Computational and Structural Biotechnology Journal 20 (2022) 4975–4983

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

PeakCNV: A multi-feature ranking algorithm-based tool for genome-wide copy number variation-association study

Mahdieh Labani [a,b,1], Ali Afrasiabi [a,1], Amin Beheshti [b,*], Nigel H. Lovell [c,d], Hamid Alinejad-Rokny [a,e,f,*]

[a] BioMedical Machine Learning Lab (BML), The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia
[b] Data Analytics Lab, School of Computing, Macquarie University, Sydney, NSW 2109, Australia
[c] The Graduate School of Biomedical Engineering (GSBmE), UNSW Sydney, Sydney, NSW 2052, Australia
[d] Tyree Institute of Health Engineering (IHealthE), UNSW Sydney, Sydney, NSW 2052, Australia
[e] UNSW Data Science Hub, The University of New South Wales, Sydney, NSW 2052, Australia
[f] Health Data Analytics Program, AI-enabled Processes (AIP) Research Centre, Macquarie University, Sydney 2109, Australia

ARTICLE INFO

Article history:
Received 15 June 2022
Received in revised form 31 August 2022
Accepted 1 September 2022
Available online 7 September 2022

Keywords:
Genetic abnormalities
Copy number variations
CNV association study
Risk genes
Artificial intelligence

ABSTRACT

Copy Number Variation (CNV) refers to a type of structural genomic alteration in which a segment of chromosome is duplicated or deleted. To date, many CNVs have been identified as causative genetic elements for several diseases and phenotypes. However, performing a CNV-based genome-wide association study is challenging due to inconsistency in length and occurrence of CNVs across different individuals under investigation. One of the most efficient strategies to address this issue is building CNV regions (genomic regions in which CNVs are overlapping - CNVRs). However, this approach is susceptible to a high false positive rate due to overlapping and co-occurring of confounding CNVRs with true positive CNVRs. Here, we develop PeakCNV that differentiates false-positive CNVRs from true positives by calculating a new metric, independence ranking score, (IR-score) via a feature ranking approach. We compared the performance of PeakCNV with other current existing tools by carrying out two case studies one using the CNV genotype data for individuals with prostate cancer (194 cases and 2,392 healthy individuals) and the second one for individuals with neurodevelopmental disorders (19,642 cases and 6,451 healthy individuals). Crucially, our benchmarking analyses on prostate cancer cohort indicated that PeakCNV identifies a fewer risk candidate CNVRs with shorter lengths compared to other tools. Importantly, these CNVRs cover a greater proportion of case over healthy individuals compared to other tools. The accuracy of PeakCNV in identifying relevant candidate CNVRs was reproducible in the case study on neurodevelopmental disorders. Using data from the FANTOM5 expression atlas and the Clinical Genomic Database, we show that the candidate CNVRs identified by PeakCNV for neurodevelopmental disorders overlap with a greater number of genes with the brain-enriched expression, and a greater number of genes that are associated with neurological conditions compared to candidate CNVRs identified by other tools. Taken together, PeakCNV outperformed current existing CNV association study tools by identifying more biologically meaningful CNVRs relevant to the phenotype of interest. PeakCNV is publicly available for the analysis of CNV-associated diseases and is accessible from https://rdrr.io/github/mahdieh1/PeakCNV.

1. Introduction

A copy number variation (CNV) is a type of structural change in the genome in which a segment of the genome is amplified or deleted [1]. To date, many CNVs have been reported in association with several phenotypes, traits, and diseases such as different types of cancers and neurodevelopmental disorders (NDD) [2,3]. For instance, duplication of the *APP* gene caused by a CNV has been shown to be associated with increasing the risk of developing Alzheimer's disease [4]. The CNV deletion or duplication at 15q11-q13 is linked to the development of Prader-Willi and Angelman syndromes [5]. A CNV deletion at 1q21.1 locus was found to be linked with the development of intellectual disability and autism [6]. Another example is a CNV deletion at 2p24.3 locus [7] which is

* Corresponding authors at: BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW Sydney (H. Alinejad Rokny); School of Computing, Macquarie University (A. Beheshti).
E-mail address: h.alinejad@ieee.org (H. Alinejad-Rokny).
[1] These authors have contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.csbj.2022.09.001

associated with increasing the risk for progressive prostate cancer. It has been also reported that the CNV duplication at 14q32.33 genomic region resulting in duplication of *IGHG3* gene contributes to high prevalence and mortality of prostate cancer [8].

Identifying risk CNVs can be used as a handle to improve the diagnosis and potentially disease management. To identify risk CNVs, a genome-wide association study needs to be performed on the genomic regions that harbor CNVs. However, one of the major obstacles in a CNV-based genome-wide association study occurs when categorizing CNVs across all cases (individuals with the phenotype of interest) and controls (healthy individuals), which is challenging because CNVs are inconsistent in sequence, size, and genomic coordinates across individuals [9]. To address this issue, one effective approach is to build CNVRs (genomic regions where CNVs are overlapping - CNVRs) prior to identifying those CNVRs statistically associated with the phenotype of interest. There are few tools currently available that utilize the abovementioned strategy as the cornerstone of their analytic pipelines. The CNVRuler tool [10] provides three different strategies including reciprocal overlap, overlapping regions and segmentation at CNV boundaries to define CNVRs and then estimates the association of each defined CNVR with the phenotype of interest. CNVRuler also offers logistic regression, linear regression, chi-squared, and fisher exact test to perform the association test. PLINK [11] performs a permutation-based one-sided overrepresentation test to carry out a CNVR-phenotype association study and determine the empirical p value for each CNVR. To determine the associated CNVRs with small size in length, Single Nucleotide Association Test CNV (SNATCNV) [12] was proposed. This tool firstly identifies genomic regions in which more often deleted/duplicated among cases than healthy individuals at a single base pair resolution using one-tailed Fisher's exact test following series of permutations. Then SNATCNV merges the resultant significantly associated base pairs if they are in close proximity to generate associated CNVRs with the phenotype under investigation. CoNVAQ [13] defines CNVRs by segmentation of genomic regions that CNVs overlapping. Then, it determines the statistically significant CNVRs with using a fisher's exact test.

Nevertheless, these existing approaches are susceptible to high false positive rates due to CNVRs which overlap or co-occur with true positive CNVRs. We here proposed a novel Artificial Intelligence (AI) based tool, PeakCNV, to correct this bias by distinguishing independent CNVR associations from that of confounding CNVRs within the same loci, resulting in identifying more accurate and biological meaningful list of CNVRs associated with phenotype of interest via a genome-wide CNV-phenotype association study.

2. Methods

PeakCNV first builds deletion and duplication CNVR maps for case and control individuals separately. Then to perform the CNV-phenotype association study with the single nucleotide resolution, it assesses the association of each base pair within deletion and duplication CNVR maps and identifies the nucleotides with deletion or duplication that are significantly over-represented for cases over controls. Then, PeakCNV identifies groups of CNVRs which have a similar association significance with their respective phenotype. Next to differentiate false positive CNVRs from true positives within each group (cluster), PeakCNV calculates a new metric, which we termed independence ranking score (IR-score) via a feature ranking algorithm. This score identifies a true positive CNVR when its significance of association is independent of any other overlapping or co-occurring CNVRs within that cluster (independent CNVRs). Lastly, PeakCNV identifies the true positive CNVRs with the highest IR-scores within each cluster as cluster representative CNVRs. The final output of the PeakCNV is a list of CNVRs with a greater probability of being true positives.

2.1. CNVR map building

PeakCNV builds deletion and duplication CNVR maps for cases and controls by mapping the genomic coordinates of deletion and duplication CNVs, respectively (Fig. 1**A1**). Then, PeakCNV estimates the probability of either deletion or duplication in cases versus controls for each base pair within these maps using a one-tailed Fisher's exact test. This provides every nucleotide (base pair) that are significantly more frequently deleted or duplicated in cases than controls (p value < 0.05 - Fig. 1**A2**). Then, it merges significant nucleotides where they are in near proximity (within one base pair distance) to identify significant CNVRs. PeakCNV performs this process for the nucleotides that are significantly deleted or duplicated in cases versus controls, separately.

2.2. Clustering process

PeakCNV uses the DBSCAN [15] clustering algorithm to identify groups of CNVRs containing CNVRs with a similar association significance with their respective phenotype. A number of clustering algorithms including hierarchical clustering gap, hierarchical clustering with Silhouette, spectral clustering, model-based clustering, and DBSCAN were tested using the Dunn Index method to determine the best performing clustering algorithm for grouping CNVRs with the similar CNV-phenotype association significance level. We performed the above-mentioned clustering algorithms for grouping CNVRs on chromosome 10 which are significantly (p value < 0.05) associated with PC. We then assessed the efficiency of each clustering algorithm using the Dunn Index method. The results indicated that DBSCAN had the highest efficiency (Supplementary Table 1). DBSCAN requires two hyperparameters including MinPts and ϵ. MinPts determines the minimum number of CNVRs required to form a cluster. Since, it has been reported that for two dimensional data, DBSCAN performs most efficiently where the hyperparameter MinPts specified to four [14], we used the same value for MinPts hyperparameter. The hyperparameter ϵ determines the maximum pairwise distance for a set of CNVRs within a given cluster. We identified the optimised value of ϵ for analysing CNVRs in each chromosome using k-distance plot [14]. The k-distance plot for each chromosome was generated by performing a k-nearest neighbours as the k specified to the same value, we used for MinPts. Two features (CNVR uniqueness and the genomic distance between CNVRs) obtained from paired wise comparison of CNVRs are used as inputs for the clustering step (Fig. 1**B and** Fig. 2). Uniqueness value refers to the number of case samples covered by a given CNVR after subtracting the common case samples between each pair of CNVRs. This value is obtained through Equation (1).

$$Uniqueness(R_i) = S_{R_i} - S_{R_i \cap R_j} \quad (1)$$

where R_i and R_j are the two CNVRs, and S_{R_i} shows the number of similar case samples covered by R_i and R_j.

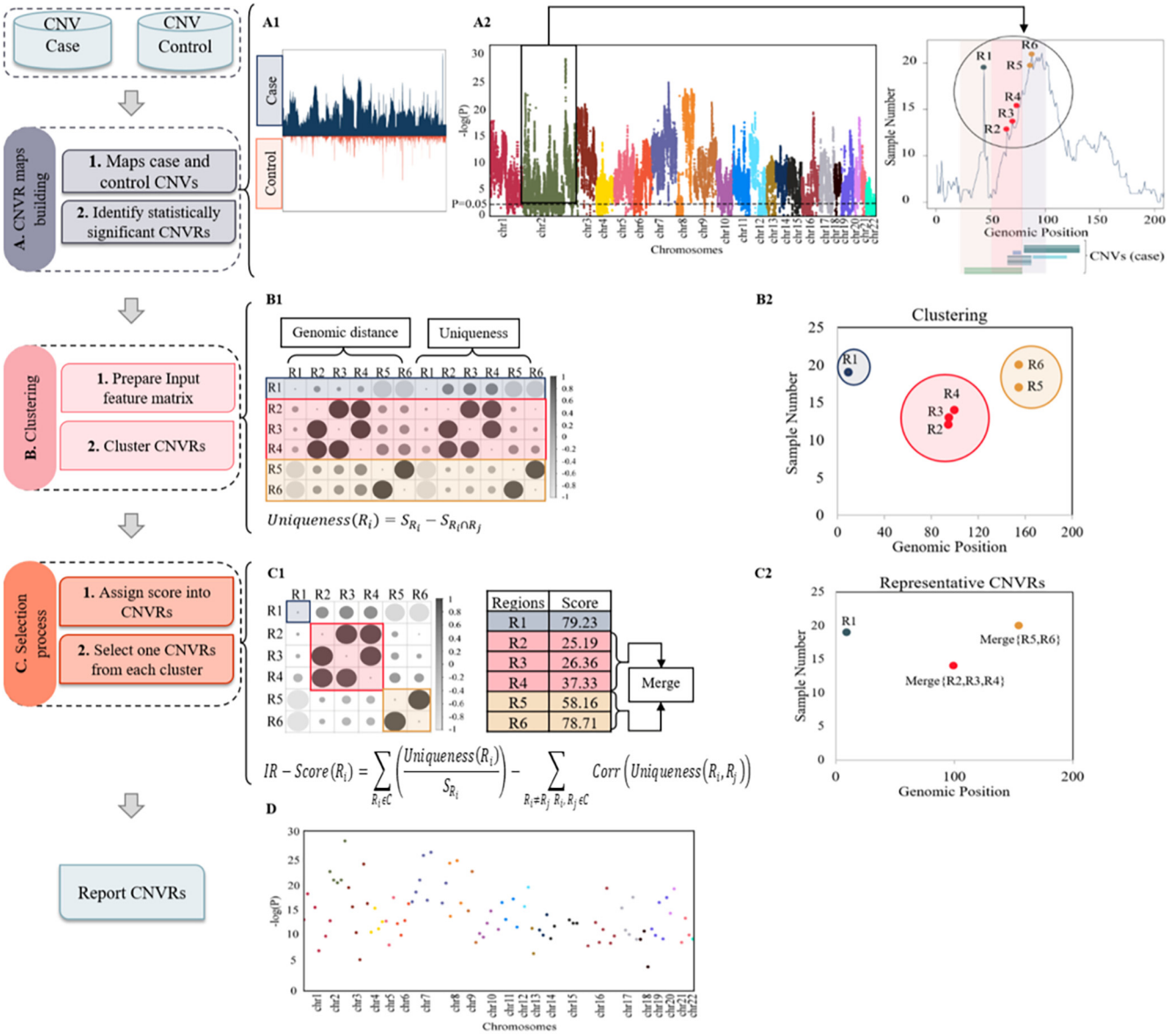

Fig. 1. The schematic workflow of PeakCNV. PeakCNV consists of three main steps; (**A**) CNVR Map Building, (**B**) Clustering and (**C**) Selection Processes. In the CNVR map building step, PeakCNV builds deletion and duplication CNVR maps for case and control individuals by merging the genomic coordinates of deletion and duplication CNVs, respectively **(A1)**. Then, PeakCNV estimates the probability of either deletion or duplication in case over control for each nucleotide within these maps using a one-tailed Fisher's exact test. This provides CNVRs that are significantly (p value < 0.05) more frequently occurring in case than control individuals **(A2)**. In the clustering step, PeakCNV groups CNVRs which have a similar association significance with the phenotype of interest using two features obtained from paired wise comparison of CNVRs: CNVR uniqueness and the genomic distance between CNVRs **(B1 and 2)**. In the selection, PeakCNV differentiates false positive CNVRs from true positives within each cluster by calculating an independence ranking score (IR-score) for each CNVR **(C1)**. Then, PeakCNV merges CNVRs with the highest IR-score in each cluster if they are in near proximity. The final output of the PeakCNV is a list of CNVRs with a greater probability of being true positives (**D**). CNV and CNVR stand for copy number variation and copy number variation region, respectively.

2.3. Selection process

In the selection process, PeakCNV identifies the most independent CNVRs within each cluster using the IR-score. Independent CNVRs are those detected in the greatest number of cases while having a minimum co-occurrence of other CNVRs (Fig. 1**C1**). PeakCNV estimates the IR-score for CNVRs by performing a feature ranking algorithm with two objectives (Equation (2)): summation and correlation of the uniqueness values of CNVRs within a cluster. The first objective indicates the absolute case sample coverage rate by a given CNVR compared to other CNVRs within a cluster, and the second objective indicates the co-occurrence (co-duplication or co-deletion) of a given CNVR with other CNVRs within a cluster. PeakCNV computes IR-score through Equation (2) for any given CNVR within each cluster to determine the most independent CNVRs within each cluster.

$$IR - Score(R_i) = \sum_{R_i \epsilon C} \left(\frac{Uniqueness(R_i)}{S_{R_i}} \right) - \sum_{R_i \neq R_j, R_i, R_j \epsilon C} Corr\left(Uniqueness(R_i, R_j)\right) \quad (2)$$

where C is the number of CNVRs in each cluster, R_i and R_j are the two randomly chosen CNVRs from the same cluster for pairwise comparison. S_{R_i} is the total number of samples for R_i and $Corr(Uniqueness(R_i, R_j))$ denotes the correlation between the uniqueness values of CNVRs within each cluster, that is calculated by a Kendall's rank correlation test [16].

Algorithm1. Clustering Process

Input *CNV*: The significant CNV list
eps : Distance threshold
Minpts : The minimum number of points required to form a cluster (default value is 5)

Output **Result:** Clustered matrix
Begin algorithm
  For each chromosome
    Compute Uniqueness matrix contains the number of samples after
    removing the common samples between each two regions.
    Compute distance matrix between each two regions
    Combine distance matrix and Uniqueness matrix as the input matrix
  End for
  For each *input* in each chr
  For each unvisited point p in $input_{chri}$
    Mark p as visited
    NeighborPts = find the neighboring points of p
    If $(length(NeighborPts) < Minpts)$
      Mark p as Noise
    Else
      C = next cluster
      Add P to cluster C
      For each point P' in *NeighborPts*
        If P' is not visited
          Mark P' as visited
            $NeighborPts'$ = find the neighboring points of p
            If $(length(NeighborPts') >= Minpts)$
$NeighborPts = NeighborPts' \cup NeighborPts$
          End If
        If P' is not yet member of any cluster
        Add P' to cluster C
      End If
  End If
  End for
  End If
End for
Return clustered matrix as the result
End for
End algorithm

Lastly, PeakCNV sorts CNVRs within each cluster by IR-score and merges regions with the highest IR-score if they are in proximity (default value is 1000 bp). A pseudo code of the clustering step is presented in Fig. 3.

Algorithm2. Selection Process

Input **Result:** CNVs with their clusters
Threshold: maximum distance for merging CNVRs (default value:1000 bp)

Output **Final list:** Selected CNV regions
Begin algorithm
For each chromosome
  C = number of clusters for each chromosome
  For each c_i in C
  For each $region_i$ in c_i
    Calculate score for each region using Equation (2).
  End for
  End for
  Sort regions based on the score
  Best region = region with maximum score
  For $each c_i in C$
  For each $region_i$ in c_i
    If $(distance(Bestregion, region_i) < Threshold)$
      $Start_{CNV} = Start_{Bestregion}$
      $End_{CNV} = End_{region_i}$
    End If
  End for
  Return Best region for each C_i
  End For
  End algorithm

3. Results and discussion

We validated the capability of PeakCNV in identifying more accurate and biological meaningful pathogenic CNVRs by performing three case studies. In the first case study, we assessed the performance of PeakCNV by analyzing a simulated ground truth CNV genotype data. We also compared the performance of PeakCNV in identifying pathogenic CNVRs for PC (case study-two) and NDD (case study-three) to that of other currently available tools; SNATCNV, PLINK, CoNVaQ and CNVRuler. SNATCNV was executed with the default parameters of "indvbased" mode, which returns significant CNVRs using a one-tailed Fisher's exact test. PLINK v1.9 was performed with "mperm" mode which identifies statistically significant CNVRs by providing empirical p values based on 50,000 null permutations. The CoNVaQ was executed in "statistical model" mode which uses a Fisher's exact test to identify significant CNVRs. The CNVRuler was carried out with "CNVR method" and "logistic regression" settings. To discriminate the significantly associated CNVRs with their respective phenotype, a p value < 0.05 was used as the threshold for statistical significance for the output of all above mentioned tools.

To interrogate and compare the biological relevance of identified risk CNVRs by our newly developed tool and other available tools, we identified genes which overlapped with identified CNVRs (CNVgenes). The reference gene list that we used here is the combination of FANTOM5 [17], Ensembl [18] and GENCODE [19] gene annotation files for hg19 genome assembly to curate a comprehensive reference gene list. The FANTOM5 gene annotation file was

Algorithm1. Clustering Process

Input CNV: The significant CNV list
eps: Distance threshold
$Minpts$: The minimum number of points required to form a cluster (default value is 5)

Output **Result:** Clustered matrix

1: **Begin algorithm**
2: **For** each chromosome
3: **Compute** Uniqueness matrix contains the number of samples after removing the common samples between each two regions.
4: **Compute** distance matrix between each two regions
5: **Combine** distance matrix and Uniqueness matrix as the input matrix
6: **End for**
7: **For** each $input$ in each chr
8: **For** each unvisited point p in $input_{chri}$
9: **Mark** p as visited
10: $NeighborPts$ = find the neighboring points of p
11: **If** ($length(NeighborPts) < Minpts$)
12: **Mark** p as Noise
13: **Else**
14: C= next cluster
15: **Add** P to cluster C
16: **For** each point P' in $NeighborPts$
17: **If** P' is not visited
18: **Mark** P' as visited
19: $NeighborPts'$= find the neighboring points of p
20: **If** ($length(NeighborPts') >= Minpts$)
21: $NeighborPts = NeighborPts' \cup NeighborPts$
22: **End If**
23: **If** P' is not yet member of any cluster
24: **Add** P' to cluster C
25: **End If**
26: **End If**
27: **End for**
28: **End If**
29: **End for**
30: **Return** clustered matrix as the result
31: **End for**
32: **End algorithm**

Fig. 2. The pseudo-code for the clustering step. First, the genomic distance and uniqueness value for each CNVR was calculated in a pairwise comparison with other CNVRs (lines 2–6). Then, values of these two features for each CNVR were merged to be used as an input in the clustering step (lines 7–31).

used as the backbone of our reference gene list, but when the gene annotation was absent from FANTOM5 these were acquired from Ensembl and GENCODE. The final reference gene list contained 82,539 genes including 58,000, 24,501 and 38 genes from FANTOM5, Ensembl and GENCODE, respectively (Supplementary Table 1). Then to identify the list of CNVgenes, we search for CNVRs and genes with overlapping genomic coordinates using bedtools v2.30.0 [20].

3.1. Case study-one: Simulated CNV genotype dataset

The ground truth simulated dataset contains the genotype data (genomic coordinate) of 29,856 number CNVs (14,275 deletion and 15,581 duplication types) for 750 case and 750 control samples. We first examined the distribution of both frequency and size of CNVs from multiple real datasets (experimentally identified CNVs) [21–23] to identify the structure underlying the CNV genotype data. According to these two obtained distributions, we generated a random set of genomic coordinates (10,000 genomic regions as CNV genomic coordinates pool) using "random" function of the bedtools package to synthesize a simulated dataset which mimics the real data architecture and structure. We took advantage of a hypergeometric distribution model to identify a set of high confident odds ratios (case to control ratio) for true and false associated CNVRs. We defined 2.3 $\leq$ case-to-control ratio as a cut-off for true associated CNVRs. The cut-off for false associated CNVRs defined as $\geq$ 0.42 case-to-control ratio. Based on these criteria we randomly selected 34,663 CNVs from CNV genomic coordinate pool. The simulated dataset contained 20 true positive phenotype-associated CNVRs (including 10 case enriched deletion-CNVRs and 10 case enriched duplication-CNVRs) and 20 false phenotype-associated CNVRs (including 10 control enriched deletion-CNVRs and 10 control enriched duplication-CNVRs). Application of PeakCNV on this simulated dataset showed that PeakCNV identifies true associated CNVRs with the 82% accuracy, 86% precision, 89% recall and 86% F-measure, respectievly. The simulated dataset is provided in Supplementary Table 1.

3.2. Case study-two: Prostate cancer (PC)

To identify CNVRs associated with susceptibility to PC, the genotype data for 11,564 CNVs (3625 deletions and 7939 duplications) from 194 patients with PC were obtained from the International Cancer Genome Consortium [21]. The CNV genotype data for 2,392 healthy individuals were also obtained from the 1000 Genomes Project [22] containing the genomic coordinates for 32,449 CNVs (22,318 deletions and 10,131 duplications). The list of CNVs used in this study is also provided in Supplementary Table 2. Fig. 4**A** indicates the statistically significant (p value < 0.05) CNVRs (both deletion and duplication CNVRs) for PC that were identified

Algorithm2. Selection Process

Input **Result:** CNVs with their clusters
Threshold: maximum distance for merging CNVRs (default value:1000bp)
Output **Final list:** Selected CNV regions
1: **Begin algorithm**
2: **For** each chromosome
3: C = number of clusters for each chromosome
4: **For** each c_i in C
5: **For** each $region_i$ in c_i
6: **Calculate** score for each region using Equation 2.
7: **End for**
8: **End for**
9: **Sort** regions based on the score
10: **Best region** = region with maximum score
11: **For** $each\ c_i\ \ in\ C$
12: **For** each $region_i$ in c_i
13: **If** $(distance(Best\ region, region_i) < Threshold)$
14: $Start_{CNV} = Start_{Best\ region}$
15: $End_{CNV} = End_{region_i}$
16: **End If**
17: **End for**
18: **Return** Best region for each C_i
19: **End For**
20: **End algorithm**

Fig. 3. The pseudo-code for selection step. PeakCNV estimates the IR-score for each CNVRs within each cluster via Equation (2) (lines 2–7). Then, PeakCNV ranks CNVRs in a descending order. PeakCNV merges CNVRs with the highest scores within each cluster if they had similar scores and were in close proximity to each other. Lastly, PeakCNV reports a list of top ranked CNVRs from each cluster (lines 8–20).

by PeakCNV and other tools. The risk CNVs identified by PeakCNV and other tools (SNATCNV, PLINK, CoNVaQ, and CNVRuler) for PC are provided in Supplementary Table 3. Fig. 4**A** also shows that clearly, the risk CNVRs identified by PeakCNV are less noisy (higher rate of true positive than false positive hits) compared to other tested tools. PeakCNV identified 291 risk CNVRs for PC with a total length of 2661.43 Mbp (Fig. 4**B**). Although PeakCNV identified a fewer number of risk-CNVRs with a shorter length for PC compared to other tools, these risk-CNVRs cover a greater proportion of cases over controls (Fig. 4**C**). Risk CNVRs identified by PeaKCNV for PC had 2.05, 1.14 and 2.04 case over control coverage rate (case-to-control ratio) for duplication, deletion and total (both duplication and deletion together) risk CNVRs which are 1.46, 1.44 and 1.19 times greater than the highest case-control ratio of risk-CNVRs identified by other tools. The list of CNVgenes identified by PeakCNV and other tools for PC is also provided in Supplementary Table 4.

3.3. Case study-three: Neurodevelopmental disorders (NDD)

The CNV genotype data for 47,143 (26,546 deletions and 20,597 duplications) and 24,859 (14,025 deletions and 10,834 duplications) CNVs for 19,644 individuals with different types of NDD and 6,452 healthy individuals was obtained from AutDB database (https://mindspec.org/autdb; download date: Sep 2018) [23] that contains multiple large CNV datasets (with different microarray versions and different 'individual-level' CNV callers). To validate the performance of PeakCNV and other tools tested here in identifying meaningful NDD-associated CNVRs; a) CNV-affected genes (genes that overlap with identified CNVRs - CNVgenes) were determined, then b) a list of nervous system specific expressing genes was used to assess the relative biological importance of identified CNVgenes within the NDD pathogenic contexts [24], and lastly c) the Clinical Genomic Database [25] was used as a source of pathogenic genes for neurological disorders to assess the association of identified CNVgenes with potentially relevant NDD pathogenic molecular processes.

Fig. 5**A** indicates the statistically significant (p value < 0.05) CNVRs (both deletion and duplication CNVRs) for NDD that were identified by PeakCNV and other tools. The risk CNVs identified by PeakCNV and other tools (SNATCNV, PLINK, CoNVaQ and CNVRuler) for NDD are provided in Supplementary Table 5. Fig. 5**A** also shows that clearly the risk CNVRs identified by PeakCNV are less noisy (higher rate of true positive than false positive hits) compared to other tested tools. PeakCNV identified 52 CNVRs for NDD with a total length of 58.56 Mbp (Fig. 5**B**). PeakCNV identified a fewer number of risk-CNVRs with a shorter length for NDD compared to other tools. However, these risk-CNVRs cover a

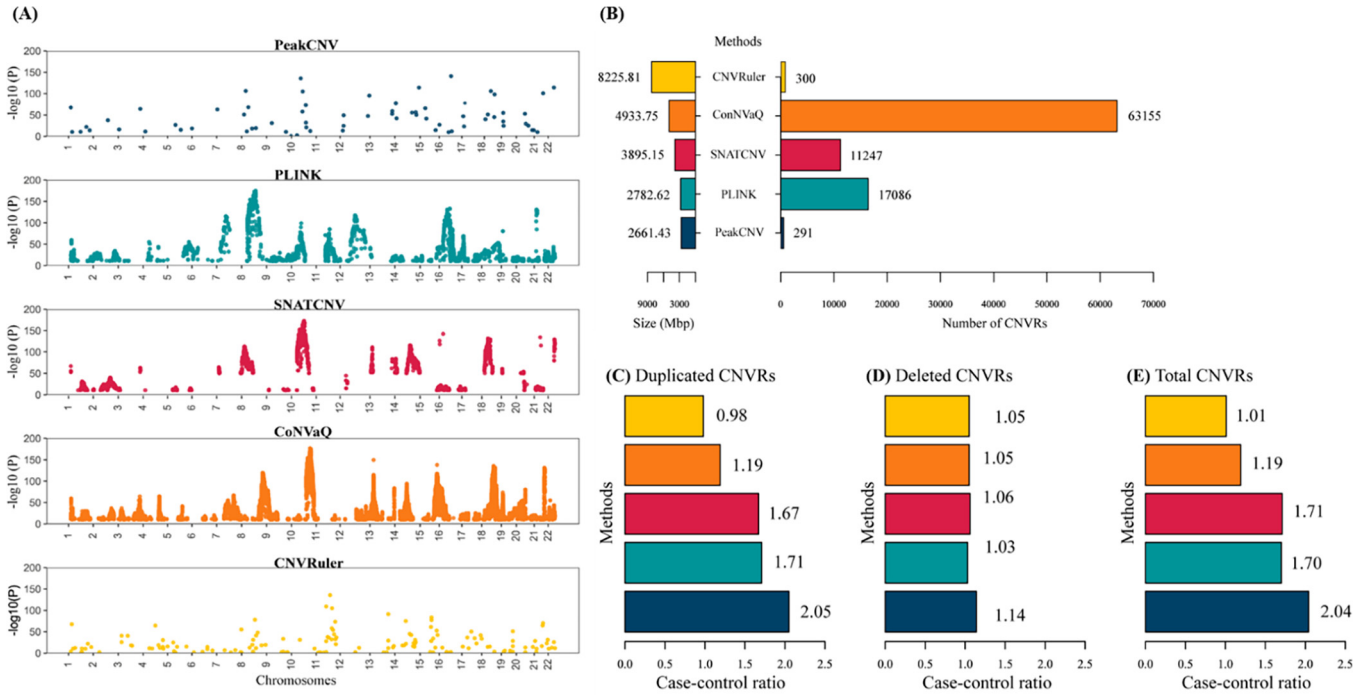

Fig. 4. The results of benchmarking of PeakCNV over other available tools using CNVs for individuals with Prostate Cancer (PC) as case study-two. **A)** Statistically significant (p value < 0.05) CNVRs (both deletion and duplication CNVRs) that were identified by PeakCNV in compared with other tools. X axis represents the genomic coordinates of CNVRs in chromosomes. Y axis represents the negative logarithm base ten of the p value. **B)** The total number of identified risk CNVRs (both deletion and duplication CNVRs together) and their genomic size in mega base pair (Mbp) by PeakCNV and other tools. (**C**), (**D**) and (**E**) represent the case-control ratio for duplication, deletion and total (both duplication and deletion together) risk CNVRs by different tools. The performance of PeakCNV in identifying more relevant CNVRs to PC was compared with PLINK, SNATCNV, CNVRuler and CoNVaQ through examining the coverage rate (case-to-control ratio) of identified risk CNVRs for case over control samples.

greater proportion of cases over controls (Fig. 5**B).** Risk CNVRs identified by PeaKCNV had 4.07 case over control coverage rate (case-to-control ratio) for NDD, which are 1.07 times greater than the highest case-control ratio of risk-CNVRs identified by other tools. Then, we identified CNV-affected genes with the similar strategy to PC. The same analysis for CNV-affected genes in PC was performed.

We then estimated the enrichment of CNVgenes identified by PeakCNV and other tools for the 508 NDD-related genes (acquired from the Clinical Genomic Database and provided in Supplementary Table 6,7) using Equation (3).

$$Enirchment\ for\ NDD-related\ genes = \frac{\frac{\#CNV\ affected\ coding\ genes\ overlapping\ with\ NDD\ associated\ coding\ genes}{\#CNV\ affected\ coding\ genes}}{\frac{\#NDD\ associated\ coding\ genes}{\#coding\ genes}} \quad (3)$$

The result shows that from 1,746 NDD risk CNVgenes identified by PeakCNV, 321 genes were putatively pathogenic for neurological disorders (representation factor 1.14, p value < 0.001), which is 1.21 times greater than the highest over-representation obtained by other tools (Fig. 5**f-h**).

We also identified the brain-specific expressing genes (BSG) using our previously published approach, with some modifications. Briefly, the expression data of 59,111 number genes from 1,897 samples related to 347 cells and tissues was obtained from FANTOM5. Then, for any given gene we first excluded samples with an expression level of less than one counts per million reads mapped (CPM), then sorted the remaining samples in descending order by their expression level. We then determined a gene as a BSG if the brain related samples were over-represented with p value threshold <0.001 for the top ranked samples (Supplementary Table 8). To estimate the enrichment of CNVgenes identified by different methods for BSGs, we used Equation (4). The list of BSG is provided in Supplementary Table 9.

$$Enrichmnet\ for\ BSG\ list = \frac{\frac{\#CNV\ affected\ genes\ overlapping\ with\ BSG}{\#CNV\ affected\ genes}}{\frac{\#BSG}{\#All\ experresing\ genes(CPM\ more\ than\ 1CPM)}} \quad (4)$$

PeakCNV identified a greater number of CNVgenes that are dominantly expressed in the NDD relevant pathogenic contexts (nervous system) compared to other tools (24 out of 831 CNV genes, representation factor 1.15, p value < 0.001), which is 1.07 times greater than the highest enrichment obtained by other tools (Fig. 5**I-K**).

4. Conclusion

The current existing tools for performing a CNVR-phenotype association study fall short to generate a false positive free list of CNVRs associated with the phenotype of interest. To address this issue, in the present study we developed PeakCNV, an AI-based tool to filter out false positive hits and keep true positive candidate CNVRs. This can enhance the efficiency of downstream analyses for diagnosis or purposes significantly. The aim of the proposed method is to identify associated CNVRs with disease using two new objectives including genomic distance and uniqueness. The results of benchmarking analysis showed that PeakCNV improves the CNVR based CNV-phenotype association study by correcting the biases arising from the overlapping and co-occurrence of confounding CNVRs, and therefore outperforms other currently existing tools by identifying more biologically meaningful candidate CNVRs.

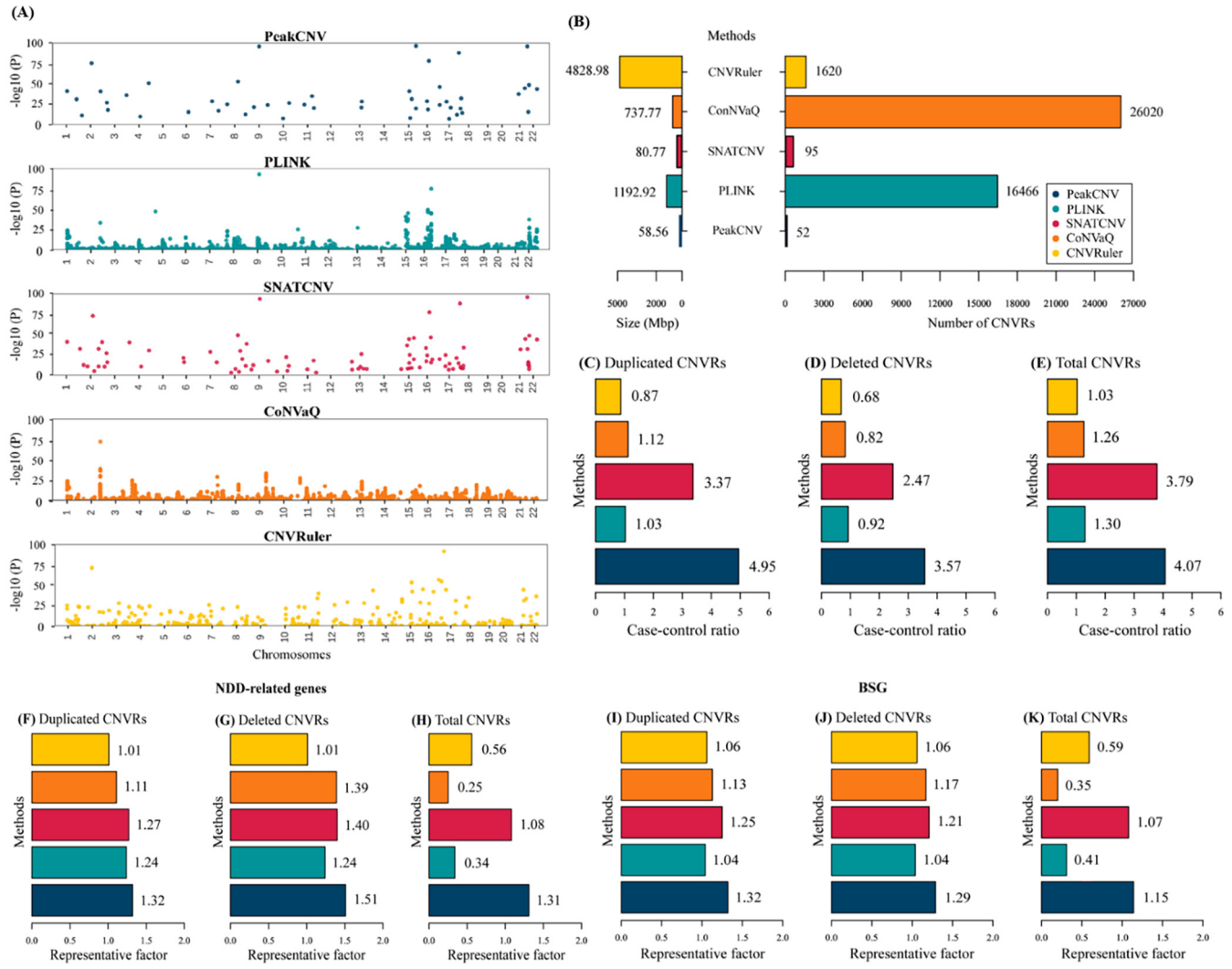

Fig. 5. The results of benchmarking of PeakCNV over other available tools using CNVs for individuals with Neurodevelopmental Disorders (NDD) as case study-three. **A)** Statistically significant (p value < 0.05) CNVRs (both deletion and duplication CNVRs) that were identified by PeakCNV and other tools. X axis represents the genomic coordinates of CNVRs in chromosomes. Y axis represents the negative logarithm base ten of the p value. **B)** The total number of identified risk CNVRs (both deletion and duplication CNVRs together) and their genomic size in mega base pair (Mbp) by PeakCNV and other tools. (**C**), (**D**) and (**E**) represent the case-control ratio for duplication, deletion and total (both duplication and deletion together) risk CNVRs by different tools, respectively. Case-control ratio is the coverage rate of identified risk CNVRs for case over control samples. (**F**), (**G**) and (**H**) represent the enrichment of CNV-affected genes (CNVgenes) in NDD-related genes for duplication, deletion and total (both duplication and deletion together) risk CNVRs identified by different tools, respectively. The NDD-related genes were obtained from the Clinical Genomic Database. (**I**), (**J**) and (**K**) represent the enrichment of CNVgenes in brain specific expressing genes for duplication, deletion and total (both duplication and deletion together) risk CNVRs identified by different tools, respectively. The performance of PeakCNV in identifying more biologically meaningful CNVRs related to NDD was compared with PLINK, SNATCNV, CNVRuler and CoNVaQ. The biological relevance of identified CNVRs was assessed by the overrepresentation of affected genes by these CNVRs for molecular pathways (brain specific expressing genes and known NDD risk genes) related to NDD pathogenesis as well as higher coverage rate for individuals with NDD over healthy individuals. Representation factor represents the magnitude of overrepresentation test.

Funding

HAR is funded by a UNSW Scientia Program Fellowship and the Australian Research Council Discovery Early Career Researcher Award (DECRA) under grant DE220101210. This study also was funded by UNSW School of Engineering GROW Program Funding to HAR. ML is supported by a Macquarie University PhD Scholarship. AA is supported by an Australian Government Research Training Program (RTP) Scholarship.

Authors' contributions

HAR and ML conceived the idea and HAR supervised the study. ML designed and implemented the PeakCNV tool. ML analyzed the results with the assistance from AA. ML generated all Supplementary Figures and Tables with the assistance from AA. ML and AA generated Figure 1. All authors contributed to the discussions. AA drafted the manuscript with the assistance from ML and HAR. HAR, NL and AB revised the manuscript. NL helped with proof reading, simulation analysis, and improving the efficiency of the toolset. All authors read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Analysis was made possible with computational resources provided by the BioMedical Machine Learning Bioinformatics Server with funding from the Australian Government and the UNSW Sydney under grant DE220101210. We acknowledge Prof. Alistair Forrest from Harry Perkins Institute of Medical Research for his help in conceptualizing the initial idea. We thank Dr. Jeremy Thomas Keane from the University of New South Wales (UNSW) for his invaluable help in proofreading the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.001.

References

[1] Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet 2011;45:203.
[2] Woodward KJ et al. Atypical nested 22q11. 2 duplications between LCR 22B and LCR 22D are associated with neurodevelopmental phenotypes including autism spectrum disorder with incomplete penetrance. Mol Genet Genomic Med 2019;7(2):e00507.
[3] Poulton C et al. A review of structural brain abnormalities in Pallister-Killian syndrome. Mol Genet Genomic Med 2018;6(1):92–8.
[4] Hooli B et al. Role of common and rare APP DNA sequence variants in Alzheimer disease. Neurology 2012;78(16):1250–7.
[5] Kalsner L, Chamberlain SJ. Prader-Willi, Angelman, and 15q11-q13 duplication syndromes. Pediatric Clinics 2015;62(3):587–606.
[6] Sønderby IE et al. 1q21. 1 distal copy number variants are associated with cerebral and cognitive alterations in humans. Transl Psychiatry 2021;11(1):1–16.
[7] Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Chang BL. Association of a germ-line copy number variation at 2p24. 3 and risk for aggressive prostate cancer. Cancer Res 2009;69(9)):2176–9.
[8] Ledet EM et al. Characterization of germline copy number variation in high-risk African American families with prostate cancer. Prostate 2013;73(6):614–23.
[9] Ionita-Laza I et al. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 2009;93(1):22–6.
[10] Kim J-H et al. CNVRuler: a copy number variation-based case–control association analysis tool. Bioinformatics 2012;28(13):1790–2.
[11] Purcell S et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Human Genet 2007;81(3):559–75.
[12] Alinejad-Rokny H, Heng JI, Forrest AR. Brain-enriched coding and long non-coding RNA genes are overrepresented in recurrent neurodevelopmental disorder CNVs. Cell Rep 2020;33(4):108307.
[13] Larsen SJ et al. CoNVaQ: a web tool for copy number variation-based association studies. BMC Genom 2018;19(1):1–9.
[14] Rahmah N, Sitanggang IS. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. IOP Conference Series: Earth And Environmental science. IOP Publishing; 2016.
[15] Ester M et al. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd 1996.
[16] Abdi H. The Kendall rank correlation coefficient. Encyclop Measure Stat 2007;2:508–10.
[17] Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 2017;543(7644):199–204.
[18] Howe KL, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl. Nucleic Acids Res 2021;49(D1):D884–91.
[19] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47(d1):D766–73.
[20] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[21] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database 2011;2011.
[22] Consortium GP. A global reference for human genetic variation. Nature 2015;526(7571):68.
[23] Pereanu W et al. AutDB: a platform to decode the genetic architecture of autism. Nucleic Acids Res 2018;46(D1):D1049–54.
[24] Afrasiabi A, Keane JT, Heng JIT, Palmer EE, Lovell NH, Alinejad-Rokny H. Quantitative neurogenetics: applications in understanding disease. Biochem Soc Trans 2021;49(4):1621–31.
[25] Solomon BD et al. Clinical genomic database. Proc Natl Acad Sci 2013;110(24):9851–5.