

1 **Template switching mechanism drives the tandem amplification of**
2 **chromosome 20q11.21 in human pluripotent stem cells**

3
4 Jason A Halliwell¹, Duncan Baker², Kim Judge³, Michael A Quail³, Karen Oliver³,
5 Emma Betteridge³, Jason Skelton³, Peter W Andrews¹, Ivana Barbaric^{1*}

6
7 ¹Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield
8 S10 2TN, UK.

9 ²Sheffield Diagnostic Genetic Services, Sheffield Children's Hospital, Sheffield S10
10 2TH, UK.

11 ³Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
12

13 *Corresponding authors

14

15 **Abstract**

16 Copy number variants (CNVs) are genomic rearrangements implicated in numerous
17 congenital and acquired diseases, including cancer. In human pluripotent stem cells
18 (PSC), the appearance of culture-acquired CNVs prompted concerns for their use in
19 regenerative medicine applications. A particularly common problem in PSC is the
20 occurrence of CNVs in the q11.21 region of chromosome 20. However, the exact
21 mechanisms of origin of this amplicon remains elusive due to the difficulty in
22 delineating its sequence and breakpoints. Here, we used long-range Nanopore
23 sequencing on two examples of this CNV, present as a duplication in one and a
24 triplication in another line. The CNVs were arranged in a head-to-tail orientation in
25 both lines, with sequences of microhomologies flanking or overlapping both the
26 proximal and distal breakpoints. These breakpoint signatures point to a specific
27 mechanism of template switching in CNV formation, with surrounding *Alu* sequences
28 likely contributing to the instability of this genomic region.

29

30

31 **Introduction**

32 Copy number variants (CNVs) are gains or losses of DNA segments ranging in size
33 from around 50bp to several megabases¹. By affecting the dosage of genes and

34 regulatory regions within the amplified or deleted sequence, CNVs underpin the
35 aetiology of many diseases from developmental disorders to cancer¹. The profound
36 effect of the CNV acquisition on cellular phenotype has been also described in
37 human pluripotent stem cells (PSC), which frequently gain a CNV located on
38 chromosome 20 in the region q11.21 upon prolonged culture²⁻⁵. Once gained, the
39 20q11.21 CNV bestows on the variant PSC attributes that provide them with a
40 growth advantage due to resistance to apoptosis^{5,6}. The 20q11.21 CNV is typically
41 gained as a tandem duplication, although PSC lines with four or five copies of this
42 CNV have been reported^{2,7}. The length of the duplicated region is also variable
43 between different lines and ranges from 0.6Mb to 4Mb^{2,7}. Nonetheless, the shared
44 overlapping region in all of the reported variants contains a dosage-sensitive gene,
45 *BCL2L1*, which was identified as the driver gene responsible for the key phenotypic
46 features of variant PSC⁵⁻⁷. The altered behaviour of PSC harbouring the 20q11.21
47 CNV, coupled with the finding that the same CNV is a genomic hallmark of some
48 cancers⁸, represents a potential impediment to the use of PSC in regenerative
49 medicine applications and necessitates an understanding of the mechanisms
50 governing the CNV appearance.

51 CNVs can arise as a consequence of DNA replication errors or during the process of
52 DNA repair, with each of the implicated mechanisms of CNV formation yielding a
53 different sequence profile within the resulting breakpoint junction¹. For example,
54 CNV formation can occur by the non-homologous end joining pathway when repair
55 of DNA double strand breaks erroneously involves ligating the broken ends of
56 different breaks instead of re-ligating the original site⁹. The editing of the broken
57 ends prior to ligation is performed without the use of a homologous template and,
58 consequently, the resulting breakpoint junctions in CNVs created by the non-
59 homologous end joining typically contain random bases with no or little homology to
60 the original sequence^{10,11}. An alternative DNA repair mechanism implicated in CNV
61 formation involves the non-allelic homologous recombination pathway, which drives
62 the recombination of non-allelic genomic regions that share high sequence similarity,
63 such as low copy repeats¹. A defining feature of CNVs arising through this
64 mechanism are long stretches of homology in the sequence flanking their
65 breakpoints¹². Finally, replication-based repair mechanisms of DNA repair, including
66 fork stalling and template switching, and microhomology-mediated break-induced

67 replication, can create CNVs by switching the nascent DNA strand from a stalled or
68 collapsed replication fork to another fork in its vicinity, thereby giving rise to an
69 insertion or a deletion of a DNA segment^{13,14}. Importantly, invasion of an alternative
70 replication fork requires a small region of homology with the complementary strand in
71 order to prime the DNA synthesis. Therefore, CNVs formed by replication-based
72 repair mechanisms are characterised by the presence of microhomology within their
73 breakpoint sequence¹⁴.

74 Although the CNV genomic sequence holds essential clues as to the mechanisms
75 governing its formation, this information is not attainable from conventionally
76 employed techniques for CNV detection, such as the CGH arrays, Fluorescent In
77 Situ Hybridisation or quantitative PCR¹⁵. By contrast, next generation sequencing
78 technology can be used to reveal the CNV sequence at the nucleotide level, with
79 increased or decreased numbers of mapped reads across genomic regions
80 indicating the presence of genomic amplifications or deletions, respectively¹⁶.
81 However, sequencing of the genome using short reads (<300bp) is ill-suited for CNV
82 detection due to the mapping ambiguity of short reads, particularly in the presence of
83 highly homologous or repetitive sequences¹⁷. Recently, the advent of long read
84 sequencing technologies allowed reads to be uniquely mapped to the reference
85 genome, thus facilitating a more effective CNV detection and identification of
86 previously cryptic CNV breakpoints¹⁸.

87 Here, we applied long-range next generation sequencing to two human PSC lines
88 that each harbour a 20q11.21 CNV, in order to delineate the CNV breakpoint
89 sequences, the orientation of the amplified segments and the genomic context
90 surrounding the CNV. The amplified segments were present in a head-to-tail
91 orientation in both of the lines and their breakpoints contained sequences of
92 microhomology, suggesting that the replication-based template switching
93 mechanisms were implicated in their genesis. Moreover, we identified *A1u* repetitive
94 sequences that intersect or flank the 20q11.21 CNV breakpoints. The presence of
95 such repetitive elements may cause inherent instability to this area of the genome,
96 making it a particular hotspot for CNV formation.

97

98 **Results**

99

100 *Detection of human PSC lines with chromosome 20q11.21 CNV*
101
102 By interphase FISH analysis, the human embryonic stem cell (ESC) line MShef7-A4,
103 a subline of MShef7^{19,20}, and the human induced pluripotent stem cell (iPSC) line
104 NCRM1²¹ each exhibited a homogeneous population of cells with a tandem
105 duplication or a triplication of the chromosome 20q11.21 region, respectively
106 (**Supplementary Fig. 1**). To identify the approximate proximal and distal breakpoint
107 position of the amplicon in each cell line (**Fig. 1**), we adapted our previously
108 published qPCR-based method for assessment of copy number of target loci and we
109 used it to assess the copy numbers of loci along the length of the q arm of
110 chromosome 20^{15,22}. In both cell lines, the proximal breakpoint was positioned
111 between the centromere and the *DEFB115* gene (**Fig. 1**). In MShef7-A4, the distal
112 breakpoint of the tandem duplication was located between the *TM9SF4* and *ASXL1*
113 genes (**Fig. 1a, b**), whereas in NCRM1 the amplicon was smaller with the distal
114 breakpoint positioned between the *TPX2* and *MYLK2* genes (**Fig. 1a, c**). In addition
115 to identifying the putative breakpoints at 20q11.21, qPCR analysis revealed the
116 presence of four copies of the amplicon in NCRM1, confirming the triplication of the
117 chromosome 20q11.21 region in this line (**Fig. 1c**).
118

119 *Nanopore sequencing reveals the chromosome 20q11.21 breakpoint in MShef7-A4*
120
121 To identify the location of the breakpoints at a single nucleotide resolution in
122 MShef7-A4 CNV and to determine the orientation of this tandem duplication, we
123 performed whole-genome Oxford Nanopore sequencing on DNA extracted from the
124 cells and aligned the sequencing reads to the hg38 human reference genome
125 assembly²³. The average read depth across chromosome 20 was 14.5 with a mean
126 read length of 15.2 kb. We noted an increased sequencing read depth along the
127 chromosome 20q11.21 relative to the rest of the chromosome (22.8 versus 14.5,
128 respectively), indicative of a change in the copy number of this region (**Fig. 2a**)^{24,25}.
129 A distinct drop in read coverage was observed at position 32,273,600 bp of the
130 chromosome 20 hg38 reference sequence (between *TPX2* and *MYLK2* genes),
131 which we surmised was to be the distal breakpoint and was in agreement with the
132 position we defined by qPCR (**Fig. 1a and 2a**). To represent reads which map to two
133 discontinuous locations in the genome, mapping algorithms “soft-clip” reads to

134 indicate that a portion of the read in question does not map to the same position as
135 the remainder of the read. Soft-clipping of reads therefore provides evidence of
136 structural variation, in our case, tandem duplication, as reads which span
137 breakpoints map to disparate regions therefore triggering soft-clipping
138 (**Supplementary Fig. 2**)^{26,27}. Furthermore, the soft-clipped proportion of the
139 sequencing read at the distal breakpoint can be used to infer the orientation of the
140 tandem duplication. We reasoned that, if the soft-clipped DNA sequence at the distal
141 breakpoint aligns to the reference genome between the centromere and *DEFB115*
142 gene, then these two distantly positioned DNA sequences must have been fused in a
143 head-to-tail orientation. However, if the soft-clipped portion of reads aligns to the
144 distal breakpoint in an inverted orientation, the duplication has occurred in a head-to-
145 head fashion. Therefore, we performed a BLAT pairwise sequence alignment of a
146 contig formed from the unmapped portion of the soft-clipped reads to identify their
147 genomic location²⁸. The contig aligned with 92% identity to a (GGAAT)n
148 microsatellite repeat in the pericentromeric region proximal of the *DEFB115* gene,
149 confirming the head-to-tail orientation of the tandem duplication (**Fig. 2b, c**). This
150 microsatellite is positioned at 31,051,509-31,107,036 bp on chromosome 20, and is
151 flanked by two unmapped regions of the reference genome. We could not locate the
152 proximal breakpoint to a single nucleotide position, which we inferred was due to the
153 breakpoint being located in a currently unmapped region of the reference genome,
154 potentially in one of the regions we observed flanking the microsatellite.

155
156 To understand the mechanism of tandem duplication in MShef7-A4, we analysed the
157 breakpoint sequences for signatures commonly observed in copy number variants.
158 For the distal breakpoint, we analysed 500 bp of the reference genome sequence
159 (hg38) surrounding the junction (**Fig. 2c**). As we were unable to locate the proximal
160 breakpoint, we used the contig of the unmapped portions of the soft-clipped reads
161 found at the distal breakpoint (**Fig. 2b, c**), which revealed a region of micro-
162 homology (AGAATCACTTAAACC) that flanked both the proximal and distal
163 breakpoint positions (**Fig. 2c**). By consulting the Dfam database of transposable
164 elements, we observed that the distal region of microhomology lies within an *A/uSz6*
165 retrotransposon that spans the distal breakpoint²⁹. These results suggest a role of
166 microhomology in the mutational mechanism of the tandem amplification of
167 chromosome 20 in the MShef7-A4 cell line.

168

169 *Break point mapping of a chromosome 20q11.21 tandem triplication*

170

171 We used the same sequencing approach to identify and analyse the breakpoints in
172 the human iPSC cell line, NCRM1, which contains a tandem triplication in the
173 20q11.21 region (**Supplementary Fig. 2**). Our Nanopore sequencing returned an
174 average read length of 19.9 kb at a mean depth of 20.3 across chromosome 20.
175 Consistent with our qPCR analysis, long-read sequencing identified a sole distal
176 breakpoint at position 31,813,288 bp between the *TPX2* and *MYLK2* genes. This
177 confirmed that both amplicon copies in NCRM1 have the same distal breakpoint
178 position. The increased read depth associated with copy number variants was
179 greater in NCRM1 (43.9) when compared with MShef7-A4, consistent with the
180 triplication indicated by our PCR and FISH analyses (**Fig. 3a**). To identify the
181 proximal breakpoint position, we performed a BLAT pairwise sequence alignment on
182 the unmapped portions of the soft-clipped reads. Our soft-clipped sequence aligned
183 with the reference genome at position 31,059,954 bp, within the same microsatellite
184 that was putatively identified as the proximal breakpoint region in MShef7-A4 (**Fig.**
185 **3b, c**). These data confirm that the tandem triplication of chromosome 20q11.21 in
186 NCRM1 has occurred in a head-to-tail orientation, and that each amplicon was of
187 equal length and contained the same breakpoint positions. Furthermore, we
188 observed a common microsatellite sequence at the proximal breakpoint in both cell
189 lines, and thus, its involvement could be complicit in the tandem amplifications that
190 commonly occur associated with chromosome 20q11.21.

191

192 To infer the mechanism involved in the tandem triplication of chromosome 20q11.21
193 in NCRM1, we interrogated the reference sequence at both the proximal and distal
194 breakpoint positions. We identified multiple regions of micro-homology (TGAA and
195 AATTGAA) that flanked both sides of the fusion junction (**Fig. 3c**). Furthermore, we
196 consulted the Dfam database of transposable elements and identified an *AluSz6*
197 element that was situated 9 bp downstream of the distal breakpoint (**Fig. 3b, c**). As
198 we were unable to find an *Alu* element at the proximal breakpoint itself, it is unlikely
199 the tandem duplication and triplication in MShef7-A4 and NCRM1, respectively, have
200 arisen through a mechanism of *Alu-Alu* recombination. Instead, we propose that the
201 *Alu* elements are sites of chromosome fragility, due to replication blockage³⁰⁻³⁴.

202 Repair of stalled and collapsed forks would then proceed through replication fork
203 switching to complementary sites of microhomology, and strand invasion upstream
204 on the same or a homologous chromosome would generate a tandem amplification
205 (**Fig. 4**).

206

207 **Discussion**

208

209 The experiments reported here have revealed the breakpoints of tandem
210 amplifications of chromosome 20q11.21 in human PSC. The distal breakpoints were
211 all found to be located in, or close to *Alu* sequences. The proximal breakpoints were
212 located in a pericentromeric microsatellite region close to 31 Mb on chromosome 20.
213 In the case of NCRM1, each amplicon of the tandem triplication was of equal length
214 with the same breakpoint positions. A detailed characterisation of the breakpoints at
215 a single nucleotide level revealed short microhomologies that flank or overlap both
216 the proximal and distal breakpoints. These breakpoint characteristics are like scars
217 left by the repair mechanism that operated on the DNA lesion.

218

219 Although rare, breakpoint microhomology of between 1-4 bp long is occasionally
220 observed with CNV formed by non-homologous end joining (NHEJ)^{35,36}. As the
221 microhomology at the breakpoints of our lines was larger than 7 bp we excluded
222 classical NHEJ as the mechanism of tandem amplification. However, alternative
223 forms of end-joining such as microhomology mediated end joining do utilize larger
224 spans of homology or microhomology³⁷⁻⁴². These mechanisms differ from classical
225 NHEJ, as they do not perform blunt-end ligation and instead utilise end-resection at
226 DNA breaks to reveal overlapping micro-homologous single stranded DNA required
227 for annealing⁴³. We eliminated alternative end-joining from the potential mutagenic
228 mechanisms, as the microhomology in both MShef7 and NCRM1 was intact and
229 tandem amplifications are not readily explained by this mechanism⁴⁴.

230

231 The tandem amplifications in MShef7 and NCRM1 had breakpoints devoid of large
232 regions of sequence homology, which ruled out mechanisms involving homologous
233 recombination such as non-allelic homologous recombination⁴⁵. However, the
234 presence of an *AluSz6* element at the distal breakpoints in both cell lines led us to
235 consider *Alu-Alu*-mediated non-allelic homologous recombination mechanism. For

236 *Alu-Alu*-mediated non-allelic homologous recombination to take place it would
237 require a second *Alu* element at the proximal breakpoint with high sequence identity
238 with the distal *Alu*⁴⁶. We found no evidence of a second *Alu* at the proximal
239 breakpoint in either of our cell lines. Despite this, the presence of *AluSz6* at distal
240 breakpoints in both cell lines suggests that it might play a role in the initiation of
241 tandem amplifications, rather than in the mechanism of mutation itself. Inverted
242 repeats, such as *Alu* elements, form hairpin loop secondary structures that can
243 impede replication, leading to fork stalling and collapse, particularly under conditions
244 of replication stress^{30-34,47-49}. It is perhaps no coincidence then, that this mechanism
245 of mutagenesis is associated with high levels of replication stress, which is a
246 characteristic of human PSC during *in vitro* culture⁵⁰⁻⁵².

247

248 The breakpoint signatures of the tandem amplifications characterised in MShef7-A4
249 and NCRM1 are consistent with the replication template switching mechanisms, fork
250 stalling and template switching and microhomology mediated break induced
251 replication, which are initiated by replication fork stalling and collapse,
252 respectively^{13,14}. In the case of fork stalling and template switching, the lagging
253 strand at the stalled fork disengages and invades another replication fork at a region
254 of microhomology. Microhomology mediated break induced replication is similar to
255 fork stalling and template switching, although following a collapsed fork the 5' end of
256 the DNA break is resected to generate a 3' single-stranded overhang that then
257 invades a template region with microhomology before replication is reinitiated. If the
258 template is upstream on the same chromosome or a homologous chromosome, a
259 tandem amplification would result (**Fig. 4a, b**)^{13,14,45,53}. Furthermore, the role of
260 microhomology mediated break induced replication and fork stalling and template
261 switching in the formation of tandem triplications has been discussed^{14,54-56}. Should
262 replication fork collapse lead to sister chromatid strand invasion at an upstream
263 region of microhomology, replication of the amplified segment will proceed. This
264 could then be followed by a second round of template switching and strand invasion
265 at the same region of microhomology, although this time into the other parental
266 homolog with replication proceeding to the distal end of the chromosome, resulting in
267 a tandem triplication (**Fig. 4a-c**)

268

269 In summary, we provide evidence from breakpoint junctions that implicate
270 replication-based repair by fork stalling and template switching and microhomology
271 mediated break induced replication as the mutational mechanism driving tandem
272 duplication in human PSC. We argue that constitutive replication stress observed
273 during the *in vitro* culture of human PSC could be driving replication fork stalling and
274 collapse at *Alu* elements that initiates these mutations. This report provides new
275 insight into the mechanisms of mutation in human PSC. The recurrent nature of
276 genetic change in human PSC is considered non-random due to the selection of
277 advantageous mutations. However, it was recently reported that mutations in human
278 PSC occur with higher frequency in non-genic regions⁵⁷. The data presented here
279 complements these findings and suggests that mutation itself may be non-random
280 but may be enriched at certain sites that can be characterised by the genomic
281 architecture. By defining these regions, it may be possible to safeguard the genome
282 stability of human PSC for their use in cell-based regenerative medicine.

283

284 **Methods**

285

286 **Human pluripotent stem cell culture.** The MShef7^{19,20} (hPSCreg:
287 <https://hpscreg.eu/cell-line/UOSE012-A>) human ESC line was derived at the
288 University of Sheffield Centre for Stem Cell Biology under the HFEA licence R0115-
289 8A (centre 0191) and HTA licence 22510. A mosaic sub-population of chromosome
290 20 variant cells was detected in a culture of MShef7, which was sub-cloned using
291 single cell deposition by FACS. The NCRM1²¹ (hPSCreg: <https://hpscreg.eu/cell-line/CRMi003-A>) human iPSC line was acquired from RUCDR Infinite Biologics and
292 was originally derived by reprogramming umbilical cord blood CD34+ cells using a
293 non-integrating episomal vector. Both cell lines were maintained in culture vessels
294 coated with a matrix of Vitronectin human recombinant protein (ThermoFisher
295 Scientific, A14700) and batch fed daily with mTeSR (STEMCELL Technologies,
296 85850). Once the cells had reached confluence, they were passaged using ReLeSR
297 (STEMCELL Technologies, 05873) according to manufacturer's guidelines.

298

300 **qPCR breakpoint determination.** DNA was extracted from cell pellets using the
301 DNeasy Blood and Tissue kit (Qiagen, 69504). DNA quantity and quality were
302 measured using a NanoPhotometer (Implen). 1µg of DNA was digested with 10 units

303 of FastDigest EcoRI enzyme (Thermo Fisher Scientific, FD0275) in FastDigest buffer
304 (Thermo Fisher Scientific, FD0275) for 5 minutes at 37°C, followed by deactivation of
305 the enzyme by incubating at 80°C for 5 minutes. qPCR was performed as previously
306 described^{15,22}, using the adapted protocol²² whereby primer sets were designed
307 along the length of the q arm of chromosome 20 (**Table 1**) to allow an estimate of the
308 amplicon length. A 10µl PCR reaction contained TaqMan Fast Universal PCR
309 mastermix (ThermoFisher Scientific, 4366072), 0.1 µM Universal probe library
310 hydrolysis probe, 0.1 µM each of the forward and reverse primers (**Table 1**) and
311 either 20ng of EcoRI-digested DNA or water only (no template control). The PCR
312 reactions were run on the QuantStudio 12K Flex Real-Time PCR System using the
313 following profile: 50°C for 2 minutes, 95°C for 10 minutes, and 40 cycles of 95°C for
314 15 seconds and 60°C for 1 minute. The copy number was determined by first
315 subtracting the average Cq values from the test sample 20q loci from the reference
316 loci (Chromosome 4p) to obtain a dCq value. The dCq for the calibrator sample at
317 the same loci was then calculated in the same way and the test sample dCq and
318 calibrator sample dCq were subtracted from one another to give ddCq. The relative
319 quantity was calculated as 2^{-ddCq} . Finally, to obtain the copy number, the relative
320 quantity value was multiplied by 2.

321

322 **Table 1.** qPCR breakpoint detection primer sets and probes²².

Gene (location) Accession Number	Primer sequences (forward and reverse)	UPL probe number
<i>RELL1</i> (4p14) NC_000004.12	tgcttgctcagaaggagctt tgggttcaggaacagagaca	12
<i>DEFB115</i> (20q11.21) 31,257,664 NM_001037730.1	tcaagcctgaacattctggtaaa cacttgtctttcccaaactc	14
<i>REM1</i> (20q11.21) 31,475,272 NM_014012.5	cccctttctcactccacaa tctgcagggggagaagtaca	46
<i>TPX2</i> (20q11.21) 31,739,101 NM_012112.4	cccccaaattcaggcctac ttaaagcaaaatccaggagtcaa	35
<i>MYLK2</i> (20q11.21) 31,819,375 NC_000020.11	ggtcaggagaacccagagtg gtctcccaggcacttcag	16

XKR7 (20q11.21) 31,968,002 NM_033118.3	gtgtcttaccggggcctatc gcctggaagggtgtgcagta	3
TM9SF4 (20q11.21) 32,109,506 NM_014742.3	taatggagccaatgccagta caaaaaccagttctgtgccttt	45
ASXL1 (20q11.21) 32,358,062 NM_015338.5	gagtgtaactgtggatggtag ctggcatatggaaccctcac	13

323

324 **Fluorescence *in situ* hybridisation (FISH) for the detection of chromosomal**
325 **variants.** Human PSC were detached from culture flasks by incubating with TrypLE
326 Express Enzyme (Fisher Scientific, 11528856) for 3 minutes at 37°C. The cells were
327 collected in DMEM/F12 basal media (D6421, Sigma Aldrich) and centrifuged at 270
328 g for 8 minutes. To the cell pellet, 1 mL of pre-warmed 37°C 0.0375 M potassium
329 chloride was added. The cells were then centrifuged at 270 g for 8 minutes, before
330 fixing the cells by adding 2 mL fixative (3 parts methanol : 1 part acetic acid, v/v), in a
331 drop-wise manner under constant agitation. FISH detection of chromosomal variants
332 was performed by Sheffield Diagnostics Genetic Service. Analysis was performed on
333 100 interphase nuclei per sample that had been probed with D20S108 (BCL2L1)
334 probe.

335

336 **DNA extraction for sequencing.** DNA was extracted from cell pellets using the
337 DNeasy Blood and Tissue kit (Qiagen, 69504). DNA quantity and quality were
338 measured using a NanoPhotometer (Implen).

339

340 **DNA sequencing.** DNA library preparation was performed using the ligation (Oxford
341 Nanopore Technologies, SQK-LSK108) or Rapid sequencing kits (Oxford Nanopore
342 Technologies, SQK-RAD004) according to the manufacturer's Genomic DNA by
343 Ligation or Rapid Sequencing protocols, respectively. The whole genome libraries
344 were sequenced using the Oxford Nanopore MinION or GridION sequencers with the
345 R9.4.1 flow cell (Oxford Nanopore Technologies, FLO-MIN106D) following the
346 manufacturer's instructions. Each flow cell yielded ~5 Gb of data.

347

348 **Data processing.** Data exported as FASTQ files were mapped to the chromosome
349 20 hg38 reference sequence using minimap2 sequence aligner (version 2-2.15)⁵⁸.
350 File management, merging, sorting and indexing was performed using Sambamba
351 (version 0.6.6) and Samtools (version 1.9)^{59,60}. Breakpoint regions were inspected
352 manually using IGV genomic viewer⁶¹ and the breakpoint location was identified
353 based on read depth and soft-clipped sequence analysis. Briefly, the aligned and
354 sorted .bam files were opened using IGV genomic viewer with soft-clipped bases
355 enabled. The distal breakpoint region identified by qPCR was inspected and the
356 breakpoint at the single nucleotide level was located by identifying a region of
357 reduced read depth with soft-clipped reads that spanned the point of reduced read
358 coverage (**Figure S2A, B**). To identify the proximal breakpoint, the soft-clipped
359 proportion of the sequencing reads at the distal breakpoint were queried using BLAT
360 sequence alignment to identify the sequence matches in the human reference
361 genome with high similarity.

362

363 **References**

364

- 365 1 Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation
366 in genomic disorders. *Nat Rev Genet* **17**, 224-238, doi:10.1038/nrg.2015.25 (2016).
- 367 2 Amps, K. *et al.* Screening ethnically diverse human embryonic stem cells identifies a
368 chromosome 20 minimal amplicon conferring growth advantage. *Nat Biotechnol* **29**,
369 1132-1144, doi:10.1038/nbt.2051 (2011).
- 370 3 Lefort, N. *et al.* Human embryonic stem cells reveal recurrent genomic instability at
371 20q11.21. *Nat Biotechnol* **26**, 1364-1366, doi:10.1038/nbt.1509 (2008).
- 372 4 Werbowetski-Ogilvie, T. E. *et al.* Characterization of human embryonic stem cells
373 with features of neoplastic progression. *Nat Biotechnol* **27**, 91-97,
374 doi:10.1038/nbt.1516 (2009).
- 375 5 Nguyen, H. T. *et al.* Gain of 20q11.21 in human embryonic stem cells improves cell
376 survival by increased expression of Bcl-xL. *Mol Hum Reprod* **20**, 168-177,
377 doi:10.1093/molehr/gat077 (2014).
- 378 6 Avery, S. *et al.* BCL-XL mediates the strong selective advantage of a 20q11.21
379 amplification commonly found in human embryonic stem cell cultures. *Stem Cell
380 Reports* **1**, 379-386, doi:10.1016/j.stemcr.2013.10.005 (2013).
- 381 7 Markouli, C. *et al.* Gain of 20q11.21 in Human Pluripotent Stem Cells Impairs TGF-
382 β -Dependent Neuroectodermal Commitment. *Stem Cell Reports* **13**, 163-176,
383 doi:10.1016/j.stemcr.2019.05.005 (2019).
- 384 8 Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human
385 cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).
- 386 9 Sishc, B. J. & Davis, A. J. The Role of the Core Non-Homologous End Joining
387 Factors in Carcinogenesis and Cancer. *Cancers (Basel)* **9**,
388 doi:10.3390/cancers9070081 (2017).

389 10 Toffolatti, L. *et al.* Investigating the mechanism of chromosomal deletion:
390 characterization of 39 deletion breakpoints in introns 47 and 48 of the human
391 dystrophin gene. *Genomics* **80**, 523-530 (2002).

392 11 Inoue, K. *et al.* Genomic rearrangements resulting in PLP1 deletion occur by
393 nonhomologous end joining and cause different dysmyelinating phenotypes in males
394 and females. *Am J Hum Genet* **71**, 838-853, doi:10.1086/342728 (2002).

395 12 Gunning, A. C. *et al.* Recurrent De Novo NAHR Reciprocal Duplications in the
396 ATAD3 Gene Cluster Cause a Neurogenetic Trait with Perturbed Cholesterol and
397 Mitochondrial Metabolism. *Am J Hum Genet* **106**, 272-279,
398 doi:10.1016/j.ajhg.2020.01.007 (2020).

399 13 Lee, J. A., Carvalho, C. M. & Lupski, J. R. A DNA replication mechanism for
400 generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**,
401 1235-1247, doi:10.1016/j.cell.2007.11.037 (2007).

402 14 Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced
403 replication model for the origin of human copy number variation. *PLoS Genet* **5**,
404 e1000327, doi:10.1371/journal.pgen.1000327 (2009).

405 15 Baker, D. *et al.* Detecting Genetic Mosaicism in Cultures of Human Pluripotent Stem
406 Cells. *Stem Cell Reports* **7**, 998-1012, doi:10.1016/j.stemcr.2016.10.003 (2016).

407 16 Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection
408 of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-1592,
409 doi:10.1101/gr.092981.109 (2009).

410 17 De Coster, W. & Van Broeckhoven, C. Newest Methods for Detecting Structural
411 Variations. *Trends Biotechnol* **37**, 973-982, doi:10.1016/j.tibtech.2019.02.003 (2019).

412 18 Chaisson, M. J., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo
413 assembly of human genomes. *Nat Rev Genet* **16**, 627-640, doi:10.1038/nrg3933
414 (2015).

415 19 Merkle, F. T. *et al.* Human pluripotent stem cells recurrently acquire and expand
416 dominant negative P53 mutations. *Nature* **545**, 229-233, doi:10.1038/nature22312
417 (2017).

418 20 Canham, M. A. *et al.* The Molecular Karyotype of 25 Clinical-Grade Human
419 Embryonic Stem Cell Lines. *Sci Rep* **5**, 17258, doi:10.1038/srep17258 (2015).

420 21 de Graaf, M. N. S. *et al.* Scalable microphysiological system to model three-
421 dimensional blood vessels. *APL Bioeng* **3**, 026105, doi:10.1063/1.5090986 (2019).

422 22 Laing, O., Halliwell, J. & Barbaric, I. Rapid PCR Assay for Detecting Common
423 Genetic Variants Arising in Human Pluripotent Stem Cell Cultures. *Curr Protoc Stem
424 Cell Biol* **49**, e83, doi:10.1002/cpsc.83 (2019).

425 23 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome
426 assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*
427 **27**, 849-864, doi:10.1101/gr.213611.116 (2017).

428 24 Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the
429 human genome. *Science* **318**, 420-426, doi:10.1126/science.1149504 (2007).

430 25 Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with
431 massively parallel sequencing. *Nat Methods* **6**, 99-103, doi:10.1038/nmeth.1276
432 (2009).

433 26 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
434 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

435 27 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
436 transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

437 28 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664,
438 doi:10.1101/gr.229202 (2002).

439 29 Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**, D81-D89, doi:10.1093/nar/gkv1272 (2015).

440 30 Lobachev, K. S. *et al.* Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* **148**, 1507-1524 (1998).

441 31 Lobachev, K. S., Gordenin, D. A. & Resnick, M. A. The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. *Cell* **108**, 183-193, doi:10.1016/s0092-8674(02)00614-1 (2002).

442 32 Narayanan, V., Mieczkowski, P. A., Kim, H. M., Petes, T. D. & Lobachev, K. S. The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. *Cell* **125**, 1283-1296, doi:10.1016/j.cell.2006.04.042 (2006).

443 33 Lobachev, K. S., Rattray, A. & Narayanan, V. Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. *Front Biosci* **12**, 4208-4220, doi:10.2741/2381 (2007).

444 34 Voineagu, I., Narayanan, V., Lobachev, K. S. & Mirkin, S. M. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc Natl Acad Sci U S A* **105**, 9936-9941, doi:10.1073/pnas.0804510105 (2008).

445 35 Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* **79**, 181-211, doi:10.1146/annurev.biochem.052308.093131 (2010).

446 36 Pannunzio, N. R., Li, S., Watanabe, G. & Lieber, M. R. Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair (Amst)* **17**, 74-80, doi:10.1016/j.dnarep.2014.02.006 (2014).

447 37 Symington, L. S. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiol Mol Biol Rev* **66**, 630-670, table of contents, doi:10.1128/mmbr.66.4.630-670.2002 (2002).

448 38 Motycka, T. A., Bessho, T., Post, S. M., Sung, P. & Tomkinson, A. E. Physical and functional interaction between the XPF/ERCC1 endonuclease and hRad52. *J Biol Chem* **279**, 13634-13639, doi:10.1074/jbc.M313779200 (2004).

449 39 Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem Sci* **40**, 701-714, doi:10.1016/j.tibs.2015.08.006 (2015).

450 40 Sinha, S., Villarreal, D., Shim, E. Y. & Lee, S. E. Risky business: Microhomology-mediated end joining. *Mutat Res* **788**, 17-24, doi:10.1016/j.mrfmmm.2015.12.005 (2016).

451 41 Wang, H. & Xu, X. Microhomology-mediated end joining: new players join the team. *Cell Biosci* **7**, 6, doi:10.1186/s13578-017-0136-8 (2017).

452 42 Black, S. J., Kashkina, E., Kent, T. & Pomerantz, R. T. DNA Polymerase θ : A Unique Multifunctional End-Joining Machine. *Genes (Basel)* **7**, doi:10.3390/genes7090067 (2016).

453 43 Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **18**, 495-506, doi:10.1038/nrm.2017.48 (2017).

454 44 Arlt, M. F., Rajendran, S., Birkeland, S. R., Wilson, T. E. & Glover, T. W. De novo CNV formation in mouse embryonic stem cells occurs in the absence of Xrcc4-dependent nonhomologous end joining. *PLoS Genet* **8**, e1002981, doi:10.1371/journal.pgen.1002981 (2012).

455 45 Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4, doi:10.1186/1755-8417-1-4 (2008).

488 46 Shaw, C. J. & Lupski, J. R. Non-recurrent 17p11.2 deletions are generated by
489 homologous and non-homologous mechanisms. *Hum Genet* **116**, 1-7,
490 doi:10.1007/s00439-004-1204-9 (2005).

491 47 Barlow, J. H. *et al.* Identification of early replicating fragile sites that contribute to
492 genome instability. *Cell* **152**, 620-632, doi:10.1016/j.cell.2013.01.006 (2013).

493 48 Mortusewicz, O., Herr, P. & Helleday, T. Early replication fragile sites: where
494 replication-transcription collisions cause genetic instability. *EMBO J* **32**, 493-495,
495 doi:10.1038/emboj.2013.20 (2013).

496 49 Arlt, M. F. *et al.* Replication stress induces genome-wide copy number changes in
497 human cells that resemble polymorphic and pathogenic variants. *Am J Hum Genet* **84**,
498 339-350, doi:10.1016/j.ajhg.2009.01.024 (2009).

499 50 Ahuja, A. K. *et al.* A short G1 phase imposes constitutive replication stress and fork
500 remodelling in mouse embryonic stem cells. *Nat Commun* **7**, 10660,
501 doi:10.1038/ncomms10660 (2016).

502 51 Halliwell, J. A. *et al.* Nucleosides rescue replication-mediated genome instability of
503 human pluripotent stem cells. *bioRxiv*, 853234, doi:10.1101/853234 (2019).

504 52 Vallabhaneni, H. *et al.* High Basal Levels of γH2AX in Human Induced Pluripotent
505 Stem Cells Are Linked to Replication-Associated DNA Damage and Repair. *Stem
Cells* **36**, 1501-1513, doi:10.1002/stem.2861 (2018).

506 53 Sahoo, T. *et al.* Concurrent triplication and uniparental isodisomy: evidence for
507 microhomology-mediated break-induced replication model for genomic
508 rearrangements. *Eur J Hum Genet* **23**, 61-66, doi:10.1038/ejhg.2014.53 (2015).

509 54 Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate
510 genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**, 849-
511 853, doi:10.1038/ng.399 (2009).

512 55 Zhang, F., Carvalho, C. M. & Lupski, J. R. Complex human chromosomal and
513 genomic rearrangements. *Trends Genet* **25**, 298-307, doi:10.1016/j.tig.2009.05.005
(2009).

514 56 Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in
515 gene copy number. *Nat Rev Genet* **10**, 551-564, doi:10.1038/nrg2593 (2009).

516 57 Thompson, O. *et al.* Low rates of mutation in clinical grade human pluripotent stem
517 cells under different culture conditions. *Nature Communications* **11**, 1528,
518 doi:10.1038/s41467-020-15271-3 (2020).

519 58 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
520 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

521 59 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
522 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

523 60 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast
524 processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034,
525 doi:10.1093/bioinformatics/btv098 (2015).

526 61 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26,
527 doi:10.1038/nbt.1754 (2011).

528 531

529 532 **Acknowledgements**

533 The authors would like to thank Matthew Parker, Emily Chambers and Mark
534 Dunning of the Sheffield Bioinformatics Core, The University of Sheffield for
535 assistance and advice with performing the data processing. This work was partly

536 funded by the European Union's Horizon 2020 research and innovation program
537 under grant agreement No. 668724 and partly by the UK Regenerative Medicine
538 Platform, MRC reference MR/R015724/1. The Wellcome Sanger Institute is grateful
539 for the Wellcome Trust general core grant number 206194.

540

541 **Author Contribution**

542 PWA and IB oversaw the project. JAH, PWA and IB devised the experimental
543 design. JAH performed the cell culture, DNA extraction, qPCR and data processing.
544 Additional help for data processing was provided by the Sheffield Bioinformatics
545 Core. DB performed interphase FISH detection of chromosome 20 amplification. KJ,
546 MAQ, KO, EB and JS performed the Nanopore library preparation and whole
547 genome sequencing. The manuscript was drafted by JAH, PWA and IB.

548

549 **Competing interest**

550 The authors declare no competing financial interests.

551

552 **Figure 1 | qPCR detection of distal breakpoint positions.** **a**, A schematic showing
553 the position and order of genes probed by qPCR along the chromosome 20q11.21.
554 Primer sets were designed to target intronic regions of the genes displayed. **b**, Copy
555 number values for the human ESC line MShef7-A4, determined by qPCR for loci
556 along the length of chromosome 20q11.21. The primer location according to the
557 hg38 reference genome are also displayed with the gene names along the X axis. **c**,
558 The qPCR determined copy number for loci along the length of chromosome
559 20q11.21 in the NCRM1 human iPSC line. The copy number of four between
560 *DEFB115* and *TPX2* indicates a triplication of this region.

561

562 **Figure 2 | Breakpoint junction detection in MShef7-A4 using Nanopore**
563 **sequencing.** **a**, Sequencing read coverage of 30 kb spanning the breakpoint
564 junction at 32,273,600 bp (chromosome 20q11.21) of the hg38 reference genome.
565 Each dot indicates the read depth at a single base pair position. The red and blue
566 lines indicate the mean read depth before and after the breakpoint position,
567 respectively **b**, Schematic of the reference genome and the tandem duplication
568 detected in MShef7-A4. Junction between genome segment A-B and B-C represents
569 the proximal and distal breakpoint, respectively. The position of genes flanking and

570 the location of the *AluSz6* in relation to the breakpoint are depicted. **c**, Reference
571 sequence spanning the distal breakpoint (B – top, green), sequence of the
572 breakpoint junction (B/B fusion – middle) and the contig sequence of the distal side of
573 the proximal breakpoint (B – bottom, blue). The regions of microhomology that flank
574 the proximal and distal breakpoint is highlighted (red).

575

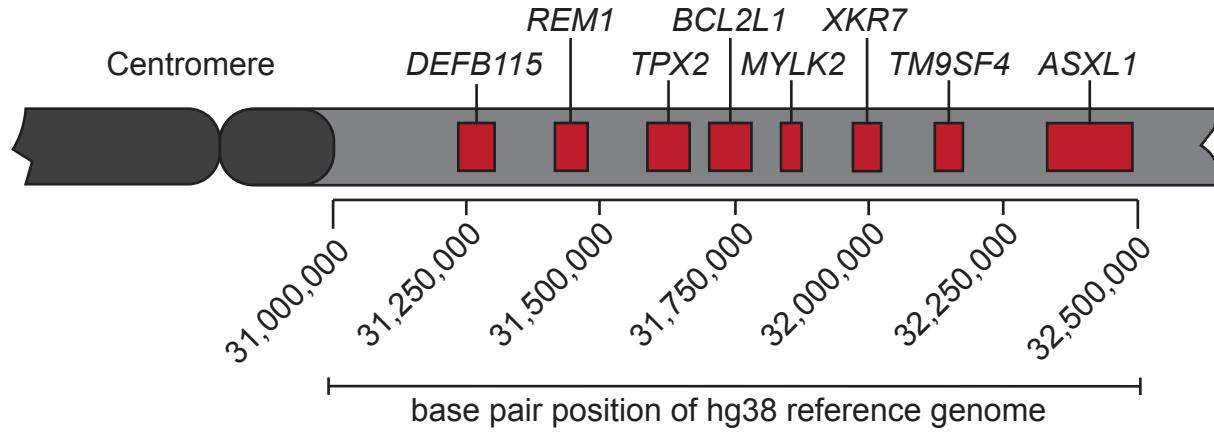
576 **Figure 3 | Breakpoint position of the tandem triplication in NCRM1. a**, Read
577 coverage of 30 kb surrounding the breakpoint junction 31,813,288 bp (chromosome
578 20q11.21) of the hg38 reference genome. The mean read depth before and after the
579 breakpoint is shown (red line and blue line, respectively). **b**, Schematic depicting the
580 reference genome and the NCRM1 tandem triplication. The distal breakpoint lies
581 between the junction of B-C and the proximal breakpoint is located on the boundary
582 of the A-B segments. The genes flanking the breakpoint, as determined by qPCR are
583 depicted. The position of the *AluSz6* identified from the Dfam database is
584 represented above the reference sequence schematic. The exact nucleotide position
585 of the proximal and distal breakpoint is written above the schematic of the tandem
586 triplication. **c**, Reference sequence spanning the distal breakpoint (B – top, green),
587 the proximal breakpoint (B – bottom, blue) and the combined amplification
588 breakpoint junction (B/B fusion – middle). The region of microhomology that flanks
589 each of the breakpoints is highlighted (red).

590

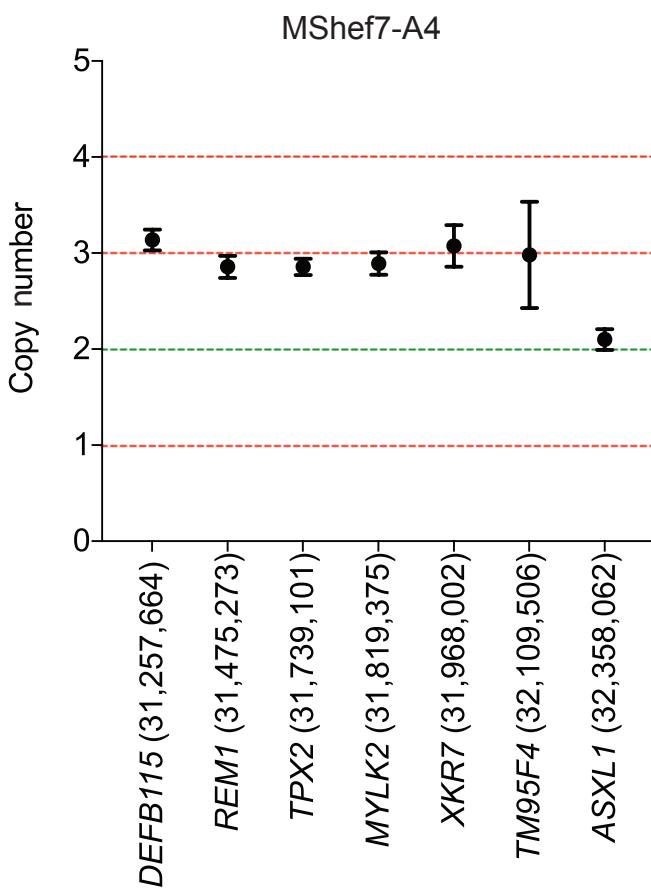
591 **Figure 4 | Replication template switching is responsible for tandem**
592 **amplification in human PSC. a**, Replication fork stalling is promoted by *Alu*
593 sequences that form hairpin loops. **b**, Replication fork repair by fork stalling and
594 template switching and/or microhomology mediated break induced replication is
595 initiated by strand invasion at a site of microhomology in the pericentromeric
596 microsatellite on the sister chromatid. Replication proceeds, duplicating 20q11.21. **c**,
597 An additional round of strand invasion and re-synthesis occurs of the other parent
598 homolog in examples of tandem triplication.

599

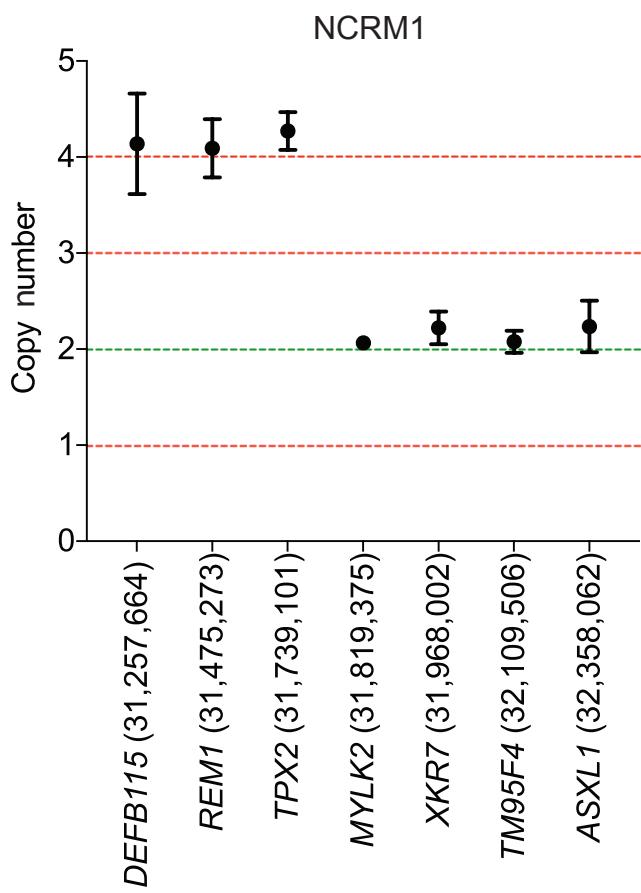
a



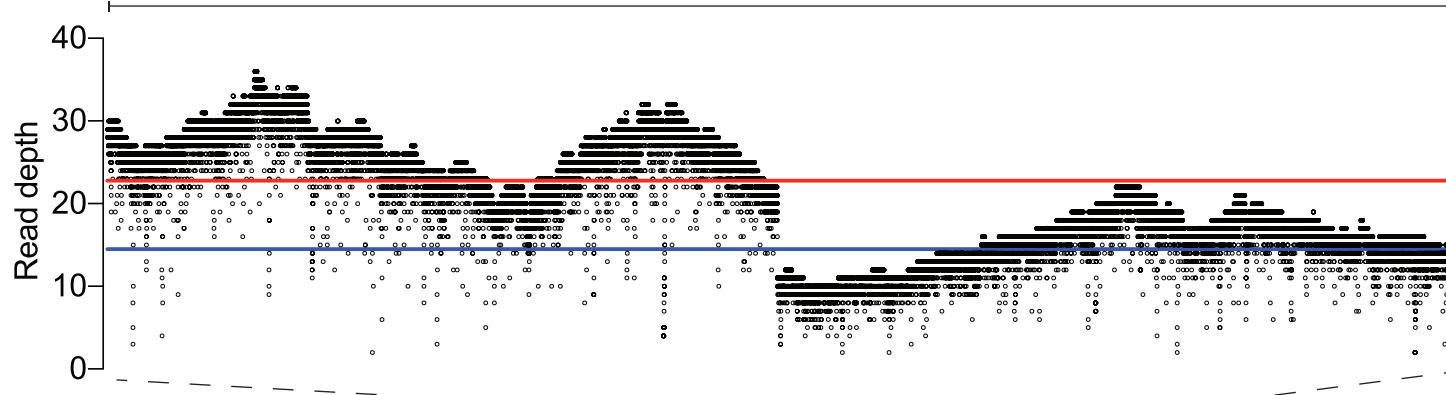
b



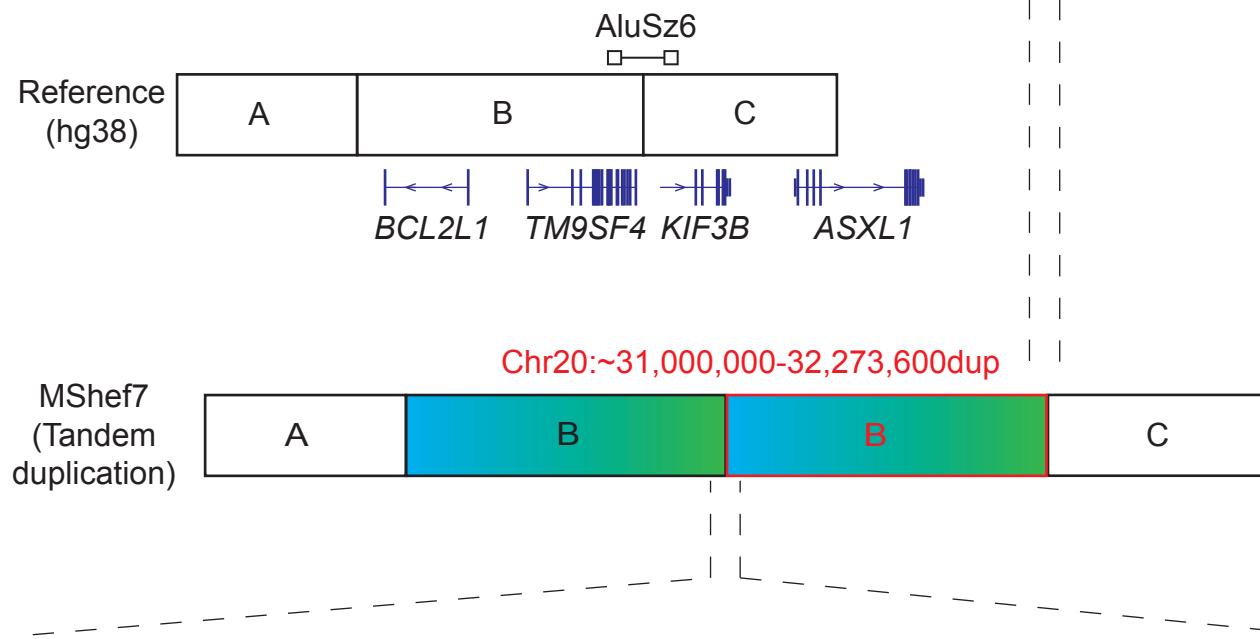
c



30 kb



b



c

Breakpoint



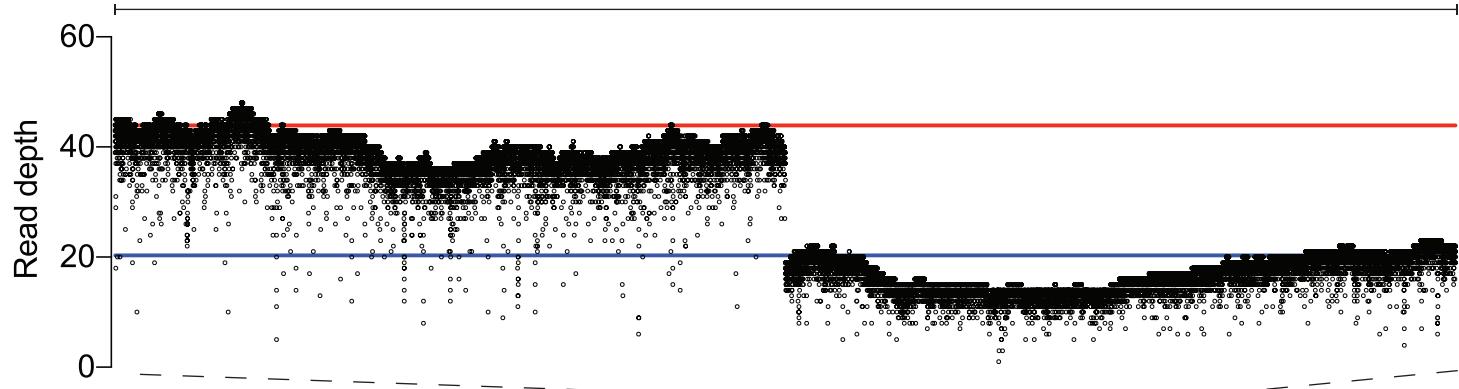
A: TGAGAATCACTTAAACCCGGGAGGT AGAGGTTGCAGTGAGCTGAGATTGC

B/B fusion: TGAGAATCACTTAAACCCGGGAGGT TTAAGAATCACTTAAACCGAAAGGAA

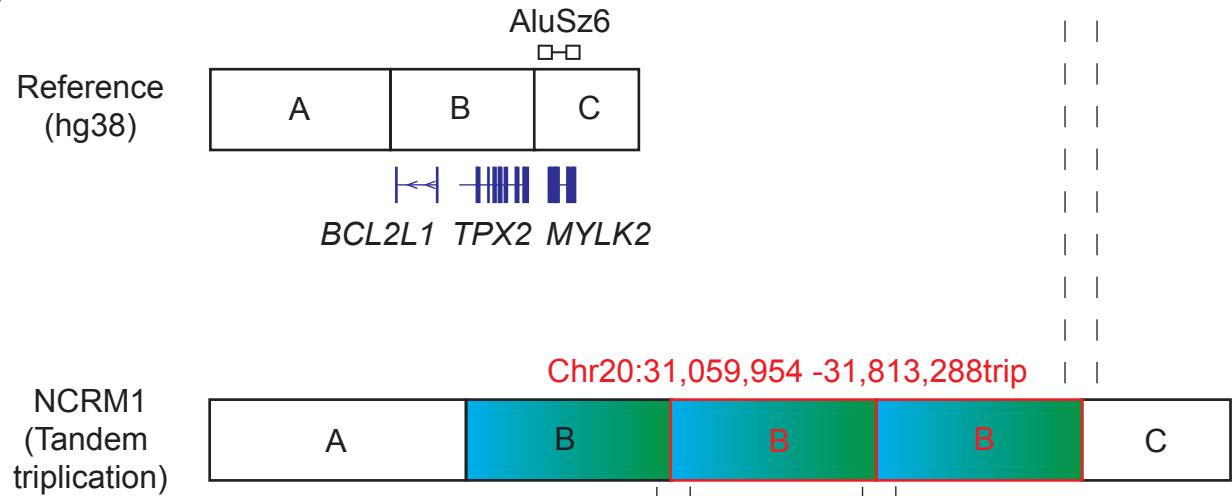
B:

TTAAGAATCACTTAAACC GAAAGGAA

30 Kb



b



c

Breakpoint

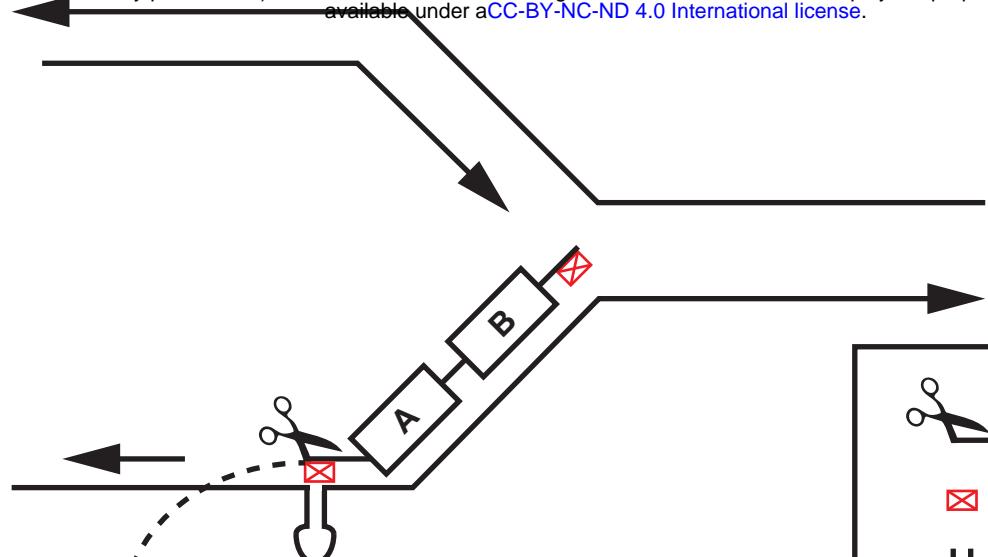


B: GGGCATTCAAGGGAAACAGAAATTG **AA**GTTTCTGGCTGGCGCAGTGGCT

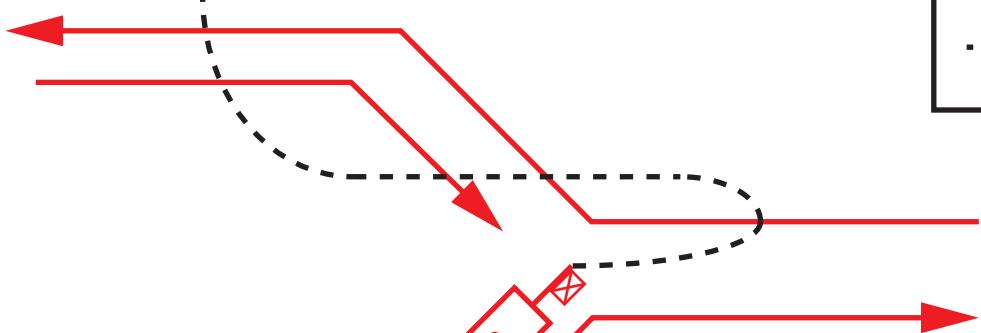
B/B fusion: GGGCATTCAAGGGAAACAGAAATTG **TG**GAATAGAATTGAATGGAAATTGAATG

B: **TG**GAATGGAAATTGAATCAACCAGAG **TG**GAATAGAATTGAAT**TG**GAATTGAATG

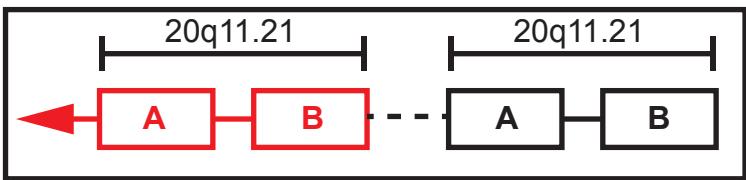
a



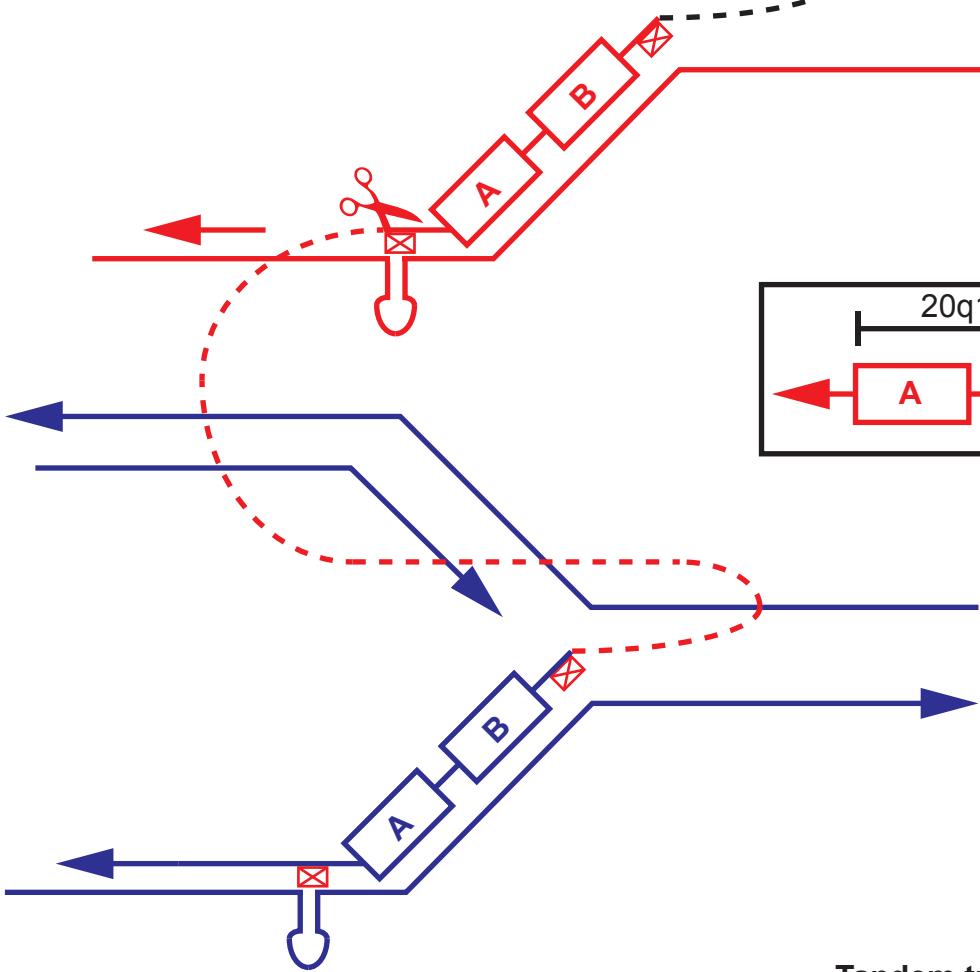
b



Tandem duplication



c



Tandem triplication

