CaBagE: a Cas9-based Background Elimination strategy for targeted, long-read DNA sequencing

Amelia Wallace[1,2], Thomas A. Sasani[3], Jordan Swanier[1], Brooke L. Gates[1], Jeff Greenland[4], Brent S. Pedersen[1,2], K-T Varley[4], Aaron R. Quinlan[1,2,3,*]

1. Department of Human Genetics, University of Utah, Salt Lake City, UT
2. Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT
3. Department of Genome Sciences, University of Washington, Seattle, WA
4. Department of Oncological Sciences, Huntsman Cancer Institute, Salt Lake City, UT
5. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
* To whom correspondence should be addressed

Abstract

A substantial fraction of the human genome is difficult to interrogate with short-read DNA sequencing technologies due to paralogy, complex haplotype structures, or tandem repeats. Long-read sequencing technologies, such as Oxford Nanopore's MinION, enable direct measurement of complex loci without introducing many of the biases inherent to short-read methods, though they suffer from relatively lower throughput. This limitation has motivated recent efforts to develop amplification-free strategies to target and enrich loci of interest for subsequent sequencing with long reads. Here, we present CaBagE, a novel method for target enrichment that is efficient and useful for sequencing large, structurally complex targets. The CaBagE method leverages the stable binding of Cas9 to its DNA target to protect desired fragments from digestion with exonuclease. Enriched DNA fragments are then sequenced with Oxford Nanopore's MinION long-read sequencing technology. Enrichment with CaBagE resulted in up to 416X coverage of target loci when tested on five genomic targets ranging from 4-20kb in length using healthy donor DNA. Four cancer gene targets were enriched in a single reaction and multiplexed on a single MinION flow cell. We further demonstrate the utility of CaBagE in two ALS patients with *C9orf72* short tandem repeat expansions to produce genotype estimates commensurate with genotypes derived from repeat-primed PCR for each individual. With CaBagE there is a physical enrichment of on-target DNA in a given sample prior to sequencing. This feature allows adaptability across sequencing platforms and potential use as an enrichment strategy for applications beyond sequencing. CaBagE is a rapid enrichment method that can illuminate regions of the 'hidden genome' underlying human disease.

Introduction

While short-read DNA sequencing technologies have enabled the discovery of genetic variants underlying numerous rare genetic disorders (1, 2), a large fraction of the human genome remains very difficult to interrogate with short-reads. These so-called "hidden" regions are difficult to sequence with short-read technologies owing to a mixture of sequence paralogy, complex haplotype structures, and tandem repeats (3, 4). Collectively these hidden regions impact over 700 genes (4).

79 Paralogous sequences consist of ancestrally duplicated genomic segments. These sequences
80 can be entire genes or segmental duplications (a duplicated sequence >1kb) and can appear in tandem
81 or interspersed throughout the genome. Due to high homology elsewhere in the genome, there is
82 ambiguity when mapping short reads to these regions. Thus, approximately 70% of segmental
83 duplications are not sequence-resolved in the human reference genome, and are simply annotated as
84 gaps (5). Polymorphic mobile element insertions are similarly difficult to map, as multiple copies exist
85 throughout the genome and yet broad phenotypic effects of this variation have been suggested (6, 7).
86 Short tandem repeats (STRs) are another class of genomic sequence that is difficult to resolve,
87 and have estimated mutation rates orders of magnitude higher than single nucleotide variation (8). Yet
88 the contribution of tandem repeats to phenotypic heterogeneity remains poorly understood due to
89 limitations in our ability to accurately detect and genotype these features. STR expansions underlie
90 over 40 developmental and neurological disorders (9), highlighting a clear need for better molecular
91 and informatics techniques to genotype these features across individuals (10). The (CCCCGG)n repeat
92 expansion in *C9orf72* segregates with up to 40% of familial amyotrophic lateral sclerosis (ALS) cases
93 (11) and is one of few established causes of the disease (12). However, sequencing through complete
94 *C9orf72* repeat expansions is difficult; therefore, diagnostics rely on laborious, semi-quantitative
95 methods such as Southern blot or repeat-primed PCR (RP-PCR). In contrast, long-read sequencing
96 (LRS) can, in principle, provide essential quantitative information such as repeat length and sequence
97 content, which may reveal connections between allelic polymorphism and clinical phenotypes such as
98 severity and age of onset.
99 Oxford Nanopore Technologies (ONT) long-read sequencing (LRS) (13) enables direct
100 measurement of loci containing complex structures without introducing biases due to amplification or
101 polymerase slippage, and permits highly accurate mapping. At the same time, native modifications to
102 DNA or RNA are preserved and can be detected concurrently with the nucleic acid sequence. While
103 higher error rates limit the accuracy of single nucleotide variant discovery compared to Illumina DNA
104 sequencing, long reads that completely span hidden genomic regions offer the potential for

105 comprehensive and accurate discovery of the structural variation therein. A recent study sequenced
106 fifteen human genomes with long reads and showed that over 80% of structural variants genotyped
107 were missed when called from Illumina data for the same subjects (14). In fact, the sensitivity of LRS
108 can greatly exceed standard next generation sequencing (NGS), particularly for large insertions
109 (>50bp) (15).

110 The ONT MinION is particularly advantageous for diagnostics, as it is affordable, portable, and
111 capable of generating reads up to 1Mb. A pressing limitation of the MinION however, is the low
112 throughput relative to other sequencing technologies (e.g., Illumina). This has motivated recent efforts
113 to enrich loci of interest for subsequent LRS without amplification, which limits target-lengths and can
114 introduce PCR bias. Many emerging methods leverage the highly specific targeting ability of the
115 CRISPR/Cas9 system, but strategies vary widely and have unique strengths and limitations related to
116 DNA input requirements, protocol execution time, target size restrictions, and efficiency (16-18).
117 CATCH was one of the first methods published and relies on pulsed-field gel electrophoresis to
118 physically isolate a DNA target of known size that is first cut at the flanks with Cas9 (17). This method
119 is amenable to very large targets (200kb) because DNA is protected from shearing in agarose plugs.
120 However, if the target length is variable or unknown, as with pathogenic repeat expansions, the
121 method suffers and amplification is often required to obtain high sequencing yields. Subsequent
122 strategies improved yield and efficiency by enriching sequencing data for target sequences without
123 physical enrichment of target DNA fragments in the sample. The nCATS method uses
124 dephosphorylation to prevent adapter ligation in sample DNA (19). Next, the 5-prime phosphates
125 flanking a target are restored using the endonuclease activity of Cas9, so that those fragments alone
126 are available for sequence adapter ligation. This method performs best for targets up to 20-30kb.
127 Most recently, ReadUntil, a computational method for real-time enrichment during sequencing, has
128 been expanded to human genomic targets (20). The method utilizes real-time sequence identification
129 to allow DNA fragments to be rejected from nanopores prior to completion of sequencing, thus
130 performing targeted sequencing without specialized library preparation. ReadUntil does not have cost

131 associated with assay design, reagents, or equipment, however rejection of fragments from pores
132 does decrease overall output from flow cells and thus reduces yield across individual targets (20).
133 Here we introduce a novel Cas9-based Background Elimination strategy, CaBagE. In contrast to
134 nCATs and ReadUntil, CaBagE physically enriches genomic DNA for specific target loci, producing
135 enrichment with comparable efficiency in terms of library preparation time and sequence output.

136 Cas9 is a single-turnover enzyme with endonuclease activity that can be easily directed to
137 specific genomic sequences using guide RNAs. The complex formed between the enzyme, its RNA
138 guide, and target DNA is very stable, and forcibly dissociates only under harsh environmental conditions
139 (21). *In vitro* studies have shown that the natural dissociation time of Cas9 from its DNA target is
140 approximately 6 hours (22). When challenged with competing proteins, Cas9 remains tightly bound in
141 most cases (23). We were therefore motivated to ask whether this property of Cas9 extends to multiple
142 progressing exonucleases. If so, one can leverage exonucleases as a means to deplete background
143 DNA and enrich for targeted loci that are bound and therefore protected by Cas9 on either side.

144 Exonucleases have previously been used to eliminate background DNA in NGS libraries. For
145 example, Nested Patch PCR protects target DNA from digestion by capping the target sequences with
146 adapters containing phosphodiester bonds (24) and ChIP-exo protocols rely on proteins bound to DNA
147 to protect the “footprint” from exonuclease activity (25). By directing Cas9 binding to either side of a
148 specific target locus, we show that the DNA flanked by Cas9 is preserved amidst extensive digestion
149 of genomic DNA by exonucleases, allowing for highly specific target enrichment without PCR. By
150 coupling Cas9-based background elimination with long-read sequencing technology, we demonstrate
151 target sequence enrichment in previously poorly characterized regions of the human genome. Further,
152 we combine this output with a computational approach that allows clustering of long-read sequence
153 alignments to yield genotypes across a pathogenic repeat expansion in *C9orf72*. This generalizable
154 molecular framework is fast, accurate, and multiplex-ready, to characterize recalcitrant yet medically
155 important genes.

156 **Results**

157 *Cas9 Background Elimination (CaBagE) targeted*

158 *sequencing strategy overview*

159 To enrich for a genomic region of interest, we developed

160 a method that uses Cas9 to selectively protect target

161 DNA from background elimination by exonucleases (**Fig**

162 **1**). First, Cas9 is targeted to both sides of a region of

163 interest using locus-specific guide RNAs. The distance

164 between the enzymes, effectively the target fragment

165 length, is highly flexible and limited only by the ability to

166 design guide RNAs flanking the target and the average

167 fragment length of source genomic DNA. Immediately

168 following Cas9 binding, Exonucleases I, III, and Lambda

169 are introduced to degrade single stranded DNA, and

170 double-stranded DNA from the 3-prime and 5-prime

171 direction, respectively. These enzymes degrade most

172 DNA present in the sample with the exception of the

173 fragments flanked by the Cas9 enzymes, namely, the

174 DNA target of interest. Heat incubation is then used to

175 inactivate the exonucleases and force dissociation of

176 the Cas9 enzyme from the target DNA. Then, the ends

177 of the target DNA fragments are available for A-tailing

178 and ligation of the sequencing adapters. Sequencing

179 libraries are prepared beginning with the adapter

180 ligation step of the ONT Cas-mediated PCR-free

181 enrichment protocol (developed for use with nCATs)

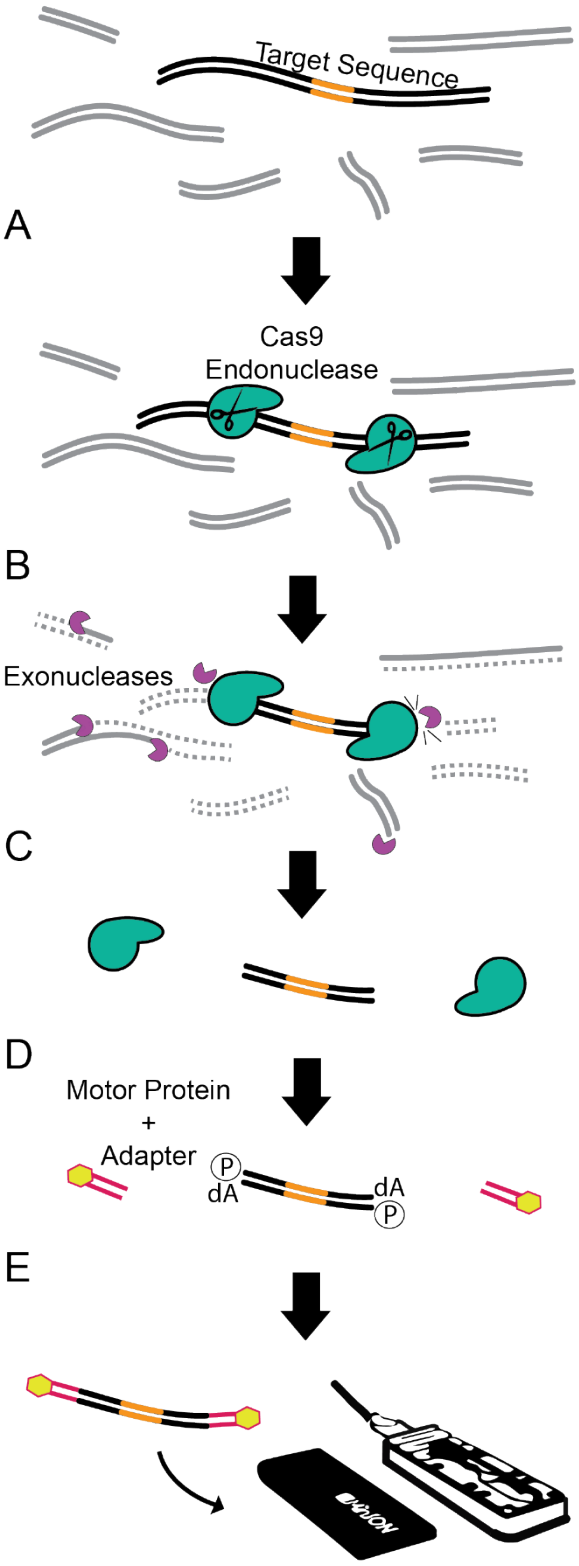


Figure 1: Schematic of Cas9 Background Elimination strategy. A) Cas9 is bound to either side of target sequence. B) Off-target DNA is digested with a combination of exonucleases. C) Heat is used to dissociate that Cas9 and inactivate the exonucleases. D) On-target fragment is available for A-tailing and sequence adapter ligation. E) Target fragments are sequenced on the MinION for 48 hours.

182 and sequenced on a single MinION flow cell for 48 hours. Target enrichment and library preparation
183 can be completed in approximately 6 hours.

184 *Cas9 prevents processive exonuclease from degrading DNA target*

185 To test whether bound Cas9 prevents DNA degradation by a combination of three processive
186 exonucleases, a 997bp synthetic double-stranded DNA gBlock (IDT) was designed to contain multiple
187 guide RNA target sites. Cas9 cleavage requires that the target DNA, which is complimentary to the
188 RNA guide, contains a 3bp protospacer adjacent motif (PAM) at its 3′ end. Cas9 binding affinity differs
189 between the PAM-proximal and distal sides of the cleavage site (22). Therefore, the gBlock was
190 designed such that flanking pairs of target sites could be in either “PAM-in” or “PAM-out” orientation,
191 where the PAM sequences contained in the paired target sites are oriented toward or away from each
192 other, respectively. (**Fig 2A**). Upon exonuclease challenge, stretches of gBlock DNA contained
193 between two bound Cas9 enzymes were protected from degradation during a 2-hour incubation, while
194 gBlock stretches not bound on both sides by Cas9 were completely degraded (**Fig 2B**). DNA was

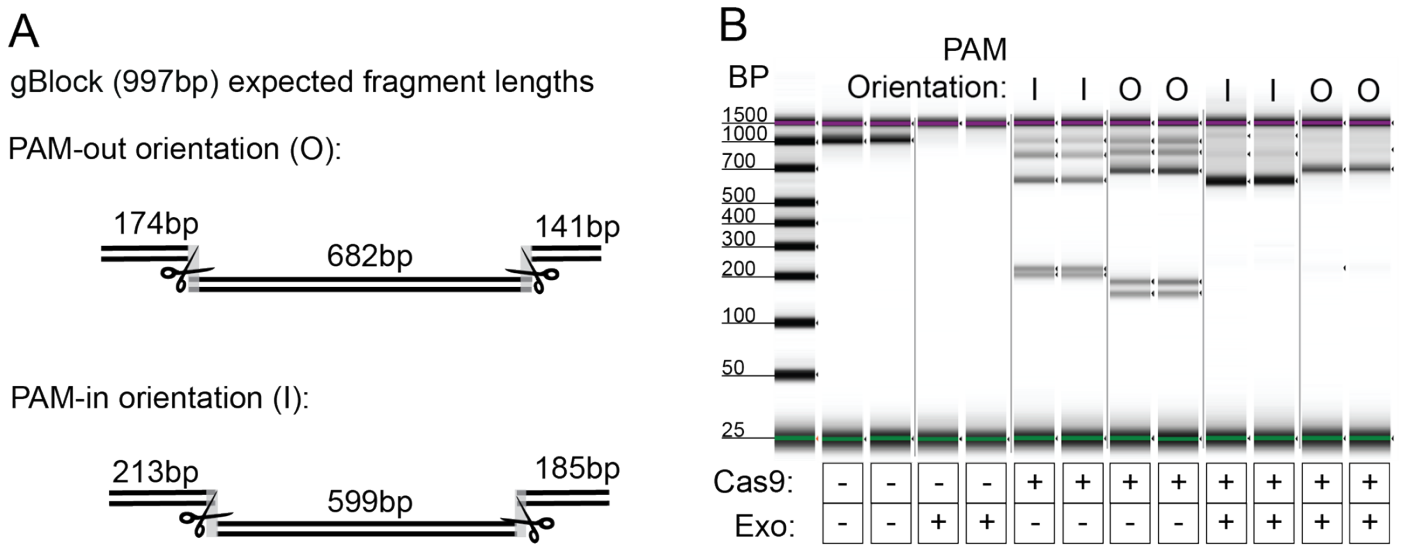


Figure 2: A. gBlock assay design for Cas9 challenge with exonuclease. gBlock contained two pairs of gRNA target sites, one with PAM-out orientation and one with PAM-in orientation. Upon Cas9 binding (depicted by scissors), each set of target sites generate 3 unique fragment lengths. B. Capillary electrophoresis results from exonuclease challenge experiment with Cas9. 15nM gBlock DNA was incubated with 40nM ribonucleoprotien complex, followed by digestion with a combination of exonucleases for 2 hours. When Cas9 is used without exonucleases, the gBlock is cut to produce expected fragment lengths. Upon challenge with exonuclease, only the fragments flanked on both sides by Cas9 remain in the sample. (I=in; O=out). Wells in gel image are re-ordered for clarity.

195 protected between two Cas9 enzymes regardless of PAM orientation. However, PAM-in orientation
196 resulted in the highest estimated concentration of the protected segment of DNA following exonuclease
197 challenge (mean PAM-in 225pg/uL, mean PAM-out 106.5pg/uL) and so was selected as the preferable
198 orientation for target enrichment. As expected, in the absence of Cas9, nearly all gBlock DNA is
199 degraded by the three exonucleases (**Fig 2B**).

200 *Yield and coverage*

201 We targeted 5 loci using the CaBagE method; guide RNAs were selected with "PAM-in" orientation and
202 are listed in **S1 Table**. As a proof of concept, we targeted loci in healthy donor DNA, including a highly
203 variable hexanucleotide repeat in *C9orf72*, and four cancer-related genes with guide RNAs previously
204 validated for PCR-free targeted sequencing (*GSTP1*, *KRT19*, *GPX1*, *SLC12A4*) (16). We multiplexed
205 up to four loci per reaction and sequenced on a single flow cell. Target enrichment and sequencing for
206 each locus was run in duplicate and runs targeted one or four loci, respectively, on a single flow cell
207 (**Table 1**). Multiplexing multiple loci on a single flow cell did not significantly impact coverage across
208 each individual locus, though coverage did vary from run-to-run.

Table 1: Results from individual CaBagE runs in DNA from healthy donors

Run ID	Total Reads[a]	Target(s) per flowcell	Target Length (bp)	On-Target Read Depth
L1R1	536,943	*C9orf72*	4,044	416
L1R2	485,412	*C9orf72*	4,044	179
L4R1	845,510	*GSTP1*	17,819	91
		KRT19	18,189	162
		GPX1	13,644	190
		SLC12A4	24,389	116
L4R2	681,142	*GSTP1*	17,819	39
		KRT19	18,189	61
		GPX1	13,644	54
		SLC12A4	24,389	63

[a]*MapQ = 60*

209 Sequence reads were aligned using MiniMap2 (26) and on-target reads were visualized with IGV (27).
210 On-target reads were considered as any reads that overlap the target region by at least 1bp and were
211 counted using samtools (28). For each target locus, on-target reads aligned to the anticipated Cas9
212 cleavage site. When sequencing across the repeat region of C9orf72 in a healthy donor, on target

213 reads spanned the entire locus, terminating at the Cas9 cleavage sites on either side and both DNA

214 strands were equally represented in the alignment data (**Fig 3**). The vast majority of off-target reads

215 were <1,000bp in length. We found that selecting for larger fragments after adapter ligation using the

216 ONT Long Fragment Buffer, which selects for fragments longer than 3kb, resulted in fewer reads overall

217 and fewer on-target reads despite target fragments being >3kb in length. For example, two independent

218 runs using the same initial DNA sample with Short Fragment Buffer and Long Fragment Buffer

219 generated 2,707,912 reads with 71 on-target and 99,191 reads with 14 on-target, respectively. As

220 expected, the Long Fragment Buffer resulted in an enrichment of longer reads and also higher

221 proportion of reads with map quality ≥ 60 (**S1 Fig**). However, due to the difference in the number of

222 on-target reads, all CaBagE runs utilize the Short Fragment Buffer. Off-target reads were typically short

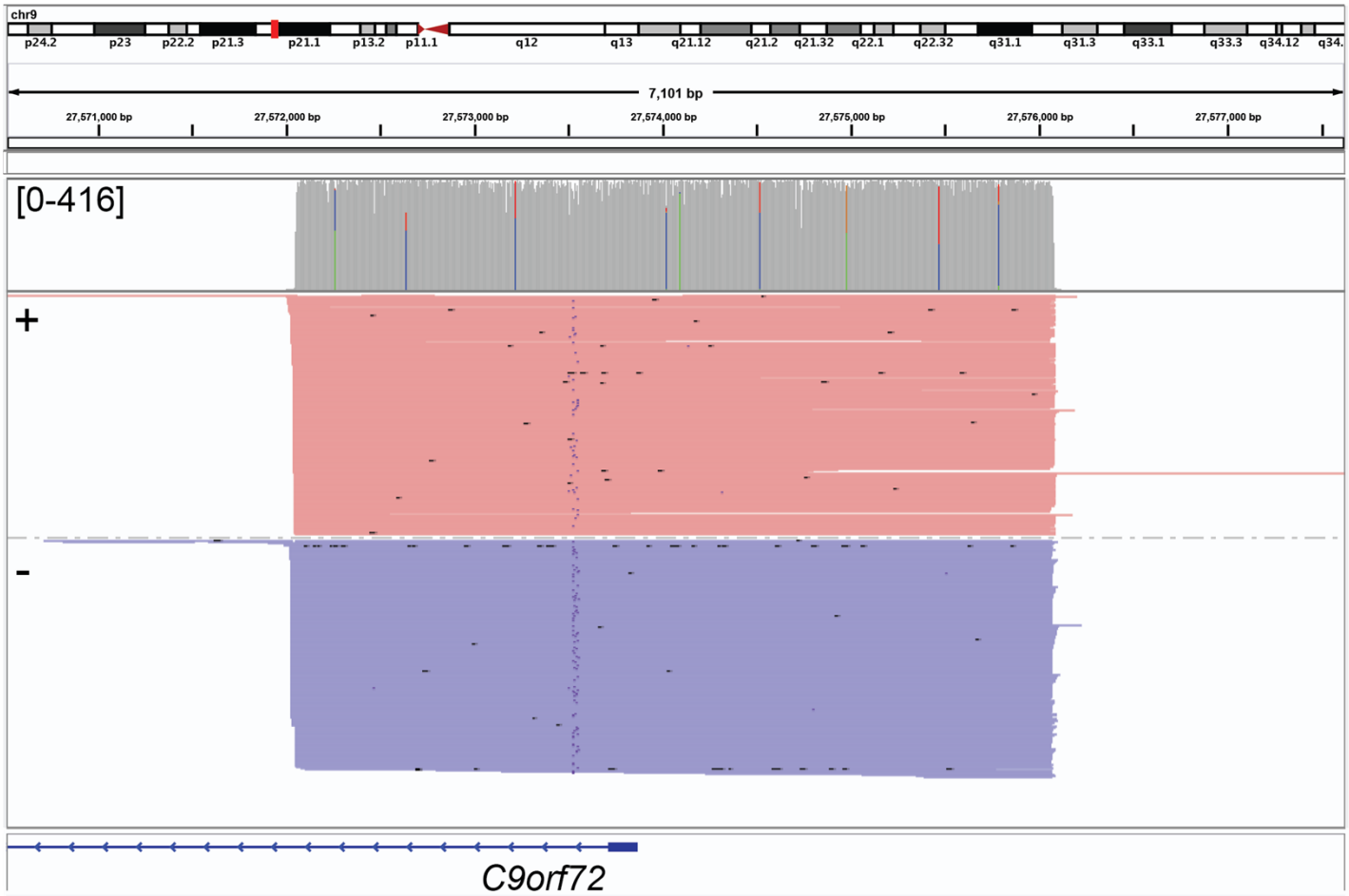


Figure 3: On-target reads (416X coverage) produced using the CaBagE target sequence enrichment strategy to capure the *C9orf72* repeat-expansion locus in a healthy individual. IGV screenshot shows aligned reads sorted by strand (plus, red; minus, blue).

223 (median length=559bp, **Fig 4A**) and randomly distributed throughout the genome, suggesting that they
224 arose primarily by incomplete exonuclease digestion rather than off-target guide RNA binding. To
225 determine whether off-target reads were enriched for other genomic features that might be
226 preferentially protected from exonuclease digestion, we tested for a statistical enrichment for overlaps
227 with G-quadruplex annotations (permutation test, p=0.97) (29, 30); further, the GC content distribution
228 centered at 39.5%, reflecting the genome average (**S2 Fig**). Ten genomic regions showed pile-ups with
229 >50X coverage, and these sites were annotated as having long chains of simple tandem repeats;
230 therefore, the pile-ups were likely the result of mapping errors. The total number of reads generated
231 from CaBagE targeted sequencing ranged from ~800,000 to 2.7 million; when restricting to reads with
232 map quality $\geq$ 60, ~40% of off-target reads are removed (**Fig 4B**).

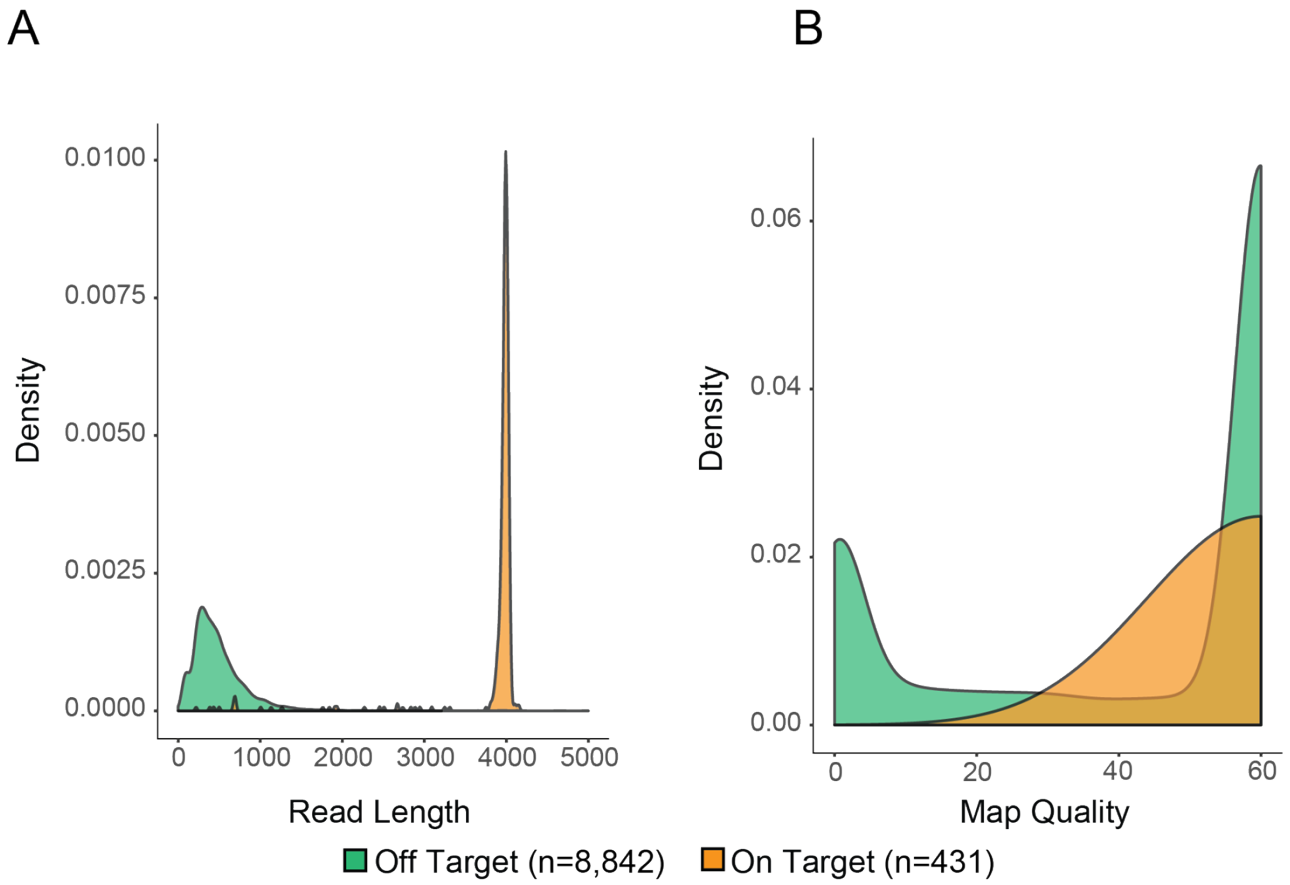


Figure 4: Characteristics of a random sample of 1% of primary alignments from off-target reads and all on-target reads from a CaBagE run enriching for a 4,044bp target in a healthy individual. A) Kernel density plot of read lengths in off- and on-target reads B) Kernel density plot of map quality scores in off- and on-target reads

233 To determine how target enrichment with CaBagE compares to nCATs in our hands, side-by-
234 side sequencing runs targeting four loci were conducted. Using identical DNA input samples,
235 concentrations, and sequencing parameters on flow cells that performed similarly during Platform QC
236 (i.e. similar number of active pores available) the on-target read depth at the target locus achieved with
237 nCATs was 2.6 to 10.7-fold higher than that of CaBagE (**S2 Table**). While the CaBagE off-target
238 sequencing rate resulting from incomplete exonuclease digestion likely contributed to its relatively lower
239 on-target yield, coverage across the targets produced by CaBagE were sufficiently high ($\geq$ 30X) for
240 locus characterization.

241 *CaBagE target enrichment produces reads that span a pathogenic repeat expansion in known carriers.*
242 To test the ability of our target enrichment strategy to sequence through disease-specific tandem repeat
243 alleles in affected individuals, we applied CaBagE to two de-identified DNA samples with known
244 *C9orf72* repeat expansions from the National Institute of Neurological Disorders and Stroke (NINDS)
245 repository at the Coriell Institute. Repeat copy numbers for these individuals were previously estimated
246 using gene specific repeat-primed PCR (RP-PCR) and gel electrophoresis (31). The upper limit of
247 detection for repeat copy number estimation using RP-PCR is ~950 copies and genotypes above 950
248 copies are denoted as EXP, for expanded (31). The PCR-based copy-number estimates for the two
249 samples' expanded alleles are 704 and EXP, respectively, where the EXP allele was beyond the upper
250 limit of detection with PCR-based methods. Targeted sequencing of the *C9orf72* repeat expansion
251 using the CaBagE method in these individuals resulted in high (>60X) depth of coverage at the target
252 locus (**Table 2**). A bias for the minus strand was observed in both NINDS ALS samples **(Fig 5**). Strand
253 bias has been previously observed when sequencing across repeats with ONT (32, 33) and can be
254 correlated with repeat length, however we observed no apparent relationship between strand and
255 repeat size. The G-rich and C-rich repeats of sense and antisense ssDNA at this locus form

Table 2: Results from CaBagE runs in known carriers of the *C9orf72* repeat expansion

Coriell ID	RP-PCR CN Estimate	Total Reads[a]	On-Target	Total Spanning Reads	Reads spanning	CaBagE CN Estimate

			Read Depth		expanded allele	
ND11386	8/704	1,490,712	115	98	21	9/749/1,893
ND13803	2/EXP	852,155	71	66	7	2/808/1,538
[a]MapQ=60						
*RP-PCR repeat-primed PCR and agarose gel electrophoresis derived genotypes from Bram *et al* (31), CN copy number						

256 different secondary structures, which may migrate through the sequencing pores at different rates (34).

257 Spanning reads were defined as reads that aligned to both the 5 prime and 3 prime flanking
258 sequence around the repeat, as well as the full repeat sequence itself. Per-read hexanucleotide repeat
259 copy number was estimated by counting the number of bases between the position in the read that
260 aligned immediately upstream of the repeat and immediately downstream, divided by six, the repeat
261 motif length. Allele-specific repeat copy numbers were estimated from subgroup means derived from a
262 Gaussian mixture model where the number of clusters was determined a priori by visually counting
263 distinct peaks from a read-length histogram. In both samples, the read-length histograms showed 3
264 populations of spanning read lengths (**Fig 5**) and triallelic repeat copy number estimates are listed in
265 **Table 2**.

266 In sample ND11386, the majority of the expanded reads supported a copy number estimate 749
267 (**Fig 5A**) and for ND13803, the majority of expanded reads supported a copy number 1,538 (**Fig 5C**),
268 consistent with the estimates derived from RP-PCR. In both samples, the largest alleles detected were
269 absent from the RP-PCR results, as they are larger than the detectable limit of the assay. Further, both
270 samples showed a strong bias to sequencing the shortest allele, representing 79% and 91% of the
271 spanning reads, respectively. This is likely an artifact of the technology sequencing shorter fragments
272 more efficiently, as has been previously observed (35-37) and the fact that longer (e.g. expanded)
273 fragments are more likely to be damaged between the flanking Cas9 binding sites, which would result
274 in failure of enrichment. The presence of the three alleles in each sample were confirmed by repeated
275 library preparation and sequencing of the same samples (**S3 Fig**). The appearance of the third alleles

276 in these samples could be artifacts of cell line transformation from which the DNA was derived. Multiple
277 populations of allele lengths have been previously observed in cell lines and was observed in ND11836
278 via Southern blot during validation of a PCR-based assay (38).

279 **Discussion**

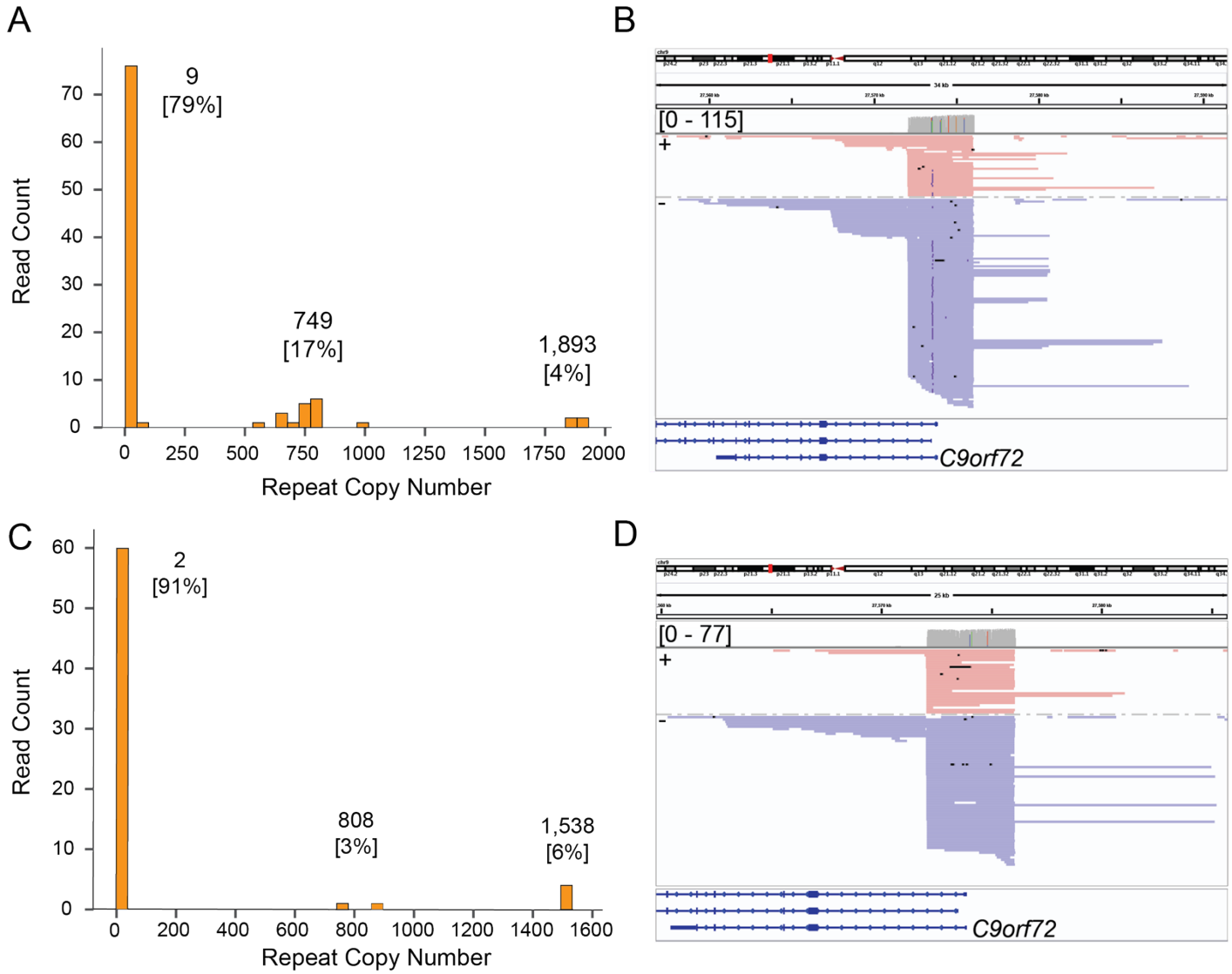


Figure 5: Targeted sequencing across repeat expansion at *C9orf72* in two ALS cases. A) Histogram of repeat copy number distribution and copy number estimates derived from a Gaussian mixture model for ND11836 (copy number, [percent of on-target reads]). B) IGV screen shot showing expanded reads across the hexanucleotide repeat for subject ND11836. C) Histogram of repeat copy number distribution and copy number estimates derived from a Gaussian mixture model for ND13803 (copy number, [percent of on-target reads]). D) IGV screen shot showing expanded reads across the hexanucleotide repeat for subject ND13803.

280 We developed a method to enrich long-read sequence data for specific target loci that is fast,
281 efficient, and amenable to the multiplexing of multiple target loci. By relying on the binding kinetics of
282 the Cas9 enzyme to its RNA-guided target, CaBagE can flexibly enrich for targets so long as most

283 fragments in the input DNA are intact between Cas9 binding sites. Therefore, to pursue very large
284 targets (>~30Kb) will likely require ultra-high molecular weight DNA, which must be obtained with
285 specialized DNA extraction methods such as agarose plugs or specialized ultra-high molecular weight
286 DNA extraction kits.

287 CaBagE performs similarly in terms of prep time and input requirements, but with a lower yield than
288 a popular competing method, nCATS (16). However, unlike nCATS and ReadUntil for amplification-
289 free targeted sequencing, the enrichment achieved from CaBagE occurs at the DNA-level, where the
290 ratio of on- to off-target DNA physically increases in the sample prior to sequencing. This unique feature
291 of CaBagE amplification-free enrichment may therefore prove useful for applications beyond long-read
292 DNA sequencing where isolating specific DNA sequence is required, for example to increase specificity
293 during size separation or for PCR-free cloning. Despite high on-target coverage, CaBagE sequences
294 off-target fragments at a high rate owing to both incomplete exonuclease digestion and the lack of a
295 selection step for long fragments. However, since an average CaBagE run yields ~1 Gb of sequence,
296 which is well under the >8 Gb typical throughput for the MinION R9.4.1 using the ligation kit, we expect
297 this high off-target rate isn't detracting from our on-target depth.

298 We demonstrated CaBagE's ability to capture pathogenic repeat-expansion alleles in two ALS
299 patients. We discovered 3 distinct read-length populations in each sample, potentially representing
300 significant mosaicism. This observation is not uncommon in studies of repeat expansions where
301 genotyping assays are performed on cell line-derived DNA (38, 39). Determining whether these 3
302 alleles were present in the blood of these patients or arose as an artifact of cell culture or sequencing
303 would require both blood and LCL-derived DNA from the same individual, which is not available for the
304 NINDS ALS Collection.

305 We note that several challenges remain in utilizing targeted long-read sequencing in the
306 identification of repeat expansions. First, longer repeat expansions have greater instability, and growing
307 and shrinking of repeat length is common and variable cell-to-cell and tissue-to-tissue in patients with
308 the *C9orf72* repeat expansion and other repeat expansion diseases (40, 41). The observation of mosaic

309 lengths of short tandem repeats in ours and previous studies poses an interesting challenge for
310 estimating repeat-length genotypes and further calls into question whether creating a consensus
311 sequence for the repeat is biologically meaningful. However, estimating a distribution of repeat lengths
312 within an individual may be of clinical relevance, wherein a greater spread may indicate instability,
313 which in turn may be correlated with pathogenesis. Second, sequencing across the repeat expansion
314 using CaBagE resulted in a strong bias in the sequencing data toward shorter alleles. Therefore, in
315 addition to needing high depth of coverage to detect the expansion, this length bias also complicates
316 the ability to accurately quantify relative clonal contributions in cases where somatic mosaicism is
317 present. Carefully extracted, high molecular weight DNA may not have as pronounced a bias, as longer
318 fragments won't be depleted in those samples. Overcoming this bias would be required for future
319 studies of mosaicism. Accurate base calling also remains a challenge using ONT technologies,
320 particularly in repeats with high GC content. We note that several reads representing the expanded
321 alleles failed base calling using Guppy and were retrieved from the "fastq_fail" folder generated by the
322 MinKNOW software. Nanosatellite is specifically designed software to accurately characterize tandem
323 repeat sequences where standard base calling algorithms perform poorly (35) and may be an effective
324 alternative for repeat sequence characterization as the performance of Guppy improves. Strand biases
325 are also exacerbated across repeats sequenced with long-read technologies (32) and should be
326 considered during repeat sequence characterization.

327 CaBagE's amplification-free targeted sequencing can be used to effectively sequence across
328 multiple, large loci on a single MinION flow cell. The method is not limited to the MinION, but should be
329 adaptable to any long-read sequencing technology. Future work to improve the method will include
330 increasing the efficiency of the exonuclease digestion and possibly adapting the method to be used for
331 tiling across much larger targets with catalytically inactive dCas9. Currently, CaBagE is a unique target
332 enrichment strategy that does not simply enrich sequencing data for specific loci, but enriches the DNA
333 sample itself without amplification, thus potentially providing utility beyond long-read sequencing. As
334 methods for DNA preparation, sequencing, and downstream data processing continue to improve,

335 targeted sequencing methods like CaBagE will become indispensable in large-scale, cost-effective
336 studies of complex structural variation.

337 **Methods**

338 *Samples. A* 997bp gBlock was designed to contain four gRNA target sites (**Supplementary Table 1**).
339 Deidentified healthy donor DNA was obtained from Promega (Human Genomic DNA: Female, G152A).
340 DNA from ALS cases (ND11836 and ND13803) were extracted from EBV transformed LCLs by from
341 the National Institute of Neurological Disorders and Stroke (NINDS) repository at the Coriell Institute.
342 DNA was pre-treated with FFPE Repair Mix from NEB (M6630S) according to manufacturer's *Protocol*
343 *for use with Other User-supplied Library Construction Reagents* to repair nicks that could result in
344 undesired target degradation by exonucleases.

345 *Guide RNA design.* Guide RNAs (sgRNA, **S1 Table**) were selected to flank up and downstream of the
346 target locus. A combination of online tools including CHOPCHOP, E-CRISP, and IDT (42-44) were
347 used to design sgRNAs with high *in silico* predicted on-target efficiency and minimal off-target effects.
348 For target loci, pairs of sgRNAs were designed such that they maintained a "PAM-in" orientation to the
349 target sequence. Preassembled gRNA comprised of crRNA and tracrRNA (IDT, Alt-R® CRISPR-Cas9
350 sgRNA, 2 nmol) sequences were purchased from IDT and resuspended in IDTE at a 10μM
351 concentration.

352 *Cas9 digestion*. The molar ratio of Cas9:gRNA:DNA target was ~10:10:1. The ribonucleoprotein
353 complex was formed by combining 150nM Cas9 enzyme with 150nM of each guide in 1X CutSmart
354 buffer (NEB) and the 23.5μL reaction was incubated at 25ºC for 10 minutes. A 40uL reaction containing
355 the RNP complex, ~15nM (3ug) human genomic DNA or 30ng of gBlock in 1x Cutsmart buffer (NEB
356 B7204) was incubated at 37ºC for 15 minutes.

357 *Exonuclease digestion.* Immediately following Cas9 digestion, 260 total units of exonucleases (Exo I
358 ([40U] NEB M0293), Exo III ([200U] NEB M0206), Lambda ([20U] NEB M0262]) diluted in 1X CutSmart
359 buffer to 10μL were added to the reaction for a final reaction volume 50uL and incubated at 37ºC for
360 two hours, followed by heat inactivation at 80ºC for 20 minutes.

361 *A-tailing.* 1µL of 10mm dATP (Zymo Research, D1005) and 1µL Taq DNA Polymerase (M0267S) were
362 added to reaction mix and incubated at 72ºC for 5 minutes.
363 *Adapter ligation.* An adapter ligation mix was prepared from the LSK-109 Ligation Sequencing Kit by
364 combining 25µL Ligation Buffer, 5µL Quick T4 Ligase (NEB E6057), 5µL Adapter Mix, and 13µL
365 nuclease-free water. The mixture was added to the previous reaction for a total volume of 100uL and
366 incubated for 10 minutes on a hula mixer at room temperature. A clean-up step was then performed
367 using 0.3X AmpureXP magnetic beads (Beckman Coulter A63881) and washed twice with 200µL of
368 Short Fragment Buffer (ONT SQK-LSK109). The final library was eluted in 16.6µL of Elution Buffer and
369 15.8µL retained.
370 *Nanopore sequencing.* Each sample was sequenced on a MinION flow cell (R9.4.1). Flow cells with
371 >800 active pores following Platform QC were primed with 800µL of Flush Buffer followed by a second
372 priming with priming mix (70µL Sequencing Buffer + 70µL nuclease-free water + 70µL Flush Buffer).
373 The final library is then immediately loaded onto the flow cell in a mixture with 26µL Sequencing Buffer,
374 9.5µL Loading Beads, and 0.5µL Sequencing Tether from the LSK-109 Ligation Kit. Sequencing was
375 performed for 48 hours using default settings with the MinKNOW software (v.19.05.0) and live base
376 calling was conducted using the high accuracy flip-flop algorithm.
377 *Sequence data alignment and QC.* All sequencing reads were aligned to the human reference GRCh38
378 using minimap2 software (45). Off-target reads with mapQ = 60 were counted using samtools v.1.9.
379 On-target depth of coverage was also measured with samtools and visualized in IGV. GC content of all
380 off target reads was calculated using samtools and awk and compared to a random sample of
381 1,000,000 intervals in the GRCH38 reference using Bedtools nuc (v2.28.0). All off-target reads were
382 also tested for enrichment with secondary structure annotations, namely G-quadruplexes, using
383 poverlap (46), which permutes a null distribution of overlapping genomic regions.
384 *Repeat copy number estimation in ALS samples.* Reads spanning the repeat expansion were
385 identified using samtools by identifying reads that aligned both 10bp upstream and downstream of the
386 repeat locus (28). Next, spanning reads were realigned (Striped Smith-Waterman, scikit-bio v.0.2.3

387 (47), Python (v.2.7) to the complex sequencing flanking the repeat expansion and the start position in
388 each read of the downstream flank alignment was subtracted from the end position of the upstream
389 flank alignment and that number was divided by 6 to estimate repeat copy number. Repeat length
390 distributions were then visualized on a histogram to determine the number of expected clusters of allele-
391 lengths, which were then fed into a Gaussian Mixture Model (scikit-learn 0.22.1 (48)) to determine
392 allele-specific repeat copy number estimates.

393 **Acknowledgements**

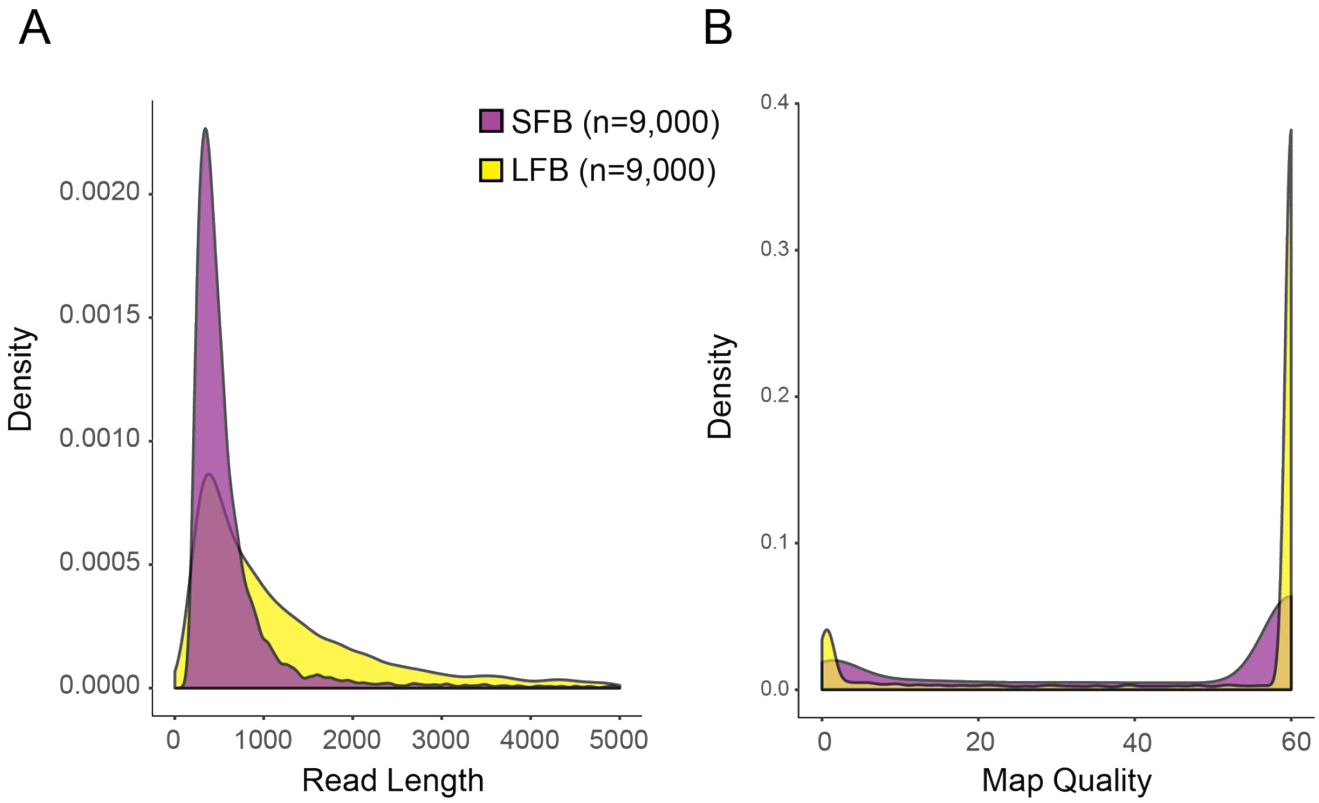
394 We would like to thank the following individuals for their expertise and effort in all aspects of this study:
395 Nels Elde for sharing laboratory space and resources, Joe Brown for hardware support, and Simone
396 Longo and Harriet Dashnow for editing the manuscript.

397 **References**
398

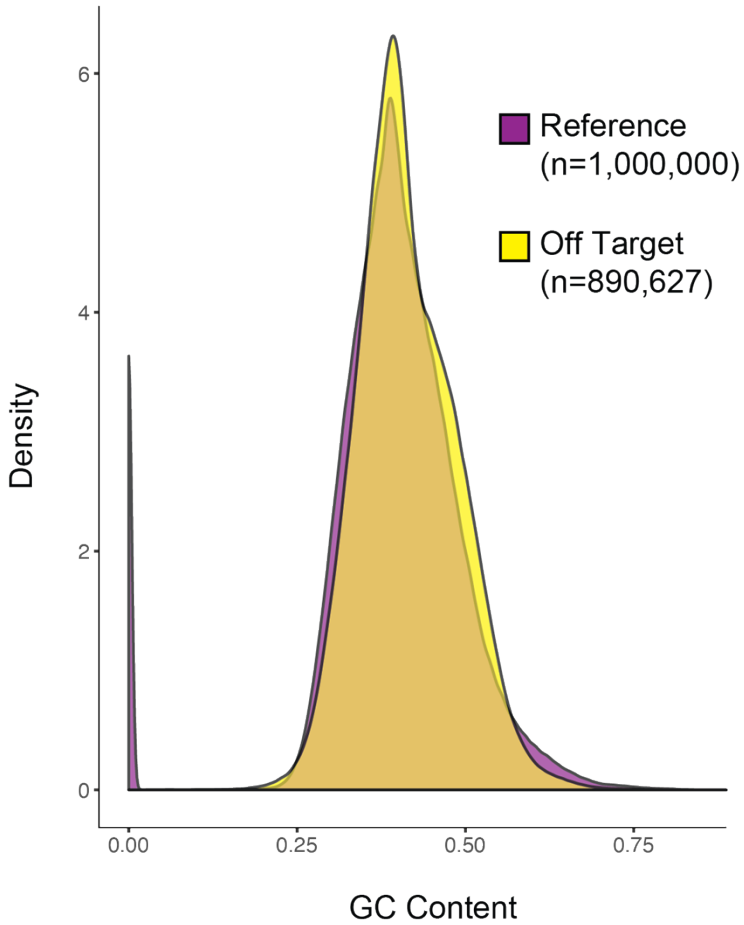
399 1. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al.
400 Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide
401 research data. Lancet. 2015;385(9975):1305-14.
402 2. Whole-genome sequencing of rare disease patients in a national healthcare system. bioRxiv.
403 2019:507244.
404 3. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, et al.
405 Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-
406 generation sequencing. Genet Med. 2016;18(12):1282-9.
407 4. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic
408 analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight.
409 Genome Biol. 2019;20(1):97.
410 5. Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, et al. Long-read
411 sequence and assembly of segmental duplications. Nat Methods. 2019;16(1):88-94.
412 6. Wallace AD, Wendt GA, Barcellos LF, de Smith AJ, Walsh KM, Metayer C, et al. To ERV Is
413 Human: A Phenotype-Wide Scan Linking Polymorphic Human Endogenous Retrovirus-K Insertions to
414 Complex Phenotypes. Front Genet. 2018;9:298.
415 7. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery
416 of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A.
417 2016;113(16):E2326-34.
418 8. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate
419 evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445-77.
420 9. Paulson H. Repeat expansion diseases. Handb Clin Neurol. 2018;147:105-23.
421 10. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common
422 principles. Nat Rev Genet. 2005;6(10):743-55.
423 11. Majounie E, Renton AE, Mok K, Dopper EG, Waite A, Rollinson S, et al. Frequency of the
424 C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and
425 frontotemporal dementia: a cross-sectional study. Lancet Neurol. 2012;11(4):323-30.

426 12. Mori K, Weng SM, Arzberger T, May S, Rentzsch K, Kremmer E, et al. The C9orf72 GGGGCC
427 repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science.
428 2013;339(6125):1335-8.
429 13. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore
430 sequencing to the genomics community. Genome Biol. 2016;17(1):239.
431 14. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al.
432 Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176(3):663-75
433 e19.
434 15. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform
435 discovery of haplotype-resolved structural variation in human genomes. Nat Commun.
436 2019;10(1):1784.
437 16. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted Nanopore
438 Sequencing with Cas9 for studies of methylation, structural variants, and mutations. bioRxiv.
439 2019:604173.
440 17. Gabrieli T, Sharim H, Fridman D, Arbib N, Michaeli Y, Ebenstein Y. Selective nanopore
441 sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH).
442 Nucleic Acids Res. 2018;46(14):e87.
443 18. Slesarev A, Viswanathan L, Tang Y, Borgschulte T, Achtien K, Razafsky D, et al.
444 CRISPR/CAS9 targeted CAPTURE of mammalian genomic regions for characterization by NGS. Sci
445 Rep. 2019;9(1):3587.
446 19. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore
447 sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433-8.
448 20. Payne A, Holmes N, Clarke T, Munro R, Debebe B, Loose M. Nanopore adaptive sequencing
449 for mixed samples, whole exome capture and targeted panels. bioRxiv. 2020:2020.02.03.926956.
450 21. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR
451 RNA-guided endonuclease Cas9. Nature. 2014;507(7490):62-7.
452 22. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed
453 genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat
454 Biotechnol. 2016;34(3):339-44.
455 23. Clarke R, Heler R, MacDougall MS, Yeo NC, Chavez A, Regan M, et al. Enhanced Bacterial
456 Immunity and Mammalian Genome Editing via RNA-Polymerase-Mediated Dislodging of Cas9 from
457 Double-Strand DNA Breaks. Mol Cell. 2018;71(1):42-55 e8.
458 24. Varley KE, Mitra RD. Nested Patch PCR enables highly multiplexed mutation discovery in
459 candidate genes. Genome Res. 2008;18(11):1844-50.
460 25. Rossi MJ, Lai WKM, Pugh BF. Simplified ChIP-exo assays. Nat Commun. 2018;9(1):2842.
461 26. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
462 2018;34(18):3094-100.
463 27. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative
464 genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
465 28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
466 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
467 29. Haiminen N, Mannila H, Terzi E. Determining significance of pairwise co-occurrences of events
468 in bursty sequences. BMC Bioinformatics. 2008;9:336.
469 30. Pedersen BB, J. . Poverlap: Simple, flexible, parallized significance testing of a pair of BED
470 files: Github; 2013 [Available from: https://github.com/brentp/poverlap.
471 31. Bram E, Javanmardi K, Nicholson K, Culp K, Thibert JR, Kemppainen J, et al. Comprehensive
472 genotyping of the C9orf72 hexanucleotide repeat region in 2095 ALS samples from the NINDS
473 collection using a two-mode, long-read PCR assay. Amyotroph Lateral Scler Frontotemporal
474 Degener. 2019;20(1-2):107-14.

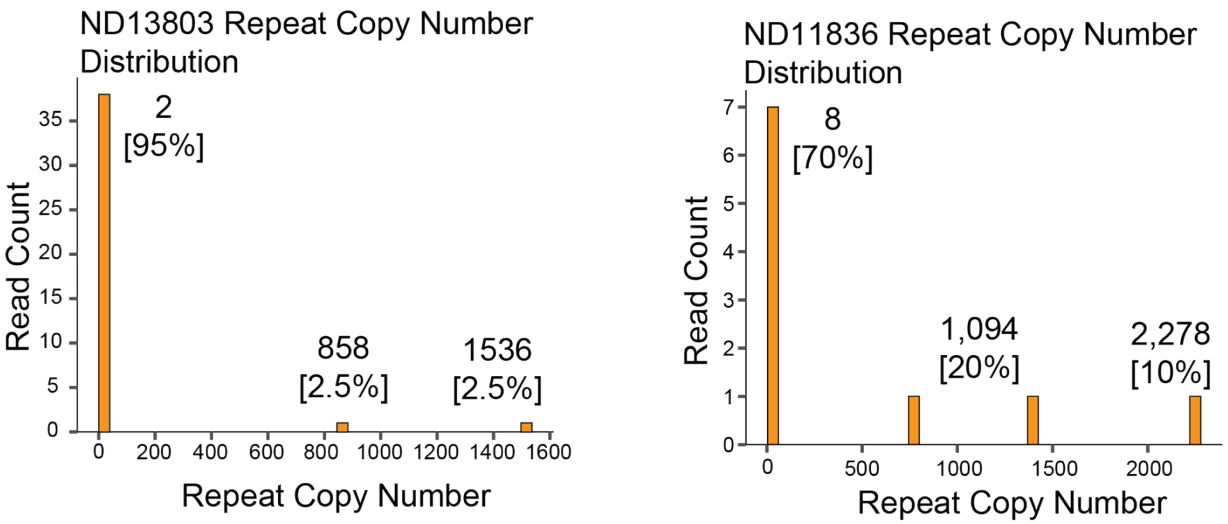
32. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019;20(1):58.
33. Flynn JM, Long M, Wing RA, Clark AG. Evolutionary Dynamics of Abundant 7-bp Satellites in the Genome of Drosophila virilis. Mol Biol Evol. 2020;37(5):1362-75.
34. Kovanda A, Zalar M, Sket P, Plavec J, Rogelj B. Anti-sense DNA d(GGCCCC)n expansions in C9ORF72 form i-motifs and protonated hairpins. Sci Rep. 2015;5:17944.
35. De Roeck A, De Coster W, Bossaerts L, Cacace R, De Pooter T, Van Dongen J, et al. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. Genome Biol. 2019;20(1):239.
36. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, Prudencio M, et al. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. Mol Neurodegener. 2018;13(1):46.
37. Giesselmann P, Brandl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. 2019;37(12):1478-81.
38. Suh E, Grando K, Van Deerlin VM. Validation of a Long-Read PCR Assay for Sensitive Detection and Sizing of C9orf72 Hexanucleotide Repeat Expansions. J Mol Diagn. 2018;20(6):871-82.
39. Bidichandani SI, Purandare SM, Taylor EE, Gumin G, Machkhas H, Harati Y, et al. Somatic sequence variation at the Friedreich ataxia locus includes complete contraction of the expanded GAA triplet repeat, significant length variation in serially passaged lymphoblasts and enhanced mutagenesis in the flanking sequence. Hum Mol Genet. 1999;8(13):2425-36.
40. van Blitterswijk M, DeJesus-Hernandez M, Niemantsverdriet E, Murray ME, Heckman MG, Diehl NN, et al. Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. Lancet Neurol. 2013;12(10):978-88.
41. Aronin N, Chase K, Young C, Sapp E, Schwarz C, Matta N, et al. CAG expansion affects the expression of mutant Huntingtin in the Huntington's disease brain. Neuron. 1995;15(5):1193-201.
42. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31(9):827-32.
43. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods. 2014;11(2):122-3.
44. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. 2019;47(W1):W171-W4.
45. Heng L. Minimap2: pairwise alignment for nucleotide sequences. arXiv. 2018.
46. Pedersen B. Poverlap GitHub2013 [Simple, flexible, parallelized significance testing of a pair of BED files.]. Available from: https://github.com/brentp/poverlap.
47. team Ts-bd. scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers 2020 [0.5.5:[Available from: http://scikit-bio.org.
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.



Supplementary Figure 1: Characteristics of a random sample of 9,000 reads produced from CaBagE enriching for a 4,044bp target. The experiment was conducted in tandem using the same sample DNA the the ONT Long Fragment Buffer (LFB) during adapter ligation and with the Short Fragment Buffer (SFB). A) Kernel density plot of read lengths in LFB and SFB reads. B) Kernel density plot of map quality scores in LFB and SFB reads



Supplementary Figure 2: GC Content distribution of all off-target reads from a single CaBagE run compared (n=890,627) to a random 1,000,000 intervals from HG38 with length equal to the mean off-target read length of the CaBagE run.



Supplementary Figure 3: Histograms of repeat copy number distributions for replicated target enrichment and sequencing across the C9ORF72 repeat expansion in two indiviudals with ALS. results confirm the presence of >2 alleles in both individuals

Supplementary Table 1: Guide RNA sequences

Target Locus	guideRNA Name	guideRNA Sequence	Target Position (GRCh38)	Target Strand
C9orf72	c9orf72_for	TAACGTAGAATAGAACCCGA	chr9:27572027	+
	c9orf72_rev	GTTCCCTGGGTGTGAGTCAA	chr9:27576071	-
GPX1	gpx1_fwd1	GCAAGGAGGGGTCAATCACC	chr3: 49352525	+
	gpx1_rev1	TAACTTCACCTAGATCCTAT	chr3: 49366169	-
GSTP1	gstp1_fwd1	CCCGATGACGCACCTCGGAG	chr11: 67576428	+
	gstp1_rev1	GAATTATAGTGATACGGAAG	chr11: 67594247	-
KRT19	krt19_fwd1	GCCCCACTGTGGACAACTCA	chr17: 41517522	+
	krt19_rev1	TACTCTCTAGCCCACCACTA	chr17: 41535711	-
SLC12A4	slc12a4_fwd1	GACGTGTCATGAGCACCCGA	chr16: 67952369	+
	slc12a4_rev1	TAGATCACTTTGACCACATC	chr16: 67976758	-

gBlock Sequence

AGCAAGTCTGTGTCATCTCGGAGCTGTGAAGCAACCAGGTCATGTCCCACAGAATGGGGAG
CACACCGACTTGCATTGCTGCCCTCATATGCAAGTCATCACCACTCTCTAGAAGCTTGGGCT
GAAATTGTGCAGTCATCTCAACACATATATCTCATCTAGCATGATCTCCTCGCCGGCAGGGA
CCGTCTCGGGTTCCTAGCGAACCCCGACTTGGTCCGCAGAAGCCGCGCGCCGCCCACCCT
CCGGCCTTCCCCCAGGCGAGGCCTCTCAGTACCCGAGGCTCCCTTTTCTCGAGCACGCAG
CGGCAGCGCTCCCATCGTGTCCACGGGAATGAGACAGCTCATGTACTGATTACGGGAAAGC
AAGGAAGAGGCCAGATTAACATCCCTTGTCCCTGGGAAGCCCGGGGCCCGGATGCAGGCA
ATTCCACCAGTCGCTAGAGGCGAAAGCCCGACACCCAGCTTCGGTCAGAGAAATGAGAGGG
AAAGTAAAAATGCGTCGCGCGCGACTCCTGAGTTCCAGAGCTTGCTACAGGCTGCGGTTGT
TTCCCTCCTTGTTTTCTTCTGGTTAATCTTTATCAGGTCTTTTCTTGTTCACCCTCAGCGAGTA
CTGTGAGAGCAAGTAGTGGGGAGAGAGGGTGGGAAAAACAAAAACACACACCTCCTAAACC
CACACCTGCTCTTGCTAGACCCCGCCCCCAAAAGAGAAGCAACCGGGCAGCAGGGACGGC
TGACACACCAAGCGTCATCTTTTACGTGGGCGGAACTTGTCGCTGTTTGACGCACCTCTCTT
TCCTAGCGGGACACCGTAGGTTACGTCTGTCTGTTTTCTATGTGCGATGACGTTTTCTCACG
AGGCTAGCGAAATGGGGCGGGGCAACTTGTCCTGTTCTTTTATCTTAAGACCCGCTCTGGA
GGAGCGTTGGCGCAATAGCGTGTGCGAACCTTAATAGGGGAGGCTGCTGGATCTGGAGAA
AGTGAAGACGATTTCGTGG

Supplementary Table 2: Side-by-side experiment comparing target

571
572

enrichment with CaBagE vs. nCATs			
Locus	CaBage Coverage	nCATs Coverage	Fold-difference
GPX1	30	322	10.73333333
GSTP1	35	93	2.657142857
KRT19	47	210	4.468085106
SLC12A4	54	302	5.592592593