OXFORD

Briefings in Bioinformatics, 2022, **23**(2), 1–8
https://doi.org/10.1093/bib/bbac049
Problem Solving Protocol

CoverageMaster: comprehensive CNV detection and visualization from NGS short reads for genetic medicine applications

Melivoia Rapti, Yassine Zouaghi, Jenny Meylan, Emmanuelle Ranza, Stylianos E. Antonarakis and Federico A. Santoni

Corresponding author: Federico A. Santoni, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland; Medigenome, Swiss Institute of Genomic Medicine, Geneva, Switzerland; Univesity of Lausanne, Lausanne, Switzerland. E-mail: federico.santoni@chuv.ch

Abstract

CoverageMaster (CoM) is a copy number variation (CNV) calling algorithm based on depth-of-coverage maps designed to detect CNVs of any size in exome [whole exome sequencing (WES)] and genome [whole genome sequencing (WGS)] data. The core of the algorithm is the compression of sequencing coverage data in a multiscale Wavelet space and the analysis through an iterative Hidden Markov Model. CoM processes WES and WGS data at nucleotide scale resolution and accurately detects and visualizes full size range CNVs, including single or partial exon deletions and duplications. The results obtained with this approach support the possibility for coverage-based CNV callers to replace probe-based methods such as array comparative genomic hybridization and multiplex ligation-dependent probe amplification in the near future.

Keywords: medical genetics, copy number variants, signal processing

Introduction

Copy number variation (CNV) is the most frequent structural alteration in the human genome. Aberrant numbers of copies of specific genes, exons or, in general, genomic regions are known to be implicated in pathogenic conditions such as Mendelian diseases and cancer [1–4]. Hence, identification of these deletion and amplification events is a primary purpose in medical genetics research. In clinical diagnostics, the identification of rare, potentially causative CNVs in a patient with a suspected genetic disorder is a long-sought objective. However, the discovery of such variants that can vary in size and copy number is a challenging task. Currently, the most commonly used high-throughput methodologies to detect clinically relevant CNVs rely on microarray-based technologies. Array comparative genomic hybridization (array CGH) offers an efficient method to detect CNVs and micro-CNVs (5Kbp < size <10Mbp) in the whole genome, but its resolution does not cover the lower size spectrum. Multiplex ligation-dependent probe amplification (MLPA) is the current golden standard to detect exon-sized CNVs but this technology can cover few exons per assay (low throughput) and its application is limited to a small number of genes [5].

In recent years, the development of next-generation sequencing (NGS) technologies of short reads has provided a standardized way for accurate coding variant analyses through whole genome sequencing (WGS) and whole exome sequencing (WES). Remarkably, this technology provides the coverage per nucleotide of clinically relevant regions of the genome. Although WGS allows for a more comprehensive overview of the entire genome with uniform coverage [6], the related sequencing costs and the computational infrastructures needed to process the raw data are still limiting its broad application in clinical practice [7]. On the other hand, WES is computationally less demanding and has reached such a high

Melivoia Rapti is PhD student at the University of Lausanne and at the Endocrinology Diabetes and Metabolism Service, CHUV Hospital, Lausanne Switzerland. She studies the application of computational pipelines to bionformatics applications.
Yassine Zouaghi is bioinformatician PhD student at the University of Lausanne and at the Endocrinology Diabetes and Metabolism Service, CHUV Hospital, Lausanne, Switzerland. His thesis focuses on finding new Congenital Hypogonadodtropic Hypogonadism causing genes through WGS.
Jenny Meylan is research technician at the Endocrinology Diabetes and Metabolism Service, CHUV Hospital, Lausanne Switzerland. She is interested on novel NGS technologies, data preparation and analysis.
Emmanuelle Ranza is clinical geneticist and CMO at Medigenome, Swiss Institute of Genomic Medicine, Geneva, Switzerland. Her research interests focus on rare genetic diseases and the improvement of diagnostic and preventive genetic services to the population.
Stylianos E. Antonarakis is professor emeritus at the University of Geneva, Switzerland, and the CEO of MediGenome, Swiss Institute of Genomic Medicine. His current research work is to identify novel genes for autosomal recessive disorders, by studying consanguineous families.
Federico A. Santoni is group leader at the University and Lausanne and at the Endocrinology Diabetes and Metabolism Service, CHUV Hospital, Lausanne, Switzerland. His research focuses on the development of computational methods for genomics and transcriptomics in the context of rare diseases, cancer and single cell multi-modal data processing.

Received: September 23, 2021. **Revised:** January 28, 2022. **Accepted:** January 31, 2022

sensitivity and specificity in variant calling to eventually become a clinical standard. Currently, WES is widely used for diagnostic purposes in many medical genetics laboratories throughout the world.

A wide range of detection algorithms have been developed to call CNVs from WGS and WES data. It is customary to define as CNVs duplication and deletions with a size >1–5 Kbp where the smaller ones are called INDELs. In the exomic space, however, the duplication/deletion of one exon (down to 100pb or less) can result in a much bigger duplication/deletion in the genomic space for the large majority of breakpoints happen in the intronic or intergenic part of the genome. Therefore, while split-reads- and gapped-reads-based algorithms [8] might be quite sensitive and precise when the breakpoints are covered (i.e. sequenced), in practice they are quite inefficient to detect exonic structural variants if the SV is bigger than the size of the read (~100–200 bp in standard WES and WGS experiments) [9]. For this reason, while waiting for long-read NGS to take over in clinical applications, read-depth-based methods [10–12] are so far considered more effective for accurate copy number detection in WES data [13]. NGS short reads are mapped to a reference sequence and the depth-of-coverage (DoC) in a genomic region is calculated by counting the number of reads that align to this region. DoC is then assumed to be proportional to the copy number of that region. In principle, DoC is sufficient for the detection of all clinically relevant CNVs, irrespectively of size and copy number and breakpoints location, promoting WGS and WES as a robust and more inclusive alternative to complementary laboratory approaches such as array CGH or MLPA.

Nevertheless, WES has technical issues that result in the generation of noisy data. First, the lack of continuity of the target regions and, second, the biases due to hybridization and sequencing processes complicate the procedure to standardize read-depth-based CNV detection [14]. As a result, current WES-based detection methods suffer from limited resolution, high false positives and false negatives calls [9].

Here, we introduce CoverageMaster (CoM), a CNV calling algorithm based on DoC maps from aligned short sequence reads from WES or WGS. CNVs are inferred with Hidden Markov Models (HMMs) at multiscale nucleotide-like levels in the Wavelet reduced space, in comparison to existing methods that utilize fixed length windows or exon averages. This approach is designed to optimize the search for CNVs of different sizes in WES and WGS data. Of note, since it is working at nucleotide resolution, CoM provides the graphical representation of the predicted CNV in all genes of interest, and, optionally, a wig formatted file compatible with UCSC Genome Browser for detailed visualization of the normalized coverage on the target genes or regions in the genomic space. We propose CoM as a potential first-line diagnostic tool in research and clinical applications.

Materials and methods

Material

The analyses reported in this study were performed on DNAs processed by WES at the Health 2030 Genome Center (https://www.health2030genome.ch/) or Medigenome (www.medigenome.ch) using Twist Human Core Exome Kit (TWIST Biosciences, San Francisco, CA, USA); NA12878 DNA has been obtained from the Coriell Institute (https://coriell.org/); sequencing was performed on Illumina HiSeq4000 or Novaseq platforms. Array CGH and MLPA were performed in GeneSupport using Agilent SurePrint G3 Human 4x180K (analyzed with Agilent CytoGenomics (V 5.1.1.15) and double checked by visual inspection) and SALSA MLPA Probemix P021 SMA (MRC Holland), respectively.

Preprocessing and transformation of exome data

CoM uses DoC maps from aligned short sequence reads to estimate CNV events. To acquire the sequence reads, the mapping is done with the standard pipeline for whole-exome or WGS data based on GATK [15], and the coverage at each nucleotide of the region of interest (ROI) is calculated and stored in tab separated COV files (format: chr nucleotide_position coverage) using samtools (samtools depth) [16]. Coverage files of a test/target plus one or more controls plus one reference coverage serve as input for the algorithm. The assumption is that control coverages are DoC maps of copy number neutral cases (diploid) or carrier of frequent CNVs in the ROI of interest. The reference set consists of a batch of coverage files from samples processed with the same technology (i.e. hybridization kit, reagents for library prep and sequencer) used to generate case and controls. First, the coverage per nucleotide per sample is normalized by the respective total number of reads. Then, mean and standard deviation of the normalized coverage values are computed over all the samples for each nucleotide.

WAVELET transform

In a genomic region of N nucleotides, the coverage of test case and control can be represented as the discrete signals $s(n)$ and $c(n)$, respectively, where n is the nucleotide number corresponding to the genomic or exonic position in the exon space [the space where covered regions (i.e. exons) are 'ligated' together]. In the ideal case, the coverage ratio $r = \frac{s}{c}$ is a non-periodic square waveform with up and down steps in correspondence of increased or decreased copy number, respectively. In order to diminish the noise induced by fast variations of the signal and, at the same time, to reduce the computational burden, the coverage ratio is compressed in the nucleotide-like space using the Discrete Wavelet Transform (DWT) equipped with the Haar basis. At scale l, the approximation and detail coefficients are $r_l, d_l, d_{l-1}, \ldots, d_0 = DWT_l(r)$. The $M =$

$N \cdot 2^{-l}$ approximation coefficients r_l are normalized to the median of the original signal and used for CNV analysis.

Multiscale CNV detection

The probability $b_j(o_m)$ of each m nucleotide-like positions of the sequence of approximation coefficients $r_l = o_1 o_2 \dots o_k o_{k+1} \dots o_M$ to be in a normal (i.e. diploid), duplicated or deleted state $s \in S \equiv \{1, \frac{3}{2}, \frac{1}{2}\}$ is defined at any scale l as a random variable with Gaussian distribution of mean s and standard deviation $\sigma(R_l)$, where $R_l(m)$ is the sequence of approximation coefficients of the reference coverage in the m -coordinates of the l -scaled nucleotide-like space.

At scale l, the indicator function (trigger) $T = argmax_s(b_s(r_l)) \neq 1$ identifies the locations of non-diploid nucleotide-like positions and masks the rest of the signal. If no location is identified, the algorithm discards this region and processes the next one.

Once the putative CNVs are identified, the Viterbi algorithm is then used to identify the most likely copy number state sequence $Q = q_1 q_2 \dots q_k q_{k+1} \dots q_M$ of the compressed genomic region, based on the corresponding sequence of observations $r_l = o_1 o_2 \dots \underline{o_k o_{k+1}} \dots o_M$. Masked observations $\underline{o_k}$ have a fixed diploid state $q_k = 1$.

More formally, if $v_t(j)$ represents the Viterbi probability that the underlying HMM is in copy number state j after seeing the first m observations and passing through the most probable state sequence $q_1 q_2 \dots q_{m-1}$, it can be shown that $v_m(j) = \mathbf{max}_{i \in S} v_{m-1}(i) \alpha_{ij} b_j(o_m)$, where $v_{m-1}(i)$ is the previous Viterbi path probability from the previous nucleotide, α_{ij} is the transition probability (here set to $a_{ij} = 5 \times 10^{-6}$ which is the probability of finding a duplication or a deletion in the human genome, calculated as the mean of the inclusive and stringent number of CNVs per nucleotide from [17]) and $b_j(o_m)$ is the observation probability given the state j as defined above.

If no putative CNV is detected at this stage, the algorithm performs a multiscale analysis by repeating the HMM phase with the masked signal transformed at scale $l - 1$. Again, in absence of CNVs, the algorithm keeps decrementing l down to, if necessary, $l = 0$ (no compression). This is computationally possible because only the relevant unmasked regions are actually inspected. Otherwise, eventual putative CNVs are saved and the algorithm proceeds to the next region.

Iteration over controls

In case more control coverages are provided, eventual putative CNVs and relative masks are stored in a temporary buffer. Following the assumption that a rare causative CNV cannot be present in any control sample, CNVs are iteratively challenged with the Multiscale CNV Detection algorithm against each control.

Generation of simulated data

Heterozygous deletions and duplications in randomly picked exonic regions have been inserted in samples BAM files using the library Pysam from Python. Briefly, a script selects a random exonic position (inter-exonic or across two or multiple exons) and, around that location, removes or duplicates half of the overlapping reads in the sample BAM files, respectively. Coverage (COV) files are then produced with *samtools depth* following the usual protocol and processed with CoM with standard parameters.

Results

CoM utilizes the representation of coverage signal ratio (case over control) in the reduced Wavelet approximation space to perform a multiscale analysis of aberrant coverage profiles, potentially underlying causative CNVs, at nucleotide resolution (Figure 1, see Methods). This approach is meant to explore a broad spectrum of CNV sizes and in particular deletions or duplications of <5 kb. At this scale, the experimental noise is caused on one hand by the particular technology used for sequencing and, for WES, DNA selection by hybridization. On the other hand, batch specific coverage distortions may occur. Intuitively, the smaller the CNV the higher the chance that the call is a false positive. To overcome this problem, CoM exploits the fact that, as all other genomic variations, clinically relevant CNVs are rare (MAF < 0.01%). Thus, it is reasonable to assume that such CNVs cannot be present in two or more independent unrelated individuals of the same batch. Following this basic principle, CoM utilizes a reference with the average coverage and standard deviation of 15–20 samples processed with the same technology (hybridization kit, reagents and sequencer). The reference provides the standard deviation per nucleotide from the expected coverage where coverage spikes are produced by reproducible experimental noise and/or recurrent CNVs. Eventually, matching CNVs in the test sample are then considered as frequent or false positives and finally discarded. Moreover, CoM pairwise compares the sample case with independent samples, used as controls, from the same batch. Spikes present in the test signal coverage and in one control sample are averaged out in the coverage ratio and consequently discarded.

In order to prove its efficiency, we tested CoM in various contexts of NGS data analysis. All samples processed here for WES were hybridized with Twist Core Exome + RefSeq Spike and sequenced with Illumina HSeq4000 or Novaseq.

Most of the published algorithms use samples form 1000 Genomes to evaluate their performances (e.g. [19]). Being the large majority of CNVs in these samples quite frequent and of no clinical relevance, this approach is not appropriate for CoM. To clarify this point, we sequenced and analyzed the exome of sample NA12878, generally considered the golden standard for this analysis [20]. Whole Genome CNV calls validated by several technologies are made available by the 1000G consortium in https://www.internationalgenome.org/

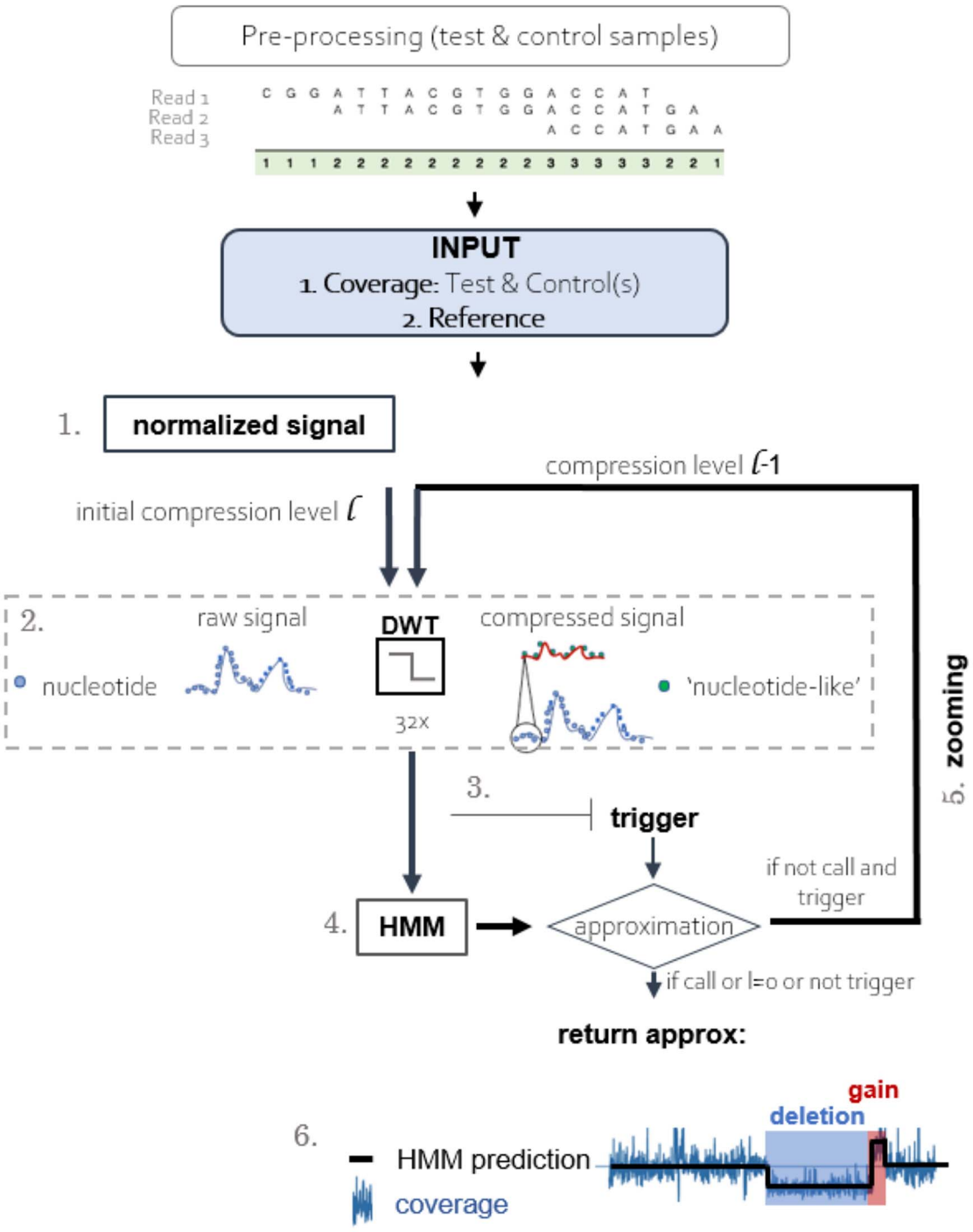

Figure 1. CoverageMaster workflow CoM is based on depth-of-coverage maps from aligned short sequence reads from WES or WGS. The normalized values of the depth-of-coverage for each nucleotide position are calculated (Step 1). The ratio of the test to control coverage signal is compressed at a specified initial scale $l(= 2^5$ by default) in the nucleotide-like space using the DWT (Step 2). For the compressed signal, an indicator detects the potential non-diploid nucleotide-like positions (Step 3). HMM is used to segment the compressed signal into regions of similar copy number and assign CNV states (Step 4). If no putative CNVs are identified, the process is repeated at scale $l - 1$ via 'zooming' (Step 5).

phase-3-structural-variant-dataset/. This sample has 45 exonic CNVs of which 34 are frequent (MAF > 5%). As expected, CoM achieved a recall of 9/45 on the full set but 9/11 on the rare CNV set (the two miss CNVs were few bases overlapping with the exonic covered region). ED identified 11/45 CNVs on the full set and 6/11 on the rare set. CODEX2 and CONTRA both detected 7/45 and 5/11 where PatternCNV scored 2/45 on the full set and 0/11 on the rare set (Supplementary Figure 1).

To design a more clinically oriented test to investigate the performance of CoM on CNVs of different size, we created a dataset of simulated WES data starting from real BAM files obtained from 10 individuals where array CGH did not previously provide any clinically

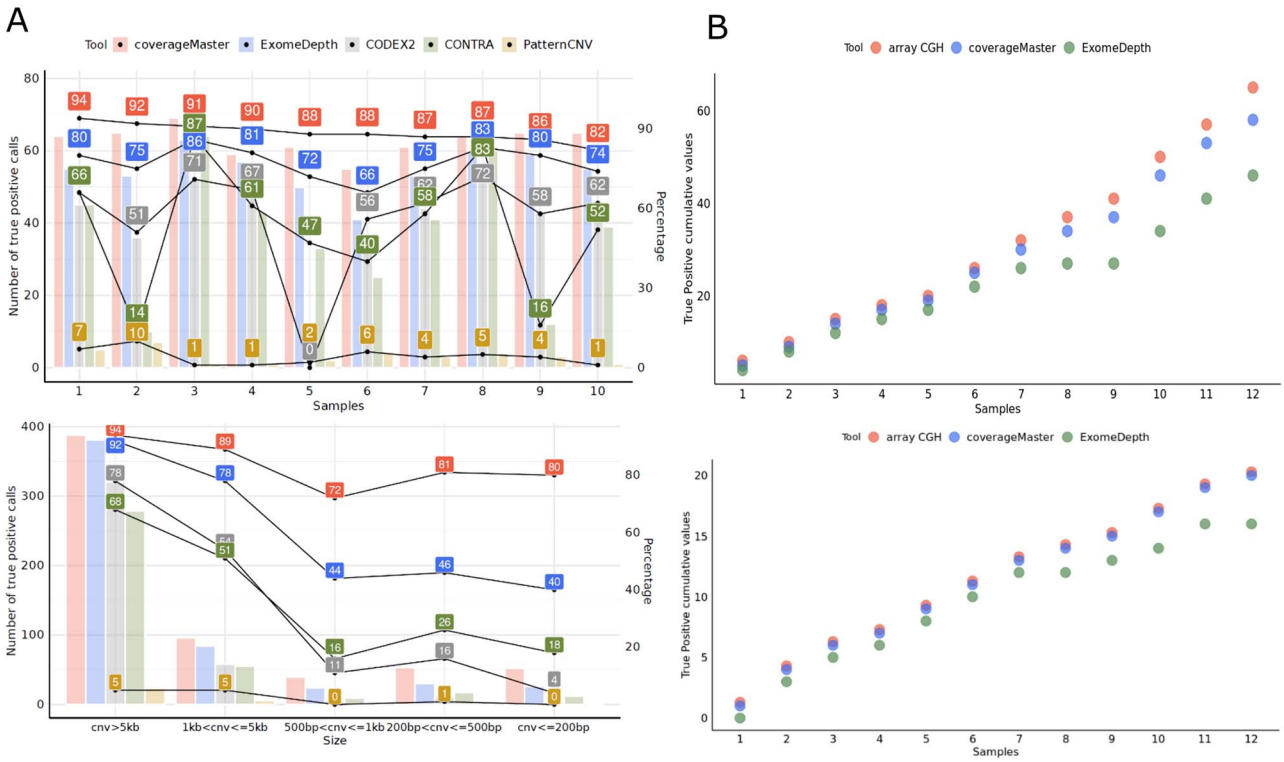

Figure 2. CNV detection in simulated data and array CGH comparison on clinical samples. **(A)** (up) Number and fraction of true calls detected by CoM (orange), ED (blue), CODEX2 (gray), CONTRA (green) and PatternCNV (yellow) in 10 samples where CNVs of various size were randomly introduced in exonic regions. (down) Number and fraction of true calls of detected CNVs by the above-mentioned tools stratified by size. **(B)** Cumulative plots of number of calls (y-axis) detected by CoM (blue) and ED (green) and the number of CNVs found by array CGH (ed) in 12 samples: (up) all calls are considered; (down) only the rare CNVs (MAF < 1%) are included.

significant call. We preferred this approach to the generation of synthetic reads as performed in other studies where they had to simulate the sequencing error model, the probability to have a single nucleotide variant, GC content effect etc. [18]. Indeed the use of real samples automatically provides all the requested features. Around 2000 heterozygous duplications and deletions of 200, 500, 1000 and 5000 base pairs were randomly introduced in the exonic regions of these samples (see Methods) and analyzed by CoM and other CNV callers such as ExomeDepth [12], CONTRA [21], PatternCNV [22] and CODEX2 [19]. The results show that CoM has the best performance with an average sensitivity of 88.5% as compared with 77% obtained by the second best performer ED (Figure 2c) and an average precision of 30% for CoM versus 16% obtained by ED with 25 control samples (with the conservative hypothesis to considering all CNV calls not overlapping with the simulated test as False Positives). It is worth to note that, in contrary to ED, CoM precision drastically increases with the number of control samples (Supplementary Figure 2). The explanation of this difference in performance between CoM and the other tools becomes evident by stratifying the CNV calls by size. It is indeed the multiscaling approach that enables CoM to keep a constant high sensitivity above 80% for all CNVs sizes in contrast to the other tools where the performance rapidly decreases with size reduction (Figure 2A).

To demonstrate further the performance of CoM in standard clinical analyses, we analyzed 12 clinical samples and compared CoM CNVs calls to standard array CGH calls (see Methods). In order to provide a point of reference, we also included the results obtained by ED given its reasonably good performance in the simulation test. In Figure 2a, the cumulative true positive values for CNVs detected by CoM and ED are reported. CoM calls coincide with almost the entire array CGH calls for each sample with the exception of some frequent benign variants discarded by CoM as they are present in most controls. In fact, when searching for CNVs with MAF < 1%, CoM identifies all CNVs detected by array CGH, in contrast to ED that detects 80% of them (Figure 2b). This result demonstrates that CoM may replace array CGH in clinical diagnostic settings.

CoM has been mainly conceived as a diagnostic support tool for clinical genetics analysis. To provide a perspective of the broad capabilities of the algorithm, we report four examples (three WES and one WGS) of solved clinical cases.

Patient 1 is a 38-year-old male with a Kallman syndrome [OMIM 308700] born from a consanguineous couple of the first degree. WGS analysis with CoM revealed a homozygous deletion of 135Kbp including the two first exons of ANOS1 that completely explain the phenotype [23]. Interestingly, WGS data can be also analyzed as WES by CoM by calculating the appropriate

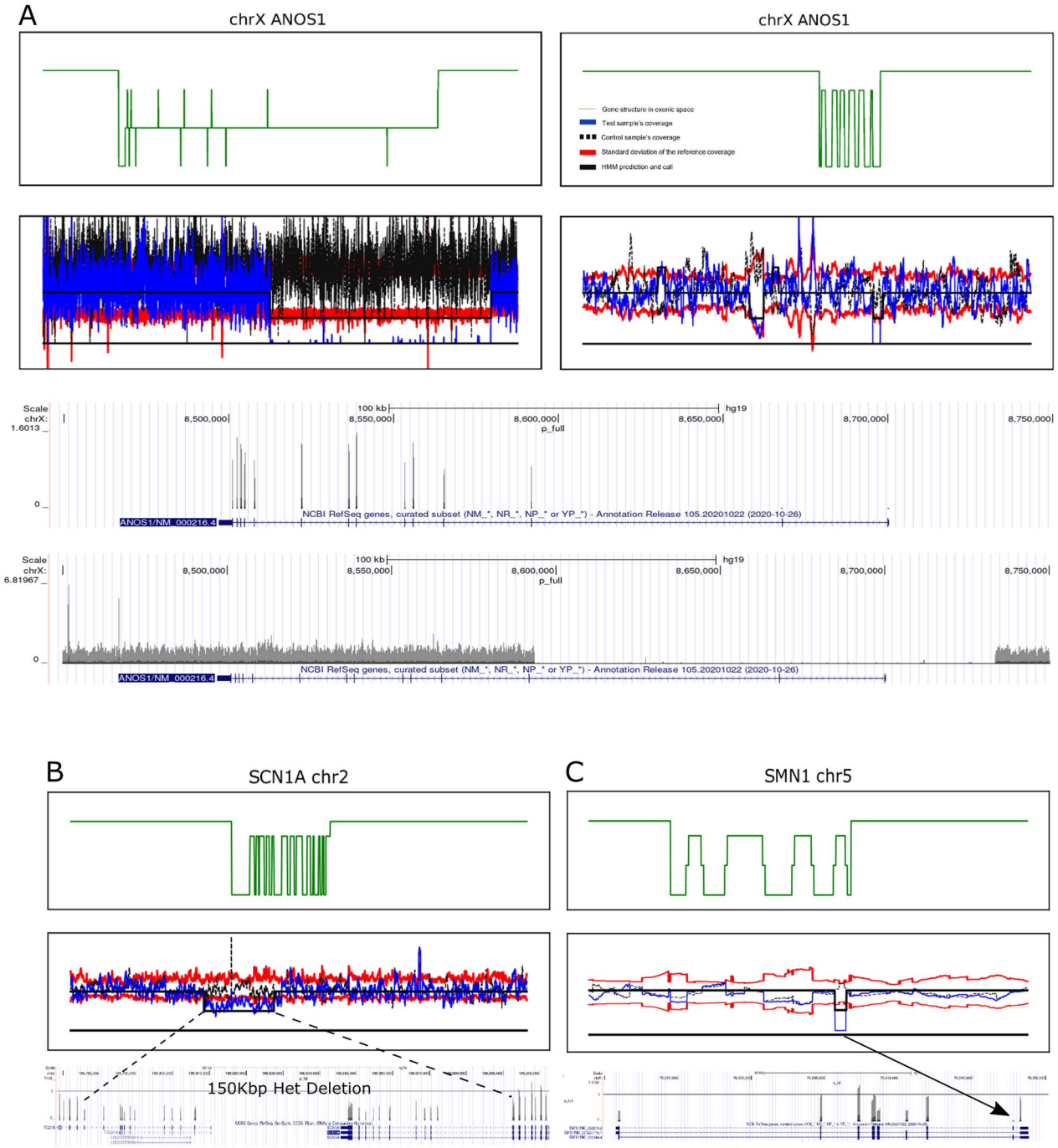

Figure 3. Examples of clinically relevant CNV identified by CoM. **(A)** In green, the collapsed exon structure of the gene of interest, up or down blocks representing one exon. Coverage profiles in exon space of test sample, control and reference (color code in the legend) are represented in the second plot. For patient 1 (see the text), the homozygous deletion of 135Kbp covering the last two exons of ANOS1 is clearly visible in the WGS analysis but less evident in the WES analysis (called by CoM but not detected by ED). Below the respective coverage as reported by CoM in the genomic space for WGS and WES data. **(B)** For patient 2 the partial heterozygous deletion of 115Kbp in SCN1A, detected from WES, is clearly visible in the exonic and genomic spaces. **(C)** Homozygous deletion of exon 7 in SMN1 in patient 3, detected in WES data, is clearly visible in the exonic and genomic spaces. It is worth noting that, in the genomic space, the coverage profile seems to show two other exons with a drop in coverage. The control, dashed line in the plot above, shows the same profile indicating a fluctuation of the coverage in this region, likely independent from the number of copies, or a common deletion.

exon coverage. From this perspective, the causative CNV appears as a full two exons deletion of <100 bp in the exonic space, detected by CoM but not by ED. Therefore, CoM can be used to perform an efficient clinical analysis of WGS data in a two-step approach: first, through a high-resolution (100–200 bp) WES profile and second, through a broad investigation of the genomic regions presenting with positive SNV calls from the first step.

Patient 2 is an, 8-year-old female child, diagnosed with drug-resistant epilepsy with febrile seizures. WES single nucleotide variant analysis did not provide any candidate on a panel of 478 genes related to epilepsy (Epilepsy MDG-1204.01, https://www.medigenome.ch/en/gene-panels/).

CoM reported a heterozygous deletion of ~120Kbp partially overlapping the last 10 exons of SCN1A (Figure 3b). The sodium channel 1A is associated with generalized epilepsy with febrile seizures, Type 2 [OMIM 604403]. Deletions in this gene are known to cause seizure disorders, ranging from early-onset isolated febrile seizures to generalized epilepsy [24].

Patient 3 is a 3-year-old female child with a suspicion of spinal muscular atrophy. WES analysis and array CGH were negative but CoM identified a homozygous deletion of the exon 7 (112 bp - Figure 3c). This deletion, confirmed by MLPA but not detected by ED, is the most frequent CNV related to SMN1-induced muscular atrophy [25] [SMA OMIM 253400]; this deletion was eventually considered as the pathogenic cause of the phenotype of the patient by the clinicians.

Discussion

CoM is an NGS coverage based CNV calling algorithm designed to work at nucleotide resolution with WES and WGS data. The capacity to analyze a given coverage signal in different scale sizes, combined with the nowadays availability of numerous controls in standard clinical batches, enables the detection of multi-sized clinically relevant deletions or duplications and in particular the detection of the so far elusive small CNVs of <5Kbp. The algorithm has been designed to reduce the analysis burden by using all available control datasets to eliminate frequent CNVs and stochastic coverage variations. We have proven the effectiveness of CoM in comparison to ExomeDepth and others broadly used *in silico* CNV callers. Performance wise, CoM is not the fastest algorithm available but in line with the state of the art (Supplementary Table 1). With 10 control samples, CoM takes 6 h to analyze a full gene panel of 20 400 genes and around 1 h to process a WES panel of 4758 clinically relevant genes from OMIM and the Clinical Genomic Database ([26], https://research.nhgri.nih.gov/CGD/) on a 16 cores machine with 32Gbyte of RAM. The analysis time, however, can be sensibly reduced by iteratively increasing the number of controls and, consequently, reducing the number of False Positives (Supplementary Figure 2). CoM, in common with all other read-depth based algorithms, is sensitive to coverage variations induced by different hybridization kits and sequencing processes. Indeed, mismatches between samples, reference and controls can lead to a consistent increase of the number of False Positives. Nowadays, this problem is less compelling given that even small labs process hundreds of WES before updating the production lines. One caveat concerns the ethnicity of patients and controls and the interpretation of CoM results. A CNV can be frequent in a specific region or population and rare elsewhere. Therefore, as for single nucleotide variants, the ethnicity of the patient must be taken into account to reach an appropriate diagnostic [27]. Future developments on CNV detection will deal with WGS as the standard technology for genetic clinical applications [28] and long reads as leading approach for SV detection [29]. We show that CoM can already be used to analyzed WGS data (Figure 3) and, in principle, there is no limitation to employ it on long-read data. A possible improvement might involve the integration of CoM CNV search and zooming process with split-read detectors to provide precise breakpoint detection for large CNVs. It is crucial information needed to understand the impact of the CNV on patient phenotype especially on cancer [30]. Of note, we are planning to apply CoM on tumor samples in the next future. Concerning WES data, CoM demonstrated to be superior to the current state-of-the-art algorithms in the detection of rare and small CNVs in simulated and clinical data and it can be a valid and inexpensive alternative to MLPA and array CGH in clinical settings.

Key Points

- CoverageMaster (CoM) is designed to identify CNVs of any size at nucleotide resolution through multiscale analysis.
- Simulated and clinical data show that CoM significantly increased CNV call sensitivity with respect to the state of the art, especially in the lower size spectrum (50–1000 bp).
- CoM can analyze whole exome or whole genome sequencing data.
- The analysis at nucleotide resolution enables the visualization of the identified CNVs in the exonic and genomic space (Genome Browser) to further support the clinical interpretation of the calls.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank Professor Nelly Pitteloud, Lucia Bartoloni and Alexia Boizot from the Endocrine Diabetes and Metabolism service of the CHUV for helping with the sample preparation and analyses, Marco Belfiore from Genesupport for providing the array CGH data and Xavier Blanc from Medigenome for constructive discussions and suggestions.

Funding

This study was supported by the EU Framework Programme for Research and Innovation Action (RIA), Horizon 2020, n°847941 (miniNO) and partially by the Swiss National Science Foudation (310030_185292) and Novartis Foundation (18AO52) to F.A.S.

Data availability

CoverageMaster is available at https://github.com/fredsanto/coverageMaster.

Contributions

M.R. contributed in developing the algorithm, performed the tools comparison and wrote the manuscript. Y.Z. performed WGS analyses. J.M. performed samples preparation for WGS sequencing and quality assessment. E.R. and S.E.A. analyzed CGH data and provided the clinical analysis of the samples reported in the study. F.A.S. designed and supervised the study, wrote the algorithm and wrote the manuscript. All authors contributed to the manuscript.

References

1. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009;**1**:62.
2. Truty R, Paul J, Kennemer M, *et al.* Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet Med* 2019;**21**:114–23.
3. Zack TI, Schumacher SE, Carter SL, *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013;**45**:1134–40.
4. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
5. Stuppia L, Antonucci I, Palka G, *et al.* Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases. *Int J Mol Sci* 2012;**13**:3245–76.
6. Rieber N, Zapatka M, Lasitschka B, *et al.* Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* 2013;**8**:e66621.
7. Marshall CR, Bick D, Belmont JW, *et al.* The medical genome initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. *Genome Med* 2020;**12**:48.
8. Shigemizu D, Miya F, Akiyama S, *et al.* IMSindel: an accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Sci Rep* 2018;**8**:5608.
9. do Nascimento F, Guimaraes KS. Copy number variations detection: unravelling the problem in tangible aspects. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:1237–50.
10. Sathirapongsasuti JF, Lee H, Horst BA, *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;**27**:2648–54.
11. Krumm N, Sudmant PH, Ko A, *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;**22**:1525–32.
12. 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* 2012;**28**:2747–54.
13. Tan R, Wang Y, Kleinstein SE, *et al.* An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* 2014;**35**:899–907.
14. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
15. Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**:11 10 11–33.
16. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
17. Zarrei M, MacDonald JR, Merico D, *et al.* A copy number variation map of the human genome. *Nat Rev Genet* 2015;**16**:172–83.
18. Xing Y, Dabney AR, Li X, *et al.* SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes. *Front Genet* 2020;**11**:82.
19. Jiang Y, Wang R, Urrutia E, *et al.* CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol* 2018;**19**:202.
20. Gordeeva V, Sharova E, Babalyan K, *et al.* Benchmarking germline CNV calling tools from exome sequencing data. *Sci Rep* 2021;**11**:14416.
21. Li J, Lupat R, Amarasinghe KC, *et al.* CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;**28**:1307–13.
22. Wang C, Evans JM, Bhagwate AV, *et al.* PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics* 2014;**30**:2678–80.
23. Franco B, Guioli S, Pragliola A, *et al.* A gene deleted in Kallmann's syndrome shares homology with neural cell adhesion and axonal path-finding molecules. *Nature* 1991;**353**:529–36.
24. Parihar R, Ganesh S. The SCN1A gene variants and epileptic encephalopathies. *J Hum Genet* 2013;**58**:573–80.
25. Ogino S, Wilson RB. Genetic testing and risk assessment for spinal muscular atrophy (SMA). *Hum Genet* 2002;**111**:477–500.
26. Solomon BD, Nguyen AD, Bear KA, *et al.* Clinical genomic database. *Proc Natl Acad Sci U S A* 2013;**110**:9851–5.
27. White SJ, Vissers LE, Geurts van Kessel A, *et al.* Variation of CNV distribution in five different ethnic populations. *Cytogenet Genome Res* 2007;**118**:19–30.
28. Stranneheim H, Lagerstedt-Robinson K, Magnusson M, *et al.* Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med* 2021;**13**:40.
29. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. *Nat Rev Genet* 2021;**22**:572–87.
30. van Belzen IAEM, Schönhuth A, Kemmeren P, *et al.* Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol* 2021;**5**:15.