OXFORD GENETICS

GENETICS, 2022, 222(4), iyac147
https://doi.org/10.1093/genetics/iyac147
Advance Access Publication Date: 28 September 2022
Investigation

BMI-CNV: a Bayesian framework for multiple genotyping platforms detection of copy number variants

Xizhi Luo [1], Guoshuai Cai [2], Alexander C. Mclain [1], Christopher I. Amos [3], Bo Cai [1,*], Feifei Xiao [4,*]

[1]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA,
[2]Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA,
[3]Department of Quantitative Sciences, Baylor College of Medicine, Houston, TX 77030, USA,
[4]Department of Biostatistics, University of Florida, Gainesville, FL 32603, USA

*Corresponding author: Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, 915 Greene St., Columbia, SC 29208, USA. Email: bocai@mailbox.sc.edu. *Corresponding author: Department of Biostatistics, College of Public Health and Health Professions and College of Medicine, University of Florida, 2004 Mowry Rd., CTRB building Room 5227, Gainesville, FL 32603, USA. Email: feifeixiao@ufl.edu.

Abstract

Whole-exome sequencing (WES) enables the detection of copy number variants (CNVs) with high resolution in protein-coding regions. However, variants in the intergenic or intragenic regions are excluded from studies. Fortunately, many of these samples have been previously sequenced by other genotyping platforms which are sparse but cover a wide range of genomic regions, such as SNP array. Moreover, conventional single sample-based methods suffer from a high false discovery rate due to prominent data noise. Therefore, methods for integrating multiple genotyping platforms and multiple samples are highly demanded for improved copy number variant detection. We developed BMI-CNV, a Bayesian Multisample and Integrative CNV (BMI-CNV) profiling method with data sequenced by both whole-exome sequencing and microarray. For the multisample integration, we identify the shared copy number variants regions across samples using a Bayesian probit stick-breaking process model coupled with a Gaussian Mixture model estimation. With extensive simulations, BMI-copy number variant outperformed existing methods with improved accuracy. In the matched data from the 1000 Genomes Project and HapMap project data, BMI-CNV also accurately detected common variants and significantly enlarged the detection spectrum of whole-exome sequencing. Further application to the data from The Research of International Cancer of Lung consortium (TRICL) identified lung cancer risk variant candidates in 17q11.2, 1p36.12, 8q23.1, and 5q22.2 regions.

Keywords: copy number variation; whole-exome sequencing; Bayesian; multiplatform integration; multisample inference

Introduction

Copy number variants (CNVs) are a major type of structural variants comprised of deletions and duplications of the genomic segments. They play an important role in complex diseases, such as cancer, muscle diseases, and neuropsychiatric diseases (Shlien and Malkin 2009; Välipakka *et al.* 2017; Takumi and Tamada 2018). In addition, recurrent and common CNVs have also been revealed to be risk factors for many diseases, such as autism spectrum disorders and schizophrenia (Moreno-De-Luca *et al.* 2010) and cardiovascular disease (Wang *et al.* 2019).

With the dramatic growth and the accompanying cost drop in sequencing technologies, massive whole-exome sequencing (WES) datasets have been generated from large-scale biomedical studies, which allows for the identification of genomic variants in functional protein-coding regions (Amos *et al.* 2017). However, exons only encompass 1% of the genome, limiting the possibility to investigate the impact of genetic variations such as CNVs located in the noncoding regions (Venter *et al.* 2001). Therefore, cohort projects often used both WES and SNP arrays for a combined genome-wide scanning and fine exome scanning. For example, the international Transdisciplinary Research In Cancer of the Lung (TRICL) consortium genotyped 2,003 subjects with both WES and SNP array data (Amos *et al.* 2017), the Alzheimer's Disease Genetics Consortium (ADGC) and the Alzheimer's Disease Sequencing Project (ADSP) (Karch *et al.* 2016; Beecham *et al.* 2017) have also collected such multiplatform data. Consequently, the demand for multiplatform (e.g. WES and SNP) integration methods that accurately study CNV in a full-coverage manner has dramatically increased but is still unfulfilled. Another challenge is that WES is subject to the nonuniform coverage of sequence reads in the assembly procedure due to the existence of short duplications or deletions, resulting in many dropped-out segments that were originally mapped to the exome. For example, Fang *et al.* (2014) found that more than 16% of the exons cannot be captured by WES experiments, losing the opportunity to detect short CNVs. Theoretically, integrating the information from SNP array can overcome this challenge and improve calling accuracy.

Similar efforts, such as iCNV, have been made by Zhou *et al.* (2018) for integrative segmentation. In iCNV, data from different platforms were first normalized and standardized and then jointly segmented by a Hidden Markov Model (Zhou *et al.* 2018). This method presented a significant boost in accuracy compared

Received: August 2, 2022. **Accepted:** September 8, 2022

to WES. However, it only used information from a single sample. As we know, technological and biological factors are prominent in real data and increase the variations and noise in data intensities, leading to unreliable findings with single-sample scanning of CNVs. Consequently, multiple sample strategies previously introduced for detecting common CNVs have great potential to improve the robustness and detection power with noisy data. Such a direction has been supported by various existing studies (Zhang *et al.* 2010; Siegmund *et al.* 2011; Klambauer *et al.* 2012; Song *et al.* 2016), but none has focused on the direction of multiplatform integration. Moreover, the most widely used WES methods, such as CODEX2 and EXCAVATOR, are also used for single-sample scanning (Magi *et al.* 2013; Jiang *et al.* 2018).

Collectively, the development of a full-spectrum CNV detection method that can achieve high performance using comprehensive information from all samples and all platforms is essential. In this study, we developed BMI-CNV, a Bayesian Multisample and Integrative CNV calling method, as the first of its kind to enhance the detection of both intergenic or intragenic and common CNVs. The significant improvement of BMI-CNV on CNV calling was demonstrated by simulation studies (Simulations showed superior performance of BMI-CNV in multiplatform integrative analysis, Simulations showed superior performance of BMI-CNV in single platform analysis) as well as the analyses of real datasets from the HapMap project (Altshuler *et al.* 2010) and the 1000 Genomes Project (Application to the 1000 Genomes Project and HapMap data) (Auton *et al.* 2015). The further application of BMI-CNV to the international TRICL dataset identified new lung cancer-associated CNVs (Integrative CNV detection and association analyses with TRICL dataset). This new method has a wide scope of applications and has great potential to be further extended to profile CNVs for whole-genome sequencing and single-cell sequencing data.

Materials and methods

Our method mainly focuses on CNV detection by integrating the SNP array and WES data, which can also be naturally applied to single platform data (Numerical simulations). The overview of the framework is shown in Fig. 1. First, WES read counts and SNP array data are integrated using a series of data integration procedures. The main segmentation algorithm consists of 2 stages: Stage I uses a Bayesian probit stick-breaking process (PSBP) method (Stage-I: shared CNV inference by Bayesian probit stick-breaking process model) coupled with a Gaussian mixture model-based initial data filtering (Supplementary A.2) to identify shared CNV regions, and Stage II uses the individual CNV calling procedure to identify sample-specific CNVs (Stage-II: Individual CNV calling).

Data description and multiplatform integration

First, we performed platform-specific normalization procedures for the original data. For WES data, with raw read depth data from test and external negative control samples, we used the exon mean read count (EMRC; Magi *et al.* 2013). EXCAVATOR2 median normalization procedure was utilized to mitigate the effects from 3 observed sources of bias: exon length, GC-content, and mappability (D'Aurizio *et al.* 2016). All negative control samples were pooled by averaging the EMRC on each exon across all samples and the logarithm between the ratio of test and pooled control samples was calculated, referred to as the log_2R intensities. These log_2R data were then processed by the lowess-scatter plot procedure to adjust read-depth differences between testing and control samples and remove coverage-dependent bias. For array data, PennCNV was used to adjust the genomic wave on genetic intensities (i.e. Log R ratio [LRR]) (Wang *et al.* 2007).

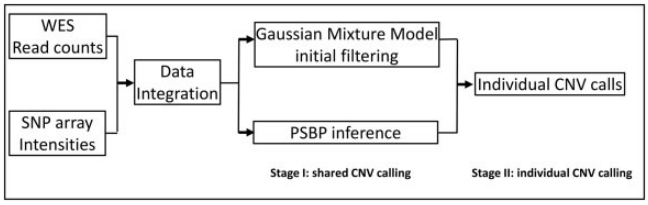

Fig. 1. Analysis workflow of BMI-CNV. BMI-CNV uses 2 inputs: (1) WES raw read count data from testing and control samples that are computed by using genotyping tools such as SAMtools; and (2) SNP array intensities. WES read counts are normalized to correct exon length, GC-content, and mappability biases. Logarithms of normalized values between testing and pooled control samples are calculated. SNP array intensities are normalized to adjust the genomic waves. The WES and SNP array data are standardized by a robust scaling approach and then integrated. For CNV calling, BMI-CNV carries out a 2-stage framework to generate CNV calls. In stage I, an initial data filtering procedure is coupled with a Bayesian PSBP method to identify shared CNV regions. In stage II, an individual CNV calling procedure is performed to call CNVs in each sample.

To combine the SNP array LRR data and the WES log_2R data, we standardized each dataset via a robust scaling approach (Rousseeuw and Croux 1993). Compared to the conventional standardization method, the robust scaling approach used median and interquartile ranges to mitigate the influence from potential outliers and signals from double deletions (Details in Supplementary A.1). The WES and SNP array data were then merged by chromosomal coordinates for joint segmentation (details in Supplementary A.1).

Notations and basic model

Let Y denote a $n \times m$ data matrix obtained from the precalling procedures described above, where Y_{ij} represented the processed genetic intensities (e.g. LRR from array or log_2R from WES) for the j—th ($j = 1, 2, \ldots, m$) marker (e.g. SNPs from array or exons from WES) in sample i ($i = 1, 2, \ldots, n$). We assumed a classic normal kernel for Y_{ij},

$$Y_{ij} \sim N(Y_{ij}|\phi_{ij}).$$

$\phi_{ij} = (\mu_{ij}, \sigma^2_{ij})$ was an unknown matrix of the underlying mean and variance. Different values of ϕ_{ij} indicated the existence of different copy number states. Five copy number states, including the deletion of single copy or double copies, diploid, duplication of a single copy or double copies were assumed in our study. We considered τ to be a change point for sample i if $\phi_{i,\tau} \neq \phi_{i,\tau+1}$. The goal is to estimate the locations of change points, CNV segments can then be generated by connecting adjacent change points.

To achieve this, we estimate change points shared by all samples and then identify their individual carriers. We use a 2-stage method: Stage I with a probit stick-breaking process (PSBP) to identify shared CNV regions across samples (Stage-I: shared CNV inference by Bayesian probit stick-breaking process model) and Stage II calling sample-specific CNVs individually (Stage-II: Individual CNV calling). We also initially filter samples without CNVs in Stage I (Supplementary A.2).

Stage-I: shared CNV inference by Bayesian probit stick-breaking process model

Let $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})^T$ denote the genetic intensity for the jth marker ($j = 1, 2, \ldots, m$) across all samples. The classic normal kernel can be rewritten as

$$\mathbf{Y}_j \sim N(\mathbf{Y}_j|\phi_j).$$

$\phi_j = (\mu_j, \sigma_j^2)$ represented the shared mean and variance scalars of position j across all samples, the parameters in which needed to be estimated. We modeled the corresponding latent means and variances using the Bayesian PSBP model (Chung and Dunson 2009; Rodriguez and Dunson 2011). The favorable shrinkage property of PSBP allows for efficiently clustering of all ϕ_j's to a small number of clusters (i.e. copy number states). Moreover, the PSBP mixture model can capture multimodal and heavy-tailed distribution, which relaxed the normality assumption of the latent means, providing more flexible scenarios for modeling the complex CNV data. Specifically, we assumed ϕ_j followed an unknown distribution $G \sim PSBP(\alpha G_0)$ with centering distribution G_0 where the shape measure α reflected how far away the random distribution is from the center. Following Rodriguez and Dunson (2011), G admitted a representation of the form:

$$\phi_j \sim G(.) = \sum_{l=1}^{L} \omega_l \delta_{\theta_l}(.) \quad (1)$$

where L represented the number of all possible copy number states (e.g. $L = 5$), $\theta_l = (\mu_l, \sigma_l^2)$ were possible distinct mean and variance specific to each copy number state ($l = 1, 2, \ldots, L$), $\delta_{\theta_l}(.)$ was a degenerate distribution at θ_l, and $\omega_l = \Phi(\alpha_l)\prod_{r<l}(1-\Phi(\alpha_r))$ represented the probability of assigning θ_l to each position where $\Phi(.)$ was the probit function and $\alpha_l \sim N(\mu_\alpha, \sigma_\alpha^2)$. Following this structure, ϕ_j was assigned to one of the $\{\theta_l\}$ based on the observed intensities across all potential carriers of the copy number state for locus j, where the carriers were initially identified using a Gaussian mixture model-based strategy described in Supplementary A.2. To simultaneously implement the variable selection and clustering procedures for the purpose of CNV detection, we further reconstructed the PSBP model (George and McCulloch 1993; Cai and Bandyopadhyay 2017):

$$\phi_j \sim \gamma_j G_{\mu=0} + (1-\gamma_j)G(.) \quad (2)$$

where the $G_{\mu=0}$ was the underlying distribution of the normal copy number states with the mean fixed at zero (i.e. diploids). $\gamma_j \sim Bernoulli(\kappa)$ was an indicator of ϕ_j being in $G_{\mu=0}$ (i.e. normal state) or not, which incorporated variable selection of the locus across samples. Specifically, when $\gamma_j = 1$, ϕ_j followed a distribution $G_{\mu=0}$; $\gamma_j = 0$ indicated a potential CNV locus following $G(.)$ defined in equation (1). Within this framework, the posterior probability of ϕ_j being $G_{\mu=0}$ or not was calculated through inference on γ_j (Supplementary A.3–A.4). In this way, each shared change-point is the position such that $\gamma_{\tau_t} \neq \gamma_{\tau_{t+1}}$ and locations of all shared change-points (i.e. $\tau = \{\tau_1, \tau_2, \ldots, \tau_T\}$) can be identified. In order to efficiently perform the posterior inference for all the parameters, we developed a Markov Chain Monte Carlo (MCMC) algorithm relying on a modification of the Gibbs sampler (Ishwaran and James 2001). With the proper choices of priors and hyperpriors, all full conditional distributions of the parameters can be analytically derived which ensured the computational speed and the convergence to the true posterior distributions (Supplementary A.3–A.4).

Stage-II: Individual CNV calling

Note that for shared CNV regions identified in the stage I, only a proportion of the samples carry the CNV for each region. We then determined the carriers for each shared CNV, that is, to call CNVs in each sample. Specifically, using the posterior mean and variance estimates specific to each copy number state (i.e. $\hat{\mu}_l$ and $\hat{\sigma}_l^2$, $l = 1, \ldots, 5$), we constructed the interval for each state as $C_l = [\hat{\mu}_l - c_{1l}\hat{\sigma}_l, \hat{\mu}_l + c_{2l}\hat{\sigma}_l]$. The individual segment would be classified into *l—th* copy number state if its segmental mean fell within one specific interval C_l. Here, C_3 represented the interval for diploid (i.e. noncarrier). Values of c_1 and c_2 should be carefully chosen according to empirical evidence about the magnitude of mean shifts of each CNV state, which may vary by genotyping platforms. In practice, to provide calibration of c_1 and c_2 and optimize the performance of our method, we will suggest users to initiate a pilot study and plot the genotyping signals of CNV segments identified under different combinations of c_1 and c_2 for visualization. True positive rate (TPR) can be calculated for each combination and the optimal choices for c_1 and c_2 will achieve the highest TPR. In our applications (Numerical simulations, Application to the 1000 Genomes Project and HapMap datasets, Analysis of the TRICL consortium case-control dataset), $c_1 = c_2 = 1.2$ were used.

Numerical simulations

To evaluate the performance of our method, we conducted simulations under various settings. Four copy number states were simulated, including single copy deletion (del.S), double copy deletion (del.D), single copy duplication (dup.S), and double copy duplication (dup.D). The CNV length (i.e. SNPs and exons) varied from 10–30 markers, 30–60 markers, and 60–100 markers. The CNV population frequency was 20%, 50%, or 100%, respectively.

First, we evaluated our method when both WES and SNP array data were available. For WES data, to generate data retaining the true noise background and exon distribution, we conducted a spike-in design (Jiang *et al.* 2018; Zhou *et al.* 2018). We started with read-depth data on chromosome 1 in 81 samples from the 1000 Genomes Project (Auton *et al.* 2015). Exons harboring CNVs identified by EXCAVATOR2 and CODEX2 and reported in the Database of Genomic Variants (DGV) were initially removed (MacDonald *et al.* 2014; D'Aurizio *et al.* 2016; Jiang *et al.* 2018). The read depth data of the remaining exons were treated as WES random noise background. We multiplied the background read depth by $c/2$, where c was sampled from a normal distribution with mean and variance provided in Supplementary A.5. For SNP array data, we utilized the similar strategy used in Xiao *et al.* (2019) to simulate intensities, which mimicked the real data from the Altshuler *et al.* (2010). 50 dispersed CNV segments of varying length and frequency were then randomly selected and spiked randomly in coding regions (i.e. exonic CNVs) or noncoding regions (i.e. intergenic or intragenic CNVs).

Our method was compared to the only existing multiplatform integrative method iCNV (Zhou *et al.* 2018). The performance of these methods was assessed by precision rate, recall rate, and F1 score (Supplementary A.5). We also evaluated the performance of our method in the single-platform mode when only WES data was available. In this scenario, only exonic CNVs were simulated and our method was compared against 3 commonly used CNV detection methods CODEX2 (Jiang *et al.* 2018), EnsembleCNV (Zhang *et al.* 2019), and EXCAVATOR2 (D'Aurizio *et al.* 2016); a multisample based method, cn. MOPS (Klambauer *et al.* 2012); and iCNV in the single-platform mode (Zhou *et al.* 2018).

Application to the 1000 Genomes Project and HapMap datasets

We analyzed the same 81 individuals with SNP array and WES data from the 1000 Genomes Project and the international

HapMap consortium. A detailed description of the experimental samples and genotyping platforms was provided in previous literature (Altshuler *et al.* 2010; Auton *et al.* 2015). Raw read counts and SNP array data were processed and normalized to generate log_2R and LRR intensities. For the WES data, we arbitrarily selected 4 normal samples from the 1000 Genomes Project as negative controls (details in Supplementary A.6). The posterior inference of BMI-CNV was based on 2,000 MCMC samples with a burn-in period of 500 iterations.

Analysis of the TRICL consortium case-control dataset

We further applied BMI-CNV to the international lung cancer consortium TRICL (Amos *et al.* 2017), which consists of 1,163 samples genotyped by both OncoArray and WES data (i.e. integrative analysis mode), and 829 samples that only had WES data (i.e. single platform analysis mode) (details in Supplementary A.6 and A.7).

CNV calls were annotated by known gene regions obtained from the University of California, Santa Cruz (UCSC) Genome Browser (Kent *et al.* 2002). A gene-based association test was performed to investigate the influence of CNVs on lung cancer susceptibility:

$$logit(P(disease = 1)) = \beta_0 + \beta_{del}DEL + \beta_{dup}DUP + \boldsymbol{\beta} + \boldsymbol{\beta}' \sum_{i=1}^{4} PC_i. \quad (3)$$

DEL and DUP were 2 indicator variables for deletions and duplications separately. We adjusted the covariates, including smoking status (ever/never), age, and gender, ancestry-related PCs representing the top principal components (Patterson *et al.* 2006). In addition to studying overall lung cancer risk, we also performed stratification analyses by histological types of lung cancer [squamous cell lung cancer (SQC) and lung adenocarcinoma (LUAD)]. The effects from deletions and duplications were tested via the Wald test, and all nominal *P*-values were adjusted by the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995).

Results

Simulations showed superior performance of BMI-CNV in multiplatform integrative analysis

We first evaluated the performance of BMI-CNV with simulated data. In various simulation settings including different CNV sizes and population frequencies, BMI-CNV outperformed iCNV in all scenarios with higher F1 scores (Fig. 2, Supplementary Table 1). iCNV tended to be conservative compared to our method, which maintained a high precision rate, although the recall rate was compromised. For example, when the simulated CNVs had a length of 30–60 markers and the population frequency was 20%, BMI-CNV had a precision at 0.70, a recall rate at 0.83, and an F1 score at 0.76. The corresponding values for iCNV were 0.99, 0.37, and 0.54, respectively. Moreover, at a fixed CNV size, the performance of BMI-CNV was improved when CNV frequencies increased from 20% to 100%, achieving the highest F1 score when all the samples were carriers, whereas the performance of the iCNV method was not sensitive to the CNV frequencies.

We further assessed the performance of BMI-CNV to detect CNVs in different regions as 25% of simulated CNVs were mainly located in the intergenic or intragenic regions which are difficult to be detected for methods using WES data alone. By integrating available SNP array data, BMI-CNV can identify most of these CNVs, and it still outperformed iCNV in all scenarios (Supplementary Table 2). For example, when the simulated CNVs had 60–100 markers and the population frequency was 20%, BMI-CNV detected 94% of the CNVs, and iCNV only detected 59%. For computational speed, our method took about 280 min to screen a chromosome with 90,739 markers from 81 samples based on 2,000 MCMC sampling runs which was 54 min for iCNV, 20 min for EXCAVATOR2, 70 min for CODEX2, and 30 min for cn. MOPs. The computation was performed on a regular laptop with an Intel Core i7 processor and 24.00 GB of RAM.

Simulations showed superior performance of BMI-CNV in single platform analysis

For single platform analysis, we assessed the performance of BMI-CNV benchmarking against existing WES methods. As a result, the performance of our method was superior in detecting medium and long CNVs reflected by the largest F1 scores (Fig. 3, Supplementary Table 3). iCNV, cn. MOPS, and EXCAVATOR2 tended to be conservative, as they achieved high precision rates but with significant sacrifice on recall rates. It is also noteworthy when the CNV size was fixed, the performance of BMI-CNV and CODEX2 were both improved with increased CNV frequencies, achieving the highest F1 scores when all the samples were carriers, and cn. MOPS tended to be conservative when CNV frequencies increased. Still, the performance of EXCAVATOR2 and iCNV were not subject to CNV frequencies, which was expected as the shared information from multiple samples was not utilized.

In conclusion, the BMI-CNV method, integrating information from multiple samples, presented evidence of superior performance in common CNV detection for both multiplatform integration and single-platform analyses. By incorporating SNP array data, BMI-CNV enabled the accurate detection of intergenic or intragenic CNVs for integrative analysis.

Application to the 1000 Genomes Project and HapMap data

We applied BMI-CNV to the public datasets from the 1000 Genomes Project and HapMap data. In total, 37,213 CNVs were identified from 81 samples (Fig. 4), 28% of which have been previously reported by DGV (MacDonald *et al.* 2014). Most CNVs tended to be short (<20 markers) and had a frequency of less than 50%. Supplementary Fig. 1 showed the summary of CNVs, which suggested no significant difference in CNV length and frequency between deletions and duplications. Moreover, by integrating the SNP array data, our method retrieved 4,418 CNVs that were located in the noncoding regions which were missed by methods using WES data alone.

A clear pattern of the signal in the shared deletion suggested high detection accuracy of our method. We illustrated a common deletion region in Supplementary Fig. 2, suggesting that 27 out of 81 samples were carriers of this variant. Besides, we explored an alternative data integration strategy that only used intronic SNPs from the array data. Compared to the main method using all SNPs, a lower concordance rate with DGV (26% vs 28%) was observed, implying that utilizing information from all SNPs slightly improved the detection accuracy (Fig. 4, Supplementary Fig. 3).

Integrative CNV detection and association analyses with TRICL dataset

With the TRICL datasets, we identified 253,183 autosomal CNVs from 1,992 samples (Fig. 5) in total. Overall, there was no

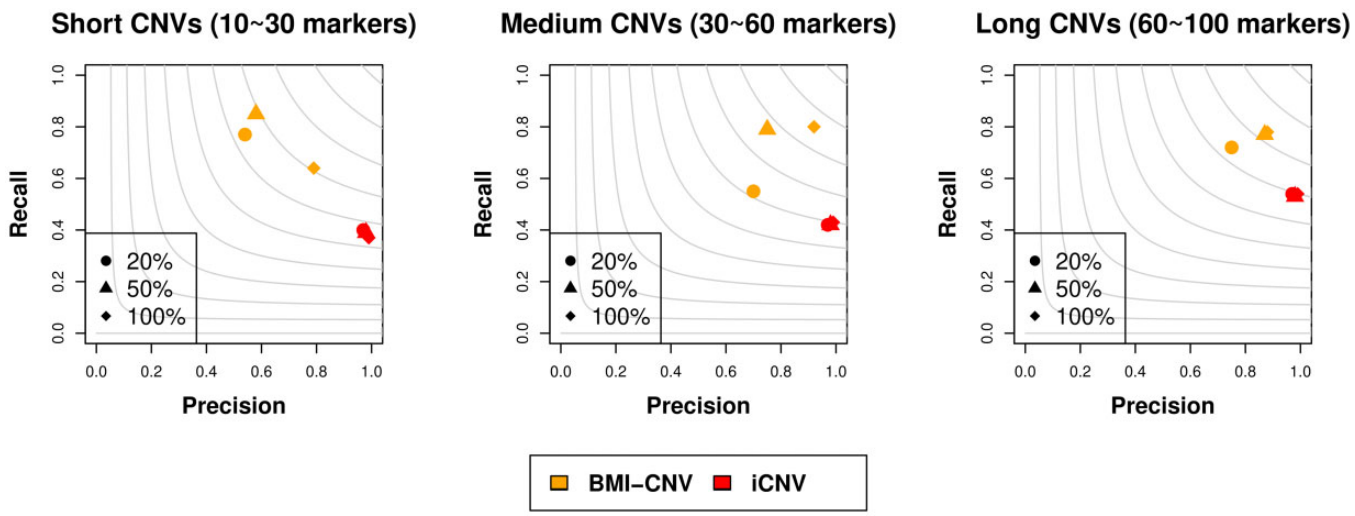

Fig. 2. Performance assessment of BMI-CNV and iCNV on simulated data in the integrative analysis. Simulated CNVs were of frequency 20%, 50%, and 100% and length 10–30 markers (short), 30–60 markers (medium), and 60–100 markers (long). The grey contours are F1 scores calculated as the harmonic mean of precision and recall rates.

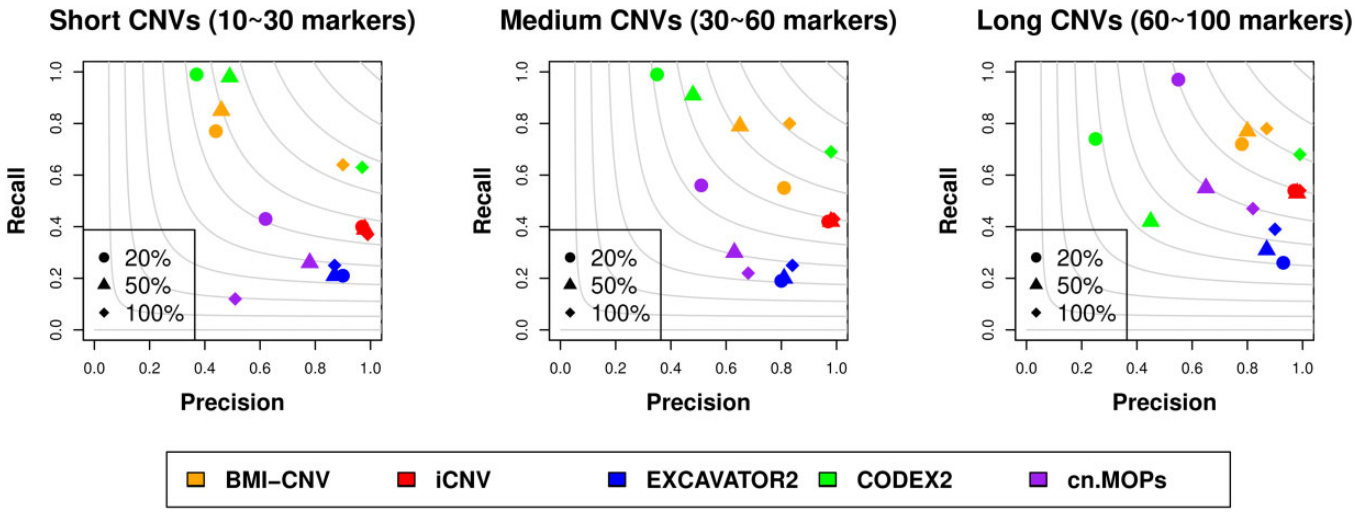

Fig. 3. Performance assessment of BMI-CNV, iCNV, EXCAVATOR2, CODEX2, cn. MOPs and EnsembleCNV on simulated data in WES analysis. Simulated CNVs are of frequency 20%, 50%, and 100% and length 10–30 markers (short), 30–60 markers (medium), and 60–100 markers (long). The grey contours are F1 scores calculated as the harmonic mean of precision and recall rates.

significant difference in the size of detected deletions or duplications (31.46 kb vs 32.51 kb), but the deletions covered more markers than the duplications (13.46 markers vs 8.72 markers) (Supplementary Table 4). No significant difference was observed in the overall proportion of deletions and duplications between cases and controls (49% vs 51% for deletions and 52% vs 48% for duplications, respectively).

The identified CNVs were mapped to 3,472 genes. Association tests in SQC subgroup highlighted the deletion gene *LGALS9* in 17q11.2 region (OR = 4.14, 95% CI = 1.65–10.38, *P*-value = 0.002), the duplication genes *HSPG2* in 1p36.12 region (OR = 4.79, 95% CI = 1.75–13.10, *P*-value = 0.002) and *EIF3E* in 8q23.1 region (OR = 2.19, 95% CI = 1.31–3.64, *P*-value = 0.003). Association in the LUAD subgroup identified the duplication gene *YTHDC2* in 5q22.2 region (OR = 2.88, 95% CI = 1.62–5.12, *P*-value = 0.0003), which was also identified in the overall lung cancer risk model by adjusting the histological subtypes as a covariate (Supplementary Table 5). The intensity plots indicated those were valid CNV segments that showed distinct data patterns from other noncarriers and adjacent regions (Supplementary Fig. 4). Although these genes became not significant after multiple comparison adjustments, they still provided potential evidence for further studies and great insight into revealing the role of CNVs in lung cancer risk.

Discussion

The importance of CNVs for elucidating the mechanism underlying many diseases has been increasingly remarked upon. Improving the accuracy of CNV detection is fundamental for downstream CNV-disease risk association and diagnostic classification. In this study, we developed a novel multiple sample-based method, BMI-CNV, to improve common CNV detection with WES data, allowing for the integration of available SNP array data. The simulation results demonstrated the desirable performance across different scenarios of CNV sizes and population frequencies. The improvement for calling long and high-frequency CNVs was the most substantial. We analyzed the WES

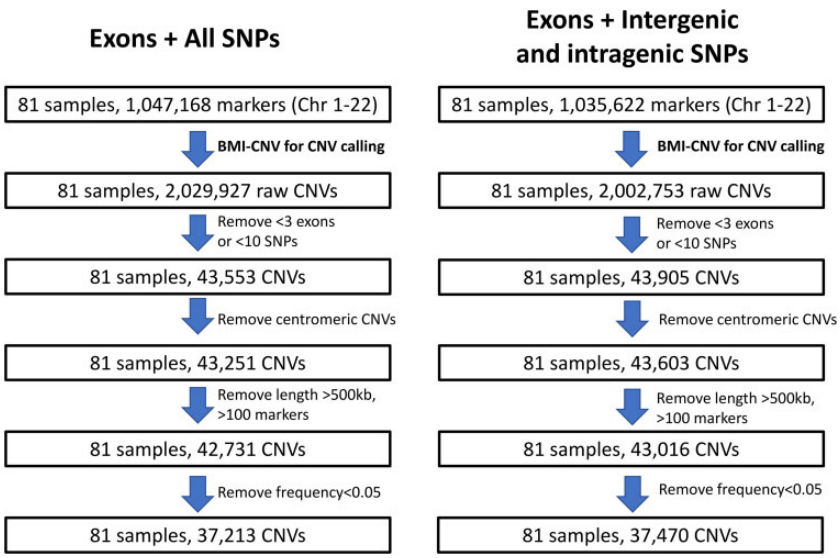

Fig. 4. Overview of the application to the 1000 Genomes Project and HapMap data. The figure outlines the study design with a brief description of quality control (QC) methods. Summary of key results includes the sample size and number of CNVs at various stages of analysis. Left: CNV calling results using all SNPs and exons; right: CNV calling results using intergenic and intragenic SNPs and exons. Chr, chromosome.

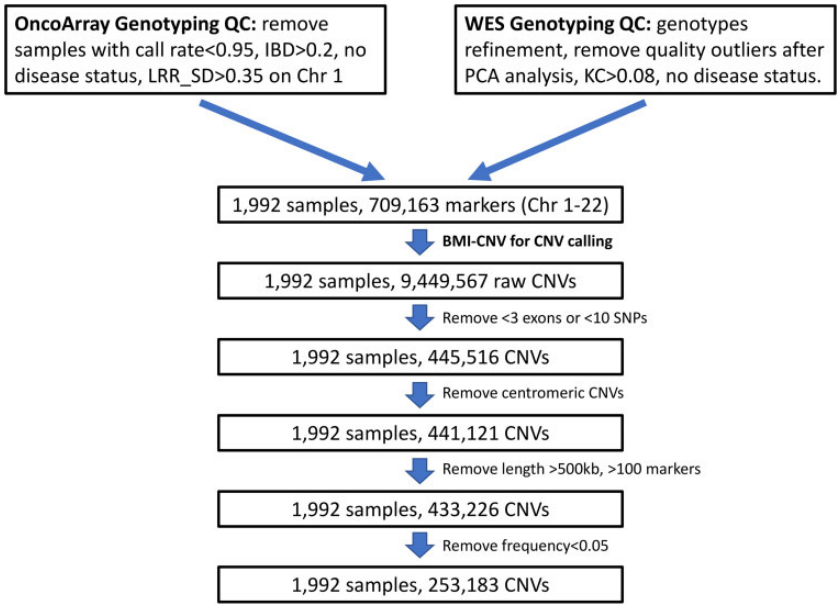

Fig. 5. Overview of the integrative analysis of the TRICL case-control study. The figure outlines the study design with a brief description of quality control (QC) steps. The summary of the key results includes the sample size and number of CNVs at various stages of analysis. IBD, identical by descent; KC, kinship coefficient; LRR SD, standard deviation of Log R ratio; Chr, chromosome; PCA, principal component analysis.

data from the 1000 Genomes Project and SNP array data previously generated by the HapMap project 3 and demonstrated the advantage of multiplatform integration over the single-platform analysis. Finally, our application of BMI-CNV to WES and OncoArray datasets of the TRICL consortium indicated potential lung cancer-associated CNVs.

Among the top associations with the TRICL study, the discovered significantly associated gene *LGALS9* was previously found to be a prognostic factor for lung cancer, low expression of which was correlated with poorer survival outcome (He *et al.* 2019). The significant amplification genes *HSPG2*, *EIF3E*, and *YTHDC2* have also been extensively found as oncogenes in diverse tumor types, including lung cancer, gastrointestinal cancers, and breast cancer (Li *et al.* 2014; Chen *et al.* 2018; Kalscheuer *et al.* 2019). Our study mainly focuses on identifying germline CNV; however, the method is also suitable for the detection of somatic aberrations, where long and recurrent copy number changes are prominent. Due to the tumor complexity and heterogeneity, it may require different strategies in data normalization and filtering procedures (Zare *et al.* 2017).

This report demonstrated the improved performance of integrative CNV detection by utilizing a multisample and multiplatform strategy. The advantages of our method in theory lie in 2 aspects. First, utilizing information across samples will dramatically reduce false positives and boost detection power. We showed that BMI-CNV presented essential advantages over other

single-sample methods in detecting common variants. The advantage was previously shown in Song *et al.* (Song *et al.* 2016), which revealed that the underlying statistical power of multi-sample methods converged to one at a faster rate than single-sample methods. Second, BMI-CNV integrates available SNP data to detect CNVs in noncoding regions, allowing for full-spectrum genomic variants investigation. Indeed, the important role of CNVs in noncoding regions has been revealed in numerous studies. For example, Kumaran *et al.* identified 1,812 breast cancer-associated CNVs mapping to noncoding regions (Kumaran *et al.* 2018). Other similar efforts in integrating data from noncoding regions have been made. EXCAVATOR2 and CopywriteR used both the targeted reads and the nonspecifically captured off-target reads (i.e. from the noncoding region) (Kuilman *et al.* 2015; D'Aurizio *et al.* 2016), in which the information is usually biased and incomplete. Our method utilizes the more complete SNP array data from the matched samples and therefore provides a more reliable and unbiased solution. Besides, iCNV assumes that those overlapping markers (i.e. exons and SNPs) share the same copy number and indeed use one platform to validate calls from the other using a single hidden Markov model (Zhou *et al.* 2018). In contrast, BMI-CNV systematically combines data sequences from multiplatforms and allows heterogenous copy number states for the overlapping markers.

In this study, we developed a Bayesian statistical framework that has several essential advantages over other modeling strategies. First, the nonparametric PSBP can relax the restrictive parametric assumption and allows flexible modeling of the complex high-throughput data. A common critique of the Bayesian method is its computational speed. In our framework, all conditional distributions implemented in the Gibbs sampling algorithm can be analytically derived, which guarantees efficient sample generation and fast computation. Second, the Bayesian framework enables great flexibility and possibility to incorporate prior relevant information such as the documented CNV hotspot information (Wang *et al.* 2007). Finally, the PSBP framework can be easily extended to accommodate the complex data dependence structure by replacing the independent weights with stochastic processes (e.g. Gaussian process) without sacrificing computational tractability (Rodriguez and Dunson 2011).

Our method presents some limitations. First, it does not detect rare CNVs, as the power will be attenuated in the existence of a large proportion of noncarriers. Second, it will have low power to detect CNVs with similar proportions of duplications and deletions in the samples, which might be less likely for germline variations. Our method may split those CNVs into several smaller deletions and duplications, as each CNV locus is equally likely to be assigned to deletion or duplication. Our study primarily focuses on systematically integrating matched WES and SNP array data, since there are many unexplored large-scale datasets that have been previously sequenced by multiple platforms (e.g. SNP array and WES). As whole genome sequencing is becoming more affordable nowadays, it is prominent to combine such data with data generated from other low-resolution platforms. Our proposed method can also be extended in such direction accompanied with the development of appropriate normalization methods. For future directions, the detection performance of BMI-CNV can be improved by combining B allele frequency (BAF) data. One strategy is to construct a composite score using both total genetic intensities and allele-specific information, similar effort has been pursued in SNP array data in a previous study of our team (Xiao *et al.* 2019). With WES, efficient derivation of allele-specific information and normalization methods will be desirable for optimal performance of the method. Moreover, our method can be extended to integrate other data types (e.g. gene expression data) and may incorporate case-control status to directly identify disease-associated CNVs in a single model.

Data availability

WES data from the 1000 Genomes Project were downloaded from the FTP site hosted at the EBI ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ (accessed 2022 June 10). SNP array data from the International HapMap project are available from the FTP site ftp://ftp.ncbi.nlm.nih.gov/hapmap (accessed 2022 June 10). The TRICL SNP array and WES data are available at dbGaP with study accession numbers: phs000877.v2.p1 and phs000878 v2.p1. BMI-CNV source code is available on GitHub at https://github.com/FeifeiXiaoUSC/BMI-CNV (accessed 2022 June 10).

Supplemental material is available at GENETICS online.

Acknowledgments

We thank the reviewers in advance for their helpful and insightful suggestions and comments. We also acknowledge the support from the international TRICL consortium.

Funding

This work of Dr. Feifei Xiao was supported in part by the U.S. National Institutes of Health (R21 HG010925) and was also partially supported by the ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina and the internal grant from the University of Florida.

Conflicts of interest

None declared.

Literature cited

Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, Casey G, Hunter DJ, Sellers TA, Gruber SB, *et al.* The OncoArray consortium: a network for understanding the genetic architecture of common cancers the OncoArray and common cancer etiology. Cancer Epidemiol Biomarkers Prev. 2017;26(1):126–135.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, *et al.*; The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

Beecham GW, Bis J, Martin E, Choi SH, DeStefano A, Van Duijn C, Fornage M, Gabriel S, Koboldt D, Larson D, *et al.* The Alzheimer's disease sequencing project: study design and sample selection. Neurol Genet. 2017;3(5):e194.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B (Methodol). 1995;57(1):289–300.

Cai B, Bandyopadhyay D. Bayesian semiparametric variable selection with applications to periodontal data. Stat Med. 2017;36(14):2251–2264.

Chen M, Wei L, Law CT, Tsang FHC, Shen J, Cheng CLH, Tsang LH, Ho DWH, Chiu DKC, Lee JMF, *et al.* RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2. Hepatology. 2018;67(6):2254–2270.

Chung Y, Dunson DB. Nonparametric Bayes conditional distribution modeling with variable selection. J Am Stat Assoc. 2009;104(488): 1646–1660.

D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using excavator2. Nucleic Acids Res. 2016;44(20): e154.

Fang H, Wu Y, Narzisi G, O'Rawe JA, Barrón LTJ, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ. Reducing indel calling errors in whole genome and exome sequencing data. Genome Med. 2014;6(10):89.

George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993;88(423):881–889.

He Y, Jia K, Dziadziuszko R, Zhao S, Zhang X, Deng J, Wang H, Hirsch FR, Zhou C. Galectin-9 in non-small cell lung cancer. Lung Cancer. 2019;136:80–85.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, *et al.* Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52.

Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc. 2001;96(453):161–173.

Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. Codex2: full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol. 2018;19(1):1–13.

Kalscheuer S, Khanna V, Kim H, Li S, Sachdev D, DeCarlo A, Yang D, Panyam J. Discovery of hspg2 (perlecan) as a therapeutic target in triple negative breast cancer. Sci Rep. 2019;9(1):11.

Karch CM, Ezerskiy LA, Bertelsen S, Goate AM; ADGC. Alzheimer's disease risk polymorphisms regulate gene expression in the zcwpw1 and the celf1 loci. PLoS One. 2016;11(2):e0148717.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012; 40(9):e69.

Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, Nevedomskaya E, Xu G, de Ruiter J, Lolkema MP, *et al.* Copywriter: DNA copy number detection from off-target sequence data. Genome Biol. 2015;16(1):1–15.

Kumaran M, Krishnan P, Cass CE, Hubaux R, Lam W, Yasui Y, Damaraju S. Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. Sci Rep. 2018;8(1):1–11.

Li Z, Lin S, Jiang T, Wang J, Lu H, Tang H, Teng M, Fan J. Overexpression of eIF3e is correlated with colon tumor development and poor prognosis. Int J Clin Exp Pathol. 2014;7:6462.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986–D992.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol. 2013;14(10):R120.

Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP, Pakula AT, Eisenhauer NJ, Uhas K, Weik L, *et al.*; GeneSTAR. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. Am J Hum Genet. 2010;87(5):618–630.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.

Rodriguez A, Dunson DB. Nonparametric bayesian models through probit stick-breaking processes. Bayesian Anal. 2011;6(1): 10.1214/11-BA605.

Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Am Stat Assoc. 1993;88(424):1273–1283.

Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1(6):62–69.

Siegmund D, Yakir B, Zhang NR. Detecting simultaneous variant intervals in aligned sequences. Ann Appl Stat. 2011;5(2A): 645–668.

Song C, Min X, Zhang H. The screening and ranking algorithm for change-points detection in multiple samples. Ann Appl Stat. 2016;10(4):2102–2129.

Takumi T, Tamada K. CNV biology in neurodevelopmental disorders. Curr Opin Neurobiol. 2018;48:183–192.

Välipakka S, Savarese M, Johari M, Sagath L, Arumilli M, Kiiski K, Sáenz A, de Munain AL, Cobo AM, Pelin K, *et al.* Copy number variation analysis increases the diagnostic yield in muscle diseases. Neurol Genet. 2017;3(6):e204.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* The sequence of the human genome. Science. 2001;291(5507):1304–1351.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11): 1665–1674.

Wang S, Zhang R, Wang T, Jiang F, Hu C, Jia W. Association of the genetic variant rs2000999 with haptoglobin and diabetic macrovascular diseases in Chinese patients with type 2 diabetes. J Diabetes Complications. 2019;33(2):178–181.

Xiao F, Luo X, Hao N, Niu YS, Xiao X, Cai G, Amos CI, Zhang H. An accurate and powerful method for copy number variation detection. Bioinformatics. 2019;35(17):2891–2898.

Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18(1):13.

Zhang NR, Siegmund DO, Ji H, Li JZ. Detecting simultaneous change-points in multiple sequences. Biometrika. 2010;97(3):631–645.

Zhang Z, Cheng H, Hong X, Di Narzo AF, Franzen O, Peng S, Ruusalepp A, Kovacic JC, Bjorkegren JL, Wang X, *et al.* EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. Nucleic Acids Res. 2019;47(7):e39.

Zhou Z, Wang W, Wang LS, Zhang NR. Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. Bioinformatics. 2018;34(14):2349–2355.

Communicating editor: Y. Li