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.

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

Olivier Quenez[1], Kevin Cassinari[1], Sophie Coutant[2], François Lecoquierre[2], Kilan Le Guennec[1], Stéphane Rousseau[1], Anne-Claire Richard[1], Stéphanie Vasseur[2], Emilie Bouvignies[2], Jacqueline Bou[2], Gwendoline Lienard[2], Sandrine Manase[2], Steeve Fourneaux[2], Nathalie Drouot[2], Virginie Nguyen-Viet[2], Myriam Vezain[2], Pascal Chambon[2], Géraldine Joly-Helas[2], Nathalie Le Meur[2], Mathieu Castelain[2], Anne Boland[3], Jean-François Deleuze[3], FREX Consortium, Isabelle Tournier[2], Françoise Charbonnier[2], Edwige Kasper[2], Gaëlle Bougeard-Denoyelle[2], Thierry Frebourg[2], Pascale Saugier-Veber[2], Stéphanie Baert-Desurmont[2], Dominique Campion[1,4], Anne Rovelet-Lecrux[1], Gaël Nicolas[1]

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

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

[3]Centre National de Recherche en Génomique Humaine, Institut de Génomique, CEA, Evry, France.

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

ABSTRACT

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

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

From GP data of 3776 samples, we reached an overall Positive Predictive Value (PPV) of 87.8%. This dataset included a complete comprehensive QMPSF comparison of 4 genes (60 exons) on which we obtained 100% sensitivity and specificity.

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

In addition, our CANOES-centered workflow on WES data allowed the detection of CNVs of any size that were missed by aCGH. Overall, switching to a NGS-only approach should be cost-effective as it allows a reduction in overall costs together with likely stable diagnostic yields.

KEYWORDS

Exome, panel, read-depth information, CANOES, CNV detection, bioinformatics, sensitivity

INTRODUCTION

Copy-number variations (CNVs) are a major cause of Mendelian disorders (Itsara et al., 2010) as well as risk factors for common diseases (Huguet et al., 2018). With the advent of next-generation sequencing (NGS), a number of software tools have been developed to detect CNVs. Whole genome sequencing (WGS) is often presented as an almost universal technique allowing the assessment of almost any type of variation, including CNVs and other structural variations. WGS may eventually be used as a first-tier diagnostics tool in the context of genetically highly heterogeneous disorders. However, the detection of structural variations from data generated using the technology of short read sequencing is still associated with a number of false positives. Such events can be detected using a plethora of bioinformatics tools based on different principles, including Depth Of Coverage (DOC) information, relative position of paired reads, split reads and DeNovo Assembly (Hehir-Kwa, Pfundt, & Veltman, 2015). Besides the development of WGS, targeted sequencing of gene panels and whole exome sequencing (WES) remain of primary use in many diagnostics and research laboratories. They are indeed still considered as more affordable and of easier access as they can be processed using usual informatics facilities accessible to most laboratories. Moreover, the input of WGS is questioning in disorders with low genetic heterogeneity and high phenotypic specificity. Hence, gene panels and WES remain largely used .

The detection of CNVs from exonic capture-based targeted sequencing solutions primarily relies on DOC information (Boeva et al., 2012; Krumm et al., 2012). Tools based on DOC information compare one sample to a reference, and predict deletions or duplications depending on the increase or decrease of the DOC as compared to the reference (figure 1). As each tool was set up and trained on a specific dataset, one of the main challenges is to evaluate the specificity and sensitivity of a given software tool on large datasets. Studies evaluating the diagnostic performances of CNV detection pipelines are scarce although they appear to be critical for their use in routine procedures.

In order to optimize CNV detection from NGS data, a classical approach consists in running multiple tools in parallel and then aggregate the results to keep a CNV as candidate only if multiple

tools called it (Collins et al., 2019). As it is more effective to do so with tools using different types of bioinformatics methods (DOC, split reads, etc.), this combinatory approach is most adapted when working on WGS, or at least if most of the intergenic or intronic regions – where breakends are more frequently found – are captured. Here, we decided to focus on one tool using the DOC approach as it still remains the most adapted one for exonic capture. In a *precision workflow* approach, we developed a workflow based on the already existing software tool CANOES (Backenroth et al., 2014). Briefly, CANOES adopts a pooling strategy to build its reference model, and uses a Hidden Markov Model to represent the DOC of this model. Lastly, it confronts the samples to the reference in order to call candidate deletions or duplications.

We performed a diagnostic performance evaluation of this workflow regarding gene panel and WES data, in two steps. First, we compared CNV calls with a gold standard, namely a comprehensive assessment by Quantitative Multiplex PCR of Short Fluorescent fragments (QMPSF) (Charbonnier et al., 2000) or array comparative genomic hybridization (aCGH), regarding targeted gene panel and WES data, respectively. Second, we implemented our workflow in our routine procedures and performed an additional evaluation of the positive predictive value of our CANOES-centered workflow using targeted confirmation of CNVs using an independent targeted technique.

MATERIAL AND METHODS

Gene panel sequencing

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

Panel 1 was set up to focus on genes involved in predisposition to colorectal cancer and digestive polyposis or Li-Fraumeni syndrome (Baert-Desurmont et al., 2018). This panel was implemented in two successive versions. V1 was used to sequence 11 genes in 2771 samples. V2 was used to sequence 15 genes (same 11 genes plus 4) in 549 samples. In both versions and for all genes, exons and introns outside repeated sequences were captured.

Panel 2 also has two successive versions and was designed to focus on two clinical indications: (i) hydrocephaly (3 genes) and (ii) Cornelia de Lange syndrome and differential diagnoses (24 genes in v1, 30 in v2). In total, 320 samples were sequenced using this panel (240 with v1, 80 with v2). For this panel, introns outside repeated sequences were captured only for two genes, namely *L1CAM* and *NIPBL*.

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

Assessment of CNV calls from gene panel data: step 1

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

Assessment of CNV calls from gene panel data: step 2

Following step 1, we implemented our CANOES-centered workflow in our routine diagnostics procedures on NGS data from all three panels (n=3311 additional samples in total). We performed confirmations of candidate CNVs using QMPSF or Multiplex Ligation-dependent Probe Amplification (MLPA) only in samples with a CANOES call. Primers used for QMPSF screening and validation are available upon request.

Whole-exome sequencing

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

Whole exomes were sequenced in the context of diverse research and diagnostics purposes (supplementary table 1). Exomes were captured using Agilent SureSelect Human All Exon kits (V1, V2 V4+UTR, V5, V5+UTR and V6). Final libraries were sequenced on an Illumina Genome Analyser GAIIX (corresponding to exomes captured with the V1, V2 or V4UTR kit, n=10), or on an Illumina HiSeq2000, 2500 or 4000 with paired ends, 76 or 100bp reads. Exome sequencing was performed in 3 sequencing centers: Integragen (Evry, France) (n=6), the French National Center of Human Genomics Research (CNRGH, Evry, France) (n=1065) and the Genome Quebec Innovation Center (Montreal, Canada) (n=128) (Kilan Le Guennec et al., 2016). Exomes were all processed through the same bioinformatics pipeline following the Broad Institute Best Practices recommendations (DePristo et al., 2011). Reads were mapped to the 1000 Genomes GRCh37 build using BWA 0.7.5a.(Li & Durbin, 2009). Picard Tools 1.101 (http://broadinstitute.github.io/picard/) was used to flag duplicate reads. We applied GATK (McKenna et al., 2010) for short insertion and deletions (indel) realignment and base quality score recalibration. All quality checks were processed as previously described (Kilan Le Guennec et al., 2016).

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

For the gold standard comparison, we analyzed data from 147 unrelated individuals with both WES and aCGH data available.

Array CGH Analysis. Oligonucleotide aCGH was performed as previously described (Rovelet-Lecrux et al., 2008). Briefly, high-resolution aCGH analysis was performed using the 1x1M Human High-Resolution Discovery Microarray Kit or the 4x180K SurePrint G3 Human CGH Microarray kit (Agilent Technologies, Santa Clara, California, USA), using standard recommended protocols. An in-house and sex-matched genomic DNA pool of at least 10 control individuals was used as reference sample. Hybridization results were analyzed with the Agilent's DNA-Analytics software (version 4.0.81, Agilent Technologies) or the Agilent Genomic Workbench (version 7.0, Agilent Technologies). Data were processed using the ADM-2 algorithm, with threshold set at 6.0 SD or 5.0 SD. CNVs of at least five or three consecutive probes were retained for analysis, respectively for the 1M and the 180K arrays.

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

In order to combine these approaches to evaluate the sensitivity of our workflow, we filtered out CNVs located in intronic and intergenic regions exclusively from the aCGH data (and on X and Y 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% (MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014).

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.

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

Following step 1, we implemented our workflow in our routine procedures. Form additional 1056 WES (supplementary table 1), we performed targeted confirmations following the detection of candidate CNVs by CANOES using QMPSF or ddPCR (Cassinari et al., 2019). We focused our confirmations on a list of 350 genes that belong to the so-called Aβ network (Campion, Pottier, Nicolas, Le Guennec, & Rovelet-Lecrux, 2016), 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.

CNV calling from NGS data using CANOES

The CANOES software tool implements an algorithm dedicated to the detection of quantitative genomic variations based on DOC information. Basically, CANOES requires DOC data for each target of the capture kit used for each of the sample that are analyzed together. It also integrates the

GC content information of each target to reduce the background variability observed in high-throughput sequencing data (Benjamini & Speed, 2012). The read depth was calculated using Bedtools (Quinlan & Hall, 2010), and the GC content was determined using the GATK suite.

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

A CANOES-centered workflow

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

As previously described, CANOES defines a statistical model for a particular sample from a judicious selection of other samples included in the analysis. The first step of our workflow consisted in the implementation of rules to select the samples that should better be analyzed together. In order to get enough material to build an efficient statistical model and following the CANOES recommendations, we always worked with at least 30 samples. Importantly, we analyzed samples with the less technical variability from each other. Practically, this consists in analyzing samples from the same run, and not to merge multiple runs if not necessary. When merging multiple

runs was inevitable (e.g., sequencing of less than 30 samples per run), we combined sequencing runs from the same platform and processed using the same technical conditions, including the same number of samples per lane in order to reduce read depth variability from each sample. Of note, CANOES is not originally set up for the analysis of CNVs on gonosomes, but we implemented modifications in the original script in order to include gonosomes in our analyses. Hence, we ran our workflow after gathering either n≥30 males or N≥30 females for the analysis of gene panels 2 and 3 that contain X-linked genes and of WES data.

Bioinformatics optimization

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

In order to gain flexibility in our analysis and to be able to add or remove samples easily, we implemented a two-step strategy consisting in (i) performing the read count step for each sample separately, and then (ii) aggregating selected samples before running CANOES. Doing so allowed, for example, intra-familial analyses including patient-parent trio approaches, where cases can be analyzed without taking related samples into account, preventing biasing the statistical model. Finally, we removed non-informative regions from our analyses. We considered a region as non-informative if more than 90% of the samples each had less than 10 reads on the target. Then, we called the CNVs using CANOES, and annotated the results using AnnotSV (Geoffroy et al., 2018) in order to get additional information about the possible effect and populations frequencies.

Nextflow integration

In order to complete our optimization of processing and analysis time, we integrated our bioinformatics pipeline into Nextflow, a data-driven workflow manager (Di Tommaso et al., 2017).

This software tool allows a quick deployment of new pipelines on different kind of computational environments, from local computers to a cloud environment. Another interest of Nextflow is to increase the performance by distributing the different steps of the workflow in regards to the computational resources available. The complete workflow, including the specific adaption of CANOES to analyze gonosomes, is available on https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-workflow.

RESULTS

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

Gene panel sequencing data

We first evaluated the performances of the CANOES tool using targeted sequencing data of a panel of 11 genes (panel 1, n=465 samples). In parallel, all samples were assessed using custom comprehensive QMPSF assessing the presence or absence of a CNV encompassing any of the 60 coding exons of 4 of these genes. We identified 14 CNVs by QMPSF (12 deletions, 2 duplications, size range: [1.556pb – 97Kpb]). All of them were accurately detected by our CANOES-based workflow from NGS data (Table 1). In addition, no additional CNV was called by CANOES, allowing us to obtain a sensitivity and a specificity of 100% (95%CI:[73.24-100]) for those 4 genes. (see supplementary table 2).

To further assess the Positive Predictive Value (PPV) of our workflow in the identification of CNVs from gene panels, we applied it to additional NGS data obtained from 3 gene panels (2222 samples from panel 1, 320 samples from panel 2, and 220 samples from panel 3). We detected 101 candidate CNVs in 98 samples and assessed their presence using either QMPSF or MLPA (Table 2). We validated 87/101 CNVs (86.13%, 95%CI:[77.50-91.94], false positive rate: 13.9%). Overall, the PPV of our workflow applied to gene panel sequencing data was 87.83% (95%CI:[80.01-92.94]). True positive calls of our workflow were 73 deletions (size range: [391pb – 1.06Mpb]) and 16 duplications (size range: [360pb – 39.4Kpb]) (see supplementary table 3). False positives were mainly deletions (10/14) and 5 of them were monoexonic.

Whole exome sequencing data

We then evaluated the performances of our workflow for the detection of CNVs from WES data. We first applied our workflow to the data obtained from 147 samples with both WES (average depth of coverage = 110x) and aCGH data available (50 samples assessed with the Agilent 1M chip and 97 samples with the Agilent 180k chip). Overall, 10 samples were removed due to a high or low number of rearrangements detected by aCGH or exome, mostly due to low DNA quality or low coverage in WES.

From aCGH data, we detected 1873 CNVs over the 137 samples remaining, of which 102 were non-polymorphic exonic CNVs. Our workflow accurately detected 89 (87.2%) of them (Table 1, supplementary table 4). Among the CNVs that were missed by our workflow, 7 were large (from 14 to 80kb) CNVs that encompassed only one (n=5) or two (n=2) targets defined by the capture kit (see figure 3).

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

Of note, an additional set of 519 candidate CNVs were detected by our CANOES-based workflow that overlapped less than 50% of segmental duplication regions but encompassed less than 3 (180 k chips) or 5 aCGH probes (1M chips). Hence, they were not reported by the CGH analysis tool and would then have been overlooked following classical aCGH data analysis. We did not perform targeted confirmation of all these candidate CNVs. Instead, with the aim to further assess the PPV of our workflow regarding exonic non-polymorphic CNVs of any size, we applied it to 1,056 additional WES performed in the context of Alzheimer disease research (with no corresponding aCGH data). We selected non-polymorphic CNVs targeting 355 genes belonging to the Aβ network involved in the pathophysiology of Alzheimer disease (Campion et al., 2016), whatever their size.

We validated 108/122 candidate CNVs (88.5%, false positive rate: 11.5%) by QMPSF (K Le Guennec et al., 2017) or ddPCR (Table 2, supplementary table 6). True positive calls of our workflow were 39 deletions (size range: [165pb – 24,2Mpb]) and 69 duplications (size range [166pb – 5,9Mpb]). Interestingly, among the 122 candidate CNVs obtained from our workflow, 75 were considered to be theoretically detectable by aCGH 1M, and 47 were considered as not detectable by aCGH 1M. Among the ones theoretically detectable by aCGH, 71 were true positives (94.6%). Among the theoretically not detectable ones, 37 were true positives (78.7%).

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

DISCUSSION

Multiple tools have been developed to detect CNVs from NGS data. As long as such tools are being implemented in diagnostic laboratories, there is a critical need to evaluate their performances. Previous studies showed a large diversity of performances, while a number was performed using simulated datasets (Roca, González-Castro, Fernández, Couce, & Fernández-Marmiesse, 2019). After having defined a CANOES-centered workflow, we applied it to three different gene panels and WES data. Overall, we reached very high detection performances following the comparison with independent techniques.

From gene panel data, we obtained a 100% sensitivity among a set of 4 genes, the copy number of all coding exons of which having been assessed prior to NGS in 465 samples. In addition, we obtained a 90.3% PPV among all genes with a CANOES call. Such high performances have previously been reported for other tools applied to small NGS panels (Fowler et al., 2016). Among 14 false positives, we observed recurrent events, which can be easily reported as so and be ignored in further analyses. We also observed false positive CNVs in regions homologous to pseudogenes. In that case, it is possible to reduce false positive calls by improving the design of the capture to reduce the chance that probes target the homologous regions, or by optimizing the alignment.

Of note, for all genes of Panel 1 and two genes of Panel 2, introns were captured in addition to exons. This might have increased the chances to detect CNVs that can be considered as small from an exon-only point of view but that can actually be much larger at the genomic level. An advantage of capturing introns might indeed be a gain in statistical power for the normalization process: increasing the number of targets may increase the robustness of the model. Among 101 CNVs detected from NGS data from all 3 panels, 75 CNVs encompassed one of these genes with intronic-plus-exonic capture. Interestingly, only 18 of these 75 CNVs encompassed a single coding exon. Such a frequency of monoexonic CNVs is not unexpected regarding mutation screens in MMR genes (monoexonic deletions accounting for 26.92 to 46.27% of all pathogenic deletions (Di Fiore et al., 2004; Taylor, Charlton, Burn, Sheridan, & Taylor, 2003; van der Klift et al., 2005), or other

rare diseases (Baker et al., 2014; David et al., 2016; Guo et al., 2019; Nicolas et al., 2014), for example. We hypothesize that all other CNVs, encompassing 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 gold standard. 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 gold standard 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 coverage). Our PPVs suggest that the vast majority are eventually true. There is no reason to think that some of the CNVs detected by CANOES only might not be as or more deleterious than CNVs detected by both techniques or exclusively by aCGH. Knowing that aCGH misses many CNVs, even using the high-sensitivity chips such as the Agilent 1M one, and even if other chip designs might increase aCGH performances on coding regions, switching to a WES-only approach for CNV detection in a diagnostic setting should not reduce the overall diagnostic yield while allowing a significant drop of costs.

As compared to aCGH, CANOES allowed the identification of CNVs of any size in regions not covered by probes but also for small CNVs including few exons. In addition, it is important to notice that the majority of CANOES false negatives were also CNVs with only few exons, which implies few targets for CANOES although non-coding probes may help detect some of them by aCGH. This decreased rate of detection of CNVs encompassing few targets has already been shown in other datasets (Miyatake et al., 2015; Samarakoon et al., 2014) and appears as a limitation inherent to DOC comparison methods.

Of note, it is possible to increase the detection of small events or events in complex regions by using the "GenotypeCNV" function of CANOES. The aim of this function is to look precisely at specific regions and call the genotype of the sample for these specific regions, however it is associated with an increase in false positive calls (David et al., 2016), as well as an increase in time and computational resources needed. In particular cases, when known core genes have already been identified in a given disorder, it is possible to combine our approach to call CNVs at the exome level and focus on specific genes using the GenotypeCNV function applied to every exon of these genes to increase the detection performances in core genes at the same time.

Of note, beyond the above-mentioned limitations of CNV detection tools from NGS data, somatic CNVs remain a challenge, both for array-based technologies and for NGS-based tools (Zare, Dow,

Monteleone, Hosny, & Nabavi, 2017). Among the CNVs detected by our workflow, at least one was considered as likely somatic, as suggested by QMPSF data. However, the sensitivity of DOC tools might remain low in this context (Zare et al., 2017).

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

ACKNOWLEDGEMENTS

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

CONFLICTS OF INTEREST

None

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., … Shen, Y. (2014). CANOES : Detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Research*, *42*(12), e97. https://doi.org/10.1093/nar/gku345

Baert-Desurmont, S., Coutant, S., Charbonnier, F., Macquere, P., Lecoquierre, F., Schwartz, M., … Tournier, I. (2018). Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. *European Journal of Human Genetics: EJHG*, *26*(11), 1597-1602. https://doi.org/10.1038/s41431-018-0207-2

Baker, M., Strongosky, A. J., Sanchez-Contreras, M. Y., Yang, S., Ferguson, W., Calne, D. B., … Rademakers, R. (2014). SLC20A2 and THAP1 deletion in familial basal ganglia calcification with dystonia. *Neurogenetics*, *15*(1), 23-30. https://doi.org/10.1007/s10048-013-0378-5

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), e72. https://doi.org/10.1093/nar/gks001

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., … Barillot, E. (2012). Control-FREEC : A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, *28*(3), 423-425. https://doi.org/10.1093/bioinformatics/btr670

Campion, D., Pottier, C., Nicolas, G., Le Guennec, K., & Rovelet-Lecrux, A. (2016). Alzheimer disease : Modeling an Aβ-centered biological network. *Molecular Psychiatry*, *21*(7), 861-871. https://doi.org/10.1038/mp.2016.38

Cassinari, K., Quenez, O., Joly-Hélas, G., Beaussire, L., Le Meur, N., Castelain, M., … Chambon, P. (2019). A Simple, Universal, and Cost-Efficient Digital PCR Method for the Targeted Analysis of Copy Number Variations. *Clinical Chemistry*, *65*(9), 1153-1160. https://doi.org/10.1373/clinchem.2019.304246

Charbonnier, F., Raux, G., Wang, Q., Drouot, N., Cordier, F., Limacher, J. M., … Frebourg, T. (2000). Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. *Cancer Research*, *60*(11), 2760-2763.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Khera, A. V., … Talkowski, M. E. (2019). *An open resource of structural variation for medical and population genetics* [Preprint]. https://doi.org/10.1101/578674

David, S., Ferreira, J., Quenez, O., Rovelet-Lecrux, A., Richard, A.-C., Vérin, M., … Nicolas, G. (2016). Identification of partial SLC20A2 deletions in primary brain calcification using whole-exome sequencing. *European Journal of Human Genetics: EJHG*, *24*(11), 1630-1634. https://doi.org/10.1038/ejhg.2016.50

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Di Fiore, F., Charbonnier, F., Martin, C., Frerot, S., Olschwang, S., Wang, Q., … Frebourg, T. (2004). Screening for genomic rearrangements of the MMR genes must be included in the routine diagnosis of HNPCC. *Journal of Medical Genetics*, *41*(1), 18-20. https://doi.org/10.1136/jmg.2003.012062

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316-319. https://doi.org/10.1038/nbt.3820

Fowler, A., Mahamdallie, S., Ruark, E., Seal, S., Ramsay, E., Clarke, M., … Rahman, N. (2016). Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Research*, *1*, 20. https://doi.org/10.12688/wellcomeopenres.10069.1

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV : An integrated tool for structural variations annotation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty304

Guo, X.-X., Su, H.-Z., Zou, X.-H., Lai, L.-L., Lu, Y.-Q., Wang, C., … Chen, W.-J. (2019). Identification of SLC20A2 deletions in patients with primary familial brain calcification. *Clinical Genetics*, *96*(1), 53-60. https://doi.org/10.1111/cge.13540

Hehir-Kwa, J. Y., Pfundt, R., & Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert Review of Molecular Diagnostics*, *15*(8), 1023-1032. https://doi.org/10.1586/14737159.2015.1053467

Huguet, G., Schramm, C., Douard, E., Jiang, L., Labbe, A., Tihy, F., … IMAGEN Consortium. (2018). Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. *JAMA Psychiatry*, *75*(5), 447-457. https://doi.org/10.1001/jamapsychiatry.2018.0039

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., & Eichler, E. E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, *20*(11), 1469-1481. https://doi.org/10.1101/gr.107680.110

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., … Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, *22*(8), 1525-1532. https://doi.org/10.1101/gr.138115.112

Le Guennec, K, Quenez, O., Nicolas, G., Wallon, D., Rousseau, S., Richard, A.-C., … Rovelet-Lecrux, A. (2017). 17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression. *Molecular Psychiatry*, *22*(8), 1119-1125. https://doi.org/10.1038/mp.2016.226

Le Guennec, Kilan, Nicolas, G., Quenez, O., Charbonnier, C., Wallon, D., Bellenguez, C., … CNR-MAJ collaborators. (2016). ABCA7 rare variants and Alzheimer disease risk. *Neurology*, *86*(23), 2134-2137. https://doi.org/10.1212/WNL.0000000000002627

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants : A curated collection of structural variation in the human genome. *Nucleic Acids Research*, *42*(Database issue), D986-992. https://doi.org/10.1093/nar/gkt958

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Miyatake, S., Koshimizu, E., Fujita, A., Fukai, R., Imagawa, E., Ohba, C., … Matsumoto, N. (2015). Detecting copy-number variations in whole-exome sequencing data using the eXome Hidden Markov Model : An « exome-first » approach. *Journal of Human Genetics*, *60*(4), 175-182. https://doi.org/10.1038/jhg.2014.124

Nicolas, G., Rovelet-Lecrux, A., Pottier, C., Martinaud, O., Wallon, D., Vernier, L., … Hannequin, D. (2014). PDGFB partial deletion : A new, rare mechanism causing brain calcification with leukoencephalopathy. *Journal of Molecular Neuroscience: MN*, *53*(2), 171-175. https://doi.org/10.1007/s12031-014-0265-z

Quinlan, A. R., & Hall, I. M. (2010). BEDTools : A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

Roca, I., González-Castro, L., Fernández, H., Couce, M. L., & Fernández-Marmiesse, A. (2019). Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutation Research*, *779*, 114-125. https://doi.org/10.1016/j.mrrev.2019.02.005

Rovelet-Lecrux, A., Deramecourt, V., Legallic, S., Maurage, C.-A., Le Ber, I., Brice, A., … Campion, D. (2008). Deletion of the progranulin gene in patients with frontotemporal lobar

degeneration or Parkinson disease. *Neurobiology of Disease*, *31*(1), 41-45. https://doi.org/10.1016/j.nbd.2008.03.004

Samarakoon, P. S., Sorte, H. S., Kristiansen, B. E., Skodje, T., Sheng, Y., Tjønnfjord, G. E., … Lyle, R. (2014). Identification of copy number variants from exome sequence data. *BMC Genomics*, *15*, 661. https://doi.org/10.1186/1471-2164-15-661

Taylor, C. F., Charlton, R. S., Burn, J., Sheridan, E., & Taylor, G. R. (2003). Genomic deletions in MSH2 or MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer : Identification of novel and recurrent deletions by MLPA. *Human Mutation*, *22*(6), 428-433. https://doi.org/10.1002/humu.10291

van der Klift, H., Wijnen, J., Wagner, A., Verkuilen, P., Tops, C., Otway, R., … Fodde, R. (2005). Molecular characterization of the spectrum of genomic deletions in the mismatch repair genes MSH2, MLH1, MSH6, and PMS2 responsible for hereditary nonpolyposis colorectal cancer (HNPCC). *Genes, Chromosomes & Cancer*, *44*(2), 123-138. https://doi.org/10.1002/gcc.20219

Zare, F., Dow, M., Monteleone, N., Hosny, A., & Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, *18*(1), 286. https://doi.org/10.1186/s12859-017-1705-x

FIGURE LEGENDS

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

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

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

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

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

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

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

Appendix. Collaborators

The FREX Consortium

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

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

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

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

Sequencing:
Delphine Bacq, Centre National de Génotypage, CEA, Evry
Anne Boland, Centre National de Génotypage, CEA, Evry
Doris Lechner, Centre National de Génotypage, CEA, Evry