






METHOD ARTICLE

An enhanced method for targeted next generation sequencing copy number variant detection using ExomeDepth [version 1; peer review: 1 approved, 1 approved with reservations]

Andrew Parrish¹, Richard Caswell ², Garan Jones¹, Christopher M. Watson ^{3,4},
Laura A. Crinnion^{3,4}, Sian Ellard ^{1,2}

¹Exeter Molecular Genetics Service, Royal Devon & Exeter NHS Foundation Trust, Exeter, EX2 5DW, UK

²Institute of Biomedical and Clinical Science, University of Exeter, Exeter, EX2 5DW, UK

³Yorkshire Regional Genetics Service, St James's University Hospital, Leeds, LS9 7TF, UK

⁴School of Medicine, University of Leeds, Leeds, LS9 7TF, UK

v1 First published: 14 Jul 2017, 2:49 (<https://doi.org/10.12688/wellcomeopenres.11548.1>)
Latest published: 14 Jul 2017, 2:49 (<https://doi.org/10.12688/wellcomeopenres.11548.1>)

Abstract

Copy number variants (CNV) are a major cause of disease, with over 30,000 reported in the DECIPHER database. To use read depth data from targeted Next Generation Sequencing (NGS) panels to identify CNVs with the highest degree of sensitivity, it is necessary to account for biases inherent in the data. GC content and ambiguous mapping due to repetitive sequence elements and pseudogenes are the principal components of technical variability. In addition, the algorithms used favour the detection of multi-exon CNVs, and rely on suitably matched normal dosage samples for comparison. We developed a calling strategy that subdivides target intervals, and uses pools of historical control samples to overcome these limitations in a clinical diagnostic laboratory. We compared our enhanced strategy with an unmodified pipeline using the R software package ExomeDepth, using a cohort of 109 heterozygous CNVs (91 deletions, 18 duplications in 26 genes), including 25 single exon CNVs. The unmodified pipeline detected 104/109 CNVs, giving a sensitivity of 89.62% to 98.49% at the 95% confidence interval. The detection of all 109 CNVs by our enhanced method demonstrates 95% confidence the sensitivity is $\geq 96.67\%$, allowing NGS read depth analysis to be used for CNV detection in a clinical diagnostic setting.

Keywords



CNV detection, NGS, Read depth, ExomeDepth




This article is included in the [Transforming Genetic Medicine Initiative \(TGMI\)](#) gateway.

Open Peer Review

Reviewer Status  

	Invited Reviewers	
	1	2
version 1 published 14 Jul 2017	 report	 report

- Dominic McMullen**, Birmingham Women's NHS Foundation Trust, Birmingham, UK
- Christian Gilissen** , Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Sian Ellard (Sian.ellard@nhs.net)

Author roles: **Parrish A:** Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Caswell R:** Investigation, Writing – Review & Editing; **Jones G:** Conceptualization, Methodology, Writing – Review & Editing; **Watson CM:** Methodology, Writing – Review & Editing; **Crinnion LA:** Investigation, Writing – Review & Editing; **Ellard S:** Conceptualization, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust [098395].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Parrish A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Parrish A, Caswell R, Jones G *et al.* **An enhanced method for targeted next generation sequencing copy number variant detection using ExomeDepth [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2017, 2:49 (<https://doi.org/10.12688/wellcomeopenres.11548.1>)

First published: 14 Jul 2017, 2:49 (<https://doi.org/10.12688/wellcomeopenres.11548.1>)

Introduction

Placed in the continuum of genetic variation between single nucleotide variation (SNV) and large-scale chromosomal changes, copy number variants (CNV) are identified as segments of DNA that have been deleted or duplicated when compared to a reference genome. Initially described as ranging in size from roughly a kilobase to less than five megabases, but increasingly referring to events greater than 50bp in size due to improved detection sensitivity (<http://dgv.tcag.ca/dgv/app/statistics?ref=GRCh37/hg19>), it is estimated that CNVs cover between 4.8 and 9.5% of the human genome¹. They are a major cause of monogenic disease, with over 30,000 CNVs reported in the [DECIPHER database](#) to date. Where CNVs include disease associated genes or their regulatory regions, they can result in a phenotypic presentation through: alteration of the copy number of dosage sensitive genes, acting in conjunction with a recessively acting mutation on the non-deleted allele, deletion of regulatory elements, or disruption of the coding sequence².

Current diagnostic testing strategies for pathogenic CNVs in patient samples principally rely on either genome-wide array comparative genomic hybridisation (aCGH), or performing locus specific multiplex ligation-dependent probe amplification (MLPA) that can interrogate dosage levels in up to 40 targeted regions per patient sample³. These approaches have both technological (lower size limit of detection, potential for false positive results, limited number of targets) and logistical (they must be run alongside Sanger sequencing for complete analysis) limitations. With the widespread use of targeted Next Generation Sequencing (NGS) panels in the diagnostic setting, new methods of CNV detection are required to allow the detection of CNVs in addition to SNV/indels from a single data set⁴.

The most applicable NGS methods to detect CNVs from hybridisation enrichment targeted gene panel data, generated during routine diagnostic testing, use read depth to infer copy number. Alternative strategies, such as split-read mapping, are often dependent on sequencing the breakpoints of the CNV⁵. While the use of read depth analysis to detect CNVs provides a single wet-laboratory workflow that addresses many of the issues encountered when using older strategies, it presents its own unique challenges, and requires validation using patient samples with known pathogenic CNVs in a clinical diagnostic setting. For highly sensitive analysis of read depth, it is necessary to account for the technical variability in relative average read depth of target regions that is not caused by the presence of a CNV. Such variation can be principally attributed to extremes of sequence GC content and ambiguous mapping of reads due to repetitive sequence elements⁶. In addition, the algorithms used to interrogate the data favour the detection of multi-exon CNVs, and rely on the use of a number of suitably matched samples without CNVs across the target regions for comparison. In the clinical diagnostic setting, it is not cost-effective to include a large number of positive controls with every batch of patient samples.

We have developed a strategy for calling CNVs using read depth in targeted NGS panels that subdivides the target intervals into smaller windows of uniform length, and uses a pool of historical samples for comparison to overcome the main limitations of a relative read

depth approach in a diagnostic testing pathway. We validated our strategy, using the R software package ExomeDepth⁷ against an unmodified calling pipeline, to assess sensitivity on a cohort of 109 samples with MLPA detected heterozygous CNVs (91 deletions, 18 duplications in 26 genes), including 25 single exon CNVs.

Methods

DNA samples from 109 patients with known CNVs identified by MLPA dosage analysis, comprising 91 deletions and 18 duplications in 26 genes and including 25 single exon CNVs, were tested ([Supplementary Table 1](#)). Samples were fragmented using either a Bioruptor (Diagenode, Liège, Belgium) or Covaris S2 (Covaris, Inc., Woburn, MA, USA), indexed for multiplexing, and hybridised to one of three custom Agilent SureSelect exon capture reagents, according to the manufacturer's instructions (Agilent Technologies, Santa Clara, CA, USA). Sequencing was performed with an Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) using 100bp paired end reads. Library preparation and sequencing was performed in one of two testing centres.

The resulting per-sample fastq files were aligned to the hg19 reference genome using the BWA mem algorithm (v0.7.5a-r405, <http://bio-bwa.sourceforge.net>), before being subjected to PCR duplicate removal using Picard (v1.140, <http://broadinstitute.github.io/picard/>) and local realignment around indels using GATK (v3.4, <https://software.broadinstitute.org/gatk/>). CNV analysis was performed using the R software package ExomeDepth (v1.1.8) (9), which compares normalised read count data between a test and an aggregate reference made up of samples from the same sequencing run to determine copy number at exon level resolution. The algorithm assumes a beta-binomial distribution of read count data, with closeness of fit measured by a correlation score. Standard and modified calling strategies were employed and assessed for sensitivity. Calls were identified by the genomic coordinates of windows implicated in the CNV. Statistical confidence in these calls was represented by the “Bayes factor”, a \log_{10} of the likelihood ratio of data for the call divided by the null.

Development of modified calling strategy

The development of a modified calling strategy integrated three alterations to the standard calling strategy. These were aimed at either improving the appropriateness of the method for use in a diagnostic setting (creation of normal dosage gender matched control pools), or the detection sensitivity (windowing of target intervals, removal of GC extremes).

Creation of normal dosage gender matched control pools. Control pools containing a minimum of 50 normal dosage samples were created by running ExomeDepth analysis to compare gender-matched samples from multiple historical batches for each capture panel over exon length intervals. Samples with known CNVs were removed, as were samples with low correlation scores (less than 90% correlation as determined by ExomeDepth), as these represent low quality samples with a markedly different coverage profile from the average, and were unlikely to be selected as controls.

The resultant R data frames containing read counts of each window for every sample for each pool were saved. These data frames were provided to ExomeDepth for selection of the

aggregate reference for CNV calling in the patient samples. This allows a large number of samples to be provided for selection of the aggregate reference, ensuring an appropriate number of gender-matched samples without CNVs are available, with no additional runtime required for each batch.

Choice of window size. To determine the optimum window size for CNV calling, BED interval files for target exons were created for a range of window sizes (200bp, 170bp, 150bp, 120bp, 100bp) for one of the three panels tested (number of samples = 24). Exon calling intervals were divided into windows of the appropriate length in one of two ways, depending on the size of the exon, such that every exon was covered by a minimum of two windows. For exons two windows or shorter in length, two windows were placed radiating out from the mid-point of the exon. For exons greater than two windows in length, a window was placed over the centre-point of the exon with additional windows added, radiating out until the exon was completely covered (see [Supplementary File 1](#) for python script used to produce interval files). The reason for adopting this strategy is that the placement of windows best matches the expected distribution of reads across the exon, such that the variability in read depth within each window is minimised.

Window GC content. The GC content of each window was calculated using the bedtools nuc module (v2.17.0). Windows with extremes of GC content (higher than 80%) were removed from

the interval file, having confirmed that this left a minimum of one window over every exon, retaining the ability to detect CNVs at this location.

Comparison between standard and modified pipelines

Following determination of the optimum window size for CNV detection, interval files and corresponding normal dosage gender matched control pools were recreated at this window size for all panels, as described above. CNV analysis was undertaken on all 109 patient samples using these files, with performance compared to the standard calling strategy (see [Supplementary File 2](#) for example Exomedept script containing all settings used).

Standard calling strategy. The standard ExomeDepth pipeline followed the strategy outlined in the vignette accompanying the package (<https://cran.r-project.org/web/packages/ExomeDepth/vignettes/ExomeDepth-vignette.pdf>). Exon length interval BED files were created for each of the three panels. Samples were interrogated for CNVs over the appropriate targets by comparison against all gender-matched samples within a sequencing run (range 1 to 14 gender-matched samples).

Results

Variable window size

The performance of the five window lengths tested is displayed in [Figure 1](#). All window lengths detected 100% of the positive control

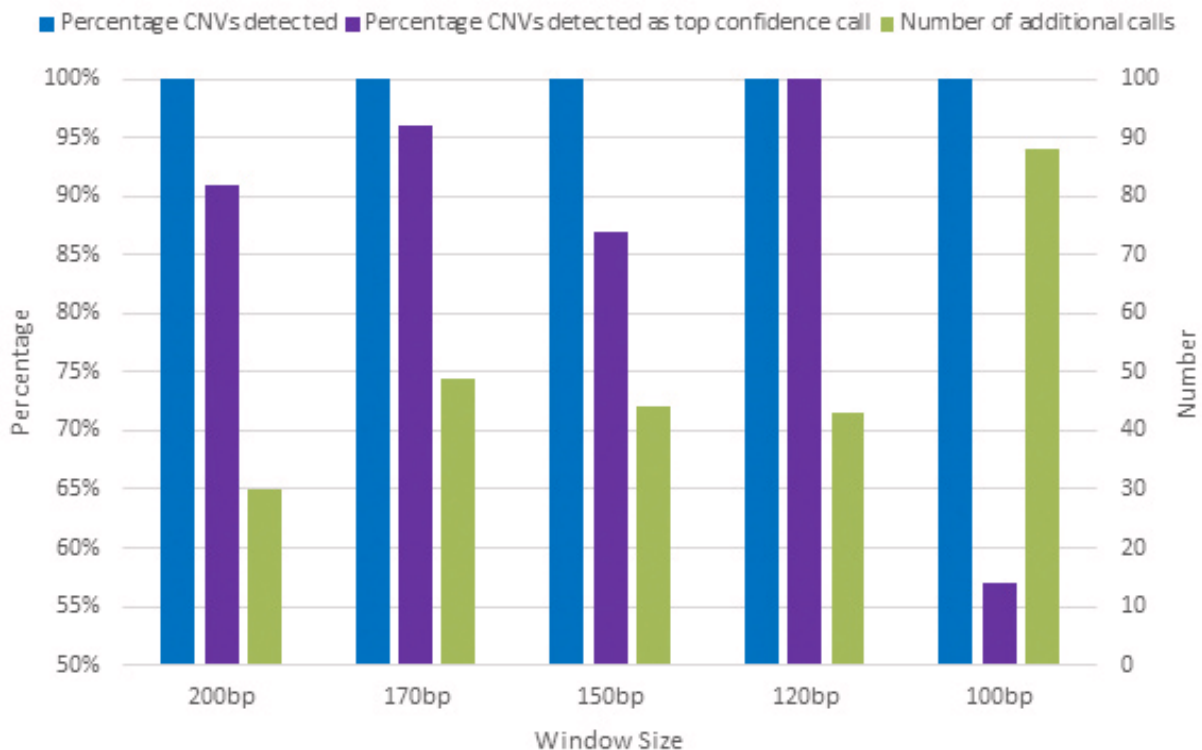


Figure 1. Performance comparison of window length. All window lengths tested detected all copy number variants (CNVs) in the controls set, although the 120bp windows performed best consistently identifying the positive control CNV as the top confidence call. The performance of the 100bp windows was considered to be the poorest, given the low confidence in the positive control CNV calls, and the greatly increased number of additional calls when compared to other window sizes tested.

CNVs. However, 120bp windows performed best for prioritisation of the known CNV, in all cases providing the highest confidence call for this CNV, while producing a limited number of additional calls. On this basis, the decision was made to use 120bp windows for future analysis.

Standard versus modified pipeline

The results of the comparison of the standard and modified calling strategies is displayed in Table 1. The standard testing pipeline detected 104/109 CNVs, giving a sensitivity of 89.62-98.49% at the 95% confidence interval. Of the five CNVs not detected using this pipeline, three were single exon events (1 duplication, 2 deletions), one only had a single sample selected as a control, and one was on the same sequencing run as a larger overlapping CNV.

Since our modified protocol detected 109/109 CNVs, we have 95% confidence that the enhanced strategy sensitivity is $\geq 96.67\%$ (96.67-100.00% at 95% CI).

Bayes factor for separation of positive calls from noise

The combination of windowing of exon length intervals and removing areas of highly variable coverage caused by high GC content decreases the number of additional calls made by the software (78 calls across all samples using 120bp windows compared to 89 using exon length intervals). The majority of the additional calls present are reported with significantly lower confidence than positive CNVs of equal length (Figure 2), raising the possibility of excluding these additional calls on the basis of the Bayes factor of the call produced by ExomeDepth.

Table 1. Detection rates of standard and modified pipelines.

Detection strategy	No. of CNVs identified	Sensitivity (95% CI)	Positive control CNVs missed
Standard	104/109	89.62-98.49%	<i>PHEX</i> Exon 12 duplication <i>PRKARIA</i> Exon 6 deletion <i>ABCC8</i> Exon 13 deletion <i>GLIS3</i> Exons 1-2 deletion <i>APC</i> Exons 7-15 deletion
Modified	109/109	96.67-100.00%	N/A

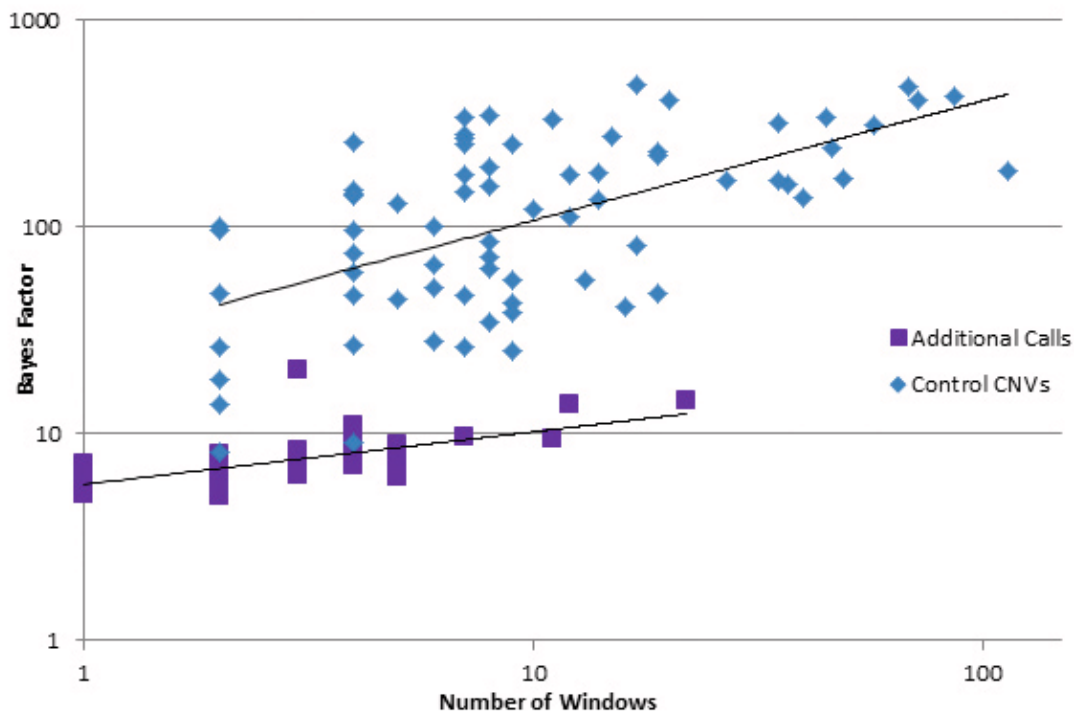


Figure 2. Bayes factors of control copy number variants (CNVs) and additional calls. While complete separation between positive control CNVs and additional calls (which have not been confirmed, and may reflect carrier status for CNVs in these patient samples) is not possible based on Bayes factor alone, it has potential to assist in the filtering of variants with other available information. This may include the presence of heterozygous variants within a called deletion, clipped reads to identify breakpoints if present, and by limiting the reporting of variants to clinically relevant genes/sub-panels.

Discussion

We have developed a highly sensitive strategy for CNV detection in targeted NGS panels, and validated this using a series of positive control samples from patients with rare genetic diseases. Since our protocol detected 109/109 CNVs, we have 95% confidence that the modified strategy sensitivity is $\geq 96.67\%$. This represents an improvement over the standard calling strategy, which detected 104/109 CNVs, giving a sensitivity of 89.62-98.49% at the 95% confidence interval.

For a highly sensitive analysis of read depth, it is necessary to account for the technical variability in relative average read depth of target regions that is not the result of the presence of a CNV. This is particularly true in the case of duplications, which result in a smaller proportional change in the number of mapped reads when compared to deletions. Strategies for combining calls from adjacent target regions, such as the hidden Markov model employed by ExomeDepth, result in comparatively lower confidence in the calling of smaller CNVs that are commonly seen in disease. By dividing exonic targets into multiple windows, we have improved the detection sensitivity of single exon CNVs, narrowing the gap in the size of sequence variation that can be detected between CNV packages and the variant callers used for SNV/indel detection. This is supported by the fact that of the five CNVs missed by the standard strategy, but detected by the modified calling pipeline, three were single exon events (*PHEX* exon 12 duplication, *PRKARIA* exon 6 deletion, *ABCC8* exon 13 deletion).

The selection of window size to use for CNV calling is clearly a critical factor when using read depth based tools such as ExomeDepth. If the window size is too large, smaller CNVs may not represent the highest confidence call in a background of variable coverage. If the window size is too small, false positive calls are likely to be introduced into the results. The performance comparison of window size in this study showed the relative highest confidence in positive CNV calls in the 120bp windows, and a two-fold increase in the number of additional calls produced by the software when using 100bp windows compared to 120bp windows. These data suggest that 120bp windows may represent the optimum size when using the read-length generated by the current sequencing technology.

The modifications made to the pipeline aimed at increasing sensitivity potentially decrease specificity. The degree to which this is true is difficult to establish, as not all targets covered by the NGS panels have been tested using MLPA to determine normal copy number. For use in a clinical diagnostic setting, where CNVs are confirmed by an alternative method prior to reporting, high sensitivity is of greater importance than high specificity to reduce the number of false negatives. It may be possible to improve specificity and separate true calls from false positives or CNVs unrelated to the clinical referral using additional data. This may incorporate the Bayes factor of the call when using ExomeDepth, but is more likely to be successful when also considering the presence of heterozygous variants within a called deletion, clipped reads to identify breakpoints if present, and limiting the reporting of variants to clinically relevant genes/sub-panels.

Read depth CNV detection strategies are dependent on uniformity of coverage across the entire capture region, so that any variation seen is indicative of genuine copy number change present in the sample. Regions of the genome to which it is difficult to unambiguously align the short reads produced by the current NGS technology (i.e. GC content extremes, pseudogene homology) interfere with this uniformity. In addition to affecting detection in these regions, this causes problems across the entire region of interest because of decreasing the correlation of the patient and reference samples used for comparison, increasing noise. Excluding those regions from analysis allows a more suitable match to be made, increasing the sensitivity.

While removing areas of GC content extremes or with a high degree of homology to pseudogenes from analysis improves the correlation between test and reference genomes, and as such the sensitivity of the method, we acknowledge it does so at the cost of the ability to detect CNVs in these areas. Although this is a limitation of any current method, which will only be possible to fully overcome when the use of long read sequencing becomes routine, the windowing of exons can ameliorate this, allowing the removal of windows with particularly high GC content, while retaining the ability to detect CNVs in the remaining areas of the exon.

When using read depth to infer copy number in NGS data, it is often recommended to compare samples within a sequencing run in order to avoid run specific technical artefacts. While this strategy can work well in large sequencing projects, it is not a feasible approach for a clinical diagnostic laboratory for a number of reasons, including cost. If overlapping CNVs are present in multiple samples, it can result in false negative results, and for calls on the sex chromosomes, it relies on having a suitable number of gender matched controls on the same run. Having made modifications to the interval files to account for the major components of variability in coverage between samples, this recommendation can be relaxed. This allows control pools of normal dosage samples from multiple runs to be used for comparison, making this method more suitable for use in a diagnostic testing pathway. This approach allowed the detection of both an *APC* exon 7-15 deletion, and a *GLIS3* exon 1-2 deletion. The *APC* exon 7-15 was missed using the standard testing strategy due to the incorporation of a sample containing a larger overlapping CNV in the aggregate reference. The aggregate reference selected by ExomeDepth for the *GLIS3* exon 1-2 deletion only contained one sample when using the standard strategy, which resulted in a poor correlation score and a false negative result.

In this study, we have used ExomeDepth as the CNV calling tool of choice. While the strategy we have suggested is yet to be tested on the multitude of other packages available, the modified approach we have developed tackles the limitations of the technology rather than the specific tool used; therefore, we anticipate that this will improve the sensitivity of the range of software available. Modification of the standard CNV analysis pipeline to address issues inherent in the targeting techniques routinely used by diagnostic laboratories allows for improved detection capability.

Ethical statement

The samples used in this study were collected for research studies or were from patients undergoing clinical diagnostic testing. The research studies were conducted in accordance with the latest version of the Declaration of Helsinki and consent was obtained from all patients or their parents. Declarations of written or oral consent are recorded in clinical notes. Samples were collected with ethical approval from the Wales Research Ethics Committee Bangor. The samples collected for clinical diagnostic testing were used for quality assurance purposes to validate the detection of CNVs by next generation sequencing.

Data availability

The individual data files used for the study are not available online, due to patient confidentiality issues of the samples collected for clinical diagnostic testing and the storage requirements of the fastq (~ 1Gb per sample) or BAM files (~ 4Gb per sample) required for the analysis of all samples. We anticipate that users will have their own datasets against which to test this strategy, but please

contact andrew.parrish@nhs.net if fastq and/or BAM files from the consented research samples used in this study provided via SFTP would be helpful.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [098395].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to acknowledge the contribution of the technical and bioinformatics staff involved in the sample preparation, in particular A. Bussell, M. Johnson, and M. Wakeling.

Supplementary material

Supplementary Table 1: Control copy number variants included in cohort.

[Click here to access the data.](#)

Supplementary File 1: Python script used to produce interval files.

[Click here to access the data.](#)

Supplementary File 2: Example Exomedepth script containing all settings used.

[Click here to access the data.](#)

References

- Zarrei M, MacDonald JR, Merico D, *et al.*: **A copy number variation map of the human genome.** *Nat Rev Genet.* 2015; **16**(3): 172–83.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome.** *Nat Rev Genet.* 2006; **7**(2): 85–97.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Schouten JP, McElgunn CJ, Waaijer R, *et al.*: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.** *Nucleic Acids Res.* [Research Support, Non-U.S. Gov't]. 2002; **30**(12): e57.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhao M, Wang Q, Wang Q, *et al.*: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.** *BMC Bioinformatics.* 2013; **14**(Suppl 11): S1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Watson CM, Crinnion LA, Tzika A, *et al.*: **Diagnostic whole genome sequencing and split-read mapping for nucleotide resolution breakpoint identification in CNTNAP2 deficiency syndrome.** *Am J Med Genet A.* [Case Reports]. 2014; **164A**(10): 2649–55.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Teo SM, Pawitan Y, Ku CS, *et al.*: **Statistical challenges associated with detecting copy number variations with next-generation sequencing.** *Bioinformatics.* [Research Support, Non-U.S. Gov't Review]. 2012; **28**(21): 2711–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Plagnol V, Curtis J, Epstein M, *et al.*: **A robust model for read count data in exome sequencing experiments and implications for copy number variant calling.** *Bioinformatics.* [Research Support, Non-U.S. Gov't]. 2012; **28**(21): 2747–54.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 31 July 2017

<https://doi.org/10.21956/wellcomeopenres.12474.r24223>

© 2017 Gilissen C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Christian Gilissen 

Department of Human Genetics, Donders Institute of Neuroscience, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Parrish et al. describe the optimization of an existing tool to identify CNVs from exome sequencing data tailored to application within diagnostic testing. They improve the existing algorithm “ExomeDepth” on three points: the reference cohort, splitting targets into smaller windows, and removal of GC rich windows. They show that with these improvements ExomeDepth is able to identify all of the 109 CNVs from a dataset of known CNVs, whereas only 104/109 were identified with the standard algorithm.

The paper is very clearly structured and written. The methods and results are fairly straightforward and clear.

1. The authors focus very much on the sensitivity of the methods, rightfully claiming that this is more important in a diagnostic setting than specificity. The authors show the number of additional calls for various choices of the window size, but not for the standard ExomeDepth method. This should be included such that the readers can verify that the optimizations are not at an large cost of specificity.
2. The authors claim that additional CNV calls in their analysis have lower quality values is clear from Figure 2. But it would be good to also quantify the difference in quality scores between Control CNVs and additional calls with a statistical measure, since the phrasing “significantly lower confidence” is used by the authors.
3. For reproducibility please include the target regions that were finally used in the analysis and the details of the calling of the control CNVs and additional calls for the final algorithm (location, type, size, Bayes factor, etc.).
4. The authors rely on a method called ExomeDepth, which seems to suggest the current study is done with exomes. Although the authors mention that this analysis is performed for a gene panel this can be a bit confusing. It would be good if the authors provided some details on the panels in the material and methods and perhaps discuss on the application of their method for whole exomes.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 28 July 2017

<https://doi.org/10.21956/wellcomeopenres.12474.r24221>

© 2017 McMullen D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Dominic McMullen

West Midlands Regional Genetics Laboratory, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

This report very clearly demonstrates that robust CNV calling from NGS data using an enhanced Exome Depth methodology is achievable without the need for running large batches and having large internal control sets.

As is evidenced this approach seems most applicable to gene panels, where sensitivity is key and will help to reduce the need for e.g MLPA as a primary CNV detection method alongside gene panel (physical and virtual) testing.

The authors state: "The modifications made to the pipeline aimed at increasing sensitivity potentially decrease specificity. The degree to which this is true is difficult to establish, as not all targets covered by

the NGS panels have been tested using MLPA to determine normal copy number. For use in a clinical diagnostic setting, where CNVs are confirmed by an alternative method prior to reporting, high sensitivity is of greater importance than high specificity to reduce the number of false negatives"

Thus for gene agnostic /genome-wide CNV detection this method may be more difficult currently to reliably implement as a replacement for e.g aCGH/SNP-array, which now both have very high specificity genome-wide. It will very be interesting to see if the Bayes factor used to separate positive calls from noise can be further novelly enhanced for this purpose.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cytogenomics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
