Integrating Genomic Correlation Structure Improves Copy Number Variations Detection

Xizhi Luo[1], Fei Qin[1], Guoshuai Cai[2], Feifei Xiao[1*]

[1]*Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA*

[2]*Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA*

**To whom correspondence should be addressed*

Corresponding to:

Feifei Xiao, Ph.D.

Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina

Discovery 449, 915 Greene St., Columbia, SC, 29208

Tel: (803) 777-8936

Email: xiaof@mailbox.sc.edu

Abstract

Copy number variation plays important roles in human complex diseases. The detection of copy number variants (CNVs) is identifying mean shift in genetic intensities to locate chromosomal breakpoints, the step of which is referred to as chromosomal segmentation. Many segmentation algorithms have been developed with a strong assumption of independent observations in the genetic loci, and they assume each locus has an equal chance to be a breakpoint (i.e., boundary of CNVs). However, this assumption is violated in the genetics perspective due to the existence of correlation among genomic positions such as linkage disequilibrium (LD). Our study showed that the LD structure is related to the location distribution of CNVs which indeed presents a non-random pattern on the genome. To generate more accurate CNVs, we therefore proposed a novel algorithm, LDcnv, that models the CNV data with its biological characteristics relating to genetic correlation (i.e., LD). To evaluate the performance of LDcnv, we conducted extensive simulations and analyzed large-scale HapMap datasets. We showed that LDcnv presents high accuracy, stability and robustness in CNV detection and higher precision in detecting short CNVs compared to existing methods. We also theoretically demonstrated the correlation structure of CNV data, which further supports the necessity of integrating biological structure in statistical methods for CNV detection. This new segmentation algorithm has a wide scope of application with next-generation sequencing data analysis and single-cell sequencing analysis.

Author Summary

Copy number variants (CNVs) refers to gains or losses of the DNA segments in comparison to a reference genome. CNVs have garnered extensive interests in recent years as they play an important role susceptibility to disorders and diseases such as autism, schizophrenia and cancer [1-7]. Although innovation in modern technology is promoting the discoveries related to CNVs, the methodology for CNV detection is still lagging, which limits the novel discoveries regarding the role of CNVs in complex diseases. In this study, we are proposing a novel segmentation algorithm, LDcnv, to accurately locate the breakpoints or boundaries of CNVs in the human genome. Instead of utilizing an independent assumption of the signal intensities as has been used in traditional segmentation algorithms, LDcnv models the correlation structure in the genome in a change-point CNV detection model, which allows for accurate and fast computation with a whole genome scan. Our study showed strong theoretical evidence of the existence of correlation structure in real CNV data, and we believe that taking this evidence into consideration will improve the power of CNV detection. Extensive simulation studies have demonstrated the advantage of the LDcnv algorithm in stability, robustness and accuracy over existing methods. We also used high-quality CNV profiles to further support the superior performance of the LDcnv algorithm over existing methods. The development of the LDcnv algorithm provides great insights for new directions in developing CNV detection tools.

Keyword: copy number variation; segmentation algorithm; linkage disequilibrium; detection accuracy; modern genetic data

1. Introduction

Copy number variants (CNVs), as a major source of genetic variation in the human genome, are gains or losses of the DNA segments in comparison to a reference genome. Recently, copy number variation has garnered considerable interest as it plays an important role in the susceptibility to disorders and diseases such as autism, schizophrenia and cancer [1-7]. It has been found that approximately 12% of the genome is subject to CNVs and nearly 80% of cancer genes harbor CNVs [8]. To date, CNV studies have been intensively conducted in many disease types which demonstrates that CNVs account for an abundance of genetic variation and play essential roles in the etiology of cancer [9-11], autoimmune diseases [12, 13] and neurological diseases [14, 15]. Specifically, about 200 CNVs have been found to be associated with breast cancer risk, among which 21 had prognostic potential [9]. Also, copy number gain of beta-defensin genes has been revealed to be associated with increased risk of psoriasis in three independent cohorts of European origin [9, 14, 16]. Two recent reports also illustrated the possible roles of CNVs in lung cancer predisposition [17, 18]. In one of these two reports, an approximately 2-fold increased risk was observed among carriers with deletion of the gene coding region of *WWOX* compared to non-carriers [17].

Although the potential clinical application of CNVs still remains uncertain, understanding the mechanisms underlying these influences will be instrumental for many basic research areas. Consequently, the detection and association of CNVs with quantitative traits and clinical phenotypes comprise critical steps toward a better understanding of disease etiology. However, due to the complexity of CNV genetics as well as numerous factors in the data generation and computational analyses that may lead to spurious associations, the discovery of CNVs in human diseases is still inadequate, which places obstacles in the path of utilizing CNVs as important biomarkers for clinical applications.

Technically, the detection of CNVs is the finding of breakpoints or boundaries of copy number regions from the genotyping signals, the step of which is called chromosomal segmentation. Change-point tests have been commonly used and implemented in many software and tools for chromosomal segmentation [19-21]. Among them, circular binary segmentation (CBS) is widely used and is based on an exhaustive test [22]. This segmentation algorithm has been widely utilized in whole exome sequencing data tools such as ExomeCNV [23] and CODEX [24]. In EXCAVATOR [25], a shifting level model segmentation algorithm was used to incorporate the distance between two genetic sites. More recently, a novel segmentation procedure was utilized in modSaRa that adopted a local search strategy and was demonstrated to be suitable for whole genome analysis with low computational complexity [26-28]. Nevertheless, all of these algorithms were developed with a strong assumption of independent observations in the genetic loci and they assume each locus has an equal chance to be a breakpoint (i.e., boundary of CNVs). However, this assumption is violated in the genetics perspective given the existence of inter-correlation among genomic positions, which is referred to as linkage disequilibrium (LD). Dictated by the presence of recombination hotspots that segment the genome into separate LD blocks, LD describes the correlated transmission of the alleles at adjacent locations in the genome. Interestingly, it was demonstrated early on that CNVs are outcomes of evolution and they originated from recombination-based processes [29]. These relationships demonstrate the possibility of the existence of CNV breakpoints located at the recombination hotspots, which violates the assumption of previous segmentation methods that assume each genetic location has an equal chance to be a CNV breakpoint. This further implies the importance of integrating the biological characteristics (i.e., LD structure) into statistical modeling for CNV detection.

Motivated by this fact, we here developed an accurate and fast segmentation algorithm by modeling the genomic correlation structure with a local search strategy for optimized

computational efficiency. Early in 1996, Kim HJ. [30] explored a likelihood ratio test for non-independent observations, yet its computational complexity with its exhaustive search largely limited its application. To investigate the performance of the newly proposed algorithm, we conducted simulation studies to investigate its performance in single nucleotide polymorphs (SNP) array studies in a variety of scenarios. In this study, we demonstrated the improved performance of the novel algorithm in array-based real data analysis by using a set of "gold standard validation sets" of CNVs from the HapMap projects [31-33]. Overall, the new algorithm presented high sensitivity and accuracy in CNV detection, especially single copy changes. This new segmentation algorithm has a wide scope of application so it can be implemented in CNV detection tools for next-generation sequencing data analysis and single-cell sequencing analysis.

2. Materials and Methods

2.1 Notations and models

We use $\mathbf{Y} = (Y_1, \dots, Y_m)^T$ to denote the genetic intensities for a sequence with m biomarkers (e.g., SNPs in array data, or exon in whole exome sequencing data). For example, $\mathbf{Y}$ may present the log R Ratio (LRR) intensities of a chromosome from SNP array data. The model we consider is a change-point method with the basic model as

$$Y_i = \mu_i + \varepsilon_i, \ i = 1,2, \dots, m. \ (1)$$

Its underlying mean $\boldsymbol{u} = (u_1, \dots, u_m)^T$ is a piecewise constant vector. A change-point method is to find change points defined as a position τ such that $\mu_\tau \neq \mu_{\tau+1}$. The locations of the change points therefore indicate the location of breakpoints or boundaries of CNVs. We assume that there are M change points in the sequence, $0 < \tau_1 < \cdots < \tau_M < m$. Considering the sparsity of CNVs across the genome, we assume m is large and M is small. The goal is therefore to estimate the number of change points, M, and the location of the change points by the location vector, $\boldsymbol{\tau} =$

$(\tau_1, \ldots, \tau_M)^T$. Many studies have worked on the problem of identifying the location of breakpoints when the *Y*'s are independent, such as CBS and screening and ranking algorithm (SaRa) [22, 26]. In this paper, to capture the biological characteristics in the process of copy number states inference, we assume the genetic intensities follow a multivariate normal distribution given the dependence structure of the genome (i.e., LD).

$$Y \sim MVN(\mu, \Sigma), \qquad (2)$$

where Σ is the covariance matrix with dimension $m \times m$. The covariance matrix (Σ) can be estimated by using the correlation matrix estimated from the samples or an exterior dataset such as samples from the 1000 Genomes project [34].

For a point x, a local diagnostic function $D(x)$ is defined as the average mean difference in the observations before and after the point.

$$D(x, w) = \sum_{k=1}^{w} Y_{x+1-k}/w - \sum_{k=1}^{w} Y_{x+k}/w, \quad (3)$$

where w is the bandwidth. The quantity of $D(x, w)$ depends on the local $2w$ data points $\tilde{Y}$: $Y_{\tau+1-w}$,..... Y_{τ} ,..... $Y_{\tau+w}$ where $\tilde{Y} \sim MVN(\tilde{\mu}, \tilde{\Sigma})$. $\tilde{\mu}$ is a sub-vector of μ with length $2w$; $\tilde{\Sigma}$ is a sub diagonal matrix of the covariance matrix Σ with dimension $2w \times 2w$, respectively. Then $D(x)$ can be rewritten as $D(x) = \vec{a}\tilde{Y}$. $\vec{a}$ is a $2w$ vector takes the form $\frac{1}{w}[\mathbb{I}_{w\times 1} \ -\mathbb{I}_{w\times 1}]$. By derivation with the linear property of multivariate normal distribution, we obtained $D(x) \sim N(\vec{a}\tilde{\mu}, \vec{a}\tilde{\Sigma}\vec{a}^T)$. It turned out that the distribution of the local diagnostic function became a univariate normal with a covariance matrix depending on the local information, $\tilde{\Sigma}$. Since both $\tilde{\mu}$ and $\tilde{\Sigma}$ are known or can be estimated, the mean and variance of $D(x)$ are functions of bandwidth w and only depend on the local sequence. As such, we proposed the algorithm, referred to LDcnv, based on a multivariate normal assumption that systematically integrated the biological characteristics into statistical modeling of the genetic intensities.

2.2 Copy number inference

After calculation of the local diagnostic statistic $D(x)$, hypothesis testing is needed to find change-point candidates by a local screening and ranking strategy [26, 35]. A similar strategy has been used in our previous work [28] which guarantees computational speed of the CNV detection method in a whole genome scan.

Providing the distribution of $D(x)$, we first define the w-local maximizer of a function. For any data point x, the interval $(x-w, x+w)$ is called the w-neighborhood of x. And, x is a w-local maximizer of function $f(\cdot)$ if f reaches the maximum at x witin the w-neighborhood of x. In other words,

$$f(x) \geq f(x') \text{for all } x' \in (x-w, x+w). \tag{4}$$

Then let $\mathcal{L}$ be the set of all local maximizers of the function $|D(x,w)|$ and we can select a subset $\widehat{\mathcal{M}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\widehat{M}}\} \subset \mathcal{L}$ by setting a threshold $|D(\hat{\tau}, w)| > \gamma$, where $\widehat{\mathcal{M}}$ and $\widehat{M}$ are the estimators for the locations and the number of change-points, respectively. To set up the threshold γ, we adopted a multiple comparison method using a false discovery rate approach (**Supplementary Text A1**). Empirical distribution of the local maximizers of the diagnostic statistic was generated mimicking the normal sequence with no change-points. As a result, the local maximizers $\widehat{\mathcal{M}}$ or, equivalently, local minimizers of p-values were selected.

Then we used the modified Bayesian Information Criteria (mBIC) to further eliminate false positives as proposed in [36]:

$$mBIC(\widetilde{M}) = \frac{n}{2} log(\hat{\sigma}^2_{\widetilde{M}}) + \tilde{J} log(n) + \frac{1}{2}\sum_{i=1}^{\widetilde{M}+1} log\left(\frac{x_{(i)}}{n} - \frac{x_{(i-1)}}{n}\right), \tag{5}$$

where $\widetilde{M}$ is all the possible values of $\widehat{M}$ and $\hat{\sigma}^2_{\widetilde{M}}$ is the maximum likelihood estimator of the variance assuming $x_1, \ldots, x_{\widetilde{M}}$ are change points. Then the final estimated number and the locations of change points are $\widehat{M}' = argmin(mBIC(\widetilde{M}))$ and $\widehat{\mathcal{M}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\widehat{M}'}\}$, respectively. For copy number inference, Gaussian mixture model-based clustering was used for copy number state classification. Each segmented region will be classified using a five-

state classification scheme (deletion of a single copy, deletion of double copies, normal/diploids, duplication of a single copy, and duplication of double copies) [37].

2.3 Numerical simulation studies

With the new proposed algorithm, we conducted extensive simulations to evaluate the performance in practice. We simulated SNP array data for demonstration of its advantage over existing algorithms.

To simulate the correlated genomic intensities, we used the first-order autoregressive (AR1) process:

$$Y_i = c + \phi Y_{i-1} + \varepsilon_i, \qquad i = 1,2, \dots n, \qquad (6)$$

where Y_i was the intensities for the i-th marker; ε_i was a Gaussian white noise process with mean zero and variance σ_ε^2; ϕ was a known coefficient that controlled the autocorrelation of the data series (for example, $|\phi| < 1$ generates a stationary sequence); c was a constant and n was the total number of markers. The underlying mean μ, variance $var(Y_i)$ and auto-covariance B_n were given as: $\mu = \frac{c}{1-\phi}$, $var(Y_i) = \frac{\sigma_\varepsilon^2}{1-\phi^2}$ and $B_n = \frac{\sigma_\varepsilon^2}{1-\phi^2}\phi^{|n|}$. One advantage of using the AR1 process is that the change of the underlying distribution of the white noise term allows one to flexibly adjust the distribution of the data, especially when the normality is not satisfied. Also, with the AR1 process, we did not need to decompose the covariance matrix , which consequently generated data much faster than the multivariate normal distribution assumption-based process.

We randomly generated LRR and B Allele frequencies (BAF) intensities for 100 sequences (i.e., chromosomes) with 20,000 markers. For each sequence, 40 dispersed CNV segments were generated, the locations of which were randomly selected and were not overlapping with each other. The mean and variance were empirical values provided by the Illumina website (https://www.illumina.com/documents). We constructed different scenarios with different

combinations of CNV sizes, status and correlation coefficient. The CNV sizes varied from 10~50 markers, 50~100 markers and 100~200 markers. The CNV status included deletion of a single copy (Del.s), deletion of double copies (Del.d), duplication of a single copy (Dup.s) and duplication of double copies (Dup.d). To investigate the CNV data with different levels of correlations, the value of ϕ. which was equivalent to the Pearson's correlation coefficient in theory, was set to be 0.1, 0.3 and 0.5, respectively. With the generated CNV data, we compared the proposed method to the performance of an independence assumption-based method, CBS, and a hidden Markov model based method, PennCNV [22, 38]. The PennCNV assumes a Markov chain; however, the LD structure is not directly incorporated in the statistical modeling. The performance of these methods was demonstrated by computing the true positive rate (TPR) and false positive rate (FPR).

2.4 Performance evaluation by application to HapMap datasets

To further assess the proposed LDcnv algorithm, we analyzed 180 healthy individuals with CNV profiles having been validated experimentally or statistically by three previous microarray studies [31-33]. The HapMap project utilized stringent genotyping quality control (QC) and merged results from multiple calling algorithms, which finally produced 856 high-quality CNV calls [33]. McCarroll et al. identified 1,320 high resolution CNV calls by joint analysis of multi-platforms data including Affymetrix SNP array, array CGH and fosmid end-sequence-pair, whereas Conrad et al. used tiling oligonucleotide array to generate a map of 11,700 CNVs, among which 8,599 were independently validated through stringent validation procedures such as quantitative PCR [31, 32]. The SNP array data were downloaded from the international HapMap 3 Consortium [33]. All individuals were Utah residents with Northern and Western European ancestry (CEU). Genotype data were generated by the genotyping platform Affymetrix Human SNP array 6.0. Specifically, a stringent QC procedure was

adopted (e.g., the CNV must overlap with 2 to 20 exons with less than 5% missing rate across all samples) to generate high-quality CNV profiles.

Using the final "gold standard validation sets", we compared the performance of the LDcnv method against PennCNV and CBS [22, 38]. To obtain high-quality CNV profiles, we excluded CNVs with less than ten markers in the calling results. Besides, we used the database of genomic variants (DGV) [17] as a reference of common variants as a quality control step to keep the high-quality CNV profile. DGV curates CNV records from 55 independent studies of clinically normal populations with 202,431 CNV regions.

These methods were assessed by the precision rate, recall rate and F1 score measures. The precision rate was defined as the ratio of identified true positives over the total number of identified CNVs. The recall rate was the ratio of identified true positives over the total number of "true CNVs" in the "gold standard validation sets". The F1 score was defined as the harmonic mean of precision and recall rate which reflected the overall accuracy. Moreover, we also evaluated the performance in subsets of the validation sets of those that were less than ten markers to assess the performance in detecting short CNVs.

2.5 Theoretical Derivation: Correlation Structure in CNV data

In this study, we hypothesized that the integration of biological characteristics will increase the accuracy of CNV detection. In this section, we therefore provide theoretical evidence to support that the genetic intensity data are presenting a correlation structure that should be deliberated in statistical modeling for CNV detection.

We start from the generation of the two SNP array intensities, LRR and BAF, which has been introduced in Wang et al. [38] and summarized here. For one SNP with two alleles defined as A and a, the raw signal intensity values are measured for each allele and then are processed with a five-step normalization procedure using the information of all SNPs. X and Y values are produced for each SNP, representing the normalized signal intensity on the A and a alleles,

respectively. Two additional measures are then calculated for each SNP, where $R = X + Y$ representing the total signal intensity, and $\theta = \frac{argtan(Y/X)}{\pi/2}$ referring to the relative allelic signal intensity ratio. As a normalized measure of total signal intensity, LRR measures the normalized total intensity of the possible alleles for a given marker, from which the magnitude of mean changes (or called jump size) are used for inference of the boundaries of CNVs (i.e., breakpoints). The LRR value for each SNP is calculated as $LRR = log_2\left(R_{observed}/R_{expected}\right)$ where $R_{expected}$ is computed from linear interpolation of canonical genotype clusters [39].

In our study, to present the correlation of two adjacent bi-allelic SNPs, we assume the reference allele and alternative allele were *A* and *a* for the first SNP, and *B* and *b* alleles for the second SNP. The total signal intensities for the two alleles are therefore $X_A + Y_A$ and $X_B + Y_B$. We assume that we observe the two reference alleles *A* and *B* with frequencies p_A and p_B in the whole population. Under the Hardy-Weinberg equilibrium assumption, the joint probability for the nine genotypes can be calculated (shown in Table 1). For example, for the genotype *AABB*, the genotype frequency will be $(p_A p_B + D_{AB})^2$ where D_{AB} is the coefficient of linkage disequilibrium between the two SNPs.

First, to calculate the correlation of LRR between the two SNPs, we need to compute the correlation of the non-linear logarithm transformation of the observed total signal intensities $log_2\left(R_{observed}/R_{expected}\right)$. The observed total signal intensities, $R_{observed}$, can be calculated from the dataset, and the value of $R_{expected}$ is a fixed value. After applying the Taylor expansions [40], the correlation of the LRR intensities can be approximately represented by the correlation of $R_{observed,A}$and $R_{observed,B}$, which is expressed by

$$\rho_{AB} = \frac{cov(X_A+Y_A, X_B+Y_B)}{\sqrt{var(X_A+Y_A)var(X_B+Y_B)}}. \qquad (7)$$

Derivation of the ρ_{AB} will show the correlation structure of the LRR intensities, which will be further discussed in the results **(Section 3.1)**.

3. Results

3.1 Theoretical proof revealed the correlation structure in intensity signals

First, to further demonstrate the necessity of integrating the correlation structure into the segmentation algorithm, we initiated a theoretical derivation to demonstrate the correlation structure of the genetic intensities (i.e., LRR) between two adjacent genomic loci. As a continued discussion of Section 2.5, we have the correlation coefficient of LRR between two loci expressed as $\rho_{AB} = \frac{cov(X_A+Y_A, X_B+Y_B)}{\sqrt{var(X_A+Y_A)var(X_B+Y_B)}}$, in which X and Y are the normalized signal intensities of the two alleles in a SNP (e.g., A and a).

For $cov(X_A + Y_A, X_B + Y_B)$, we obtained

$$cov(X_A + Y_A, X_B + Y_B) = cov(X_A, X_B) + cov(X_A, Y_B) + cov(Y_A, X_B) + cov(Y_A, Y_B). \quad (8)$$

As $cov(X_A, X_B) = E(X_A X_B) - E(X_A)E(X_B)$, the expected values of the normalized signal intensities X_A, X_B and the expected values of their product need to be derived. We assume the joint probability density function to be $f_{X_A,X_B}(x_A, x_B)$ which are bivariate normal distributions conditional on the genotype G:

$$f_{X_A,X_B}(x_A, x_B) = \sum_{k=1}^{4} f_{X_A,X_B}(x_A, x_B|G)\, P(G = G_k), \quad (9)$$

where $G = [AABB, AABb, AaBB, AaBb]^T$ is the vector of genotypes that contain alleles A and B. After mathematical derivation (detailed in **Supplementary Text A2**), the covariance between the two normalized signal intensities can be formulated as:

$$cov(X_A, X_B) = \sum_{k=1}^{4} E(X_A|G_k)E(X_B|G_k)[P(G = G_k) - q(G = G_k)]. \quad (10)$$

$q(G = G_k)$ is the genotype frequency under the condition that the two loci are in LD. For example, $q(AABB) = p_A^2 p_B^2$. The expression of all the other genotype frequencies $P(G = G_k)$ can be found in Table 1.

Similarly, we can derive the other three terms in equation (8) and then obtain $cov(X_A + Y_A, X_B + Y_B)$ as

$$\sum_{i=1}^{4}\sum_{k=1}^{4} E(X_A|G_{ik})^{I_{1i}}E(X_B|G_{ik})^{I_{2i}}E(Y_A|G_{ik})^{1-I_{1i}}E(Y_B|G_{ik})^{1-I_{2i}}[P(G_{ik}) - q(G_{ik})], \quad (11)$$

where $G_{1.} = [AABB, AABb, AaBB, AaBb]^T$, $G_{2.} = [AABb, AAbb, AaBb, Aabb]^T$, $G_{3.} = [AaBB, AaBB, aaBB, aaBb]^T$ and $G_{4.} = [AaBb, Aabb, aaBb, aabb]^T$. I_{1i} and I_{2i} are indicator functions of whether X_A and X_B contribute to the bivariate density in equation (9). $I_{1i} = 1$, if $i = 1$ or 2. $I_{2i} = 1$, if $i = 1$ or 3. Otherwise, $I_{1i} = I_{2i} = 0$.

Combining results from equation (11) and the expression of the denominator of ρ_{AB} in equation (7) from **Supplementary Text A3**, the correlation of LRR between the two loci can be defined as:

$$\rho_{AB} = \frac{\sum_{i=1}^{4}\sum_{k=1}^{4} E(X_A|G_{ik})^{I_{1i}}E(X_B|G_{ik})^{I_{2i}}E(Y_A|G_{ik})^{1-I_{1i}}E(Y_B|G_{ik})^{1-I_{2i}}[P(G_{ik}) - q(G_{ik})]}{\sqrt{\pi_1 var(X_A) + \pi_2 var(Y_A)}\sqrt{\pi_3 var(X_B) + \pi_4 var(Y_B)}}. \quad (12)$$

According to expression of Equation (12), the correlation of LRR depends on the correlation of the two SNPs which was measured by the LD coefficient D_{AB}. For example, $P(G_{ik}) - q(G_{ik}) = \left(p_A p_B + D_{AB}\right)^2 - p_A^2 p_B^2$ for genotype $AABB$ ($i = 1, k = 1$). $\rho_{AB} = 0$ if $D_{AB} = 0$.

In summary of this section, we found that the correlation of LRR between two loci are related to the coefficient of LD measure, although the relationship does not admit a simple format.

3.2 Real data shows that that CNV locations are related to the Genomic Structure

To explore the relationship between CNV locations and LD structure, we utilized the high-quality CNV profile from the international HapMap 3 Consortium which merged probe-intensity data from both Affymetrix and Illumina arrays [33]. The CNV profile set contained

856 records for 1,184 individuals. We randomly selected 300 high-quality CNVs and mapped them to the LD block maps (Figure 1). Obviously, most of these CNVs are located outside of the LD blocks (across block, hybrid or random), with only 2.0% residing within LD blocks (inter-block) (Supplementary Table 1). Among the CNV types that do not involve LD structure (i.e., across block, inter-block and hybrid), there were only 10.7% of the CNVs were located within blocks. These results implicated that CNVs are not randomly distributed across the genome, and their distribution is closely related to the local LD structure. Such results motivated the development of the LDcnv algorithm and are consistent with the theory derivation of the correlations structure in SNP array data (**Section 3.1**).

3. 3 Simulation studies show improved performance of LDcnv

First, we used the simulated data to evaluate the performance of the LDcnv method in SNP array analysis under a variety of scenarios: (1) different correlation level; (2) different CNV sizes; and (3) different CNV status (see Methods). For data with moderate correlation coefficient ($\phi = 0.3$) that is assumed to be close to real data, the LDcnv method presented a consistent power gain in detecting CNVs from single copy duplication/deletions (i.e., Dup.s and Del.s) to double copy changes (i.e., Dup.d and Del.d) (Table 2). The performance of the LDcnv method was obviously superior to the other two methods when the CNVs had small jump sizes (Dup.s and Del.s), especially for short CNVs (<200 markers). For example, when the CNVs had a length between 10-50 markers and the CNV status was single copy duplication (Dup.s), the LDcnv method had a TPR at 0.88, while the FPR was 0.12. The corresponding TPRs and FPRs for PennCNV were 0.84 and 0.11; and 0.69, 0.32 for CBS, respectively. When the CNV size increased from 10-50 markers to 100-200 markers, the LDcnv methods presented stable estimations of CNVs, whereas the other two methods had diminished power. A similar pattern was observed when the correlation increased from 0.1 to 0.5 (Table 3).

In conclusion, the LDcnv method which integrated the correlation structure in the model, largely presented overall high accuracy, stability and robustness in CNV detection, especially for detection of CNVs with small jump sizes.

3. 4 Application to the HapMap datasets

We further applied the LDcnv method to a real data study in comparison with CBS and PennCNV [22, 38]. Using the DGV as a common variant reference database, 79.65% of the CNVs identified by LDcnv have been reported as common variants that are not diseases relevant which implicated the validity of the CNV calls from our method.

With the validation sets from the three datasets (including HapMap3, Conrad et al., and McCarrol et al.), the total number of "true" CNVs included in the three sets were 19,936, 121,453 and 11,961, separately. Among them, 10,005, 98,387 and 5,277 were short CNVs with length less than ten markers. The overall accuracy of the LDcnv method was greater than that of the two other methods in all three validation sets (Table 4, Figure 2). Specifically, LDcnv presented higher recall rate or sensitivity that it detected more true positives. For detection of short CNVs (Table 5, Figure 3), the LDcnv method was comparable to the other methods in all three validation sets and it presented obviously higher precision or specificity in detecting short CNVs. It is noteworthy that CBS was the most sensitive one that detected the highest number of variants which was consistent with previous findings showing that CBS was good at calling the exact boundaries of CNVs [44], but such high recall rate came at the expense of precision. CBS also has the lowest computational speed among these three methods. As expected, PennCNV was the most conservative one that detected the lowest number therefore presented the lowest prevision rate (Table 5). These results further demonstrated that the integration of correlation structure significantly improved the overall performance of CNV detection.

4. Discussion

CNVs play important roles in human complex diseases [1-7].While numerical CNV detection tools have been developed for modern genotyping technologies, current detection methods are using segmentation algorithms that mainly focus on detecting random signals and do not fully address the CNV-specific challenges. These challenges include the multiple natural features of a CNV, including dependent random noise signals and small jump sizes of the breakpoints. In this work, we introduce a correlation-based segmentation algorithm for CNV detection analysis that accommodates the non-independence nature of the genetic intensities. Simulation and real data analyses suggested that the LDcnv algorithm presented stable performance across different scenarios of CNV sizes, states and correlation coefficients, and it had a better or comparable accuracy compared to the independence assumption methods. The largest power gain tended to occur when CNVs were short and with small jump sizes, e.g., the duplication of a single copy.

This is the first report that demonstrates the promise of improved CNV detection by integrating the biological characteristics (i.e., LD structure) into statistical modeling. Utilization of these biological characteristics of CNVs improved accuracy and boosted power with the noisy and complex data. Instead of assuming equal weight on each chromosomal site, the LDcnv algorithm tends to put more weight on recombination hotspots which are more likely to be CNVs. For example, those CNVs with specific LD block structures missed by traditional algorithms will be identified by the new algorithms. This approach will also provide a valuable knowledgebase for CNV detection methodology development to integrate the genomic structure that delivers comprehensive information among genes.

In this study, we first presented the theoretical derivation of the correlation structure of the genetic intensity data from SNP array data. Well-developed array-based CNV analytical tools are usually based on segmentation and smoothing of LRR and BAF [38]. We found that the array-based LD structure, which was computed from the genotype frequencies, can be reflected

in the correlation structure of the genetic intensity data. We stated that the correlation between two loci in the genetic intensity will depend on the LD coefficient computed from SNP allele frequencies. This evidence provides strong evidence of the existence of genomic correlation structure in the CNV data.

By implementing the correlation structure in the statistical model, we showed that the LDcnv algorithm presented essential advantages over the other independence assumption methods (e.g., CBS and PennCNV). These advantages are demonstrated by the simulation studies and the HapMap project with the "gold standard validation sets". The superiority of the LDcnv algorithm over PennCNV was further demonstrated, especially in detecting short CNVs, which is the most difficult copy number states to be detected due to the embedded undetectable signal in the random noises. A possible explanation for this phenomenon is that short CNVs tend to have more evident correlation structure when they are located within an LD block. Such a characteristic cannot be easily captured by the hidden Markov model adopted in PennCNV, which assumes a constant level of dependence across the genome.

Indeed, the clinical relevance of small CNVs has been demonstrated in many studies in recently years. For example, Reza et al. [41] investigated a cohort of 714 patients with neurodevelopment disorders and verified the diagnostic importance of small CNVs. However, due to the noise of genotyping data, small segments are usually very difficult to distinguish from the normal noise signals. As such, the LDcnv algorithm may serve as an important tool for detecting small CNVs. Our result in simulations and the analyses of the HapMap CEU samples showed that LDcnv was capable of capturing those signals.

With LDcnv, we address the non-independence noise signal assumption by introducing a covariance matrix in the statistical modeling. To retain the covariance structure in the model, we can either use the correlation matrix estimated from the samples in the data or LD-based computation from reference samples (i.e., samples from the 1000 Genomes project). The

advantage of the data-based estimation of the covariance relies on its feasibility and simplicity; however, the covariance might be data specific and the computational concerns will be encountered for large sample sizes. In contrast, the LD-based estimates with information coming from the population level might be more stable but be susceptible to a specific population substructure. As discussed in Mathew et al. [42], an alternative way is to use the map functions (e.g., the Haldane function) in an exponential function to estimate the covariance structure on each chromosome.

Moreover, the current study of the LDcnv algorithm is mainly focuses on its application to SNP array data, but it has great potential to be implemented in the whole exome sequencing (WES) or whole genome sequencing (WGS) data analysis. The main challenges for sequencing data come from the high level of biases and artifacts effects, which therefore require a well-designed normalization procedure. Another difficulty is that the WES data are count data that requires discrete variable distribution. It has been suggested that read counts are not appropriately modeled by the normal assumption model, even after a commonly used log-based transformation is applied [43]. To allow the implementation of the LDcnv algorithm, the modelling framework and segmentation procedure need to be adjusted accordingly in NGS data analysis.

Through theoretical derivation, it is the first report in which we showed the relationship between the SNP-based LD coefficient and the correlation coefficient in the intensity data. We have demonstrated that the LD structure can be reflected in the genetic intensity data (i.e., LRR). However, the correlation structure of the other important source of information, the BAF intensities, was not clear and not easily constructed. The theoretical study of BAF and the implementation of this information requires future studies. In addition, we used 300 high-quality CNV data from HapMap to demonstrate that CNVs were not randomly distributed across the genome. One possible explanation might be that the LD blocks under study were not

long enough to contain a CNV. Rigorous examinations of this assumption with shorter CNVs should be further studied in the future. Still, the LDcnv algorithm may open a door for integrating biological characteristics in CNV detection methodology development with change-point methods.

Supporting Information

Supplementary Text A1. This section includes the detailed information about the FDR approach that is used in the LDcnv algorithm.

Supplementary Text A2-A3. This section includes the detailed derivation of components in the correlation structure in CNV data.

Supplementary Tables. The supplementary tables are attached to support the study are attached.

Acknowledgements

We thank the reviewers in advance for their thoughtful and insightful comments.

Funding

The preparation of this manuscript was supported by the National Science Foundation (NSF) grant DMS1722562 to Dr. Feifei Xiao.

Author Contributions

The conception and design of this study were developed by FX. The theoretical derivation and methodology development were conducted by XL and FX. The data analysis and interpretation were conducted by XL, FX, CG and FQ. The writing, review and manuscript revision were conducted by FX, XL and CG.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466(7304):368-72. doi: 10.1038/nature09146. PubMed PMID: 20531469; PubMed Central PMCID: PMCPMC3021798.
2. Chung BH, Tao VQ, Tso WW. Copy number variation and autism: new insights and clinical implications. J Formos Med Assoc. 2014;113(7):400-8. doi: 10.1016/j.jfma.2013.01.005. PubMed PMID: 24961180.
3. Castellani CA, Awamleh Z, Melka MG, O'Reilly RL, Singh SM. Copy number variation distribution in six monozygotic twin pairs discordant for schizophrenia. Twin Res Hum Genet. 2014;17(2):108-20. doi: 10.1017/thg.2014.6. PubMed PMID: 24556202.
4. O'Dushlaine C, Ripke S, Ruderfer DM, Hamilton SP, Fava M, Iosifescu DV, et al. Rare copy number variation in treatment-resistant major depressive disorder. Biol Psychiatry. 2014;76(7):536-41. doi: 10.1016/j.biopsych.2013.10.028. PubMed PMID: 24529801; PubMed Central PMCID: PMCPMC4104153.
5. Fanale D, Iovanna JL, Calvo EL, Berthezene P, Belleau P, Dagorn JC, et al. Analysis of germline gene copy number variants of patients with sporadic pancreatic adenocarcinoma reveals specific variations. Oncology. 2013;85(5):306-11. doi: 10.1159/000354737. PubMed PMID: 24217364.
6. Al-Sukhni W, Joe S, Lionel AC, Zwingerman N, Zogopoulos G, Marshall CR, et al. Identification of germline genomic copy number variation in familial pancreatic cancer. Hum Genet. 2012;131(9):1481-94. Epub 2012/06/06. doi: 10.1007/s00439-012-1183-1. PubMed PMID: 22665139; PubMed Central PMCID: PMCPMC3808836.
7. Walker LC, Wiggins GA, Pearson JF. The Role of Constitutional Copy Number Variants in Breast Cancer. Microarrays (Basel). 2015;4(3):407-23. doi: 10.3390/microarrays4030407. PubMed PMID: 27600231; PubMed Central PMCID: PMCPMC4996380.
8. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11(5):R52. Epub 2010/05/21. doi: 10.1186/gb-2010-11-5-r52. PubMed PMID: 20482838; PubMed Central PMCID: PMCPMC2898065.
9. Kumaran M, Cass CE, Graham K, Mackey JR, Hubaux R, Lam W, et al. Germline copy number variations are associated with breast cancer risk and prognosis. Sci Rep. 2017;7(1):14621. Epub 2017/11/09. doi: 10.1038/s41598-017-14799-7. PubMed PMID: 29116104.
10. Chen W, Yuan L, Cai Y, Chen X, Chi Y, Wei P, et al. Identification of chromosomal copy number variations and novel candidate loci in hereditary nonpolyposis colorectal cancer with mismatch repair proficiency. Genomics. 2013;102(1):27-34. Epub 2013/02/26. doi: 10.1016/j.ygeno.2013.02.003. PubMed PMID: 23434627.
11. Lin CH, Lin JK, Chang SC, Chang YH, Chang HM, Liu JH, et al. Molecular profile and copy number analysis of sporadic colorectal cancer in Taiwan. J Biomed Sci. 2011;18:36. Epub 2011/06/08. doi: 10.1186/1423-0127-18-36. PubMed PMID: 21645411; PubMed Central PMCID: PMCPMC3123622.
12. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49(1):27-35. Epub 2016/11/22. doi: 10.1038/ng.3725. PubMed PMID: 27869829; PubMed Central PMCID: PMCPMC5737772.
13. Li N, Zhang J, Liao D, Yang L, Wang Y, Hou S. Association between C4, C4A, and C4B copy number variations and susceptibility to autoimmune diseases: a meta-analysis. Sci Rep. 2017;7:42628. Epub 2017/02/17. doi: 10.1038/srep42628. PubMed PMID: 28205620; PubMed Central PMCID: PMCPMC5311832.
14. Stuart PE, Huffmeier U, Nair RP, Palla R, Tejasvi T, Schalkwijk J, et al. Association of beta-defensin copy number and psoriasis in three cohorts of European origin. J Invest Dermatol.

2012;132(10):2407-13. Epub 2012/06/29. doi: 10.1038/jid.2012.191. PubMed PMID: 22739795; PubMed Central PMCID: PMCPMC3447111.
15. Hou S, Qi J, Liao D, Zhang Q, Fang J, Zhou Y, et al. Copy number variations of complement component C4 are associated with Behcet's disease but not with ankylosing spondylitis associated with acute anterior uveitis. Arthritis Rheum. 2013;65(11):2963-70. Epub 2013/08/07. doi: 10.1002/art.38116. PubMed PMID: 23918728.
16. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet. 2008;40(1):23-5. Epub 2007/12/07. doi: 10.1038/ng.2007.48. PubMed PMID: 18059266; PubMed Central PMCID: PMCPMC2447885.
17. Yang L, Liu B, Huang B, Deng J, Li H, Yu B, et al. A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. Hum Mol Genet. 2013;22(9):1886-94. doi: 10.1093/hmg/ddt019. PubMed PMID: 23339925.
18. Li X, Chen X, Hu G, Liu Y, Zhang Z, Wang P, et al. Combined analysis with copy number variation identifies risk loci in lung cancer. Biomed Res Int. 2014;2014:469103. Epub 2014/08/06. doi: 10.1155/2014/469103. PubMed PMID: 25093167; PubMed Central PMCID: PMCPMC4100386.
19. Gai X, Perin JC, Murphy K, O'Hara R, D'Arcy M, Wenocur A, et al. CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. Bmc Bioinformatics. 2010;11:74. Epub 2010/02/06. doi: 10.1186/1471-2105-11-74. PubMed PMID: 20132550; PubMed Central PMCID: PMCPMC2827374.
20. Deng X. SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. Bmc Bioinformatics. 2011;12:267. Epub 2011/07/01. doi: 10.1186/1471-2105-12-267. PubMed PMID: 21714929; PubMed Central PMCID: PMCPMC3148209.
21. Darvishi K. Application of Nexus copy number software for CNV detection and analysis. Curr Protoc Hum Genet. 2010;Chapter 4:Unit 4 14 1-28. Epub 2010/04/08. doi: 10.1002/0471142905.hg0414s65. PubMed PMID: 20373515.
22. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72. Epub 2004/10/12. doi: 10.1093/biostatistics/kxh008. PubMed PMID: 15475419.
23. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27(19):2648-54. Epub 2011/08/11. doi: 10.1093/bioinformatics/btr462. PubMed PMID: 21828086; PubMed Central PMCID: PMCPMC3179661.
24. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 2015;43(6):e39. Epub 2015/01/27. doi: 10.1093/nar/gku1363. PubMed PMID: 25618849; PubMed Central PMCID: PMCPMC4381046.
25. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol. 2013;14(10):R120. Epub 2013/11/01. doi: 10.1186/gb-2013-14-10-r120. PubMed PMID: 24172663; PubMed Central PMCID: PMCPMC4053953.
26. Niu YS, Zhang H. The screening and ranking algorithm to detect DNA copy number variations. Annals of Applied Statistics. 2012;6(3):1306-26. doi: 10.1214/12-AOAS539SUPP. PubMed PMID: 24069112; PubMed Central PMCID: PMCPMC3779928.
27. Xiao F, Luo X, Hao N, Niu YS, Xiao X, Cai G, et al. An Accurate and Powerful Method for Copy Number Variation Detection. Bioinformatics. 2019. Epub 2019/01/17. doi: 10.1093/bioinformatics/bty1041. PubMed PMID: 30649252.
28. Xiao F, Niu Y, Hao N, Xu Y, Jin Z, Zhang H. modSaRa: a computationally efficient R package for CNV identification. Bioinformatics. 2017;33(15):2384-5. Epub 2017/04/30. doi: 10.1093/bioinformatics/btx212. PubMed PMID: 28453611.

29. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451-81. Epub 2009/09/01. doi: 10.1146/annurev.genom.9.081307.164217. PubMed PMID: 19715442; PubMed Central PMCID: PMCPMC4472309.
30. Kim HJ. Change-point detection for correlated observations. Stat Sinica. 1996;6(1):275-87. doi: Doi 10.1007/Bf02535741. PubMed PMID: WOS:A1996TT17400017.
31. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40(10):1166-74. Epub 2008/09/09. doi: 10.1038/ng.238. PubMed PMID: 18776908.
32. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12. Epub 2009/10/09. doi: 10.1038/nature08516. PubMed PMID: 19812545; PubMed Central PMCID: PMCPMC3330748.
33. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8. Epub 2010/09/03. doi: 10.1038/nature09298. PubMed PMID: 20811451; PubMed Central PMCID: PMCPMC3173859.
34. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. Epub 2015/10/04. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.
35. Niu YS, Hao N, Zhang HP. Multiple Change-Point Detection: A Selective Overview. Stat Sci. 2016;31(4):611-23. doi: 10.1214/16-Sts587. PubMed PMID: WOS:000392894500018.
36. Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics. 2007;63(1):22-32. Epub 2007/04/24. doi: 10.1111/j.1541-0420.2006.00662.x. PubMed PMID: 17447926.
37. Xiao F, Min X, Zhang H. Modified screening and ranking algorithm for copy number variation detection. Bioinformatics. 2015;31(9):1341-8. doi: 10.1093/bioinformatics/btu850. PubMed PMID: 25542927; PubMed Central PMCID: PMCPMC4410664.
38. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. 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):1665-74. Epub 2007/10/09. doi: 10.1101/gr.6861907. PubMed PMID: 17921354; PubMed Central PMCID: PMCPMC2045149.
39. Peiffer DA, Le JM, Steemers FJ, Chang WH, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Research. 2006;16(9):1136-48. doi: 10.1101/gr.5402306. PubMed PMID: WOS:000240238600008.
40. Benaroya H, Han SM. Probability models in engineering and science. Boca Raton: Taylor & Francis; 2005. xvi, 732 p. p.
41. Asadollahi R, Oneda B, Joset P, Azzarello-Burri S, Bartholdi D, Steindl K, et al. The clinical significance of small copy number variants in neurodevelopmental disorders. J Med Genet. 2014;51(10):677-88. Epub 2014/08/12. doi: 10.1136/jmedgenet-2014-102588. PubMed PMID: 25106414; PubMed Central PMCID: PMCPMC4173859.
42. Mathew B, Leon J, Sillanpaa MJ. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity (Edinb). 2018;120(4):356-68. Epub 2017/12/15. doi: 10.1038/s41437-017-0023-4. PubMed PMID: 29238077; PubMed Central PMCID: PMCPMC5842222.
43. 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):202. Epub 2018/11/28. doi: 10.1186/s13059-018-1578-y. PubMed PMID: 30477554; PubMed Central PMCID: PMCPMC6260772.

44. Dellinger, Andrew E et al. “Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays.” Nucleic acids research vol. 38,9 (2010): e105. doi:10.1093/nar/gkq040

Table 1: Joint genotype probabilities for two diallelic loci. The joint genotype probabilities were calculated under the Hardy-Weinberg equilibrium assumption. *A* and *a* are reference and alternate alleles in the first locus, p_A is the probability of the reference allele; *B* and *b* are the reference and alternate alleles in the second locus, p_B is the probability of the reference allele; D_{AB} is the coefficient of linkage disequilibrium between two loci.

Locus 1	Locus 2	Probabilities
AA	BB	$(p_A p_B + D_{AB})^2$
AA	Bb	$2(p_A p_B + D_{AB})(p_A(1-p_B) - D_{AB})$
AA	bb	$(p_A(1-p_B) - D_{AB})^2$
Aa	BB	$2(p_A(1-p_B) - D_{AB})\big((1-p_A)p_B - D_{AB}\big)$
Aa	Bb	$2(p_A p_B + D_{AB})\big((1-p_A)(1-p_B) + D_{AB}\big) + 2(p_A(1-p_B) - D_{AB})\big((1-p_A)p_B - D_{AB}\big)$
Aa	bb	$2(p_A(1-p_B) - D_{AB})\big((1-p_A)(1-p_B) + D_{AB}\big)$
aa	BB	$\big((1-p_A)p_B - D_{AB}\big)^2$
aa	Bb	$2\big((1-p_A)p_B - D_{AB}\big)\big((1-p_A)(1-p_B) + D_{AB}\big)$
aa	bb	$\big((1-p_A)(1-p_B) + D_{AB}\big)^2$

Table 2: Summary of CNV calls on simulated data at ϕ =0.3 from all methods. True positive rates (TPRs) and false positive rates (FPRs) of LDcnv, PennCNV and CBS with different CNV states and CNV sizes, the autoregressive coefficient (ϕ) was fixed at ϕ =0.3 which was corresponding to Pearson's correlation coefficient at 0.3. Del.d: deletion of double copies; Del.s: deletion of single copy; Dup.s: duplication of single copy; Dup.d: duplication of double copies.

CNV State	Method	CNV length (markers)					
		10~50		50~100		100~200	
		TPR	FPR	TPR	FPR	TPR	FPR
Del.d	LDcnv	0.99	<0.01	0.97	<0.01	0.99	0.01
	PennCNV	1.00	<0.01	1.00	<0.01	1.00	<0.01
	CBS	1.00	0.04	1.00	0.07	1.00	0.09
Del.s	LDcnv	0.99	0.02	0.99	0.02	0.99	0.03
	PennCNV	0.96	0.03	0.95	0.04	0.86	0.11
	CBS	0.98	0.03	0.98	0.04	0.98	0.05
Dup.s	LDcnv	0.97	0.03	0.94	0.03	0.93	0.05
	PennCNV	0.91	0.07	0.92	0.08	0.87	0.11
	CBS	0.85	0.11	0.88	0.12	0.88	0.13
Dup.d	LDcnv	1.00	<0.01	1.00	<0.01	0.99	<0.01
	PennCNV	1.00	<0.01	0.99	<0.01	0.99	0.01
	CBS	1.00	0.01	1.00	0.02	1.00	0.04

Table 3: Summary of CNV calls on simulated data at ϕ =0.5 from all methods. True positive rates (TPRs) and false positive rates (FPRs) of LDcnv, PennCNV and CBS with different CNV states and CNV sizes, the autoregressive coefficient (ϕ) was fixed at ϕ =0.5 which was corresponding to Pearson's correlation coefficient at 0.5. Del.d: deletion of double copies; Del.s: deletion of single copy; Dup.s: duplication of single copy; Dup.d: duplication of double copies.

CNV State	Method	CNV length (markers)					
		10~50		50~100		100~200	
		TPR	FPR	TPR	FPR	TPR	FPR
Del.d	LDcnv	0.99	0.01	0.96	0.01	0.99	0.03
	PennCNV	0.99	0.01	0.99	0.01	1.00	<0.01
	CBS	1.00	0.23	1.00	0.44	1.00	0.64
Del.s	LDcnv	0.96	0.06	0.95	0.08	0.96	0.09
	PennCNV	0.89	0.05	0.88	0.09	0.83	0.14
	CBS	0.94	0.26	0.94	0.46	0.95	0.62
Dup.s	LDcnv	0.88	0.12	0.89	0.09	0.91	0.11
	PennCNV	0.84	0.11	0.86	0.13	0.83	0.18
	CBS	0.69	0.32	0.80	0.57	0.79	0.76
Dup.d	LDcnv	0.99	0.01	0.96	0.01	0.99	0.03
	PennCNV	0.99	0.01	0.99	0.01	1.00	<0.01
	CBS	1.00	0.23	1.00	0.44	1.00	0.64

Table 4. Overall assessment of CNV calling on the HapMap project dataset. Performance assessment of CNV calls from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV and CBS on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The recall rate was defined as the ratio of identified true positives over the total number of "true CNVs". The F1 score was calculated as harmonic mean of precision rate and recall rate. TP: True positives among the detected CNVs.

	HapMap3				Conrad				MCC			
	TP	Precision	Recall	F1	TP	Precision	Recall	F1	TP	Precision	Recall	F1
LDcnv	4463	52.72%	22.30%	31.42	5888	64.99%	4.84%	9.02	2861	63.53%	23.91%	34.75
PennCNV	3760	53.23%	18.81%	27.85	4880	64.56%	4.01%	7.56	2640	66.48%	22.07%	33.14
CBS	4044	55.41%	20.18%	29.69	5099	65.02%	4.19%	7.88	2572	64.94%	21.40%	32.30

Table 5. Assessment of calling performance in short CNVs on the HapMap project dataset. Performance assessment on detecting short CNVs (<10 markers) from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV and CBS on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The recall rate was defined as the ratio of identified true positives over the total number of "true CNVs". The F1 score was calculated as harmonic mean of precision rate and recall rate. TP: True positives among the detected CNVs.

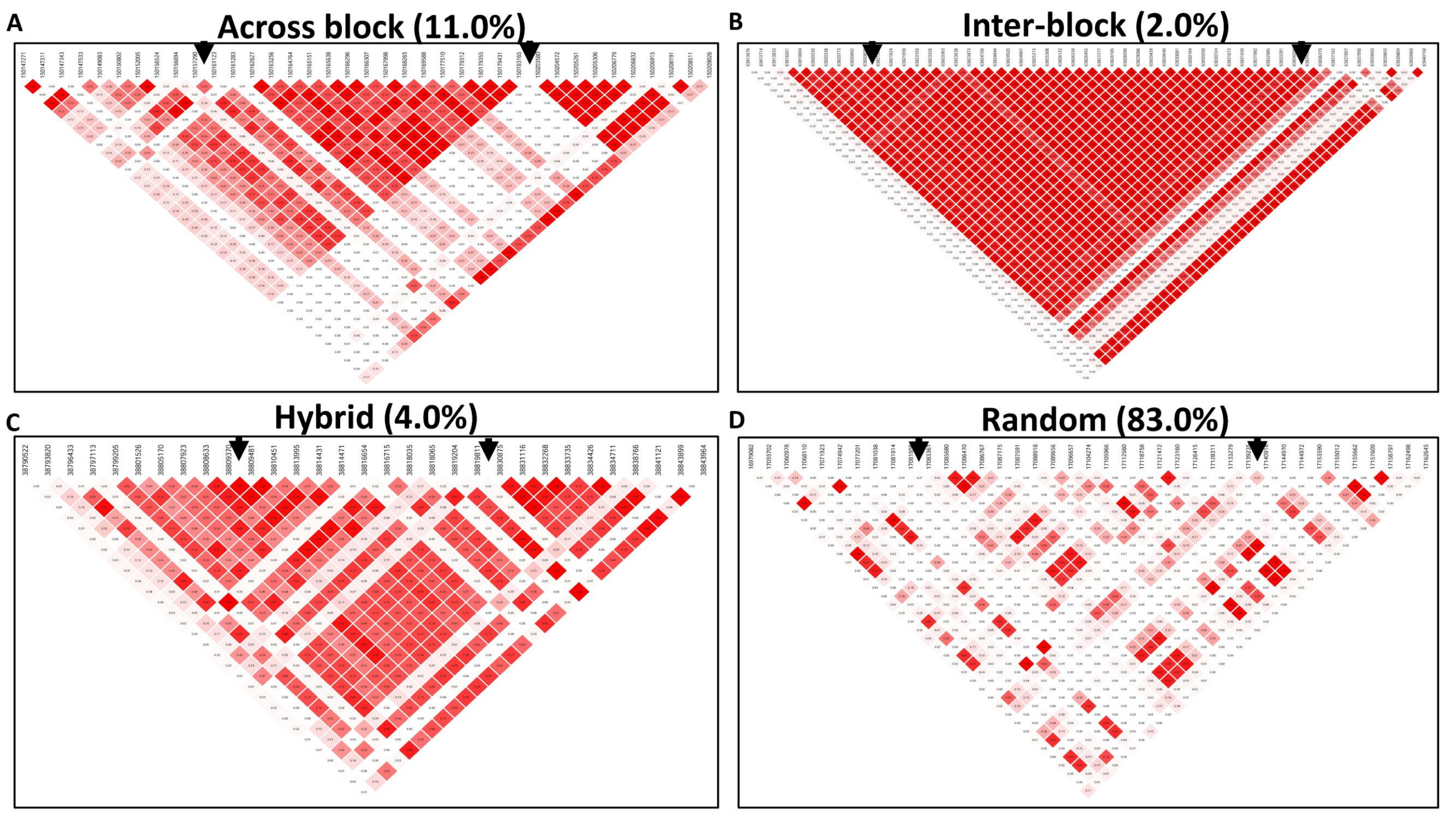
	HapMap3				Conrad				MCC			
	TP	Precision	Recall	F1	TP	Precision	Recall	F1	TP	Precision	Recall	F1
LDcnv	963	10.40%	9.62%	10.00	1757	16.39%	1.78%	3.22	703	14.16%	13.32%	13.72
PennCNV	177	2.65%	1.76%	2.12	698	8.78%	0.70%	1.31	206	5.33%	3.90%	4.51
CBS	1279	8.48%	12.78%	10.19	2933	16.63%	2.98%	5.05	850	10.59%	16.10%	12.78

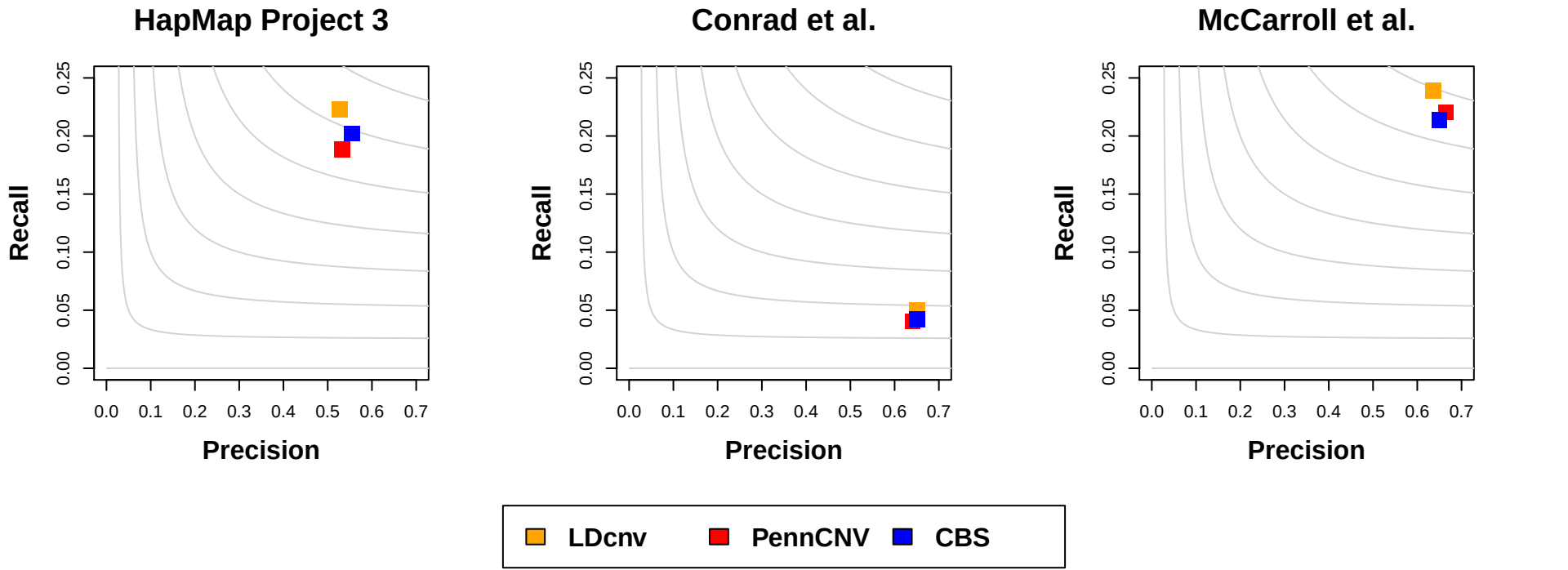
Figure legends

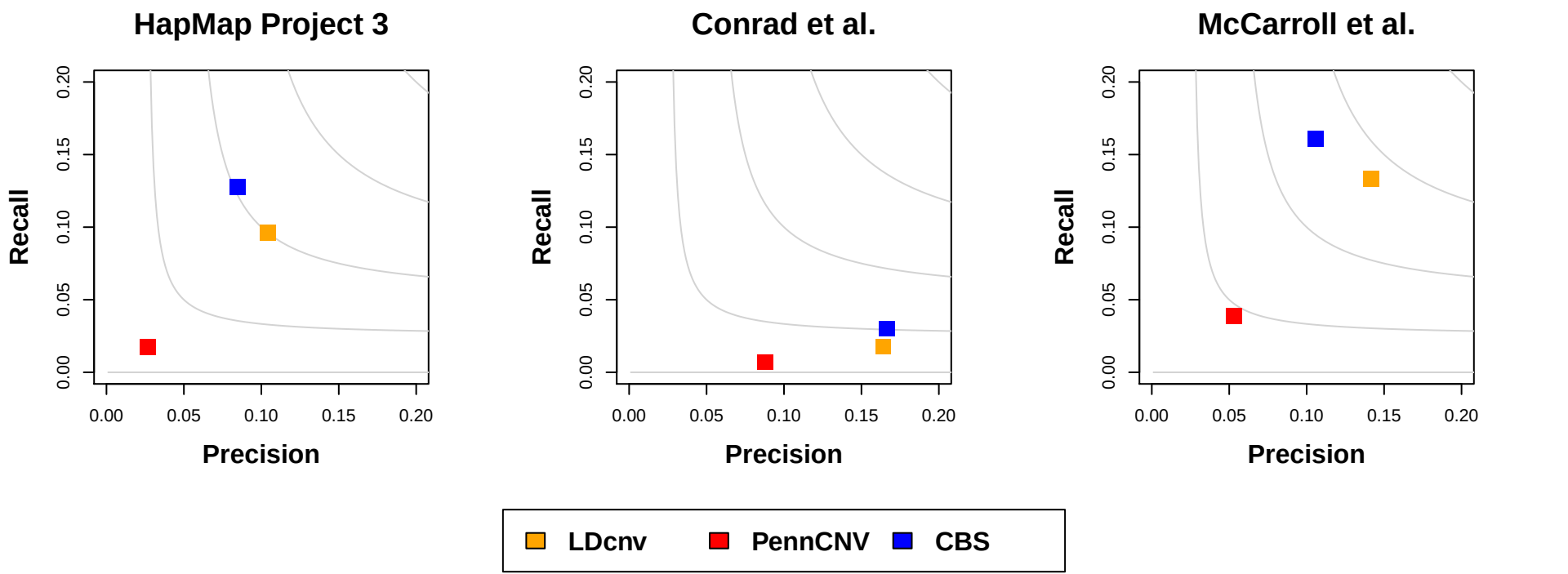
Figure 1. Four classifications of the CNV locations in the LD genome map. The graphs summarized the frequency of CNV types with the existing high-quality CNVs from the HapMap phase 3 project. (a) Across block: CNVs spanning at least one LD blocks, (b) Inter-block: CNVs locate within a LD block, (c) Hybrid: only one breakpoint locating within LD block, and (d) Random: CNVs locating in the area with weak or no LD structure. The black arrows in each plot note the start and end points of the CNV.

Figure 2. Assessment of CNV calls generated by LDcnv, PennCNV and CBS with validation CNV calls. Performance of the LDcnv, PennCNV and CBS methods in detection validated CNVs from (a) HapMap 3 (b) Conrad et al (c) McCarroll et al. The grey contours are F1 scores calculated as the harmonic mean of precision rate and recall rate.

Figure 3. Assessment of CNV calls calling performance in short CNVs on the HapMap project dataset. Performance assessment on detecting short CNVs (<10 markers) from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV and CBS on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The grey contours are F1 scores calculated as the harmonic mean of precision rate and recall rate.


A
Across block (11.0%)
B
Inter-block (2.0%)
C
Hybrid (4.0%)
D
Random (83.0%)


HapMap Project 3
Conrad et al.
McCarroll et al.
Recall
Precision
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
LDcnv
PennCNV
CBS


HapMap Project 3
Conrad et al.
McCarroll et al.
Recall
Precision
0.00
0.05
0.10
0.15
0.20
LDcnv
PennCNV
CBS