1 The Brazilian Rare Genomes Project: validation of whole 2 genome sequencing for rare diseases diagnosis

3 **Running head:** The Brazilian Rare Genomes Project

4

5 Antonio Victor Campos Coelho¹, Bruna Mascaro Cordeiro de Azevedo¹, Danielle
6 Ribeiro Lucon¹, Maria Soares Nóbrega¹, Rodrigo de Souza Reis¹, Rodrigo Bertollo de
7 Alexandre¹, Livia Maria Silva Moura¹, Gustavo Santos de Oliveira¹, Rafael Lucas Muniz
8 Guedes¹, Marcel Pinheiro Caraciolo¹, Nuria Bengala Zurro¹, Murilo Castro Cervato¹,
9 João Bosco de Oliveira Filho^{1,*}

10 1 Hospital Israelita Albert Einstein, Av. Albert Einstein 627, São Paulo, São Paulo,
11 Brazil.

12 * Corresponding author

13 João Bosco de Oliveira Filho.

14 Hospital Israelita Albert Einstein, Av. Albert Einstein 627, São Paulo, São Paulo, Brazil.

15 joao.bosco@einstein.br

16

17 **Funding**

18 This study was funded by a PROADI/SUS grant, from the Albert Einstein Hospital and
19 The Brazilian Ministry of Health.

20 **Abstract**

21 Rare diseases affect 3.2 to 13.2 million individuals in Brazil. The Brazilian Rare
 22 Genomes Project is envisioned to further the implementation of genomic medicine into
 23 the Brazilian public healthcare system. Here we report the results of the validation of a
 24 whole genome sequencing (WGS) procedure for implementation in a clinical
 25 laboratory. In addition, we report data quality for the first 1,200 real world patients
 26 sequenced. For the validation, we sequenced a well characterized group of 76
 27 samples, including seven gold standard genomes, using a PCR-free WGS protocol on
 28 Illumina Novaseq 6000 equipment. We compared the observed variant calls with their
 29 expected calls, observing good concordance for single nucleotide variants (SNVs;
 30 mean F-measure = 99.82%) and indels (mean F-measure = 99.57%). Copy number
 31 variants and structural variants events detection performances were as expected (F-
 32 measures 96.6% and 90.3%, respectively). Our protocol presented excellent intra- and
 33 inter-assay reproducibility, with coefficients of variation ranging between 0.03% and
 34 0.20% and 0.02% and 0.09%, respectively. Limitations of the procedure include the
 35 inability to confidently detect variants such as uniparental disomy, balanced
 36 translocations, repeat expansion variants and low-level mosaicism. In summary, the
 37 observed performance of the test was in accordance with that seen in the best centers
 38 worldwide. The Rare Genomes Project is an important initiative to improve Brazil's
 39 general population access to the innovative WGS technology which has the potential to
 40 reduce the time until diagnosis of rare diseases, bringing pivotal improvements for the
 41 quality of life of the affected individuals.

42

43 Introduction

44 Rare diseases represent a group of over 9,000 disorders affecting an estimated
45 114 to 470 million patients globally (1.5% to 6.2% of the global population) ¹. Rare
46 diseases with genetic etiology are the leading cause of death in children and the
47 diagnosis is challenging, but studies showed that early genetic testing leads to clear
48 benefits by reducing the time until diagnosis, leading to better choice of therapeutic
49 interventions, improving couples' confidence in having children again and reducing
50 healthcare costs ².

51 The human genome was first mapped through the Human Genome Project
52 (HGP), an extensive international collaboration over 13 years ³. Important advances in
53 sequencing technology, such as the development of next-generation sequencing
54 (NGS), have enabled the sequencing of a complete genome within hours, at a fraction
55 of the initial cost, which resulted in the generation of a large amount of data, and a
56 widespread application for diagnosis and research ⁴.

57 NGS for DNA sequencing encompasses several approaches: whole genome
58 (WGS), whole exome (WES) and targeted (panel) sequencing. With WGS, it is possible
59 to read approximately all 3 billion base pairs of human genome ⁵. The falling cost,
60 increasing ease of application and comprehensive nature of WGS is making it the ideal
61 tool for routine use in rare disease diagnosis.

62 WGS can overcome many of the technical limitations of other NGS approaches,
63 including uneven coverage and low sensitivity for the detection of copy number,
64 structural and expansion repeat variants ⁶. In addition, it enables the identification of
65 noncoding and mitochondrial variants ⁷. In fact, many studies have shown that WGS
66 has high diagnostic yield and that early molecular diagnosis improves outcomes and
67 reduces healthcare costs ^{8,9}.

68 The WGS workflow can be divided in three major steps: wet laboratory sample
69 processing, bioinformatics analyses for variant calling and annotation, and correlation

70 of the clinical and molecular findings, resulting in a medical report. The implementation
71 of WGS in clinical laboratories thus requires critical assay design, validation and
72 implementation of quality control measures according to specific guidelines
73 recommendations, to ensure adequate performance prior to use in diagnostic routine ¹⁰,
74 ¹¹.

75 The Brazilian Rare Genomics Project envisions to further the implementation of
76 genomic medicine into the Brazilian national public healthcare system (SUS),
77 complementing current policies, significantly improving the diagnostic capacity for rare
78 disorders.

79 Here we report the results of the development and validation of a PCR-free
80 WGS protocol for clinical use in the project, including wet-lab workflow and
81 bioinformatics pipelines. In addition, we document the protocol performance in the first
82 1,200 samples sequenced.

83

84 **Materials and methods**

85 *Sample selection and test scope*

86 Our validation dataset was composed of 76 samples (Supplementary Table 1).
87 Among them, 22 were international reference samples purchased from Coriell Life
88 Sciences (Philadelphia, PA, USA) for benchmarking and validation, including seven
89 reference samples from Genome in a Bottle Consortium (GiaB) ¹². The remaining 54
90 are samples previously characterized by other methodologies: single nucleotide
91 polymorphism (SNP) array, array comparative genomic hybridization (aCGH),
92 conventional karyotyping, or fluorescence *in-situ* hybridization (FISH). We intended to
93 detect and report single nucleotide variants – SNVs, insertion/deletions – indels, copy
94 number variants – CNVs (large deletions and duplications, chromosomal aneuploidy),
95 and structural variants – SVs (inversions, translocations), as well as mitochondrial

96 SNVs. Repeat expansions and mosaicism were not included in the scope of this first
97 phase of validation.

98 Samples were sequenced across three independent routines. We selected two
99 benchmark samples to assess reproducibility: the reference sample NA24385 was
100 replicated into one routine for intra-assay reproducibility evaluation whereas NA24694
101 was included in all three routines for inter-assay reproducibility evaluation. Routines
102 were independently performed by different operators.

103

104 *Research ethics statement*

105 This study adheres to the Declaration of Helsinki principles for research in
106 human beings and was approved by the Hospital Israelita Albert Einstein's Research
107 Ethics Committee (São Paulo, Brazil. Protocol number: CAAE 29567220.4.1001.0071).
108 All individuals provided written consent for WGS testing.

109

110 *DNA extraction, quantification, and fragmentation*

111 DNA was extracted from whole blood samples using QIAasympyphony DNA Mini
112 Kit at QIAasympyphony automated system (both Qiagen, Valencia, CA, USA). The
113 extracted DNA was eluted into a final volume of 90 µl with an elution buffer. Genomic
114 DNA purity was evaluated using NanoDrop 2000 (thresholds: 260/280 ratio $\approx$ 1.8 and
115 260/230 ratio between 1.8 and 2.2). DNA quantification was performed with Qubit® 4
116 fluorometer using the Qubit® dsDNA HS assay (both Life Technologies, Carlsbad, CA,
117 USA).

118 Genomic DNA was fragmented into 350 bp inserts by Covaris ME220
119 ultrasonicator (Covaris, Woburn, MA, USA), with the following treatment settings –
120 DNA input: 1 µg (final volume of 55 µl, completed with resuspension buffer); peak

121 incident power: 50 W, duty factor: 20%, cycles per burst: 200, duration: 65 seconds,
122 and temperature set point: 20 °C.

123

124 *Whole Genome Sequencing library preparation*

125 The paired-ends sequencing libraries were prepared using 50 µl of the
126 fragmented DNA solution (1 µg DNA final) as input and Illumina TruSeq® DNA PCR-
127 Free Library Prep protocol HS (Illumina Inc., San Diego, CA, USA) for Whole Genome
128 Sequencing reagent kit, following the manufacturers' instructions. Briefly, the protocol
129 steps were: (1) Cleanup of fragmented DNA, (2) Repair ends and selection of library
130 size, (3) Removal of large DNA fragments, (4) Removal of small DNA fragments, (5) 3'-
131 ends adenylation (6) Adapter ligation, and (7) Cleanup of not-ligated fragments.

132

133 *Library quality control*

134 For quality control of adapter-ligated fragment sizes, libraries were diluted 1:5
135 with water and 2 µl were evaluated in the automated electrophoresis analysis
136 TapeStation System, with D1000 High Screen Tape (Agilent Technologies, Santa
137 Clara, CA, USA). High-quality (ideal) libraries displayed only one peak around 900 bp
138 (equivalent to $\approx$ 470 bp fragments due to the forked structures of adapter-ligated
139 fragments) and peak molarity $\geq$ 300 pM. Good-quality libraries had peak molarity
140 between 200 and 300 pM. Libraries with peak molarity $\leq$ 200 pM were rejected and
141 preparation was repeated.

142

143 *Library pooling and quantification*

144 We optimized the protocol for the pooling of a maximum of 28 sample libraries
145 for sequencing on each NovaSeq® 6000's S4 flow cell. Briefly, each library was
146 quantified with Qubit® 4 fluorometer then normalized to 7 nM in a final volume of 11

147 μL . Then, we pooled the 28 libraries into a final volume of 308 μL ($28 \times 11 = 308 \mu\text{L}$).
 148 Next, starting with 5 μL of the pooled solution as input, we performed two dilutions in a
 149 resuspension buffer (1:10 and 1:100, reaching the final 1:1,000 concentration). Four μL
 150 of the diluted pooled solution were used for real time quantitative polymerase chain
 151 reaction (qPCR) on ABI 7500 real time platform (Thermo Fisher Scientific, Waltham,
 152 MA, USA) using KAPA Library Quantification Kit (Roche, Pleasanton, CA, USA). The
 153 qPCR was performed in triplicate. In each qPCR run, six KAPA DNA standards with
 154 defined concentrations were included to produce a standard quantification curve. With
 155 the mean cycle threshold (CT) of the diluted samples, we calculated the concentration
 156 of the pooled libraries solutions via linear regression, while correcting for the size-
 157 difference of the KAPA standards in relation to the adapter-ligated fragments (452 bp
 158 versus 470 bp). Each pooled library was then normalized to 3 nM final concentration.

159 The pooled libraries were then spiked with 1.9 μL of 2.5 nM PhiX Control v3
 160 reagent (Illumina Inc., San Diego, CA, USA). The pooled libraries were then denatured
 161 with 77 μL of fresh 0.2N NaOH solution, followed by homogenization by vortex (1800
 162 RPM for 1 minute), centrifugation at $280 \times g$ for 1 minute and incubation at room
 163 temperature for 8 minutes. Then, 78 μL of 400 mM Tris-HCl (pH 8.0) solution was
 164 added to the libraries pool to neutralize the NaOH. Once again, the pooled libraries
 165 solution was homogenized by vortex (1800 RPM for 1 minute) and centrifuged at $280 \times$
 166 g for 1 minute. The full volume (466.9 μL) of the PhiX-spiked denatured library pool
 167 solution was then transferred to NovaSeq® 6000 Reagent Kit tube and proceeded to
 168 sequencing.

169

170 *Sequencing*

171 We performed sequencing with NovaSeq® 6000 platform using S4 flow cells
 172 with 300 cycles (150 for forward reads and 150 for reverse reads). Usually, each round
 173 of sequencing was composed by 28 pooled samples as described above, using both

174 flow cells available (total 56 samples per run). Desirable sequencing quality metrics
175 were cluster passing filter > 70% and flow cell occupation > 70%.

176

177 *Bioinformatics pipeline and quality control metrics*

178 The raw sequencing files (base call file, BCL format) were converted to FASTQ
179 format and demultiplexed in a single step using Illumina's *bcl2fastq* program ¹³.
180 Illumina's DRAGEN pipeline version 3.6.3 was used to perform all alignment and
181 variant call (SNVs, indels, CNVs, SVs) steps. Quality control metrics are provided
182 during each DRAGEN run.

183 Desirable alignment quality metrics were percentage of bases that meet Q30
184 score > 90%, 20X minimum coverage for both whole genome and autosomes,
185 uniformity of coverage $\geq$ 80%, median insert size > 300 bp, percentage of mapped
186 reads > 98%, percentage of chimeric (supplementary) reads < 5%, DNA contamination
187 $\leq$ 2%, and percentage of mapped reads marked as duplicate < 10%. Some of these
188 thresholds were adopted from recommendations published elsewhere ¹⁰.

189 The DRAGEN-generated Variant Call Format (VCF) files were validated to
190 ensure they had the correct format, and sample- and variant-specific quality metrics
191 were also calculated. Each sample was assessed to ensure that percent autosomal
192 callability was > 95% as suggested elsewhere ¹⁰.

193 High-quality variants were those which passed Variant Quality Score
194 Recalibration (VQSR) filter, had read depth (RD) $\geq$ 10 and genotype quality (GQ) $\geq$ 20
195 in at least 80% of the individuals in the sample, their alternative alleles were present in
196 at least one individual with RD $\geq$ 10 and GQ $\geq$ 20, and were not located into locations
197 with high multiallelic variation (more than four alleles, includes non-pseudoautosomal
198 region of X and Y chromosomes).

199 Functional annotation of the variants was performed with a proprietary tool,
200 Varstation (<https://varsomics.com/varstation/>) developed by Hospital Israelita Albert

201 Einstein (HIAE). The VCF files were uploaded into the service, whose workflow is
202 based on ANNOVAR ¹⁴. The variants are then classified according to international
203 good practices on genetic variants analyses and guidelines from the American College
204 of Medical Genetics (ACMG) ¹⁵ and the Association for Molecular Pathology (AMP) ¹⁶.

205

206 *Data analysis*

207 The 76 samples were separated into two different analytical groups. The first
208 group included seven sequencing libraries corresponding to GiaB benchmark samples
209 (NA12878, NA24385, NA24149, NA24143, NA24631, NA24694, and NA24695). The
210 second group included the remaining 69 samples, i.e., the remaining 15 GiaB samples
211 and the 54 in-house characterized samples.

212 The first group was analysed by comparison of the VCF files generated by our
213 Bionformatics pipeline with reference VCF files provided by GiaB (version NISTv3.3.2).
214 Each sample had an accompanying BED file with high-confidence regions coordinates.
215 We performed the comparison through *vcfeval* software (Real Time Genomics,
216 Hamilton, New Zealand) ¹⁷. Briefly, *vcfeval* quantifies the number of true positives (the
217 variant call is present in both the reference file and our file), false positives (the variant
218 call is absent in the reference file but present in our file) and false negatives (the
219 variant call is present in the reference file but absent in our file). We then calculated the
220 precision, sensitivity (recall), and F-measures with those numbers. Additionally, we
221 stratified each file by SNVs and indels coordinates. In this step we calculated the mean
222 of each metric mentioned above alongside 95% confidence intervals (95% CI).

223 The second analytical group samples were evaluated manually to assess the
224 performance of not only SNVs detection, but also for CNVs and SVs, by comparing the
225 pipeline output with the in-house annotation or the GiaB annotation, depending on the
226 sample origin. Supplementary Table 1 contains a list of the samples used, quality
227 metrics and a summary of expected and observed variant calls.

228

229 **Results**

230 The mean sequencing yield considering all routines was 2.84 TB of data per S4
231 flow cell. Mean %Q30 score was 92.60% $\pm$ 1.36%, mean genomic coverage 38.96X $\pm$
232 10.37X and mean uniformity was 96.31% $\pm$ 0.25%. Mean mitochondrial coverage was
233 7650.97X $\pm$ 5559.1X.

234 Variant calls from our WGS procedure yielded very high concordance with the
235 reference samples. For SNVs, the mean F-measure (n=7 reference GiaB samples) was
236 99.82% (95% CI = 99.44% - 100.0%), whereas for indels of any length was 99.57%
237 (95% CI = 99.29% - 99.85%) (Table 1, Supplementary Table 2, Supplementary Figure
238 1).

239 Our procedure worked best for small indels with length between 1 and 5 base-
240 pairs (bp) (mean F-measure = 99.63%, 95% CI = 99.36% - 99.89%). Six to 15 bp
241 indels yielded mean F-measure = 99.27%, 95% CI = 98.87% - 99.68% and 16-bp or
242 more indels yielded mean F-measure = 98.13%, 95% CI = 97.44% - 98.83% (Table 1).

243 Our optimized WGS protocol presented excellent intra- and inter-assay
244 reproducibility. Regarding SNVs, the intra-assay coefficient of variation (CV) of the F-
245 measures was 0.04%, whereas the inter-assay was 0.03%. Regarding indels, the intra-
246 assay F-measures CV was 0.16% whereas the inter-assay CV was 0.07% (Table 2).

247 The pathogenic/likely pathogenic variant profile of the 54 in-house
248 characterized samples included: 12 SNVs (eight missense, two nonsense, one splicing
249 acceptor and another splicing donor), 65 large deletions (lengths ranging between 538
250 bp – 53,247,491 bp), including 29 loss of heterozygosity (LOH) regions identified by
251 SNP array (lengths ranging between 812,863 bp and 72,740,279 bp); 22 large
252 duplications (ranging between 6147 bp and 95,325,642 bp), three events of trisomy
253 (chromosomes 13, 15 or 21), one insertion, four inversions, 10 translocations, two

Robertsonian translocations and a single event of uniparental disomy, totalling 120 events.

All SNVs were correctly detected by our variant call procedure (F-measure = 100.0%). The detection of the single event of uniparental disomy failed (Table 3). The CNV and SV events detection performances were overall good (F-measures 96.6% and 90.3% respectively).

Currently, we have sequenced over 2,000 among 3,000 enrolled patients with rare diseases or hereditary cancer syndromes with our optimized WGS protocol. Sequencing and alignment metrics are available for about 1,200 samples and have been consistently of high quality, compatible with clinical diagnostic routine (Supplementary Table 3). For example, cross-individual contamination is virtually non-existent (mean 0.008% $\pm$ 0.11), the mean uniformity of coverage is 96.4% $\pm$ 0.26%, the median genome coverage is 36.5X, the mean percentage of bases with quality score Q30 or more is 91.3% $\pm$ 3.6% and mean genome callability is 96.3% $\pm$ 1.2%. Of those, over 300 patients have received a diagnostic report, with approximately 37% presenting a definitive molecular diagnosis, with the detection of a pathogenic/likely pathogenic variant compatible with the patient's phenotype.

271

272 Discussion

The diagnosis of patients with rare disorders is currently a lengthy process, taking on average four or more years. Early adoption of WGS could be beneficial, shortening the diagnostic odyssey¹⁸. A recent meta-analysis of 37 studies involving children with genetic diseases showed that WGS testing had higher clinical utility (see¹⁹ for a definition) than chromosomal microarray, and an accompanying meta-regression showed that the odds of diagnosis through WGS increased by 16% each year, possibly due to methodological improvements (the meta-analysis included WGS studies published between 2015 and 2017)²⁰. Other studies reported the discovery of

281 three new genetic conditions in children with complex phenotypes ²¹ and clinical utility
282 for children undergoing intensive care ²². Therefore, WGS testing is a powerful tool with
283 clear benefits for the diagnosis of rare diseases.

284 Brazil is the only country with a population larger than 100 million people which
285 has a public, universal and free of charge health care system ²³. Provisioning of a rapid
286 but cost effective genomic testing strategy within a national healthcare service in order
287 to deliver equity of access is nevertheless challenging ²⁴, especially with a system of
288 this magnitude.

289 As with all intensive care procedures, balancing the extreme stress of families
290 with the complexities of an informed consent requires skill and sensitivity while
291 incomplete understanding of a new technology creates uncertainty for care providers ²⁵⁻
292 ²⁷. Genomic medicine has the capacity to revolutionize the healthcare of an individual
293 with a rare disease or cancer by offering prompt and accurate diagnosis, risk
294 stratification based upon genotype and the capacity for personalized treatments. A
295 recent application of WGS coupled to rare diseases diagnosis in a national context (the
296 UK 100K Genomes Project) revealed a remarkable benefit to routine healthcare ²⁸. A
297 meta-analysis of psychological outcomes suggested no harm following WGS result
298 disclosure and even an overall trend for a decrease in anxiety ²⁹.

299 We developed and validated a comprehensive WGS workflow with an optimized
300 laboratory turnaround time coupled with cutting edge Bioinformatics pipeline for variant
301 calling, functional annotation and classification. Our procedure was performed following
302 important benchmarking guidelines ^{30, 31} and yielded excellent performance. The key
303 step for a robust validation is the careful sample selection. Using a set composed of
304 reference benchmark samples, which have millions of completely validated variants of
305 different types, and in-house characterized or purchased samples for more complex
306 variants such as structural, mitochondrial and LOH events is of paramount importance.
307 In addition, the validation of detection of hard-to-detect variant types, such as repeat
308 expansions, variants in genes with pseudogenes or homologous genes and low-level

309 mosaicism requires even further steps, including additional samples, possibly on a
310 gene-by-gene basis¹⁰.

311 The challenge in assessing and interpreting variants is complex, and we
312 acknowledge some limitations of our protocol. We did not assess repeat expansion
313 variants, tandem duplications, mitochondrial genome heteroplasmy, mosaicism, and
314 processed pseudogene insertions. Moreover, we did not have in the sample set
315 deletions or duplications smaller than 538 bp. Therefore, the detection sensitivity of
316 CNVs with less than 538 bp may differ from the one reported here. We plan on
317 validating the detection of some of these variant types in the near future, in order to
318 improve the test robustness, sensitivity, specificity and limits of detection.

319

320 **Conclusion**

321 Large-scale WGS projects are important initiatives to expand the population's
322 access to robust genomic technologies, such as NGS. The validation of our WGS
323 workflow is the first step for this achievement and has the potential to reduce time until
324 diagnosis of patients with rare diseases, improving the affected individuals and their
325 family's quality of life. Also, considering the high diversity and miscegenation of our
326 population, The Rare Genomes Project is fundamental for the creation of a disease-
327 related variants databases, collaborating with the future of precision medicine.

328

329 **Supplementary Figure 1 caption**

330 Supplementary Figure 1. Circular plot displaying variant calls in the seven gold-
331 standard GiaB samples. They are ordered according to chromosome (circular sectors)
332 and genomic position. Each library corresponds to one circumference of said sectors.
333 Yellow points/top third of the sector: true positive calls, red points/middle third of the

334 sector: false positive calls, blue points/bottom third of the sector: false negative calls.
 335 Left: result for SNVs, right: result for indels.

336

337 References

- 338 1. Ferreira CR: The burden of rare diseases. *Am J Med Genet A* 2019, 179:885-
 339 892.
- 340 2. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM,
 341 Thiruvahindrapuram B, Merico D, Jobling R, Nalpathamkalam T, Pellecchia G,
 342 Sung WWL, Wang Z, Bikangaga P, Boelman C, Carter MT, Cordeiro D,
 343 Cytrynbaum C, Dell SD, Dhir P, Dowling JJ, Heon E, Hewson S, Hiraki L, Inbar-
 344 Feigenberg M, Klatt R, Kronick J, Laxer RM, Licht C, MacDonald H, Mercimek-
 345 Andrews S, Mendoza-Londono R, Piscione T, Schneider R, Schulze A,
 346 Silverman E, Siriwardena K, Snead OC, Sondheimer N, Sutherland J, Vincent
 347 A, Wasserman JD, Weksberg R, Shuman C, Carew C, Szego MJ, Hayeems
 348 RZ, Basran R, Stavropoulos DJ, Ray PN, Bowdin S, Meyn MS, Cohn RD,
 349 Scherer SW, Marshall CR: Improved diagnostic yield compared with targeted
 350 gene sequencing panels suggests a role for whole-genome sequencing as a
 351 first-tier genetic test. *Genet Med* 2018, 20:435-443.
- 352 3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K,
 353 Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A,
 354 Howland J, Kann L, Lehoczy J, LeVine R, McEwan P, McKernan K, Meldrim J,
 355 Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R,
 356 Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A,
 357 Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee
 358 C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I,
 359 Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A,
 360 Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC,
 361 Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson
 362 RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT,
 363 Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL,
 364 Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton
 365 SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T,
 366 Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M,
 367 Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives
 368 CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL,
 369 Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T,
 370 Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin
 371 W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith
 372 DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J,
 373 Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang
 374 J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel
 375 NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox
 376 DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S,
 377 Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J,
 378 Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H,
 379 Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A,
 380 Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC,
 381 Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS,
 382 Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H,

383 Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S,
384 Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A,
385 Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz
386 J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J,
387 Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh
388 RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos
389 A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen
390 YJ, Szustakowki J, International Human Genome Sequencing C: Initial
391 sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

392 4. Wetterstrand KA: DNA Sequencing Costs: Data from the NHGRI Genome
393 Sequencing Program (GSP). National Human Genome Research Institute,
394 2020.

395 5. Nagarajan N, Pop M: Sequence assembly demystified. *Nature reviews*
396 *Genetics* 2013, 14:157-167.

397 6. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L,
398 Boisson B, Casanova J-L, Abel L: Whole-genome sequencing is more powerful
399 than whole-exome sequencing for detecting exome variants. *Proceedings of the*
400 *National Academy of Sciences* 2015, 112:5473-5478.

401 7. Bick D, Jones M, Taylor SL, Taft RJ, Belmont J: Case for genome sequencing
402 in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet*
403 2019, 56:783-791.

404 8. Howell KB, Eggers S, Dalziel K, Riseley J, Mandelstam S, Myers CT, McMahon
405 JM, Schneider A, Carvill GL, Mefford HC, Victorian Severe Epilepsy of Infancy
406 Study G, Scheffer IE, Harvey AS: A population-based cost-effectiveness study
407 of early genetic testing in severe epilepsies of infancy. *Epilepsia* 2018, 59:1177-
408 1187.

409 9. Vissers L, van Nimwegen KJM, Schieving JH, Kamsteeg EJ, Kleefstra T,
410 Yntema HG, Pfundt R, van der Wilt GJ, Krabbenborg L, Brunner HG, van der
411 Burg S, Grutters J, Veltman JA, Willemsen M: A clinical utility study of exome
412 sequencing versus conventional genetic testing in pediatric neurology. *Genet*
413 *Med* 2017, 19:1055-1063.

414 10. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM,
415 Rowsey R, Klee EW, Liu P, Worthey EA, Jobanputra V, Dimmock D, Kearney
416 HM, Bick D, Kulkarni S, Taylor SL, Belmont JW, Stavropoulos DJ, Lennon NJ,
417 Medical Genome I: Best practices for the analytical validation of clinical whole-
418 genome sequencing intended for the diagnosis of germline disease. *npj*
419 *Genomic Medicine* 2020, 5:47.

420 11. Barra GB, Júnior NG, Filho JBO: Lista de Orientação em Diagnóstico
421 Molecular. Segunda versão, 2018.

422 12. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y,
423 Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F,
424 Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z,
425 Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC,
426 Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J,
427 Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P,
428 Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS,
429 Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M: Extensive sequencing of
430 seven human genomes to characterize benchmark reference materials.
431 *Scientific Data* 2016, 3:160025.

432 13. Illumina Inc.: bcl2fastq2 Conversion Software v2.20. San Diego, California,
433 USA: Illumina Inc., 2019.

434 14. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic
435 variants from high-throughput sequencing data. *Nucleic Acids Research* 2010,
436 38:e164-e164.

15. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee ALQA: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015, 17:405-424.
16. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN: Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn* 2017, 19:4-23.
17. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, Zook JM, Trigg L, De La Vega FM: Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* 2015:023754.
18. Wu AC, McMahon P, Lu C: Ending the Diagnostic Odyssey-Is Whole-Genome Sequencing the Answer? *JAMA Pediatr* 2020, 174:821-822.
19. ACMG Board of Directors: Clinical utility of genetic and genomic services: a position statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine* 2015, 17:505-507.
20. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, Kingsmore SF: Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Medicine* 2018, 3:16.
21. Costain G, Walker S, Marano M, Veenma D, Snell M, Curtis M, Luca S, Buera J, Arje D, Reuter MS, Thiruvahindrapuram B, Trost B, Sung WWL, Yuen RKC, Chitayat D, Mendoza-Londono R, Stavropoulos DJ, Scherer SW, Marshall CR, Cohn RD, Cohen E, Orkin J, Meyn MS, Hayeems RZ: Genome Sequencing as a Diagnostic Test in Children With Unexplained Medical Complexity. *JAMA Netw Open* 2020, 3:e2018109.
22. Sanford EF, Clark MM, Farnaes L, Williams MR, Perry JC, Ingulli EG, Sweeney NM, Doshi A, Gold JJ, Briggs B, Bainbridge MN, Feddock M, Watkins K, Chowdhury S, Nahas SA, Dimmock DP, Kingsmore SF, Coufal NG: Rapid Whole Genome Sequencing Has Clinical Utility in Children in the PICU. *Pediatr Crit Care Med* 2019, 20:1007-1020.
23. Castro MC, Massuda A, Almeida G, Menezes-Filho NA, Andrade MV, de Souza Noronha KVM, Rocha R, Macinko J, Hone T, Tasca R, Giovanela L, Malik AM, Werneck H, Fachini LA, Atun R: Brazil's unified health system: the first 30 years and prospects for the future. *Lancet* 2019, 394:345-356.
24. Berg JS, Agrawal PB, Bailey DB, Jr., Beggs AH, Brenner SE, Brower AM, Cakici JA, Ceyhan-Birsoy O, Chan K, Chen F, Currier RJ, Dukhovny D, Green RC, Harris-Wai J, Holm IA, Iglesias B, Joseph G, Kingsmore SF, Koenig BA, Kwok PY, Lantos J, Leeder SJ, Lewis MA, McGuire AL, Milko LV, Mooney SD, Parad RB, Pereira S, Petrikin J, Powell BC, Powell CM, Puck JM, Rehm HL, Risch N, Roche M, Shieh JT, Veeraraghavan N, Watson MS, Willig L, Yu TW, Urv T, Wise AL: Newborn Sequencing in Genomic Medicine and Public Health. *Pediatrics* 2017, 139.
25. Tan N, Amendola LM, O'Daniel JM, Burt A, Horike-Pyne MJ, Boshe L, Henderson GE, Rini C, Roche MI, Hisama FM, Burke W, Wilfond B, Jarvik GP: Is "incidental finding" the best term?: a study of patients' preferences. *Genet Med* 2017, 19:176-181.
26. Braverman G, Shapiro ZE, Bernstein JA: Ethical Issues in Contemporary Clinical Genetics. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 2018, 2:81-90.

- 491 27. Char DS, Lee SS, Magnus D, Cho M: Anticipating uncertainty and irrevocable
492 decisions: provider perspectives on implementing whole-genome sequencing in
493 critically ill children with heart disease. *Genet Med* 2018, 20:1455-1461.
- 494 28. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, Allen HL,
495 Sanchis-Juan A, Frontini M, Thys C, Stephens J, Mapeta R, Burren OS,
496 Downes K, Haimel M, Tuna S, Deevi SVV, Aitman TJ, Bennett DL, Calleja P,
497 Carss K, Caulfield MJ, Chinnery PF, Dixon PH, Gale DP, James R, Koziell A,
498 Laffan MA, Levine AP, Maher ER, Markus HS, Morales J, Morrell NW, Mumford
499 AD, Ormondroyd E, Rankin S, Rendon A, Richardson S, Roberts I, Roy NBA,
500 Saleem MA, Smith KGC, Stark H, Tan RYY, Themistocleous AC, Thrasher AJ,
501 Watkins H, Webster AR, Wilkins MR, Williamson C, Whitworth J, Humphray S,
502 Bentley DR, Nihir BioResource for the GP, Kingston N, Walker N, Bradley JR,
503 Ashford S, Penkett CJ, Freson K, Stirrups KE, Raymond FL, Ouwehand WH:
504 Whole-genome sequencing of patients with rare diseases in a national health
505 system. *Nature* 2020, 583:96-102.
- 506 29. Robinson JO, Wynn J, Biesecker B, Biesecker LG, Bernhardt B, Brothers KB,
507 Chung WK, Christensen KD, Green RC, McGuire AL, Hart MR, Griesemer I,
508 Patrick DL, Rini C, Veenstra D, Cronin AM, Gray SW: Psychological outcomes
509 related to exome and genome sequencing result disclosure: a meta-analysis of
510 seven Clinical Sequencing Exploratory Research (CSER) Consortium studies.
511 *Genetics in Medicine* 2019, 21:2781-2790.
- 512 30. Krusche P, Trigg L, Boutros PC, Mason CE, La Vega FMD, Moore BL,
513 Gonzalez-Porta M, Eberle MA, Tezak Z, Labadibi S, Truty R, Asimenos G,
514 Funke B, Fleharty M, Chapman BA, Salit M, Zook JM: Best Practices for
515 Benchmarking Germline Small Variant Calls in Human Genomes. *bioRxiv*
516 2018:270157.
- 517 31. Koboldt DC: Best practices for variant calling in clinical sequencing. *Genome*
518 *Med* 2020, 12:91.

Table 1. Quality metrics. Seven Genome in a Bottle Consortium gold standard samples were whole-genome sequenced and variant call was performed with our Varstation pipeline. The variant call files were then compared with the gold standard files through *vcfeval* program, which quantified true positives, false positives, and false negatives according to variants captured by each sample BED file with coordinates of high-confidence regions. Precision, Sensitivity and F-measure were then calculated.

Target	Metric	Mean	Standard deviation	95% confidence interval	
				Lower bound	Upper bound
SNVs	Precision	0.9986	0.0011	0.9965	1.0000
	Sensitivity	0.9979	0.0029	0.9922	1.0000
	F-measure	0.9982	0.0020	0.9944	1.0000
Indels, overall	Precision	0.9961	0.0008	0.9944	0.9977
	Sensitivity	0.9954	0.0021	0.9912	0.9995
	F-measure	0.9957	0.0014	0.9929	0.9985
Indels, 1 to 5 bp	Precision	0.9965	0.0008	0.9949	0.9981
	Sensitivity	0.9961	0.0019	0.9923	0.9998
	F-measure	0.9963	0.0013	0.9936	0.9989
Indels, 6 to 15 bp	Precision	0.9939	0.0015	0.9910	0.9968
	Sensitivity	0.9916	0.0028	0.9861	0.9971
	F-measure	0.9927	0.0021	0.9887	0.9968
Indels, ≥ 16 bp	Precision	0.9832	0.0055	0.9725	0.9939
	Sensitivity	0.9795	0.0082	0.9634	0.9955
	F-measure	0.9813	0.0036	0.9744	0.9883

526

527

Table 2. Reproducibility. The benchmark sample NA24385 was selected for intra-assay reproducibility evaluation whereas NA24694 was included in all three independent routines for inter-assay reproducibility evaluation. Coefficients of variation (CV) of quality metrics are reported.

Reproducibility	Samples	SNVs			Indels		
		Precision	Sensitivity	F-measure	Precision	Sensitivity	F-measure
Intra-assay	NA24385	0.9970	0.9918	0.9944	0.9969	0.9948	0.9959
	NA24385-2	0.9963	0.9914	0.9939	0.9952	0.9920	0.9936
	<i>Mean</i>	<i>0.9967</i>	<i>0.9916</i>	<i>0.9941</i>	<i>0.9961</i>	<i>0.9934</i>	<i>0.9947</i>
	<i>SD</i>	<i>0.0005</i>	<i>0.0003</i>	<i>0.0004</i>	<i>0.0012</i>	<i>0.0020</i>	<i>0.0016</i>
	CV (%)	0.0501	0.0322	0.0411	0.1221	0.1991	0.1607
Inter-assay	NA24694	0.9992	0.9993	0.9993	0.9968	0.9980	0.9974
	NA24694-2	0.9994	0.9993	0.9993	0.9974	0.9986	0.9980
	NA24694-3	0.9988	0.9989	0.9988	0.9963	0.9968	0.9966
	<i>Mean</i>	<i>0.9991</i>	<i>0.9991</i>	<i>0.9991</i>	<i>0.9968</i>	<i>0.9978</i>	<i>0.9973</i>
	<i>SD</i>	<i>0.0003</i>	<i>0.0002</i>	<i>0.0003</i>	<i>0.0005</i>	<i>0.0009</i>	<i>0.0007</i>
	CV (%)	0.0312	0.0235	0.0271	0.0551	0.0909	0.0726

SD – standard deviation, CV – coefficient of variation.

Table 3. Quality metrics of the variant call procedure performed on 69 samples, including 54 in-house characterized samples by other methodologies. The seven GiaB gold-standard samples are not considered here, see Table 1. Also see Supplementary Table 1 for a breakdown of expected and observed variant calls (analysis group 2 rows).

Variant	True positives (TP)	False negatives (FN)	Precision	Sensitivity	F-measure
SNVs					
Missense	8	0	1.0000	1.0000	1.0000
Nonsense	2	0	1.0000	1.0000	1.0000
Splicing acceptor	1	0	1.0000	1.0000	1.0000
Splicing donor	1	0	1.0000	1.0000	1.0000
Overall	12	0	1.0000	1.0000	1.0000
CNVs					
Deletions	60	5	1.0000	0.9231	0.9600
Duplications	22	0	1.0000	1.0000	1.0000
Trisomy 13	1	0	1.0000	1.0000	1.0000
Trisomy 15	0	1	Not calculated	0.0000	0.0000
Trisomy 21	1	0	1.0000	1.0000	1.0000
Overall	84	6	1.0000	0.9333	0.9655
SVs					
Insertions	1	0	1.0000	1.0000	1.0000
Inversions	4	0	1.0000	1.0000	1.0000
Robertsonian translocations	0	2	Not calculated	0.0000	0.0000
Translocations	9	1	1.0000	0.9000	0.9474
Overall	14	3	1.0000	0.8235	0.9032
Uniparental (UPD) disomy	0	1	Not calculated	0.0000	0.0000
SNVs + CNVs + SVs + UPD	110	10	1.0000	0.9167	0.9565