2016;11(3):e0151963.

354   22. Li H. Aligning sequence reads, clone sequences and assembly

355   23. contigs with BWA-MEM. arXiv. 2013;arXiv:1303.3997.

15

356 24. Okonechnikov K, Golosova O, Fursov M, UGENE team. Unipro UGENE: a

357 unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166-7.

358 25. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz

359 G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.

360

361

362 **Author Contributions**

363 SM and HM designed the study and collected experimental materials.

364 SM, SN, MU, HM, and TI analyzed and interpreted the data. SM drafted the

365 original manuscript.

366

367 **Competing interests**

368 The authors report no disclosures relevant to the manuscript.

369

370 **Web Resources**

371 LAST: http://last.cbrc.jp

372 BWA: http://bio-bwa.sourceforge.net

373 Ape: http://biologylabs.utah.edu/jorgensen/wayned/ape/

374 UGENE: http://ugene.net

375 The URL for the human whole genome sequence data for NA12878 used in this

376 study is as follows:

377 https://github.com/nanopore-wgs-consortium/NA12878

378

379

16

Funding

This study was supported by MEXT-Supported Program for the Strategic

Research Foundation at Private Universities (to S.N. and M.T.U.). This work

was supported by JSPS KAKENHI Grant Number JP15K19477 (to H.M).

Figure legends

Figure 1

(a) Vector map of RP11-242C23 generated using Ape software. EcoRI sites are

shown. The D4Z4 array with 13 repeats and flanking regions was excised using

EcoRI digestion, yielding a 49877-bp product. (b) Agarose gel electrophoresis

of the EcoRI-digested vector DNA. Arrow shows the band of the 49877-bp D4Z4

array.

Figure 2

Mapped reads were visualized using IGV software. Coverage of reads is shown

on the upper part of the IGV image. Scheme shows the 13 D4Z4 repeats with

flanking sequences. The bottom scheme shows the enlarged last D4Z4 repeat

with the pLAM region (haplotype A). This region encodes pathogenic *DUX4*.

Figure 3

Nanopore sequence of the pLAM region. Exon 3 of *DUX4*, the 3′ UTR of the

gene, and polyA signal were determined with an accuracy of 100%. The upper

sequence is the reference and the bottom shows the nanopore sequence.

17

404

405   Figure 4

406   The consensus sequence of the last D4Z4 repeat and pLAM region obtained by

407   whole human genome nanopore sequencing. The upper sequence is the

408   reference and the lower sequence is the publicly available rel4 data set.

Figure 1
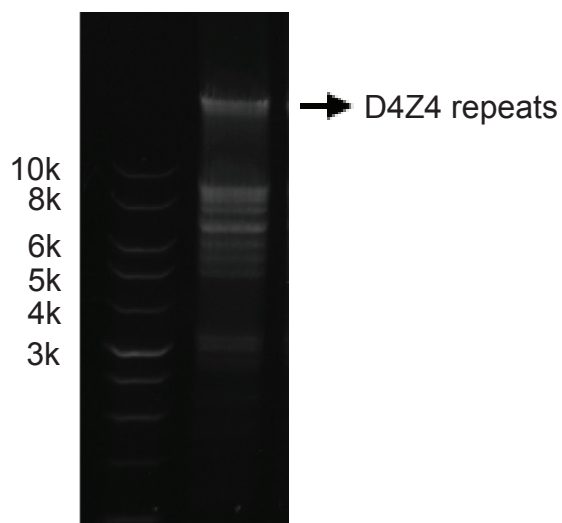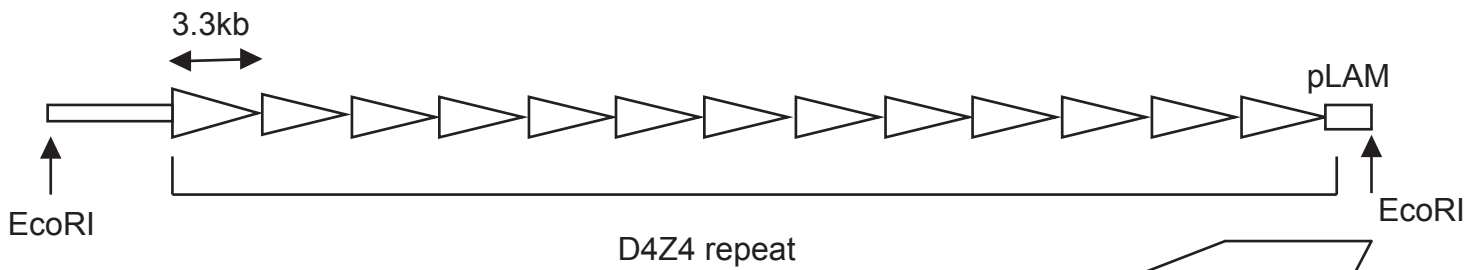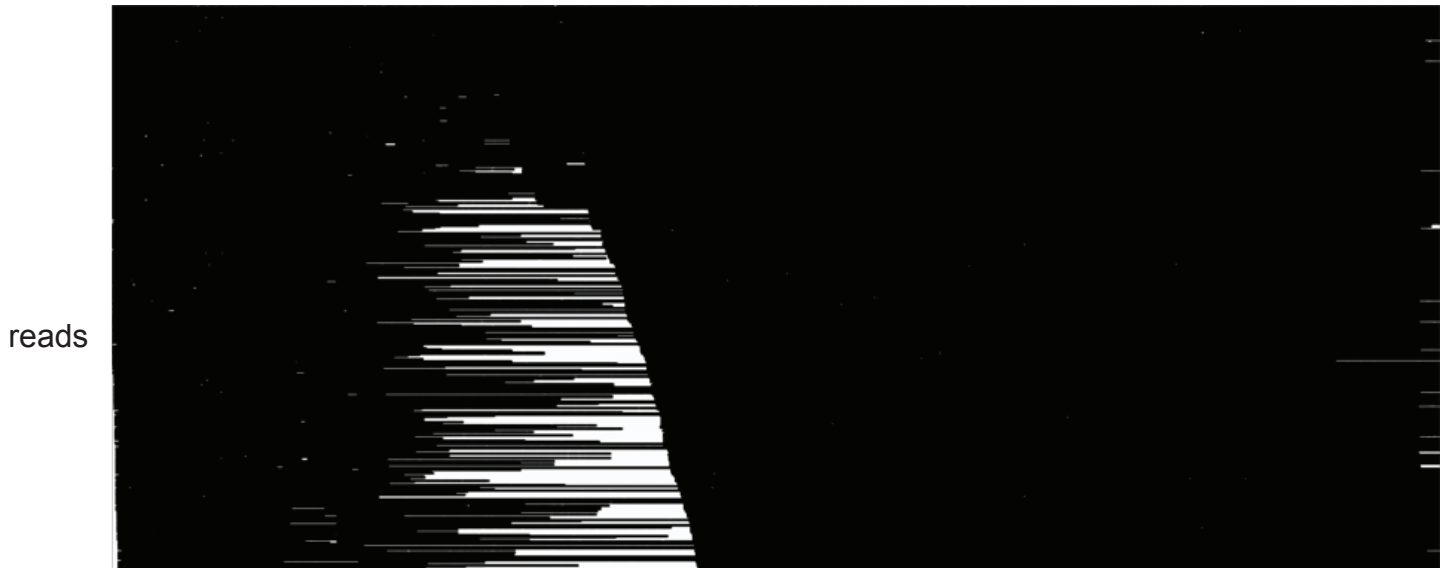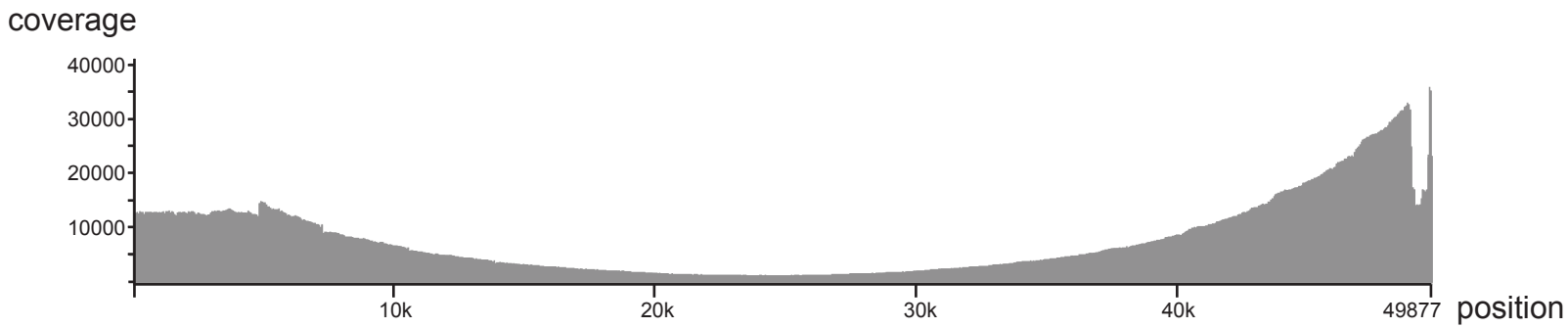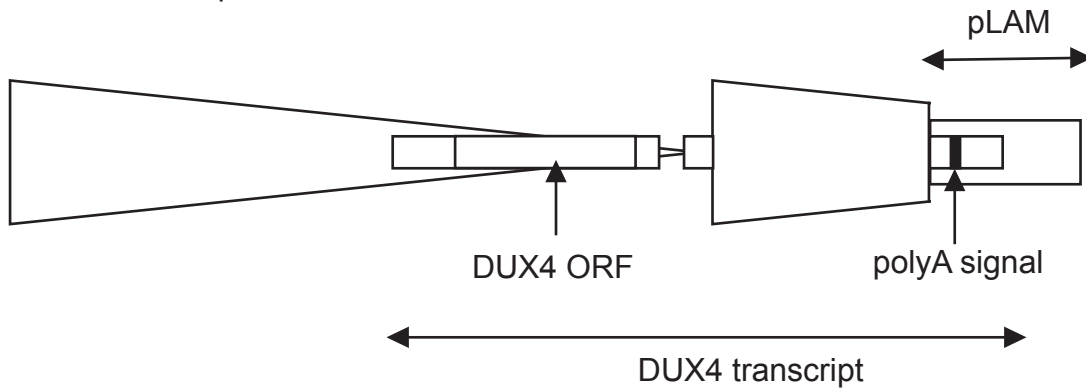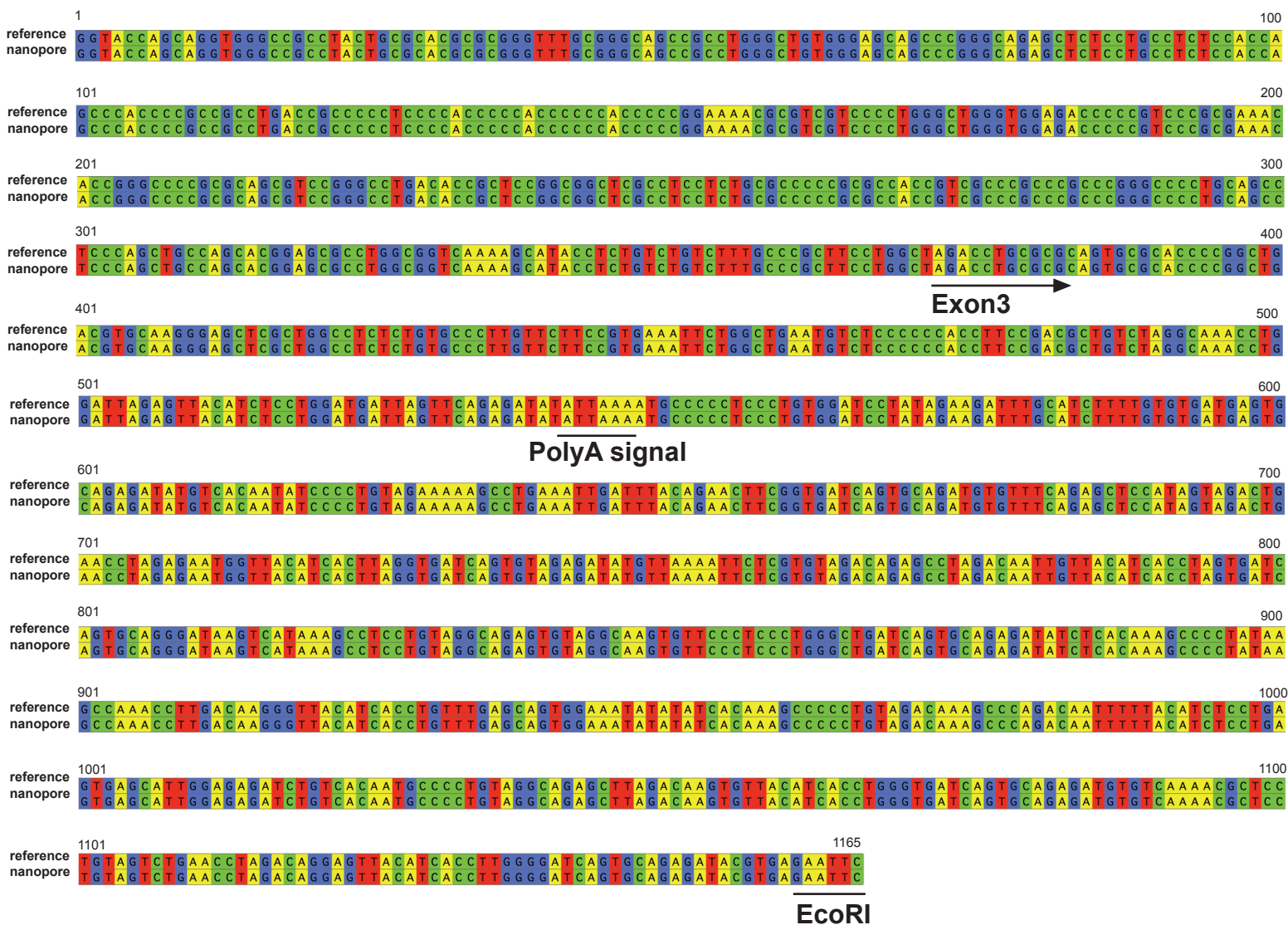
a

b
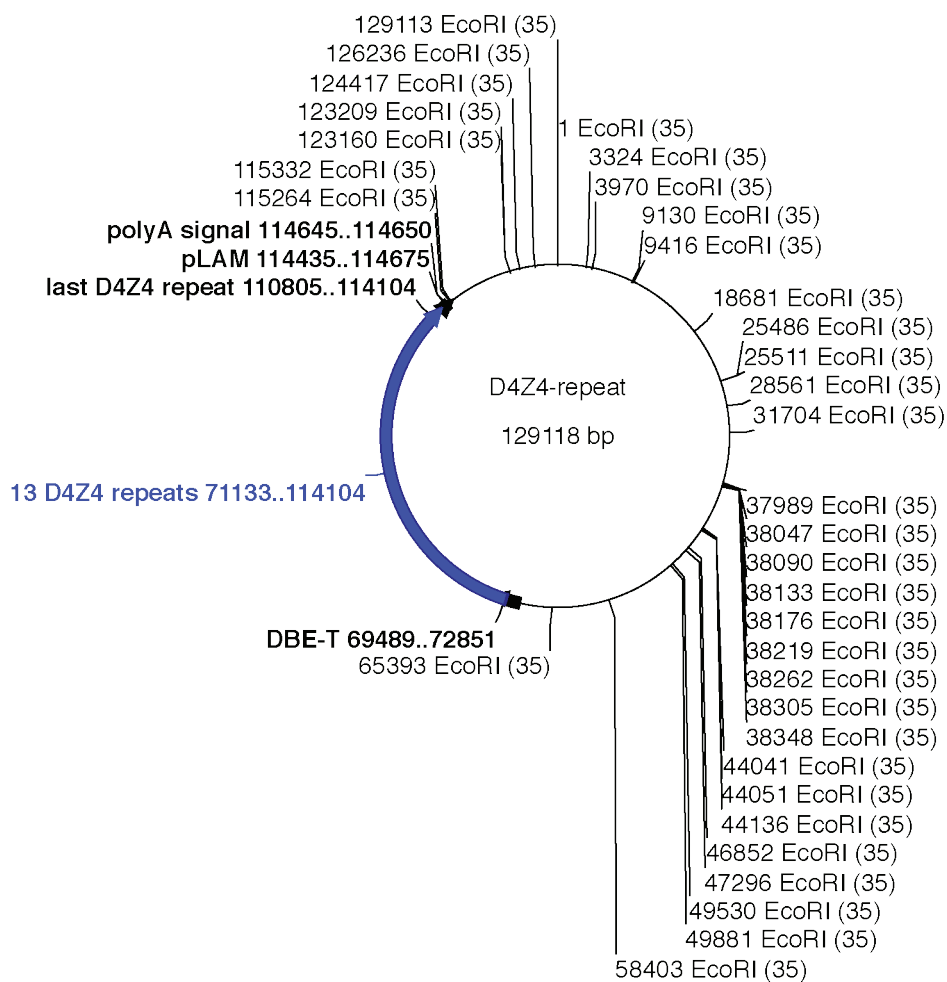
Figure 2

Figure 3

Figure 4

The last D4Z4

pLAM