



Detection of copy number variations from NGS data using read depth information: a diagnostic performance evaluation

Olivier Quenez, Kevin Cassinari, Sophie Coutant, Francois Lecoquierre, Kilan Le Guennec, Stéphane Rousseau, Anne-Claire Richard, Stéphanie Vasseur, Emilie Bouvignies, Jacqueline Bou, et al.

► To cite this version:

Olivier Quenez, Kevin Cassinari, Sophie Coutant, Francois Lecoquierre, Kilan Le Guennec, et al.. Detection of copy number variations from NGS data using read depth information: a diagnostic performance evaluation. 2019. hal-02317979v2

HAL Id: hal-02317979

<https://normandie-univ.hal.science/hal-02317979v2>

Preprint submitted on 21 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Detection of copy number variations from NGS data using read depth information: a
2 diagnostic performance evaluation

3 Running title : Evaluation of a CANOES-centered workflow

4

5 Olivier Quenez¹, Kevin Cassinari¹, Sophie Coutant², François Lecoquierre², Kilan Le Guennec¹,
6 Stéphane Rousseau¹, Anne-Claire Richard¹, Stéphanie Vasseur², Emilie Bouvignies², Jacqueline
7 Bou², Gwendoline Lienard², Sandrine Manase², Steeve Fourneaux², Nathalie Drouot², Virginie
8 Nguyen-Viet², Myriam Vezain², Pascal Chambon², Géraldine Joly-Helas², Nathalie Le Meur²,
9 Mathieu Castelain², Anne Boland³, Jean-François Deleuze³, FREX Consortium, Isabelle Tournier²,
10 Françoise Charbonnier², Edwige Kasper², Gaëlle Bougeard², Thierry Frebourg², Pascale Saugier-
11 Veber², Stéphanie Baert-Desurmont², Dominique Champion^{1,4}, Anne Rovelet-Lecrux¹, Gaël Nicolas¹

12 ¹Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and
13 CNR-MAJ, Normandy Center for Genomic and Personalized Medicine, Rouen, France.

14 ²Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics,
15 Normandy Center for Genomic and Personalized Medicine, Rouen, France

16 ³Centre National de Recherche en Génomique Humaine, Institut de Génomique, CEA, Evry, France

17 ⁴Department of Research, Centre hospitalier du Rouvray, Sotteville-lès-Rouen, France

18

19 Corresponding author : Gaël Nicolas, Inserm U1245, Faculté de médecine, 22, boulevard Gambetta,
20 76183 Rouen, tel. 0033 235 14 83 08, e-mail: gaelnicolas@hotmail.com

21

22 This study received fundings from Clinical Research Hospital Program from the French Ministry of
23 Health (GMAJ, PHRC 2008/067), the JPND PERADES and France Génomique. This study was co-
24 supported by the Centre National de Référence Malades Alzheimer Jeunes (CNR-MAJ), European
25 Union and Région Normandie. Europe gets involved in Normandie with the European Regional
26 Development Fund (ERDF).

27 **ABSTRACT**

28 The detection of Copy Number Variations (CNVs) from NGS data is under-exploited as chip-based
29 or targeted techniques are still commonly used. We assessed the performances of a workflow
30 centered on CANOES, a bioinformatics tool based on read depth information.

31 We applied our workflow to gene panel (GP) and Whole Exome Sequencing (WES) data, and
32 compared CNV calls to Quantitative Multiplex PCR of Short Fluorescent fragments (QMSPF) or
33 array Comparative Genomic Hybridization (aCGH) results.

34 From GP data of 3,776 samples, we reached an overall Positive Predictive Value (PPV) of 87.8%.

35 This dataset included a complete comprehensive QMPSF comparison of 4 genes (60 exons) on
36 which we obtained 100% sensitivity and specificity.

37 From WES data, we first compared 137 samples to aCGH and filtered comparable events (exonic
38 CNVs encompassing enough aCGH probes) and obtained an 87.25% sensitivity. The overall PPV
39 was 86.4% following the targeted confirmation of candidate CNVs from 1,056 additional WES.

40 In addition, our CANOES-centered workflow on WES data allowed the detection of CNVs of any
41 size that were missed by aCGH. Overall, switching to a NGS-only approach should be cost-
42 effective as it allows a reduction in overall costs together with likely stable diagnostic yields. Our
43 bioinformatics pipeline is available at : [https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-](https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-workflow)
44 workflow.

45

46

47

48 **KEYWORDS**

49 Exome, panel, CANOES, CNV detection, bioinformatics, sensitivity

50 **INTRODUCTION**

51 Copy-number variations (CNVs) are a major cause of Mendelian disorders (1) as well as risk
52 factors for common diseases (2). With the advent of next-generation sequencing (NGS), a number
53 of software tools have been developed to detect CNVs. Whole genome sequencing (WGS) is often
54 presented as an almost universal technique allowing the assessment of almost any type of variation,
55 including CNVs and other structural variations. WGS may eventually be used as a first-tier
56 diagnostics tool in the context of genetically highly heterogeneous disorders. However, the
57 detection of structural variations from data generated using the technology of short read sequencing
58 is still associated with a number of false positives. Such events can be detected using a plethora of
59 bioinformatics tools based on different principles, including Depth Of Coverage (DOC)
60 information, relative position of paired reads, split reads and DeNovo Assembly (3). Besides the
61 development of WGS, targeted sequencing of gene panels and whole exome sequencing (WES)
62 remain of primary use in many diagnostics and research laboratories. They are indeed still
63 considered as more affordable and of easier access as they can be processed using usual informatics
64 facilities accessible to most laboratories. Moreover, the input of WGS is questioning in disorders
65 with low genetic heterogeneity and high phenotypic specificity. Hence, gene panels and WES
66 remain largely used .

67 The detection of CNVs from exonic capture-based targeted sequencing solutions primarily relies on
68 DOC information (4,5). Tools based on DOC information compare one sample to a reference, and
69 predict deletions or duplications depending on the increase or decrease of the DOC as compared to
70 the reference (figure 1). As each tool was set up and trained on a specific dataset, one of the main
71 challenges is to evaluate the specificity and sensitivity of a given software tool on large datasets.
72 Studies evaluating the diagnostic performances of CNV detection pipelines are scarce although they
73 appear to be critical for their use in routine procedures.

74 In order to optimize CNV detection from NGS data, a classical approach consists in running
75 multiple tools in parallel and then aggregate the results to keep a CNV as candidate only if multiple
76 tools called it (6). As it is more effective to do so with tools using different types of bioinformatics
77 methods (DOC, split reads, etc.), this combinatory approach is most adapted when working on
78 WGS, or at least if most of the intergenic or intronic regions – where breakends are more frequently
79 found – are captured. Here, we decided to focus on one tool using the DOC approach as it still
80 remains the most adapted one for exonic capture. In a *precision workflow* approach, we developed a
81 workflow based on the already existing software tool CANOES (7). Briefly, CANOES adopts a
82 pooling strategy to build its reference model, and uses a Hidden Markov Model to represent the
83 DOC of this model. Lastly, it confronts the samples to the reference in order to call candidate
84 deletions or duplications.

85 We performed a diagnostic performance evaluation of this workflow regarding gene panel and WES
86 data, in two steps. First, we compared CNV calls with a reference technique, namely a
87 comprehensive assessment by Quantitative Multiplex PCR of Short Fluorescent fragments
88 (QMPSF) (8) or array comparative genomic hybridization (aCGH), regarding targeted gene panel
89 and WES data, respectively. Second, we implemented our workflow in our routine procedures and
90 performed an additional evaluation of the positive predictive value of our CANOES-centered
91 workflow using targeted confirmation of CNVs using an independent targeted technique.

92

93

94 **MATERIAL AND METHODS**

95 **Gene panel sequencing**

96 In order to evaluate our workflow, we analyzed data from three gene panels (for detailed
97 information, see supplementary table 1). Patients provided informed written consent for genetic
98 analyses in a diagnostics setting.

99 Panel 1 was set up to focus on genes involved in predisposition to colorectal cancer and digestive
100 polyposis or Li-Fraumeni syndrome (9). This panel was implemented in two successive versions.
101 V1 was used to sequence 11 genes in 2,771 samples. V2 was used to sequence 15 genes (same 11
102 genes plus 4) in 549 samples. In both versions and for all genes, exons and introns outside repeated
103 sequences were captured.

104 Panel 2 also has two successive versions and was designed to focus on two clinical indications: (i)
105 hydrocephaly (3 genes) and (ii) Cornelia de Lange syndrome and differential diagnoses (24 genes in
106 v1, 30 in v2). In total, 320 samples were sequenced using this panel (240 with v1, 80 with v2). For
107 this panel, introns outside repeated sequences were captured only for two genes, namely *LICAM*
108 and *NIPBL*.

109 Panel 3 was designed to focus on genes involved in non-specific Intellectual Disability. It has been
110 used to analyses 220 samples and is composed of 48 genes (coding regions only). The list of genes
111 is available upon request.

112

113 **Assessment of CNV calls from gene panel data: step 1**

114 For the comparison to a reference technique, we used data obtained from samples for which both
115 NGS (panel 1, v1) and comprehensive QMPSPF screening data were available (n=465). This
116 QMSPF assessment included all 60 exons of 4 genes from this panel (*APC*, *MSH2*, *MSH6*, *MLH1*)
117 and was applied to all 465 samples.

118

119 **Assessment of CNV calls from gene panel data: step 2**

120 Following step 1, we implemented our CANOES-centered workflow in our routine diagnostics
121 procedures on NGS data from all three panels (n=3,311 additional samples in total). We performed
122 confirmations of candidate CNVs using QMPSF or Multiplex Ligation-dependent Probe
123 Amplification (MLPA) only in samples with a CANOES call. Primers used for QMPSF screening
124 and validation are available upon request.

125

126 **Whole-exome sequencing**

127 Patients provided informed written consent for genetic analyses either in a diagnostics or in a
128 research setting, following the approval by our ethics committee.

129 Whole exomes were sequenced in the context of diverse research and diagnostics purposes
130 (supplementary table 1). Exomes were captured using Agilent SureSelect Human All Exon kits (V1,
131 V2 V4+UTR, V5, V5+UTR and V6) (Agilent technologies, Santa Clara, CA, USA). Final libraries
132 were sequenced on an Illumina Genome Analyser GAIIX (corresponding to exomes captured with
133 the V1, V2 or V4UTR kit, n=10), or on an Illumina HiSeq2000, 2500 or 4000 with paired ends, 76
134 or 100bp reads (Illumina, San Diego, Ca, USA). Exome sequencing was performed in 3 sequencing
135 centers: Integragen (Evry, France) (n=6), the French National Center of Human Genomics Research
136 (CNRGH, Evry, France) (n=1,065) and the Genome Quebec Innovation Center (Montreal, Canada)
137 (n=128) (10). Exomes were all processed through the same bioinformatics pipeline following the
138 Broad Institute Best Practices recommendations (11). Reads were mapped to the 1000 Genomes
139 GRCh37 build using BWA 0.7.5a.(12). Picard Tools 1.101 (<http://broadinstitute.github.io/picard/>)
140 was used to flag duplicate reads. We applied GATK (13) for short insertion and deletions (indel)
141 realignment and base quality score recalibration. All quality checks were processed as previously
142 described (10).

143

144 **Assessment of CNV calls from whole exome sequencing data: step 1**

145 For the comparison to a reference technique, we analyzed data from 147 unrelated individuals with
146 both WES and aCGH data available.

147 Array CGH Analysis. Oligonucleotide aCGH was performed as previously described (14). Briefly,
148 high-resolution aCGH analysis was performed using the 1x1M Human High-Resolution Discovery
149 Microarray Kit or the 4x180K SurePrint G3 Human CGH Microarray kit (Agilent Technologies,
150 Santa Clara, California, USA), using standard recommended protocols. An in-house and sex-
151 matched genomic DNA pool of at least 10 control individuals was used as reference sample.
152 Hybridization results were analyzed with the Agilent's DNA-Analytics software (version 4.0.81,
153 Agilent Technologies) or the Agilent Genomic Workbench (version 7.0, Agilent Technologies). Data
154 were processed using the ADM-2 algorithm, with threshold set at 6.0 SD or 5.0 SD. CNVs of at
155 least five or three consecutive probes were retained for analysis, respectively for the 1M and the
156 180K arrays.

157 WES/aCGH comparison. Array CGH enables the detection of genome-wide rearrangements thanks
158 to the measurement of the deviation of the fluorescent signal of the patient as compared to a control
159 DNA. The number of probes depends of the type of chip that is used (here, Agilent 1M or 180K).
160 The threshold to consider a deletion or a duplication was set to the deviation of 5 or 3 consecutive
161 probes respectively. This restricts the detection to CNVs of 8kb or for 20kb Agilent 1M and
162 Agilent180K chips, respectively, on average. On the contrary, as CANOES analysis is based on
163 WES data, it is strictly restricted to CNVs covering exonic sequences, but it can detect CNVs as
164 small as one single exon.

165 In order to combine these approaches to evaluate the sensitivity of our workflow, we filtered out
166 CNVs located in intronic and intergenic regions exclusively from the aCGH data (and on X and Y
167 chromosomes for the samples processed without gonosome CNV calling). Moreover, as CANOES

analysis is based on the calculation of a mean and variance of coverage on a given genomic region, the detection of polymorphic rearrangements is very uncertain. For that reason, we also filtered out all polymorphic CNVs from aCGH data. We defined as polymorphic a CNV that overlaps at least at 70% with CNVs reported in the Gold Standard section of the Database of Genomic Variants with a frequency superior to 1% (15).

Regarding the evaluation of the positive predictive value of our workflow, we restricted our analysis to candidate non-polymorphic CNVs detected from WES data (i) that are theoretically detectable by aCGH as they encompass at least 3 or 5 probes, depending on the chip used and (ii) that do not overlap with segmental duplication regions among >50% of the CANOES target regions.

As most aCGH data were processed using the hg18 genome as reference, we used the liftOver tool from UCSC (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>) to establish the correspondence to hg19. If there were no lift over possibility, we manually checked genes encompassing CNVs.

180

181 **Assessment of CNV calls from whole exome sequencing data: step 2**

Following step 1, we implemented our workflow in our routine procedures. From additional 1056 WES (supplementary table 1), we performed targeted confirmations following the detection of candidate CNVs by CANOES using QMPSF or ddPCR (16). We focused our confirmations on a list of 350 genes that belong to the so-called A β network (17), as all the samples used at this step were sequenced in the context of Alzheimer disease research. This list of genes was built thanks to literature curation on Alzheimer pathophysiology, independently of any genomic information. Candidate CNVs were selected for targeted confirmation if (i) they encompassed genes belonging to this network, and (ii) they were not polymorphic i.e. with a frequency below 1% in our dataset.

Primers used for QMPSF or ddPCR validation are available upon request.

191

192 **CNV calling from NGS data using CANOES**

193 The CANOES software tool implements an algorithm dedicated to the detection of quantitative
194 genomic variations based on DOC information. Basically, CANOES requires DOC data for each
195 target of the capture kit used for each of the sample that are analyzed together. It also integrates the
196 GC content information of each target to reduce the background variability observed in high-
197 throughput sequencing data (18). The read depth was calculated using Bedtools (19), and the GC
198 content was determined using the GATK suite.

199 CANOES builds its statistical reference model from a subset of the samples included in the same
200 analysis (at least 30 samples are recommended). To obtain the best possible fit, CANOES selects
201 the samples that are the most correlated to the currently analyzed sample. This allows the detection
202 of small CNVs, but also reduces the detection susceptibility of recurrent events. CANOES uses a
203 Hidden Markov Model to represent the variability of the DOC distribution built from the selected
204 samples. Then, it uses the Viterbi algorithm to assign deletions, duplications or normal regions.
205 After the calling step, a 'Not Applicable' (NA) score is attributed to all CNVs from samples carrying
206 more than 50 rearrangements. Such samples are usually characterized by higher or lower average
207 read depth and cannot be compared to the reference model. All CNVs assigned with an NA score
208 were thus removed from further analyses. As CANOES used the capture kit definition to detect
209 CNVs, boundaries of events were defined by the start position of the first target and the end position
210 of the last target detected as deviated in comparison with the model.

211

212 **A CANOES-centered workflow**

213 To optimize CANOES performances, we focused on two different approaches, a methodological
214 approach in sample selection and a bioinformatics approach (Figure 2).

215 As previously described, CANOES defines a statistical model for a particular sample from a
216 judicious selection of other samples included in the analysis. The first step of our workflow
217 consisted in the implementation of rules to select the samples that should better be analyzed

218 together. In order to get enough material to build an efficient statistical model and following the
219 CANOES recommendations, we always worked with at least 30 samples. Importantly, we analyzed
220 samples with the less technical variability from each other. Practically, this consists in analyzing
221 samples from the same run, and not to merge multiple runs if not necessary. When merging multiple
222 runs was inevitable (e.g., sequencing of less than 30 samples per run), we combined sequencing
223 runs from the same platform and processed using the same technical conditions, including the same
224 number of samples per lane in order to reduce read depth variability from each sample. Of note,
225 CANOES is not originally set up for the analysis of CNVs on gonosomes, but we implemented
226 modifications in the original script in order to include gonosomes in our analyses. Hence, we ran
227 our workflow after gathering either $n \geq 30$ males or $N \geq 30$ females for the analysis of gene panels 2
228 and 3 that contain X-linked genes and of WES data.

229

230 **Bioinformatics optimization**

231 The first step consisted in the modification of the target definition from the capture kit information.
232 We decided to merge close targets (less than 30 pb) if they covered the same exon. Concerning gene
233 panels that include introns, we decided to split large targets that include both intronic and exonic
234 regions.

235 In order to gain flexibility in our analysis and to be able to add or remove samples easily, we
236 implemented a two-step strategy consisting in (i) performing the read count step for each sample
237 separately, and then (ii) aggregating selected samples before running CANOES. Doing so allowed,
238 for example, intra-familial analyses including patient-parent trio approaches, where cases can be
239 analyzed without taking related samples into account, preventing biasing the statistical model.
240 Finally, we removed non-informative regions from our analyses. We considered a region as non-
241 informative if more than 90% of the samples each had less than 10 reads on the target. Then, we

242 called the CNVs using CANOES, and annotated the results using AnnotSV (20) in order to get
243 additional information about the possible effect and populations frequencies.

244

245 **Nextflow integration**

246 In order to complete our optimization of processing and analysis time, we integrated our
247 bioinformatics pipeline into Nextflow, a data-driven workflow manager (21). This software tool
248 allows a quick deployment of new pipelines on different kind of computational environments, from
249 local computers to a cloud environment. Another interest of Nextflow is to increase the performance
250 by distributing the different steps of the workflow in regards to the computational resources
251 available. The complete workflow, including the specific adaption of CANOES to analyze
252 gonosomes, is available on <https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-workflow>.

253

254 **RESULTS**

255 After building a workflow centered on the CANOES tool, we assessed its performances in the
256 context of (i) gene panel NGS data and (ii) WES data, both generated following capture and
257 Illumina short read sequencing.

258

259 **Gene panel sequencing data**

260 We first evaluated the performances of the CANOES tool using targeted sequencing data of a panel
261 of 11 genes (panel 1, n=465 samples). In parallel, all samples were assessed using custom
262 comprehensive QMPSF assessing the presence or absence of a CNV encompassing any of the 60
263 coding exons of 4 of these genes. We identified 14 CNVs by QMPSF (12 deletions, 2 duplications,
264 size range: [1,556pb – 97Kpb]). All of them were accurately detected by our CANOES-based
265 workflow from NGS data (Table 1). In addition, no additional CNV was called by CANOES,

266 allowing us to obtain a sensitivity and a specificity of 100% (95%CI:[73.24-100]) for those 4 genes.
267 (see supplementary table 2).
268 To further assess the Positive Predictive Value (PPV) of our workflow in the identification of CNVs
269 from gene panels, we applied it to additional NGS data obtained from 3 gene panels (2,222 samples
270 from panel 1, 320 samples from panel 2, and 220 samples from panel 3). We detected 101 candidate
271 CNVs in 98 samples and assessed their presence using either QMPSF or MLPA (Table 2). We
272 validated 87/101 CNVs (86.13%, 95%CI:[77.50-91.94], false positive rate: 13.9%). Overall, the
273 PPV of our workflow applied to gene panel sequencing data was 87.83% (95%CI:[80.01-92.94]).
274 True positive calls of our workflow were 73 deletions (size range: [391pb – 1.06Mpb]) and 16
275 duplications (size range: [360pb – 39.4Kpb]) (see supplementary table 3). False positives were
276 mainly deletions (10/14) and 5 of them were monoexonic.

277

278 **Whole exome sequencing data**

279 We then evaluated the performances of our workflow for the detection of CNVs from WES data.
280 We first applied our workflow to the data obtained from 147 samples with both WES (average
281 depth of coverage = 110x) and aCGH data available (50 samples assessed with the Agilent 1M chip
282 and 97 samples with the Agilent 180k chip). Overall, 10 samples were removed due to a high or low
283 number of rearrangements detected by aCGH or exome, mostly due to low DNA quality or low
284 coverage in WES.
285 From aCGH data, we detected 1,873 CNVs over the 137 samples remaining, of which 102 were
286 non-polymorphic exonic CNVs. Our workflow accurately detected 89 (87.2%) of them (Table 1,
287 supplementary table 4). Among the CNVs that were missed by our workflow, 7 were large (from 14
288 to 80kb) CNVs that encompassed only one (n=5) or two (n=2) targets defined by the capture kit
289 (see figure 3).

290 In order to determine the PPV of our workflow from WES data, we selected 223 CNVs called by
291 our workflow and (i) theoretically detectable by aCGH as encompassing at least 3 (180 k chips) or 5
292 (1M chips) probes and (ii) which did not overlap with segmental duplication regions for more than
293 50% of the CANOES targets. Of them, 190 (85.2%) CNVs were confirmed as true positives
294 following aCGH data assessment (Table 1, supplementary table 5).

295 Of note, an additional set of 519 candidate CNVs were detected by our CANOES-based workflow
296 that overlapped less than 50% of segmental duplication regions but encompassed less than 3 (180 k
297 chips) or 5 aCGH probes (1M chips). Hence, they were not reported by the CGH analysis tool and
298 would then have been overlooked following classical aCGH data analysis. We did not perform
299 targeted confirmation of all these candidate CNVs. Instead, with the aim to further assess the PPV
300 of our workflow regarding exonic non-polymorphic CNVs of any size, we applied it to 1,056
301 additional WES performed in the context of Alzheimer disease research (with no corresponding
302 aCGH data). We selected non-polymorphic CNVs targeting 355 genes belonging to the A β network
303 involved in the pathophysiology of Alzheimer disease (17), whatever their size. We validated
304 108/122 candidate CNVs (88.5%, false positive rate: 11.5%) by QMPFSF (22) or ddPCR (Table 2,
305 supplementary table 6). True positive calls of our workflow were 39 deletions (size range: [165pb –
306 24,2Mpb]) and 69 duplications (size range [166pb – 5,9Mpb]). Interestingly, among the 122
307 candidate CNVs obtained from our workflow, 75 were considered to be theoretically detectable by
308 aCGH 1M, and 47 were considered as not detectable by aCGH 1M. Among the ones theoretically
309 detectable by aCGH, 71 were true positives (94.6%). Among the theoretically not detectable ones,
310 37 were true positives (78.7%).

311 Overall, the PPV of our CANOES-based workflow was 86.3% from WES data after taking into
312 account results from step 1 and step 2 altogether.

313

314 **DISCUSSION**

315 Multiple tools have been developed to detect CNVs from NGS data. As long as such tools are being
316 implemented in diagnostic laboratories, there is a critical need to evaluate their performances.
317 Previous studies showed a large diversity of performances, while a number was performed using
318 simulated datasets (23). After having defined a CANOES-centered workflow, we applied it to three
319 different gene panels and WES data. Overall, we reached very high detection performances
320 following the comparison with independent techniques.

321 From gene panel data, we obtained a 100% sensitivity among a set of 4 genes, the copy number of
322 all coding exons of which having been assessed prior to NGS in 465 samples. In addition, we
323 obtained a 90.3% PPV among all genes with a CANOES call. Such high performances have
324 previously been reported for other tools applied to small NGS panels (24). Among 14 false
325 positives, we observed recurrent events, which can be easily reported as so and be ignored in further
326 analyses. We also observed false positive CNVs in regions homologous to pseudogenes. In that
327 case, it is possible to reduce false positive calls by improving the design of the capture to reduce the
328 chance that probes target the homologous regions, or by optimizing the alignment.

329 Of note, for all genes of Panel 1 and two genes of Panel 2, introns were captured in addition to
330 exons. This might have increased the chances to detect CNVs that can be considered as small from
331 an exon-only point of view but that can actually be much larger at the genomic level. An advantage
332 of capturing introns might indeed be a gain in statistical power for the normalization process:
333 increasing the number of targets may increase the robustness of the model. Among 101 CNVs
334 detected from NGS data from all 3 panels, 75 CNVs encompassed one of these genes with intronic-
335 plus-exonic capture. Interestingly, only 18 of these 75 CNVs encompassed a single coding exon.
336 Such a frequency of monoexonic CNVs is not unexpected regarding mutation screens in MMR
337 genes (monoexonic deletions accounting for 26.92 to 46.27% of all pathogenic deletions (25–27), or
338 other rare diseases (28–31), for example. We hypothesize that all other CNVs, encompassing
339 multiple targets, would probably have been easily detected, had the introns been excluded from the

capture design. Further analyses may be required to better assess the performances of our workflow from single exon CNVs and the effect of including introns or not in the capture design. The observed higher rate of false positives in CNV calls encompassing genes without introns captured (22.22%) may also require further assessments,

We used here a *precision workflow* approach, focusing on the optimization of one tool based on DOC. Interestingly, as some of our genes included non-coding sequences in gene panels, these specific exonic-plus-intronic captures could provide us the possibility to apply complementary tools using different approaches, like the ones developed for WGS. This can indeed increase both detection performances of CNVs and the spectrum of structural variants that can be detectable in these data.

Of note, all our panels included multiple genes. We do not expect that a design including a single gene, even with its intronic sequences, would reach the sufficient number of targets for CANOES to build a robust model.

We also applied our workflow to multiple WES datasets and reached an overall PPV of 86.38 % (95%CI:[82.19 – 89.72]). As for gene panel CNV detection, a confirmation by an independent technique is hence still required following the detection of a candidate CNV from WES data, although this high value allows a limited number of molecular confirmations. One of the major features usually required to apply a new technique in a diagnostic workflow is a high sensitivity as compared to a reference technique. Here, we reached a sensitivity of 87.25% (95%CI:[78.84 – 82.77]). Although the sensitivity was not 100%, it is important to notice that aCGH is considered as reference here although the spectrum of events that can be detected is still limited. When comparing our results to aCGH data, it appeared that we missed fewer events than the potential number of true positive CNVs that were missed by aCGH itself. Indeed, from aCGH data, we missed 13 CNVs, but our analyses called 519 candidate CNVs from corresponding WES data and which were theoretically undetectable by aCGH (i.e. either small CNVs or in regions with no aCGH probes

365 coverage). Our PPVs suggest that the vast majority are eventually true. There is no reason to think
366 that some of the CNVs detected by CANOES only might not be as or more deleterious than CNVs
367 detected by both techniques or exclusively by aCGH. Knowing that aCGH misses many CNVs,
368 even using the high-sensitivity chips such as the Agilent 1M one, and even if other chip designs
369 might increase aCGH performances on coding regions, switching to a WES-only approach for CNV
370 detection in a diagnostic setting should not reduce the overall diagnostic yield while allowing a
371 significant drop of costs.

372 As compared to aCGH, CANOES allowed the identification of CNVs of any size in regions not
373 covered by probes but also for small CNVs including few exons. In addition, it is important to
374 notice that the majority of CANOES false negatives were also CNVs with only few exons, which
375 implies few targets for CANOES although non-coding probes may help detect some of them by
376 aCGH. This decreased rate of detection of CNVs encompassing few targets has already been shown
377 in other datasets (32,33) and appears as a limitation inherent to DOC comparison methods.

378 Of note, it is possible to increase the detection of small events or events in complex regions by
379 using the “GenotypeCNV” function of CANOES. The aim of this function is to look precisely at
380 specific regions and call the genotype of the sample for these specific regions, however it is
381 associated with an increase in false positive calls (29), as well as an increase in time and
382 computational resources needed. In particular cases, when known core genes have already been
383 identified in a given disorder, it is possible to combine our approach to call CNVs at the exome
384 level and focus on specific genes using the GenotypeCNV function applied to every exon of these
385 genes to increase the detection performances in core genes at the same time.

386 Of note, beyond the above-mentioned limitations of CNV detection tools from NGS data, somatic
387 CNVs remain a challenge, both for array-based technologies and for NGS-based tools (34). Among
388 the CNVs detected by our workflow, at least one was considered as likely somatic, as suggested by
389 QMPSF data. However, the sensitivity of DOC tools might remain low in this context (34).

390 In conclusion, we performed an evaluation of the performances of a CNV detection workflow based
391 on read depth comparison from capture-prepared NGS data, one of the most popular methods for
392 NGS in research and diagnostic settings. We highlight very high sensitivity and positive predictive
393 value, for both NGS gene panel and whole exome sequencing. Although the sensitivity was not
394 perfect for WES data as compared to aCGH, a number of additional true calls were not detected by
395 the so-called reference technique. This highlights the absence of a genuine gold standard up to now.
396 Overall, we consider that switching to a NGS-only approach is cost-effective as it allows a
397 reduction in overall costs together with likely stable diagnostic yields.

398

399 **ACKNOWLEDGEMENTS**

400 This study received fundings from Clinical Research Hospital Program from the French Ministry of
401 Health (GMAJ, PHRC 2008/067), the JPND PERADES and France Génomique. This study was co-
402 supported by the Centre National de Référence Malades Alzheimer Jeunes (CNR-MAJ), European
403 Union and Région Normandie. Europe gets involved in Normandie with the European Regional
404 Development Fund (ERDF).

405

406 **CONFLICTS OF INTEREST**

407 None

408

409 REFERENCES

- 410 1. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, et al. De novo rates and
411 selection of large copy number variation. *Genome Res.* 2010 Nov;20(11):1469–81.
- 412 2. Huguet G, Schramm C, Douard E, Jiang L, Labbe A, Tihy F, et al. Measuring and Estimating
413 the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based
414 Samples. *JAMA Psychiatry.* 2018 01;75(5):447–57.
- 415 3. Hehir-Kwa JY, Pfundt R, Veltman JA. Exome sequencing and whole genome sequencing for
416 the detection of copy number variation. *Expert Rev Mol Diagn.* 2015;15(8):1023–32.
- 417 4. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC:
418 a tool for assessing copy number and allelic content using next-generation sequencing data.
419 *Bioinforma Oxf Engl.* 2012 Feb 1;28(3):423–5.
- 420 5. Krumm N, Sudmant PH, Ko A, O’Roak BJ, Malig M, Coe BP, et al. Copy number variation
421 detection and genotyping from exome sequence data. *Genome Res.* 2012 Aug;22(8):1525–32.
- 422 6. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera AV, et al. An open resource of
423 structural variation for medical and population genetics [Internet]. *Genomics*; 2019 Mar [cited
424 2019 Oct 9]. Available from: <http://biorxiv.org/lookup/doi/10.1101/578674>
- 425 7. Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, et al. CANOES:
426 detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.*
427 2014 Jul;42(12):e97.
- 428 8. Charbonnier F, Raux G, Wang Q, Drouot N, Cordier F, Limacher JM, et al. Detection of exon
429 deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal
430 cancer families using multiplex polymerase chain reaction of short fluorescent fragments.
431 *Cancer Res.* 2000 Jun 1;60(11):2760–3.
- 432 9. Baert-Desurmont S, Coutant S, Charbonnier F, Macquere P, Lecoquierre F, Schwartz M, et al.
433 Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic
434 and intronic sequences of panel genes. *Eur J Hum Genet EJHG.* 2018;26(11):1597–602.
- 435 10. Le Guennec K, Nicolas G, Quenez O, Charbonnier C, Wallon D, Bellenguez C, et al. ABCA7
436 rare variants and Alzheimer disease risk. *Neurology.* 2016 Jun 7;86(23):2134–7.
- 437 11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for
438 variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.*
439 2011 May;43(5):491–8.
- 440 12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
441 *Bioinforma Oxf Engl.* 2009 Jul 15;25(14):1754–60.
- 442 13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
443 Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing
444 data. *Genome Res.* 2010 Sep 1;20(9):1297–303.

- 445 14. Rovelet-Lecrux A, Deramecourt V, Legallic S, Maurage C-A, Le Ber I, Brice A, et al. Deletion
446 of the progranulin gene in patients with frontotemporal lobar degeneration or Parkinson
447 disease. *Neurobiol Dis.* 2008 Jul;31(1):41–5.
- 448 15. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic
449 Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*
450 2014 Jan;42(Database issue):D986-992.
- 451 16. Cassinari K, Quenez O, Joly-Hélas G, Beaussire L, Le Meur N, Castelain M, et al. A Simple,
452 Universal, and Cost-Efficient Digital PCR Method for the Targeted Analysis of Copy Number
453 Variations. *Clin Chem.* 2019 Sep;65(9):1153–60.
- 454 17. Champion D, Pottier C, Nicolas G, Le Guennec K, Rovelet-Lecrux A. Alzheimer disease:
455 modeling an A β -centered biological network. *Mol Psychiatry.* 2016;21(7):861–71.
- 456 18. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput
457 sequencing. *Nucleic Acids Res.* 2012 May;40(10):e72.
- 458 19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
459 *Bioinforma Oxf Engl.* 2010 Mar 15;26(6):841–2.
- 460 20. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated
461 tool for structural variations annotation. Berger B, editor. *Bioinformatics* [Internet]. 2018 Apr
462 14 [cited 2018 Oct 2]; Available from: [https://academic.oup.com/bioinformatics/advance-](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty304/4970516)
463 [article/doi/10.1093/bioinformatics/bty304/4970516](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty304/4970516)
- 464 21. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables
465 reproducible computational workflows. *Nat Biotechnol.* 2017 Apr;35(4):316–9.
- 466 22. Le Guennec K, Quenez O, Nicolas G, Wallon D, Rousseau S, Richard A-C, et al. 17q21.31
467 duplication causes prominent tau-related dementia with increased MAPT expression. *Mol*
468 *Psychiatry.* 2017 Aug;22(8):1119–25.
- 469 23. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access
470 copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res.*
471 2019 Mar;779:114–25.
- 472 24. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical
473 detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome*
474 *Open Res.* 2016 Nov 25;1:20.
- 475 25. Di Fiore F, Charbonnier F, Martin C, Frerot S, Olschwang S, Wang Q, et al. Screening for
476 genomic rearrangements of the MMR genes must be included in the routine diagnosis of
477 HNPCC. *J Med Genet.* 2004 Jan;41(1):18–20.
- 478 26. Taylor CF, Charlton RS, Burn J, Sheridan E, Taylor GR. Genomic deletions in MSH2 or
479 MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer: identification of
480 novel and recurrent deletions by MLPA. *Hum Mutat.* 2003 Dec;22(6):428–33.
- 481 27. van der Klift H, Wijnen J, Wagner A, Verkuilen P, Tops C, Otway R, et al. Molecular
482 characterization of the spectrum of genomic deletions in the mismatch repair genes MSH2,

- 483 MLH1, MSH6, and PMS2 responsible for hereditary nonpolyposis colorectal cancer
484 (HNPCC). *Genes Chromosomes Cancer*. 2005 Oct;44(2):123–38.
- 485 28. Baker M, Strongosky AJ, Sanchez-Contreras MY, Yang S, Ferguson W, Calne DB, et al.
486 SLC20A2 and THAP1 deletion in familial basal ganglia calcification with dystonia.
487 *Neurogenetics*. 2014 Mar;15(1):23–30.
- 488 29. David S, Ferreira J, Quenez O, Rovelet-Lecrux A, Richard A-C, V  rin M, et al. Identification
489 of partial SLC20A2 deletions in primary brain calcification using whole-exome sequencing.
490 *Eur J Hum Genet EJHG*. 2016;24(11):1630–4.
- 491 30. Guo X-X, Su H-Z, Zou X-H, Lai L-L, Lu Y-Q, Wang C, et al. Identification of SLC20A2
492 deletions in patients with primary familial brain calcification. *Clin Genet*. 2019 Jul;96(1):53–
493 60.
- 494 31. Nicolas G, Rovelet-Lecrux A, Pottier C, Martinaud O, Wallon D, Vernier L, et al. PDGFB
495 partial deletion: a new, rare mechanism causing brain calcification with leukoencephalopathy.
496 *J Mol Neurosci MN*. 2014 Jun;53(2):171–5.
- 497 32. Miyatake S, Koshimizu E, Fujita A, Fukai R, Imagawa E, Ohba C, et al. Detecting copy-
498 number variations in whole-exome sequencing data using the eXome Hidden Markov Model:
499 an ‘exome-first’ approach. *J Hum Genet*. 2015 Apr;60(4):175–82.
- 500 33. Samarakoon PS, Sorte HS, Kristiansen BE, Skodje T, Sheng Y, Tj  nnfjord GE, et al.
501 Identification of copy number variants from exome sequence data. *BMC Genomics*. 2014 Aug
502 7;15:661.
- 503 34. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation
504 detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*. 2017
505 May 31;18(1):286.

506
507

508

509 **FIGURE LEGENDS**

510 **Figure 1. Principles of Depth Of Coverage (DOC) comparison.** Schematic distribution of reads
511 among three different samples over 5 sequenced exons. **(A)** absence of any CNV. **(B)** Duplication of
512 two exons (2 and 3). **(C)** Deletion of exon 4. In order to call those CNVs, software tools have to
513 establish a reference. Some tools compare paired data from the same patient, *e.g.* tumor tissue
514 against germline, while others build their reference from a pool of samples and then compare a
515 given sample to this reference, as the CANOES tool used in our workflow.

516

517 **Figure 2. CANOES-centered workflow.** File (square) with their format in parenthesis, and
518 process (rounded) constituting the workflow. From the original capture kit definition, we merge
519 closed target from the same exon, then do in parallel the DOC and the GC content estimation. We
520 regroup DOC individual files depending on the project, sequencing batch, unrelated samples, and
521 remove non-informative regions. The last steps consist in CNV calling using CANOES and
522 annotation with annotSV.

523

524 **Figure 3. Example of a CNV detected by aCGH but missed by the CANOES-centered**
525 **workflow.**

526 A CNV (highlight region) detected by a-CGH encompassing multiple CGH probes (1M probes
527 array, in gray) but only one target from the SureSelect V5 capture kit. Of note, this deletion would
528 have been missed by using a 180k probes array CGH (in black).

529

530 **Figure 4. Example of CNVs detected by the CANOES-centered workflow from WES data but**
531 **missed by aCGH.**

532 A. The highlighted region represents the CNV called by the CANOES-centered workflow,
533 encompassing one exon of *RHCE*.

534 B. View of the same region from DNA-Analytics (aCGH data 1M) in the same patient. This deletion
535 was not called following aCGH data analysis as the number of deviated probes did not reach the
536 threshold for calling. However, as 3 probes (in white) were deviated, this allows the confirmation of
537 the deletion of the region.

538

539

540 **Appendix. Collaborators** 541 **The FREX Consortium**

542

543 **Principal Investigators:**

544 Emmanuelle Génin (chair), Inserm UMR1078, CHRU, Univ Brest, Brest, France
545 Dominique Campion, Inserm UMR1079, Faculté de Médecine, Rouen, France
546 Jean-François Dartigues, Inserm UMR1219, Univ Bordeaux, France
547 Jean-François Deleuze, Centre National de Génotypage, CEA, Fondation
548 Jean Dausset-CEPH, Evry, France
549 Jean-Charles Lambert, Inserm UMR1167, Institut Pasteur, Lille, France
550 Richard Redon, Inserm UMR 1087 / CNRS UMR 6291, l'institut du thorax, Nantes, France

551

552 **Collaborators:**

553 **Bioinformatics group:**

554 Thomas Ludwig (chair), Inserm UMR1078, CHRU, Univ Brest, Brest
555 Benjamin Grenier-Boley, Inserm UMR1167, Institut Pasteur, Lille
556 Sébastien Letort, Inserm UMR1078, CHRU, Univ Brest, Brest
557 Pierre Lindenbaum, Inserm UMR 1087 / CNRS UMR 6291, l'institut du thorax, Nantes
558 Vincent Meyer, Centre National de Génotypage, CEA, Evry
559 Olivier Quenez, Inserm UMR1079, Faculté de Médecine, Rouen

560

561 **Statistical genetics group:**

562 Christian Dina (chair), Inserm UMR 1087/CNRS UMR 6291, l'institut du thorax, Nantes
563 Céline Bellenguez, Inserm UMR1167, Institut Pasteur, Lille²³
564 Camille Charbonnier-Le Clézio, Inserm UMR1079, Faculté de Médecine, Rouen
565 Joanna Giemza, Inserm UMR 1087 / CNRS UMR 6291, l'institut du thorax, Nantes

566

567 **Data collection:**

568 Stéphanie Chatel, Inserm UMR 1087 / CNRS UMR 6291, l'institut du thorax, Nantes
569 Claude Férec, Inserm UMR1078, CHRU, Univ Brest
570 Hervé Le Marec, Inserm UMR 1087 / CNRS UMR 6291, l'institut du thorax, Nantes
571 Luc Letenneur, Inserm UMR1219, Univ Bordeaux
572 Gaël Nicolas, Inserm UMR1079, Faculté de Médecine, Rouen
573 Karen Rouault, Inserm UMR1078, CHRU, Univ Brest

574

575 **Sequencing:**

576 Delphine Bacq, Centre National de Génotypage, CEA, Evry
577 Anne Boland, Centre National de Génotypage, CEA, Evry

578 Doris Lechner, Centre National de Génomique, CEA, Evry
579
580
581