

# 1 Benchmarking small variant detection with ONT reveals high performance in 2 challenging regions

3 Peter L. Møller<sup>1</sup>, Guillaume Holley<sup>2</sup>, Doruk Beyter<sup>2</sup>, Mette Nyegaard<sup>1</sup>, Bjarni V. Halldórsson<sup>2,3</sup>

4 1. Department of Biomedicine, Aarhus University, Høegh-Guldbergs Gade 10, 8000 Aarhus C, Denmark

5 2. deCODE genetics/Amgen Inc., Sturlugata 8, 102, Reykjavík, Iceland

6 3. School of Technology, Reykjavik University, 102, Reykjavik, Iceland

## 7 Abstract

8 **Background:** The development of long read sequencing (LRS) has led to greater access to the human genome. LRS  
9 produces long read lengths at the cost of high error rates and has shown to be more useful in calling structural  
10 variants than short read sequencing (SRS) data. In this paper we evaluate how to use LRS data from Oxford  
11 Nanopore Technologies (ONT) to call small variants in regions in- and outside the reach of SRS.

12 **Results:** Calling single nucleotide polymorphisms (SNPs) with ONT data has comparable accuracy to Illumina when  
13 evaluating against the Genome in a Bottle truth set v4.2. In the major histocompatibility complex (MHC) and  
14 regions where mapping short reads is difficult, the F-measure of ONT calls exceeds those of short reads by 2-4%  
15 when sequence coverage is 20X or greater.

16 We develop recommendations for how to perform small variant calling with LRS data and improve current  
17 approaches to the difficult regions by re-genotyping variants to increase the F-measure from 97.24% to 98.78%.  
18 Furthermore, we show how LRS can call variants in genomic regions inaccessible to SRS, including medically  
19 relevant genes such as *STRC* and *CFC1B*.

20 **Conclusions:** Although small variant calling in LRS data is still immature, current methods are clearly useful in  
21 difficult and inaccessible regions of the genome, enabling variant calling in medically relevant genes not accessible  
22 to SRS.

## 23 Introduction

24 The field of genomics is constantly evolving as developments in sequencing technology allow greater access to  
25 genomic variation. Since the turn of the century, short read sequencing (SRS) has led to tremendous insight into  
26 the human genome, with SRS becoming an integral part of diagnostics [1]. Currently, SRS is almost synonymous  
27 with Illumina sequencing, with read lengths around 150 bp and error rates from 0.1-1% depending on platform and  
28 protocol [2].

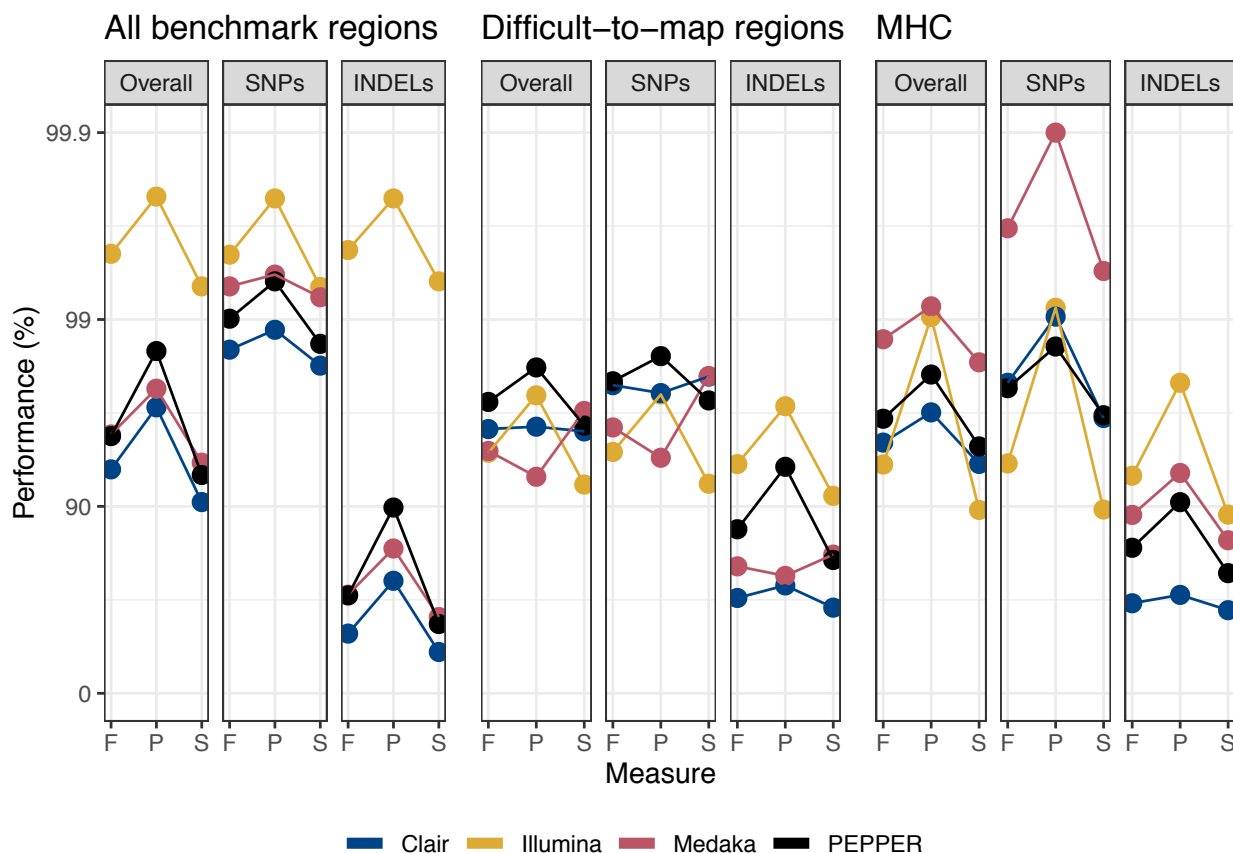
29 In the past decade we have seen the emergence of long read sequencing (LRS), with Pacific Biosciences' (PacBio)  
30 single-molecule real-time (SMRT) technology in 2011 and Oxford Nanopore Technologies (ONT) in 2014 [3]. Both  
31 technologies were initially plagued by high error rates (10-15%) [4], making variant calling very challenging. PacBio  
32 solved this issue with the introduction of the circular consensus sequencing (CCS) protocol, producing high fidelity  
33 (HiFi) reads with lengths of 10-20 kb and error rates of 0.2% [5]. The drawback of this approach is a highly reduced  
34 output, with a single SMRT Cell 8M producing 15-25 Gb of HiFi data [6]. Meanwhile, ONT is still error prone but  
35 provides far longer reads, typically 10-100 kb, and outputting 50-100 Gb of data per PromethION flow cell [6]. ONT  
36 also offers an ultra-long read protocol consistently producing reads exceeding 100 kb at the cost of decreased  
37 output [7]. Considering the PromethION flow cell being slightly cheaper than the SMRT Cell, this results in a much  
38 lower cost per base for ONT data [8].

39 The development of LRS has made previously inaccessible regions of the genome available for study [9]. These  
40 regions were described by Ebbert et al. as "dark", due to low coverage ( $\leq 5$  reads) or low mapping quality ( $\geq 90\%$   
41 reads with MAPQ  $< 10$ ). They found that PacBio reduced the percentage of dark bases by 58.2% for all gene bodies  
42 and 77.7% for coding sequence (CDS). In comparison ONT reduced the percentage of dark bases by 77.9% in all  
43 gene bodies and by 95.6% in the CDS. Recently, the Telomere-to-Telomere (T2T) consortium also showed the  
44 strength of ultra-long ONT reads, using them as part of their effort to create a complete assembly of the human  
45 CHM13hTERT cell line, underlining the importance of read length in accessing dark regions. [10,11].  
46 Further proof of the usefulness of LRS was the genome in a bottle (GIAB) truth set v4.1, which expanded the high  
47 confidence regions to 92.2% of the genome compared to 85.4% in v3.3.2 using PacBio HiFi reads [12].

48 Here we chose to evaluate variant discovery across the genome based on ONT data, as the technology provides  
49 the greatest access to the dark regions at the lowest cost per base. We test some of the most recent variant  
50 callers, namely Medaka, a diploid-aware neural network developed by ONT [13]; Clair, a deep neural network  
51 based variant caller [14] and P.E.P.P.E.R./DeepVariant (“PEPPER” going forward), a deep neural network polisher  
52 and caller [15,16] presented in the PrecisionFDA Truth Challenge V2.

## 53 Results

54 Truth set benchmarking reveals inconsistent performance across different evaluation regions  
55 We analyzed data from the publicly available Ashkenazim trio (HG002, HG003, HG004). Variant calls were  
56 evaluated against the GIAB v4.2 truth set released in relation to the PrecisionFDA Truth Challenge V2 capturing  
57 92.2% of the genome. Illumina data was benchmarked with DeepVariant [16] to establish baseline performance of  
58 a known caller with SRS data. The HG002 truth set has been widely used for model training, we therefore use the  
59 HG003 and HG004 truth sets for evaluation and report their average.  
60 The Illumina data had 60X coverage for all individuals, while coverage for ONT data varied from 50X for HG002  
61 (8.81% error rate) to 80X for HG003 (7.82% error rate) and HG004 (8.24% error rate).



62

63 *Figure 1. Performance metrics measured against the Genome in a Bottle v4.2 truth set. All benchmark regions: Complete truth*  
 64 *set. MHC: Intersect of major histocompatibility complex and truth set. Difficult-to-map regions: Intersect of the truth set with*  
 65 *segmental duplications and regions where 100 bp read pairs have  $\leq 2$  mismatches and  $\leq 1$  indel difference from another part*  
 66 *of the genome. MHC: Major histocompatibility complex; F: F-measure; P: Precision; S: Sensitivity*

67 As seen in Figure 1, the best variant calling performance across all benchmark regions is achieved using Illumina  
 68 data. This is no surprise, as short reads were the foundation of previous versions of the truth set and most of the  
 69 human genome is sufficiently unique to map short reads unambiguously. When stratifying performance by variant  
 70 type, we see a more detailed picture. This highlights decent SNP calling with ONT data, with both Medaka and  
 71 PEPPER F-measure surpassing 99%, while Clair achieves 98.55%. Meanwhile, indel detection with ONT lags severely

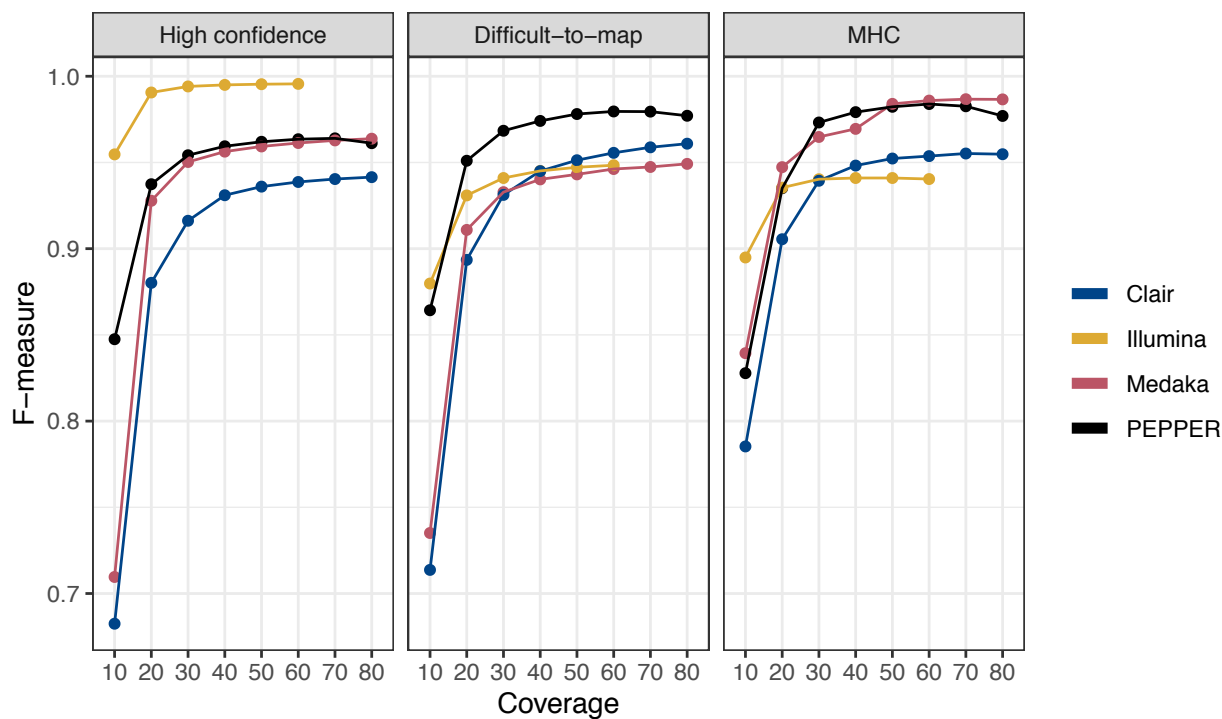
72 behind. For Illumina we see an F-measure of 99.58%, while Medaka, the best performing ONT caller, achieves  
73 70.34%.

74 Analyzing performance in more complex regions, such as the difficult-to-map regions and the major  
75 histocompatibility complex (MHC), reveals the benefit of LRS. In the difficult-to-map regions (145 mb), overall ONT  
76 performance surpasses Illumina with an F-measure of 97.24% using PEPPER, while Illumina reaches 94.84%. A  
77 similar picture is seen in the MHC (4.6 mb), except the overall best performance is achieved by Medaka, having an  
78 F-measure of 98.73%, while Illumina reaches 94.04%. In both the MHC and the difficult-to-map regions Illumina is  
79 5-8% better for calling indels than the best ONT caller.

80 Subsampling reveals high performance from 30X coverage

81 As 80X ONT whole genome sequencing is not necessarily feasible for large scale experiments we benchmarked  
82 performance at 10X coverage increments (Figure 2). This highlighted the need for at least 20X ONT to surpass SRS  
83 performance in the difficult-to-map regions, while 30X was necessary in the MHC to achieve a meaningful  
84 improvement. Interestingly, this also showed PEPPER as the better choice for ONT data across most depths and  
85 regions, despite the author recommendation of 50-80X coverage [15].

86 PEPPER with 30X coverage resulted in an overall f-measure of 95.42%, while doubling the coverage increased  
87 performance less than 1%. Finally, we observed a slight dip in PEPPER performance above 60X coverage.



88

89 *Figure 2. Variant calling performance as a function of sequence coverage. F-measure was determined between 10 and 80X*

90 *coverage in 10X increments for each evaluation region. Numbers are average of HG003 and HG004. MHC: Major*

91 *histocompatibility complex.*

92 Mendelian concordance decreases outside high confidence regions

93 Evaluating the performance of variant callers in the 7.8% of the genome not included in the GIAB truth set is more

94 difficult. Here we look at two measurements; 1) the total number of variants as an indicator of sensitivity and 2)

95 the Mendelian concordance as an indicator of precision. Mendelian concordance is a commonly used metric, when

96 no truth set exists [17], while the number of variants is important to avoid overestimating the performance of

97 conservative callers. To achieve consistent Mendelian concordance calculations for each setup, we only benchmark

98 variant callers supporting gVCF output (DeepVariant for Illumina, PEPPER for ONT).

99 As seen in Table 1, variant calling in the high confidence (HC) regions is very consistent, with both technologies

100 resulting in Mendelian concordance above 99%. At the same time, the total number of variants in ONT data

101 displays a shortcoming of Mendelian concordance. The call set is missing approximately 200,000 indels, caused by

102 low sensitivity (57.41%, Figure 1), but maintains high concordance due to consistently missing indels. In the

103 complementary (Comp) regions the mendelian concordance decreases for both technologies. Stratifying by variant

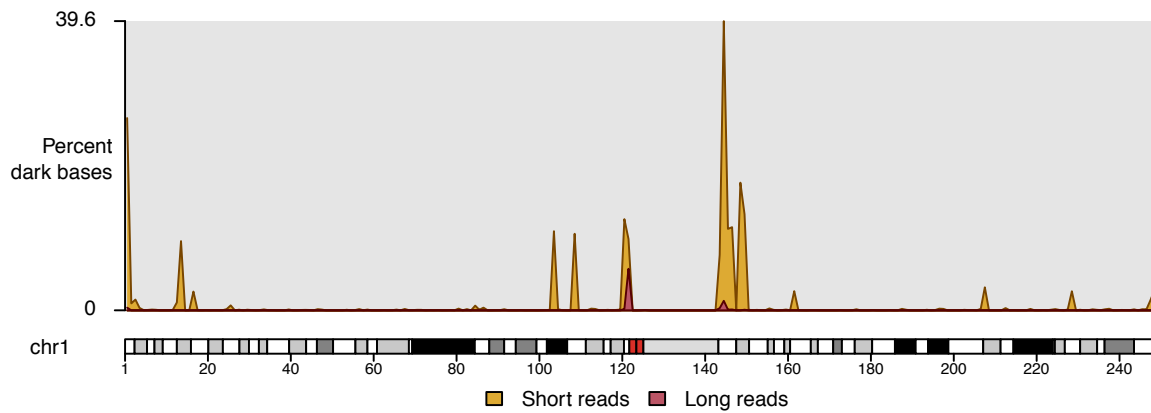
104 type shows the concordance of ONT SNPs to be within 2% of Illumina, while calling 90,000 more SNPs, presumably  
105 due to greater access to the traditionally dark regions of the genome.

106 *Table 1. Total number of variants called in HG002 and their Mendelian concordance*

Data	Region (variant type)	Variants	MC (%)
Illumina	HC (SNPs + indels)	3,872,611	99.87
	Comp (SNPs + indels)	734,056	95.00
	Comp (SNPs)	382,568	94.59
	Comp (indels)	335,519	95.81
ONT	HC (SNPs + indels)	3,662,447	99.40
	Comp (SNPs + indels)	584,514	91.24
	Comp (SNPs)	469,811	92.62
	Comp (indels)	109,753	85.62

107 *MC: Mendelian concordance; HC: High confidence; Comp: Complementary regions; SNPs: Single nucleotide*  
108 *polymorphisms; ONT: Oxford Nanopore Technologies.*

109 Long reads reveal 22 mb of dark genome including medically relevant genes  
110 We identified dark regions for both short and long reads, subsequently identifying regions uniquely dark to either  
111 technology. Here we adapted the dark region definition from [9] described previously. This approach found 22 mb  
112 of the genome dark only to short reads, while, surprisingly, 1.5 mb was solely dark to long reads (Figure 3). These  
113 regions were spread across 32607 sites, ranging from 1 to 103,863 bp (median 155 bp) in size, with the 1324  
114 largest regions (3,335 bp and above) making up 50% of the dark bases.



115

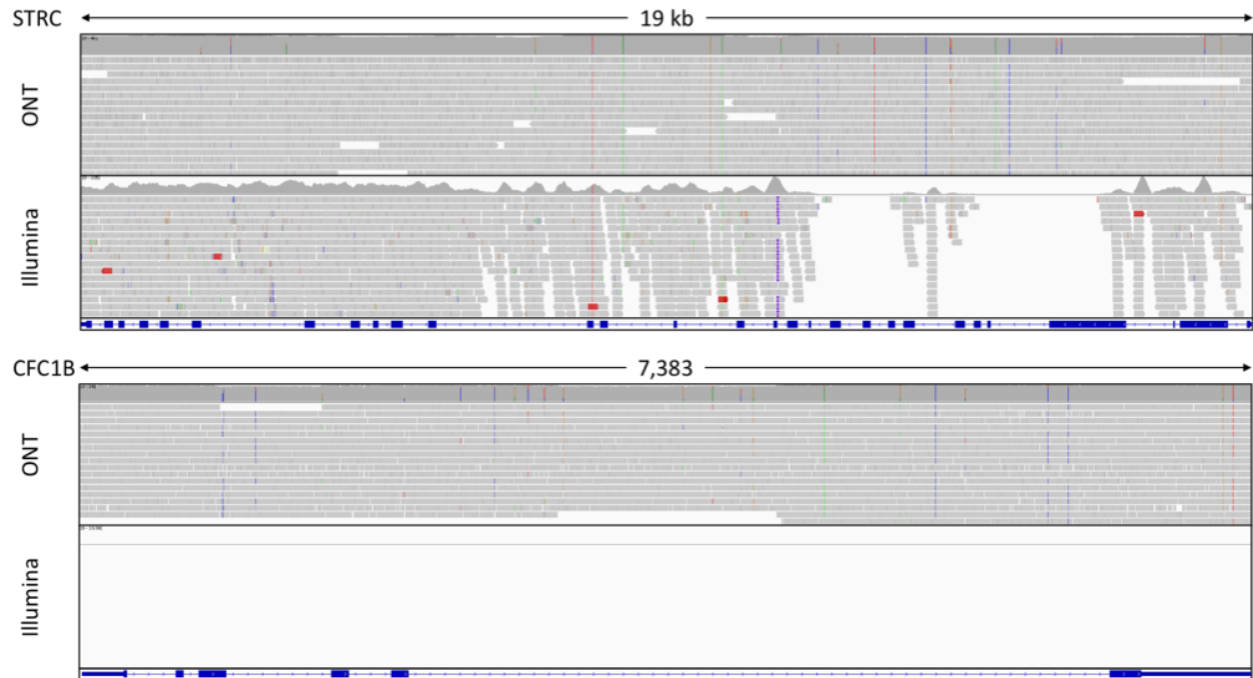
116 *Figure 3. Dark regions of chromosome 1. Percentage of bases in 1 mb windows which are dark to only one technology.*

117 Looking at genes previously defined as medically relevant but challenging to short-read technologies [18], we find  
118 618 of 4,773 genes where some bases can only be reached by ONT and 49 genes where at least 10% of the gene is  
119 only reachable with ONT. The same approach for all genes in the Ensembl database [19] identified 2,336 of 19,190  
120 genes with any bases only reached by ONT and 453 genes above the 10% threshold. Figure 4 shows two medically  
121 relevant genes, *STRC* and *CFC1B*, of which 23.39% and 100% of the genes can only be accessed with ONT. A  
122 comparison to PacBio HiFi found 12 mb of the genome, including *CFC1B*, to be dark to PacBio Hifi data but not  
123 ONT.

124 Intersecting the short read dark regions with the PEPPER variants identified 54,000 variants in HG002, of which  
125 48,000 were SNPs. This number corresponds to half of the 90,000 SNPs identified by ONT and not Illumina.

126 Analysis of these variants will however be difficult as Mendelian concordance is only 86.87%. Furthermore, the  
127 PEPPER call set contained an extra 25,000 events in these regions, which were not called as variants. Upon visual  
128 inspection several events were missing genotypes (./.) or called as homozygous reference (0/0) despite high  
129 coverage and variant allele frequencies above 50%. We assume this is caused by DeepVariant (the final step of  
130 PEPPER) sometimes recognizing SNP-dense regions as segmental duplications, leading it to call homozygous  
131 reference [20].





132

133 *Figure 4. Coverage of the medically relevant genes STRC and CFC1B. Top panel shows the coverage of the STRC gene. Bottom*

134 *panel shows the coverage of the CFC1B gene.*

135 Re-genotyping variants outside the high confidence regions increase mendelian concordance

136 As our original calls in the dark regions showed low Mendelian consistency, we re-genotyped the joint VCF files

137 from PEPPER using Whatshap [21]. Whatshap can take an input VCF + BAM file to compute haplotype-aware

138 genotypes for each event in the VCF file based on a Hidden Markov Model [22]. Using this approach on all events

139 in the dark regions increased the number of variants by approximately 15,000 while simultaneously improving the

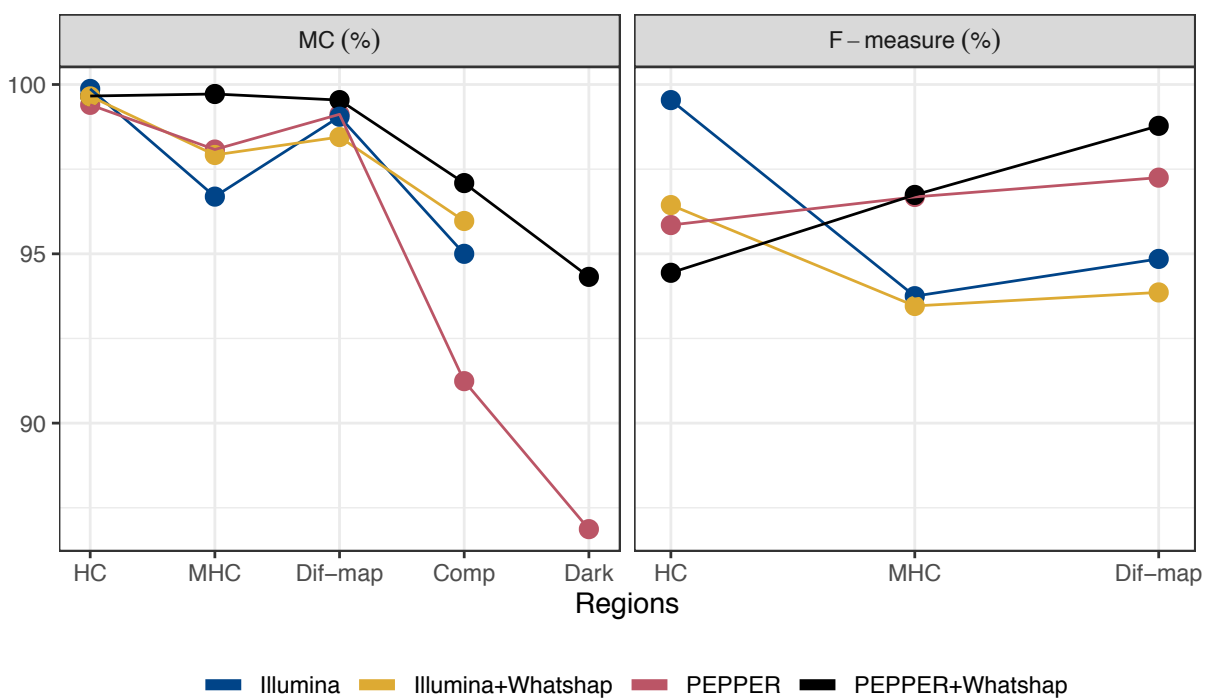
140 Mendelian concordance by more than 7% (Table 2).

141 *Table 2. Number of variants and Mendelian concordance of PEPPER and re-genotyped PEPPER in the dark regions.*

Variants	MC (%)	Re-genotyping
54,363	86.87	No
71,866	94.32	Yes

142 *MC: Mendelian concordance*

143 Extending the re-genotyping to each subset of the genome (HC, MHC, difficult-to-map, complementary, dark  
144 regions) showed increased Mendelian concordance for all regions (Figure 5). For the high confidence, MHC and  
145 difficult-to-map regions we also re-calculated the F-measure of the new HG003 and HG004 genotypes. Re-  
146 genotyping resulted in 1.5% decreased F-measure in the high confidence regions. However, in the MHC, the F-  
147 measure increased marginally, while in the difficult-to-map regions it improved from 97.25% to 98.78%. Further,  
148 re-genotyping improved the consistency of variant calling between samples in the MHC. In this region, the F-  
149 measure for both Illumina and ONT varied 2-4% between HG003 and HG004, while in other regions it varied 0-  
150 0.2%. Re-genotyping reduced the F-measure difference in the MHC to 0.2-2%.  
151 For the Illumina data, the F-measure decreased for all regions (Figure 5), while re-genotyping had mixed effects on  
152 the Mendelian concordance with slight increases in the MHC and complementary regions and decreases in the HC  
153 and difficult-to-map regions.



154  
155 *Figure 5. The effect of re-genotyping on mendelian concordance and F-measure. For PEPPER we see an increase in Mendelian*  
156 *concordance for all regions, while the F-measure is decreased in the HC, but increased in the MHC and difficult-to-map regions.*  
157 *For Illumina we see improved Mendelian concordance in the MHC and complementary regions, while it is decreased in the HC*

158 *and difficult-to-map regions, the F-measure is decreased in all regions. HC: High confidence; MHC: Major histocompatibility*  
159 *complex; Dif-map: Difficult-to-map; Comp: Complementary regions (outside HC, excluding centromere and sex chromosomes).*

## 160 Discussion

161 We have analyzed multiple variant calling approaches in ONT data from the publicly available Ashkenazim trio  
162 (HG002, HG003, HG004). This has shown very consistent SNP calling, with both Medaka and PEPPER exceeding  
163 99% F-measure when evaluating against the latest release of the GIAB truth sets. Limiting evaluations to more  
164 challenging regions of the truth sets shows performance 2-4% higher than SRS with as little as 20X coverage.  
165 No single variant calling approach was the best across all regions and coverages, but the consistency of PEPPER  
166 makes it a good default choice for analysis up to at least 50X coverage, while we begin to see diminishing results at  
167 60X and above. Another benefit of PEPPER compared to other variant calling options is the ability to output gVCF  
168 format, which is easily processed with GLnexus [23,24] to create joint VCF files.  
169 Evaluating performance outside the high confidence regions we observed decreased Mendelian concordance,  
170 which was expected as these regions are generally more challenging. Re-genotyping calls in these regions with  
171 Whatsap improved Mendelian concordance by almost 6%, resulting in ONT surpassing SRS. This approach also  
172 improved the F-measure of ONT data in both the MHC and the difficult-to-map regions. A similar approach for SRS  
173 was not beneficial, highlighting how some of the behavior of DeepVariant is good for short reads but at times  
174 detrimental to long reads. Re-genotyping PEPPER variant calls in the MHC (highest SNP-density region) also  
175 improved the consistency of variant calling, while maintaining a similar average F-measure, which in our opinion is  
176 to be preferred.  
177 Finally, we show that ONT can reach an additional 22 mb of the genome, finding more than 50,000 variants, which  
178 are completely inaccessible to SRS. As shown, these regions include medically relevant genes like *CFC1B*, which can  
179 only be reached with the very long reads from ONT or potentially PacBio continuous long reads (CLR).  
180 During this work, we have observed new releases of almost every software used, including the Guppy basecaller  
181 used to generate the initial sequence files. It is therefore not unreasonable to expect improved performance of  
182 ONT data in the future, both through better basecalling but also variant calling. Meanwhile, SRS has reached a

183 point where future improvements will be minimal, leading us to believe that the current performance difference in  
184 regions accessible to both methods will become smaller.  
185 The primary issue for ONT is now indel performance, which we have not put substantial focus on in this paper.  
186 Current results are inferior to SRS in all evaluation regions by 5-30%, making it an obvious focus point for future  
187 development.

## 188 Conclusion

189 For researchers whose primary interest is small variants inside the high confidence regions, SRS is both cheaper,  
190 better and easier to work with. ONT sequencing technology has already been shown to be useful for structural  
191 variant detection [25] and methylation calling [26]. We now show that ONT is beneficial for small variant calling in  
192 the MHC, the difficult-to-map regions and regions outside the high confidence regions, in particular we find 22 mb  
193 accessible only to ONT.

194 In the challenging regions of the genome, ONT outperforms SRS in SNP calling, helping researchers gain access to  
195 genomic regions and genes which are otherwise completely dark. Here we find PEPPER to be the best performing  
196 variant caller, without access to very high coverage data (>60X). Further, we advise a re-genotyping step, as it  
197 improves consistency and performance of variant calling in these regions.

198 As a technology, ONT is still quite immature, making it a challenge to utilize to its full potential. At the same time,  
199 this immaturity promises greater performance and easier use in the future, if developments continue at the  
200 current pace.

## 201 Methods

202 Data preparation:

203 GRCh38 aligned Illumina BAM files from the Ashkenazim trio (HG002, HG003, HG004) were downloaded and used  
204 as is. ONT and PacBio HiFi FASTQ files were downloaded and aligned to GRCh38 using Minimap2 v2.14 [27] utilizing  
205 the presets for each technology (“-ax map-ont” and “-ax asm20”, respectively). Sorting and indexing was  
206 performed with SAMtools v1.9 [28].

207 All Illumina data had a coverage of 60X, ONT ranged from 50X (HG002) to 80X (HG003, HG004), while PacBio had  
208 35X coverage for all.

209 Variant calling:

210 Illumina:

211 A singularity image of DeepVariant v0.10.0 was created using 'singularity pull deepvariant\_0.10.0.simg  
212 docker://google/deepvariant:0.10.0'. Variant calling was performed with standard parameters (--model\_type  
213 WGS) using the singularity exec command, outputting both VCF and gVCF files.

214 ONT:

215 Medaka v1.0.3 was run using medaka\_variant, calling variants by chromosome before combining VCF files with  
216 BCFtools v1.9 [28].

217 Clair v2.0.6 was run using the callVarBam module with the pretrained ONT model, variants were called by  
218 chromosome.

219 A singularity image of PEPPER/DeepVariant was created using 'singularity pull  
220 docker://kishwars/pepper\_deepvariant\_cpu:latest' (Image from 15/6/2020). Variants were called by modifying the  
221 run\_pepper\_deepvariant.sh script to include gVCF output and storing the PEPPER models outside the image to  
222 ensure write permission. Variant calling was performed in 'Run-time' mode.

223 For both data types gVCF output from each individual was joined using GLnexus v1.2.7. A singularity image was  
224 created using 'singularity pull glnexus\_v1.2.7.simg docker://quay.io/mlin/glnexus:v1.2.7'. The image was executed  
225 with --config DeepVariantWGS.

226 Re-genotyping

227 Whatsap v1.0 was used to re-genotype the joint VCFs from GLnexus, running Whatsap genotype on individual  
228 chromosome with the --indel flag. VCF files were re-combined using BCFtools concat. Whatsap genotype required  
229 the presence of a read group tag and sample in the bam header. This was added using SAMtools with the  
230 command 'samtools addreplacerg -r "ID:HG002\tSM:HG002"'

231 Evaluation

232 Truth sets:

233 Version 4.2 of truth sets (VCF + BED files) were downloaded for each individual and used with RTG tools v3.10.1

234 [29].

235 MHC and difficult-to-map regions:

236 BED files of the major histocompatibility complex (MHC) and difficult-to-map regions, as defined by GIAB, were

237 downloaded and used as is.

238 Outside high confidence regions:

239 A complementary BED file of the HG002 high confidence regions was created using BEDTools v2.18.2 [30] to

240 extract variants for mendelian concordance testing. From this BED file we subtracted the centromere regions

241 (UCSC table browser > Mammal > Human > GRCh38 > All tables > centromeres) as well as the X and Y

242 chromosome, as these regions were too noisy and unsuited for Mendelian concordance, respectively.

243 Dark regions and medically relevant genes:

244 For Illumina, ONT and PacBio, dark regions were computed from the BAM files of the trio. Dark regions were

245 defined as regions of at least 30 bp, with an average coverage below 5X or less than 10% of reads having a

246 mapping quality at or above 10. The centromere regions were subtracted due to noise. Finally, for Illumina and

247 PacBio we created BED files of dark regions reachable by ONT by subtracting the ONT BED file from each.

248 The overlaps between dark regions and genes were computed by intersecting the gene coordinates with the dark

249 regions using BEDTools.

250 RTG Tools:

251 To benchmark against the truth sets we evaluated each VCF file using the vcfEval module of RTG Tools, using the

252 truth VCF as baseline (-b) and the truth BED to define the regions (--bed-regions). The evaluation region (-e) was

253 defined using either the truth BED, the MHC BED or the difficult-to-map BED.

254 Benchmarking outside the truth sets was achieved by intersecting the joint VCF file from GLnexus with the “outside

255 high confidence regions” BED, followed by RTG Tools mendelian to compute Mendelian concordance rates with

256 both mother, father and mother+father. The same approach was used for other BED files when reporting

257 Mendelian concordance for other regions.

258 Visualization:

259 Visualizations were made with R v3.6 [31], ggplot2 [32], karyoploteR [33], inlmisc [34] and IGV v2.8 [35].

260 Availability of data and materials

261 Illumina sequencing data and evaluation BED files are made available by the GIAB consortium [36–38] from their

262 FTP server [39]. ONT and PacBio data are available through the FDA precision challenge [40]. Centromere BED file

263 can be downloaded from UCSC [41,42] at [43]. The precomputed Clair model is available at [44]. Ensembl release

264 98 is available from their FTP server [45].

## 265 **Declarations**

266 Competing interests

267 G.H., D.B. and B.V.H. are employees of deCODE Genetics/Amgen, Inc.

268 Authors' contributions

269 P.L.M, G.H., D.B. and B.V.H. designed the experiments. P.L.M. performed all variant calling and benchmarking. G.H.

270 developed the method for detection of dark regions. P.L.M wrote the initial version of the manuscript, and G.H.,

271 D.B., M.N. and B.V.H. contributed to subsequent versions. All authors reviewed and approved the final version.

272 Acknowledgements

273 The authors would like to thank our colleagues from deCODE genetics and Amgen Inc. for their helpful feedback.

274 We would also like to thank the individuals who provided a biological sample to Genome in a Bottle.

## 275 **References**

276 1. Rexach J, Lee H, Martinez-Agosto JA, Németh AH, Fogel BL. Clinical application of next-generation sequencing to  
277 the practice of neurology. *Lancet Neurol.* 2019;18:492–503.

278 2. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and  
279 subclonal mutations. *Nat Rev Genet.* 2018;19:269–85.

- 280 3. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read  
281 sequencing data analysis. *Genome Biol.* 2020;21:30.
- 282 4. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet.*  
283 *England*; 2018;34:666–81.
- 284 5. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-  
285 read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37:1155–62.
- 286 6. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet.*  
287 *England*; 2020;21:597–614.
- 288 7. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human  
289 genome with ultra-long reads. *Nat Biotechnol.* 2018;36:338–45.
- 290 8. UCDavis. UCDavis sequencing rates [Internet]. [cited 2020 Oct 13]. Available from:  
291 <https://dnatech.genomecenter.ucdavis.edu/uc-prices/>
- 292 9. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and  
293 camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* 2019;20:97.
- 294 10. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovych MA, Koren S, et al. The structure, function, and evolution of  
295 a complete human chromosome 8. *bioRxiv.* 2020;
- 296 11. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a  
297 complete human X chromosome. *Nature.* 2020;585:79–84.
- 298 12. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, et al. Benchmarking challenging small variants with  
299 linked and long reads. *bioRxiv.* 2020;
- 300 13. Oxford Nanopore Technologies. Medaka [Internet]. [cited 2020 Oct 13]. Available from:  
301 <https://github.com/nanoporetech/medaka>
- 302 14. Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, et al. Exploring the limit of using a deep neural  
303 network on pileup data for germline variant calling. *Nat Mach Intell.* 2020;2:220–7.
- 304 15. Shafin K, Pesout T, Jain M, Paten B. P.E.P.P.E.R. [Internet]. [cited 2020 Oct 13]. Available from:  
305 <https://github.com/kishwarshafin/pepper>
- 306 16. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant



- 307 caller using deep neural networks. *Nat Biotechnol.* 2018;36:983–7.
- 308 17. Toptas BÇ, Rakocevic G, Kómar P, Kural D. Comparing complex variants in family trios. *Bioinformatics.*  
309 2018;34:4241–7.
- 310 18. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, et al. Navigating highly  
311 homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet*  
312 *Med. United States;* 2016;18:1282–9.
- 313 19. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. *Nucleic Acids Res.*  
314 2020;48:D682–8.
- 315 20. Github. DeepVariant issue 266 [Internet]. [cited 2020 Oct 13]. Available from:  
316 <https://github.com/google/deepvariant/issues/266>
- 317 21. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: Weighted Haplotype  
318 Assembly for Future-Generation Sequencing Reads. *J Comput Biol.* Mary Ann Liebert, Inc., publishers;  
319 2015;22:498–509.
- 320 22. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware diplotyping from noisy long reads.  
321 *Genome Biol.* 2019;20:116.
- 322 23. Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, et al. GLnexus: joint variant calling for large cohort  
323 sequencing. *bioRxiv.* 2018;
- 324 24. Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant  
325 and GLnexus. *bioRxiv.* 2020;
- 326 25. Beyter D, Ingimundardottir H, Eggertsson HP, Bjornsson E, Kristmundsdottir S, Mehringer S, et al. Long read  
327 sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. *bioRxiv.* 2019;
- 328 26. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using  
329 nanopore sequencing. *Nat Methods. United States;* 2017;14:407–10.
- 330 27. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–100.
- 331 28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and  
332 SAMtools. *Bioinformatics.* 2009;25:2078–9.
- 333 29. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Comparing Variant Call Files for

- 334 Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. bioRxiv. 2015;
- 335 30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*.
- 336 2010;26:841–2.
- 337 31. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R
- 338 Foundation for Statistical Computing; 2019. Available from: <https://www.r-project.org/>
- 339 32. Hadley W. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016.
- 340 33. Gel B, Serra E. karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary
- 341 data. *Bioinformatics*. 2017;33:3088–90.
- 342 34. Fisher JC. inlmisc---Miscellaneous functions for the U.S. Geological Survey Idaho National Laboratory Project
- 343 Office. Reston, Va.; 2020.
- 344 35. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer.
- 345 *Nat Biotechnol*. 2011;29:24–6.
- 346 36. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets
- 347 provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32:246–51.
- 348 37. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately
- 349 benchmarking small variant and reference calls. *Nat Biotechnol*. 2019;37:561–6.
- 350 38. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes
- 351 to characterize benchmark reference materials. *Sci Data*. 2016;3:160025.
- 352 39. Genome In A Bottle. GIAB FTP [Internet]. [cited 2020 Oct 20]. Available from: [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/)
- 353 [trace.ncbi.nlm.nih.gov/giab/ftp/](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/)
- 354 40. PrecisionFDA. Truth Challenge V2 [Internet]. [cited 2020 Oct 15]. Available from:
- 355 <https://precision.fda.gov/challenges/10/>
- 356 41. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data
- 357 retrieval tool. *Nucleic Acids Res*. 2004;32:D493-6.
- 358 42. UCSC. UCSC Genome Browser [Internet]. [cited 2020 Oct 15]. Available from: <https://genome.ucsc.edu/>
- 359 43. UCSC. Centromere BED [Internet]. [cited 2020 Oct 15]. Available from:
- 360 <http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/centromeres.txt.gz>

- 361 44. Clair. Clair ONT model [Internet]. [cited 2020 Oct 15]. Available from:  
362 [http://www.bio8.cs.hku.hk/clair\\_models/ont/122HD34.tar](http://www.bio8.cs.hku.hk/clair_models/ont/122HD34.tar)
- 363 45. Ensembl. Ensembl release 98 [Internet]. [cited 2020 Oct 16]. Available from:  
364 [ftp://ftp.ensembl.org/pub/release-98/gff3/homo\\_sapiens/Homo\\_sapiens.GRCh38.98.gff3.gz](ftp://ftp.ensembl.org/pub/release-98/gff3/homo_sapiens/Homo_sapiens.GRCh38.98.gff3.gz)