rBioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

Integrative DNA copy number detection and genotyping from sequencing and array-based platforms

Zilu Zhou[1] , Weixin Wang[2], Li-San Wang[2], and Nancy Ruonan Zhang[3, *]

[1]Graduate Group in Genomics and Computational Biology, University of Pennsylvania

[2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania

[3]Department of Statistics, The Wharton School, University of Pennsylvania

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Copy number variations (CNVs) are gains and losses of DNA segments that are highly associated with multiple diseases. Previously, many large disease cohorts perform CNV analysis with data from SNP-array. However, with the cost of Next Generation Sequencing (NGS) dropped dramatically, those large cohorts redo their CNV analysis with technologies such as whole exome sequencing (WES) and whole genome sequencing (WGS), where both SNP-array data and NGS data are available. An integrated analysis is expected to improve resolution and accuracy, yet there is no tool that is able to combine data from multiple platforms, nor has there been a systematic study of the expected accuracy improvement achievable by integration.

Results: Here we propose a statistical framework, integrated Copy Number Variation caller (iCNV), which can be applied to various study design: WES only, WGS only, SNP array only, or any combination of SNP and sequencing data. The distinguishing feature of iCNV includes adaptation of platform specific normalization, utilization of allele specific reads from sequencing and integration of matched NGS and SNP-array data by a Hidden Markov Model (HMM). We, first, compare CNV calling between joint platforms versus simply performing intersection or union of the two platforms and show that joint platform increases sensitivity and robustness. We further compare CNV detection power between single platform versus joint platforms by an in silico spike-in study and reveal that detection power increases greatly by adding array to WES, while not too much by adding array to WGS. Finally, we compare iCNV to other commonly used CNV calling methods on a WGS data set with pedigree information, which shows that iCNV improves upon the existing methods.

Availability:

Contact: nzh@wharton.upenn.edu, zhouzilu@mail.med.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Copy number variations (CNV) are large chunks of DNA that have been deleted or duplicated during evolution, leading to polymorphisms in their numbers of copies in the observed population. Studies have shown that CNV is an important type of variation in the human genome, some of which playing key roles in disease susceptibility. Accurate identification and genotyping of CNV is important for population genetic and disease studies, and can lead to improved understanding of disease mechanisms and discovery of drug targets. To detect and profile CNV, earlier studies relied on array-based technologies such as array comparative genome hybridization (CGH) or single-nucleotide polymorphism (SNP) genotyping arrays, while in recent years, next generation sequencing (NGS) technology has allowed for high resolution CNV profiling. With the drop

of sequencing costs, many large-scale genetic studies have adopted whole exome sequencing (WES) and/or whole genome sequencing (WGS) to profile genetic variation in large cohorts. Often, these cohorts were previously studied using array-based technologies. For example, Alzheimer's Disease Sequencing Project (ADSP) is an ongoing study that has 578 samples with both WGS and SNP-array data and 10913 individuals with both WES and SNP-array data; Alzheimer's Disease Genetics Consortium (ADGC) is another ongoing study that has 3084 samples with both WES and SNP-array data. It is yet unclear in such studies how to combine data from multiple platforms, and how such multi-platform integration can improve accuracy.

There is also room for improvement in the detection of CNV from NGS data alone. Sequencing data is subject to multiple sources of experimental noise such as GC bias and batch effects. Numerous CNV detection tools have been developed for sequencing data, but they often make contradicting detections on the same data set. CNV detection on SNP-array platforms has improved accuracy when B-allele frequency is used, but few of the CNV detection tools currently available for sequencing data make use of this feature.

Here we propose integrated Copy Number Variation caller (iCNV), a statistical framework for CNV detection that can be applied to multiple study designs: WES only, WGS only, SNP array only, or any combination of SNP and sequencing data. Compared to using only NGS data, iCNV improves copy number detection accuracy in three ways: (1) utilization of B allele frequency, and (2) integration of sample matched SNP-array data when available, and (3) adoption of platform-specific normalization. iCNV produces cross-platform joint segmentations of each sample's genome into deleted, duplicated, and normal regions, and further infers integer copy numbers in deletion and duplication regions.

To test iCNV, we first compare CNV calling between joint platforms versus simply performing intersection or union of the two platforms. Result suggests that our joint method has higher sensitivity and robustness. We further compare CNV detection power between single platform versus joint platforms by an in silico spike-in study, and find substantial increase in sensitivity when adding SNP array to WES, while the increase is almost negligible when adding SNP array to WGS. We also compared iCNV to other commonly used CNV calling methods on a WGS data set with pedigree information, where the degree of familial sharing indicates that iCNV the existing methods.

2 Methods

2.1 Overview of pipeline

Fig. 1 shows an overview of iCNV analysis pipeline. Input data depends on experiment design: When both SNP array and NGS data are available, the input includes (i) SNP log R ratio (LRR) and (ii) B allele frequency (BAF), which quantify, respectively, relative probe intensity and allele proportion, and (iii) sequencing mapped reads (BAM file). This pipeline is simplified when data from only one platform is available (Supplementary Figs. 1 and 2). For sequencing data, iCNV also requires target positions (BED file) for read depth normalization. In WES, the targets are exons, while for WGS, iCNV bins the genome and treats each bin as a target (the default bin size is 1kb). iCNV first performs cross-sample bias correction using CODEX and computes a Poisson log-likelihood ratio (PLR) for each target. This step also generates a filtered BED file that removes target positions that fails quality checks. Heterozygous SNPs are detected and BAFs computed within filtered target regions using SAMTOOLS. Integrated CNV detection is achieved through a hidden Markov model (HMM) that treats the array intensity, array BAF, sequencing PLR and sequencing BAF as observed emissions from a hidden copy number state. The HMM segments the genome of each sample into regions of homogeneous copy number and outputs an integrated Z-score for each position that summarizes the evidence for an abnormal copy number at that position. Finally, integer-valued copy numbers are estimated in regions of high absolute Z-score, utilizing information from all platforms.

2.2 Platform specific normalization

Due to the heterogeneity in noise between platforms, we adopt a platform specific normalization step, which aims to remove platform-specific technical biases. For sequencing data, we apply CODEX normalization (see supplement for detail) that removes biases related to target length, mappability, GC content and other latent systematic factors such as capture efficiency and amplification bias. All of these systematic factors are prevalent in sequencing protocols (REF) and detrimental to CNV detection. CODEX normalization results in a PLR representing relative intensity of the target. Sequencing BAFs of heterozygous variants in filtered targets are called by SAMTOOLs (detail see supplementation). As for array data, SNP log R ratio and BAF are standard output from genome studio, which measure relative probe intensity and allele proportion respectively.

To bring SNP array intensity ($x_{SNP,j}$) and sequencing coverage-derived PLR ($x_{WES,j}$ and/or $x_{WGS,j}$) to the same scale, we standardize each to produce a normalized intensity score:

$$y_{pj} = \frac{(x_{pj} - \frac{\sum_j x_{pj}}{K_p})}{\sigma_i},\ \sigma_i = \sum_j \left(x_{ij} - \sum_j x_{pj}\right)^2 / K_p \qquad (1)$$

where i represents sample, p represents the platform and $j \in \{1 \dots K_p\}$ represents number of targets in the platform.

2.3 HMM models

After normalization, the normalized intensity score and BAFs from all platforms are analyzed by the integrated Hidden Markov Model (HMM), shown in Figure 2, which integrates evidence from all platforms and produces a joint segmentation. The HMM model has the following key features: (i) Overlapping targets (e.g. exons, bins, or SNPs) share the same underlying copy number, even if their boundaries are not identical. (ii) Each sequencing derived target (e.g. exon or genome bin) can have multiple BAFs, if multiple heterozygous SNP loci are detected in the target. In such scenarios, the BAFs are assumed independent. (iii) Exons/genome bins that don't overlap with any heterozygous SNPs are assigned BAF value 0, because the fact that these regions are homogeneous is informative. (iv) For a hidden state c_l, there are three possible values: diploid, deletion and duplication, i.e. $c_l \in \{del, dip, dup\}$. Specific integer copy numbers are inferred post-segmentation, see Section 2.5.

Transition probabilities between hidden states rely on genomic distance traversed. Both XHMM and PennCNV use a distance-dependent exponential attenuation factor $f(d) = e^{-d/D}$, which we also adopt in our model. D is set to the mean distance between targets (default 100kb). Supplementary Fig. 3 illustrates the relationship between distance and transition probability.

There are two types of emission distributions in the HMM, one for normalized intensity score; the other for BAF. These emission distributions are shown in Supplementary Figure 3. The emission probability

distribution of coverage- and LRR-derived normalized intensity score is a mixture of three normal centered at $-\mu_1$, 0, $+\mu_3$ that share the same standard deviation σ. Previously, XHMM and PennCNV used a symmetric Gaussian mixture centered at $-\mu$, 0, and $+\mu$ with standard deviation 1 representing deletion, diploid and duplication. We found that empirically, the Z-score distribution isn't perfectly symmetric around zero and that its variance varies across individuals (Supplementary Fig. 4). In particular, the mean Z-score for duplications deviates less from 0 than that for deletion's. This motivates the differing absolute deviations $-\mu_1$ (default value -3) and $+\mu_3$ (default value 2) with shared standard deviation σ, estimated by Baum-Welch for each sample. The emission probability of BAF is more straightforward, and is a mixture of truncated normal distributions shown in Supplementary Figure 3.

2.4 Parameter estimation and quality metrics

To estimate the HMM parameters, the Baum-Welch (EM) algorithm is used to maximize the likelihood of observed data. The Viterbi algorithm is used to infer the most likely path given estimated parameters. The detailed formulas are in the supplement.

To further quantify CNV calling quality, we calculate an integrated Z-score for each hidden state as follows:

$$Z_l = e_l\sqrt{-\log\big(p(c_l = dip|y)\big)},$$

$$e_l = \begin{cases} -1, p(c_l = del|y) > \ p(c_l = dup|y) \\ +1, p(c_l = del|y) \leq \ p(c_l = dup|y) \end{cases}. \quad (2)$$

The score is positive if the hidden state has higher conditional probability of being a duplication than a deletion, and negative sign otherwise. This quality score, in a single value, integrates information from intensity/coverage and BAFs across multiple platforms, allowing for easy visualization and straightforward quality assessment.

2.5 Copy number inference

To infer the integer copy number, we use a maximum likelihood procedure. For each CNV region, we infer deletion copy number and duplication copy number separately. Deletion can only be 0 or 1 copy. For duplication, we only characterize two cases: 3 copies or greater than 3 copies. We find it very difficult to separate copy numbers 4 or greater because (1) such events are rare, thus making it hard to infer the mean of their distribution; (2) BAF distributions among 4, 5 or more copies are too similar. The likelihood calculations are shown in the appendix. We assign the maximum likelihood copy number to each CNV region.

2.6 Design of spike-in

In silico spike-in is a useful way to assess sensitivity. Spike-in differs from simulation in that signals are inserted into real data matrices which retain the true noise structure and probe distribution, thus giving more realistic projections of detection power. In our spike-in design, we start by removing CNV regions detected by our program from original datasets. We then add CNV signals randomly to the presumed diploid region with lengths ranging from 100bp to 500kb. Exons, bins or SNPs that overlap with the added CNV regions have their intensities and BAFs changed according to the simple and standard model described in the supplement. As a result, not all of the spike-in CNVs are detectable, especially when the data comes from WES and SNP arrays with low-resolution target set. iCNV is then applied to the spike-in dataset, using the single-platform mode for each platform as well as the integrated multi-platform mode combining the platforms. Results are compared with the underlining truth for sensitivity assessment.

2.7 Samples and datasets

To evaluate the accuracy of iCNV, and also to serve as illustration, we analyze two set of samples from the Alzheimer's Disease Sequencing Project. The first set contains 38 unrelated individuals with SNP array and WES data. The second set comprises of 75 related individuals with SNP array and WGS data. The family structure in the second data set allows the comparison between detection methods.

3 Results

3.1 Comparison between WES and array joint platform calls versus single platform calls

We first apply iCNV with default parameters to 38 individuals with both WES and SNP array data. We analyzed the data in three ways: Joint segmentation using both WES and array, segmentation using WES alone, and segmentation using array alone. To illustrate the relationship between the integrated HMM Z-score and the raw data input values, Figure 3 shows all of the input values along with the HMM Z-score for chromosome 22 of one typical sample. The heatmap shows detected CNV regions and integrated HMM Z-scores for that specific individual. Regions of the genome with low intensity/coverage and an enrichment of 0/1 BAFs have negative Z-scores, indicating putative deletion events; regions with high intensity/coverage and an enrichment of BAFs at duplication levels (see supplement) tend to have a positive Z-score, indicating putative duplication.

Table 1. Summary of called CNV* (WES, Array, 38 samples, Chr22)

CNV type	WES+Array**	WES	Array	WES ∩ SNP	WES ∪ SNP
Deletion	1364	152	1134	1252	1278
Duplication	381	124	144	252	268

* This table measure the number of hidden state, i.e. SNP or exon/bin
** Integrated CNV calling with data from both platforms

Table 1 compares the integrated analysis with a simple intersection or union of results from a separate analysis of each individual platform. Details of how the union and intersections are performed are given in the supplement. In this data set, we find that an integrated analysis yields more deletion and duplications than ad hoc union of individual results from the two platforms. Compared to a simple intersection of calls from the two platforms, iCNV rescues signals that are modest in one platform while strong in another platform; an example is given in Supplementary Fig 5. Even though a simple intersection gives the most stringent call set, 87% of integrated iCNV calls have overlap with the intersection call set, implying strong confidence of the iCNV result (Supplementary Table 1). On the other hand, by conventional wisdom it seems that taking a simple union of single platform calls should increase sensitivity but also increase false positive rate. However, in this case we detect more CNVs by integrated method than by simple union. Of the integrated call set, 8.94% are not present in the simple union, whereas of the union call set, 12.04% are not present in the integrated results. A signal that is moderate in both

platforms would be present in the integrated call set but not in the union call set. A signal that is only present in one platform but absent in the other would be present in the union call set but not detected during integration. It is evident that by combining the two platforms, probe density is increased, thus improving CNV detection power.

To assess the power improvement achieved by combining platforms, we conducted an in silico spike-in study. We define a detection to be a true positive if there is overlap between the detected region and a known spike-in CNV location. Power is defined as the percentage of spiked-in CNVs that are detected. The result shows that combining SNP array with WES indeed increases power for all but the largest CNVs (Fig. 5a).

3.2 Comparison between WGS and array platform calls versus individual platform calls

We next apply iCNV to 75 individuals with both WGS and array data, performing a similar analysis as above. Figure 4 shows, for a region of Chromosome 22, a heatmap of the integrated HMM Z-scores output by iCNV run on joint platform mode. Regions reported by iCNV as deletions and duplications are also shown. In this data set, the joint method detects more CNVs than a simple intersect of a separate analysis of the two platforms, and less than their simple union (Table 2). Compared to WGS, SNP array has much lower resolution and thus detects significantly fewer CNVs. More than 71% of the SNP calls and 96% of the WGS calls overlap with joint platform calls (Supplementary Table 2). On the other hand, around 92% of the joint platform calls overlap with calls made on WGS alone. Since WGS already has very good coverage across the genome, there seems to be, as expected, a much smaller power gain achievable by adding SNP arrays to the joint calling procedure.

We conduct another in silico spike-in study to study the power gains from adding array data to WGS. The result shows that WGS has comparable power to the integrated method, performing only slightly worse (Fig. 5b). However, for the detection of small CNVs in SNP-dense regions, SNP-array does add valuable information.

Table 2. Summary of called CNV* (WGS, Array, 75 samples, Chr22)

CNV type	WGS+Array**	WGS	Array	WGS ∩ SNP	WGS ∪ SNP
Deletion	9572	9840	825	9235	10368
Duplication	5902	6069	48	5490	6090

* This table measure the number of hidden state, i.e. SNP or exon/bin
** Integrated CNV calling with data from both platforms

3.3 CNV validation by pedigree relationships

To further evaluate performance, we utilize pedigree data that is available for the 75 individuals with WGS and array data. CNV detections that are true positives are more likely to be shared between related individuals than between unrelated individuals. Based on this fact, we can compute the CNV sharing frequency f at hidden state l between related individuals, defined as

$$f_l = \frac{\sum_{m \in S} |c_{i \in m,l} = c_{i' \in m,l} \neq diploid|}{\#\ of\ individuals} \quad (3)$$

where S represents the set of families, against the cohort call frequency, defined simply as the fraction of individuals where this CNV is detected among the 75 individuals analyzed. As a baseline, we compute the expected sharing frequency under random permutation of family labels, which represent the null scenario of a random detection. Enrichment of detections above the permutation-derived mean is evidence for enrichment of true positives.

A comparison of calls made by integrating WGS and SNP array versus calls made by a separate analysis of each platform alone is shown in the supplement. Based on this metric, there is no detectable gain of adding SNP array to WGS, which is expected given our analysis in Section 3.2 which showed very little power gain from adding SNP array to WGS and a large overlap between the integrated and WGS results on this data set.

This pedigree-based analysis also allows us to compare iCNV, run on WGS-only mode, to other WGS-based CNV detection methods including CNVnator and cn.MOPS. CNVnator and cn.MOPS were run with default parameters and the same bin size (1kb) as iCNV. Results are shown in Figure 6 and Table 3. To quantify the enrichment of familial sharing, we compute a sharing enrichment score by subtracting from the observed familial sharing frequency its permutation mean and then dividing the difference by its permutation standard deviation; see supplement for details. Larger sharing enrichment score indicates stronger within family enrichment while zero indicates no enrichment. As shown in Fig 6, cn.MOPS and iCNV show significant enrichment of familial sharing as compared to CNVnator. The enrichment scores of the iCNV call set are significantly higher than CNVnator (one-side t.test, pvalue= 1.28×10^{-7}) and cn.MOPS (one-side t.test, pvalue= 1.54×10^{-2}).

Table 3. Total length of CNV (WGS, 75 samples, Chr22)

CNV type	iCNV	Cn.MOPS	CNVnator
Deletion*	842	707	935
Duplication*	213	1485	1686

* Size showed in kilobase

3.4 CNV genotype inference

An integer copy numbers is assigned to each CNV region after the detection step. The assignment is based on maximizing a likelihood function that quantifies the probability of the observed normalized intensities and BAFs for each copy number state. The data input (normalized intensities and BAFs from all platforms) are shown in Figure 7, along with their copy number assignments shown as contours. The marginal densities of the normalized intensity values for each platform seem to be well modeled by a mixture of normals with platform-specific mean and variance. The BAF is much noisier and do not show any platform specific trend. Thus, the likelihood model we use is based on a mixture of normal for the normalized intensities, with platform and copy number specific means determined by an initial K-means clustering step, and a mixture of truncated normal for the BAF with pre-fixed means and standard deviations. The maximum likelihood copy number state is assigned to each segment.

4 Discussion

We have proposed a method, iCNV, to improve CNV detection and genotyping accuracy using high throughput sequencing data, allowing for integration of SNP-array data. The distinguishing features of iCNV

compared with existing methods are as follows: *(i)*iCNV adopts CODEX to improve the normalization of sequencing data, removing biases due to target length, mappability, GC content and other latent systematic factors; *(ii)* iCNV utilizes B-allele frequency information from sequencing data, which is valuable for CNV detection and exact copy number inference; *(iii)* Array data, if available, are combine with sequencing data to allow more sensitive and robust CNV detection than either platform alone; (*v*) iCNV outputs a Z-score from an integrated HMM that summarizes evidence across multiple platforms, allowing for easy visualization and quality assessment; *(vi)* Even though we combine cross-platform data for CNV detection, we use platform specific parameter for exact copy number estimation and thus minimize noise effect due to platform specific latent variables.

How much does SNP-array data add to NGS data for CNV detection? Our results, based on spike-ins and pedigree-based quality evaluations, show that SNP-arrays give a significant boost in accuracy to WES but relatively little gains for WGS. For CNV detection and genotyping using WGS alone, we compared iCNV against other read depth based CNV detection methods including cn.MOPS and CNVnator. CNVs detected by iCNV have higher within-family sharing than the other methods being compared, suggesting a higher accuracy. Collectively, iCNV provides a systematic framework for single and cross platform CNV detection.

Acknowledgements

Funding

ZZ has been supported by training grant ###, …, NRZ has been supported by 2 R01 HG006137.

Conflict of Interest: none declared.

References

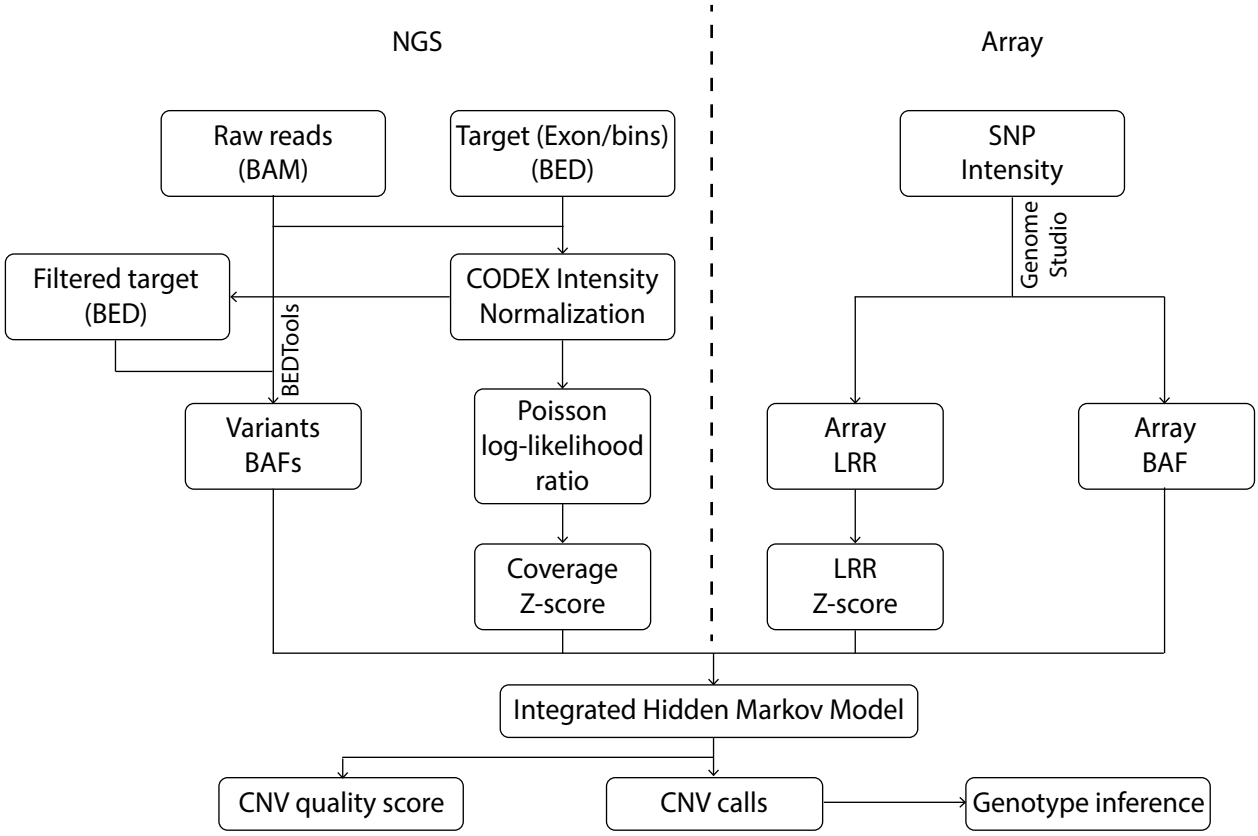

Figure 1. A flowchart ouling the procedures of iCNV in data normalization, CNV calling and genotyping from NGS and array data. For NGS data, the first step is to normalize intensity data and to calculate a Poisson log-likelihood ratio (PLR) using CODEX. The variants BAF are then calculated with filtered target position using BEDTools. We further convert PLR to a coverage Z-score. For array data, GenomeStudio calculate a log R ratio - further normalized to a Z-score - and BAF from raw SNP intensity data. The integrated Hidden Markov Model takes the inputs from NGS and array generating CNV calls and quality score. We further infer their genotypes.

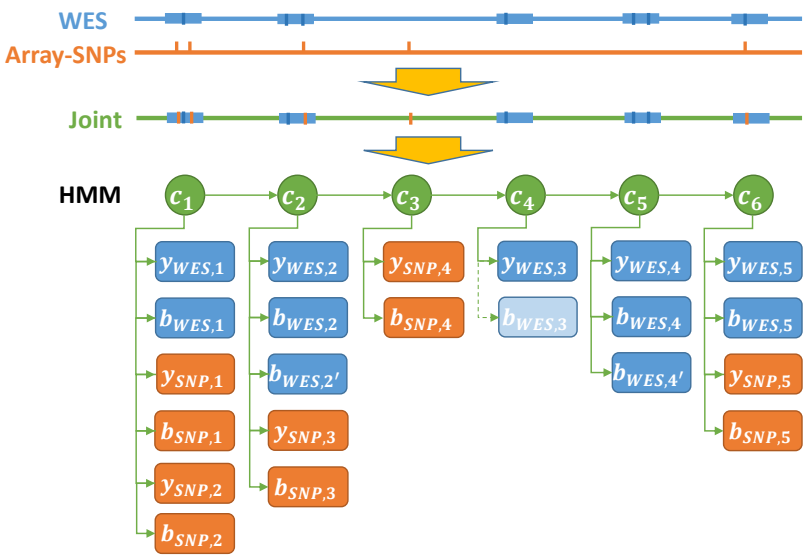

Figure 2. An illustration of the integrated Hidden Markov Model with array and WES data. c_i indicates hidden states at position i; $y_{SNP/WES,i}$ indicates observation of Z-score at position i; $b_{SNP/WES,i}$ indicates observation of BAF at position i. Specifically in hidden state c_4, since there is no variants BAF from WES, we assign 0.

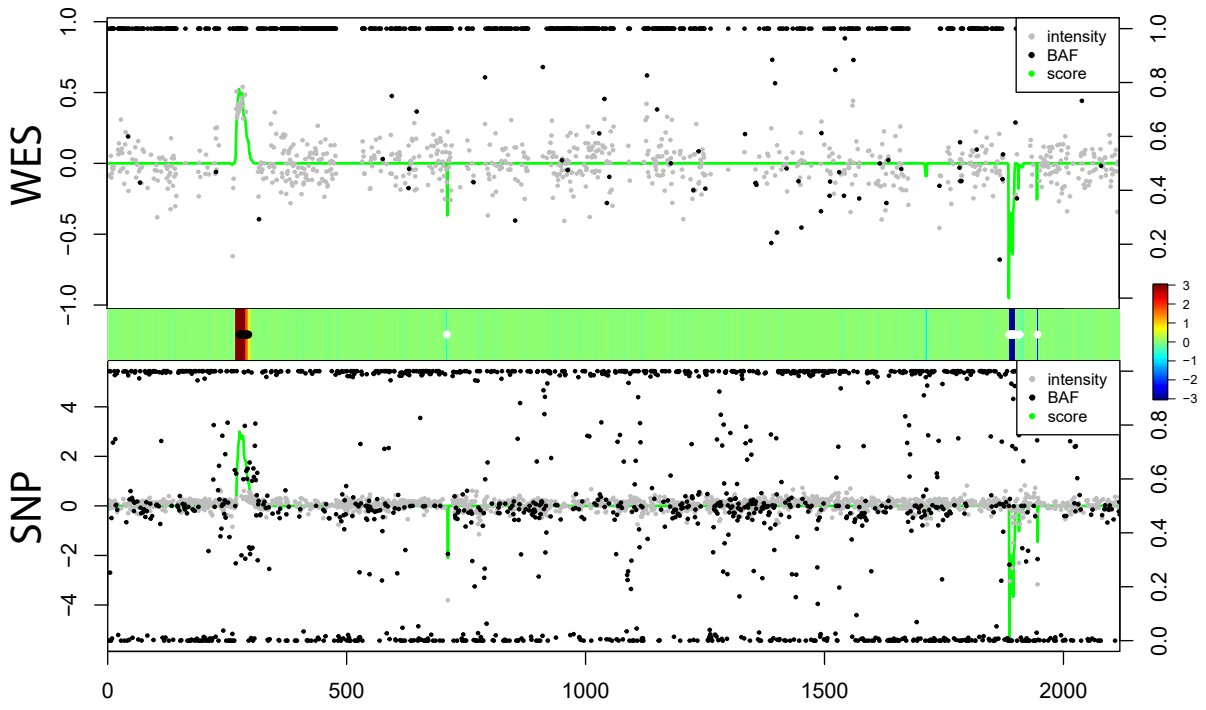

Figure 3. Case wise illustration of relationship between score (green line), Z-score (grey dots) and BAF (black dots) in WES and SNP array. Heat map in the middle indicates score and CNV calling (white dots: deletion; black dots: duplication).

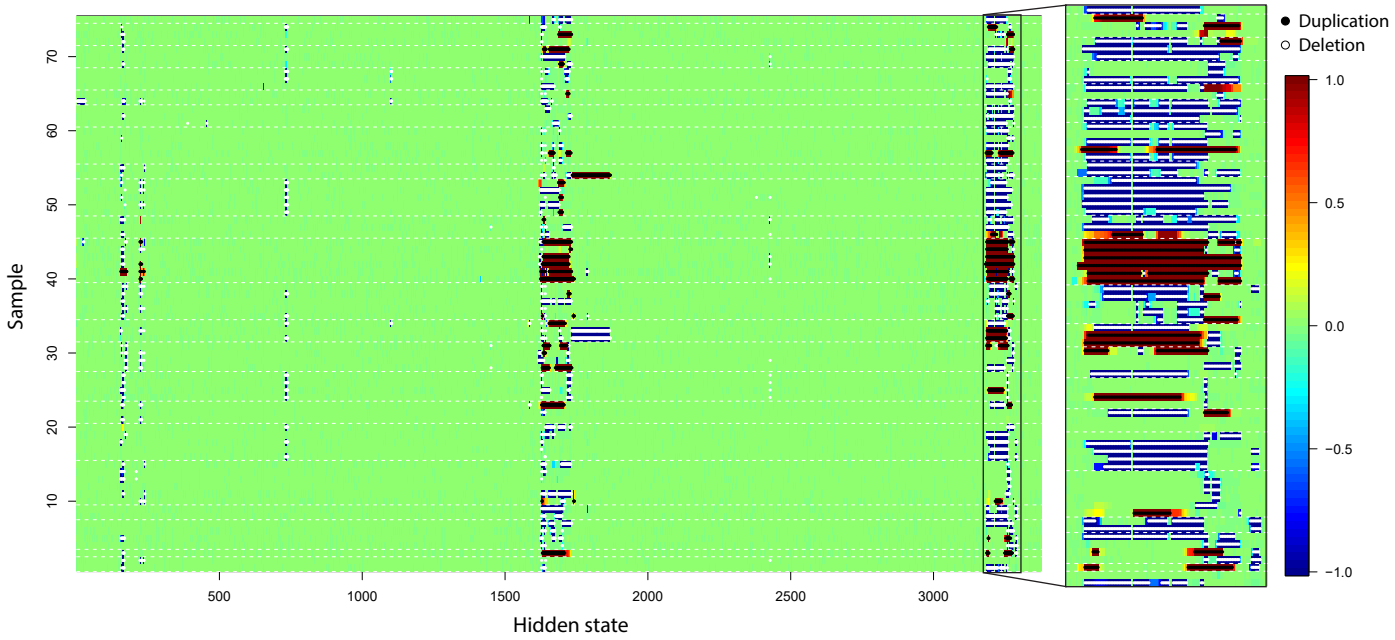

Figure 4. CNV score at each hidden state from 75 related individuals with SNP array and WGS data on Chr22 CNV hot region (17000-20800kb). Region (20320-20720kb) is enlarged for detailed visualization.

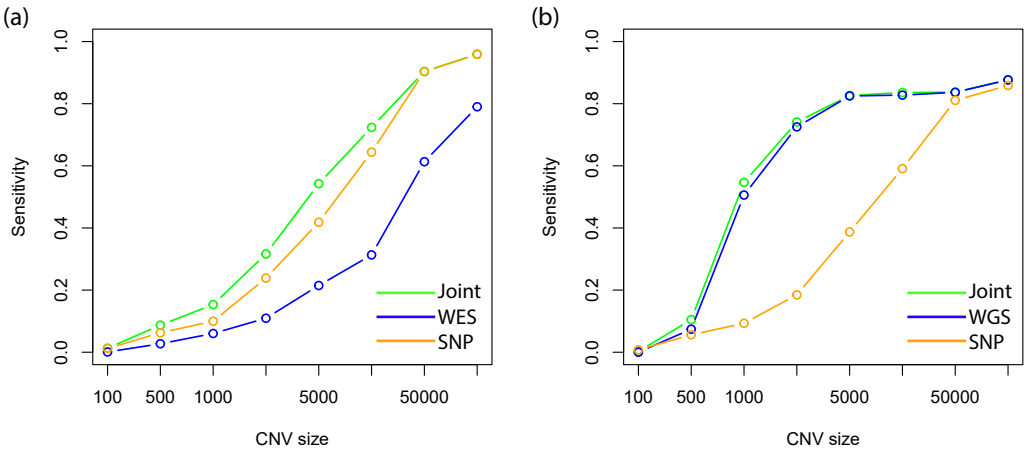

Figure 5. Relationship between power (or sensitivity) and size of CNV using spike-in in (a) WES and SNP array scheme and (b) WGS and SNP array scheme. Joint means integrated analysis using iCNV with both WES and SNP array data.

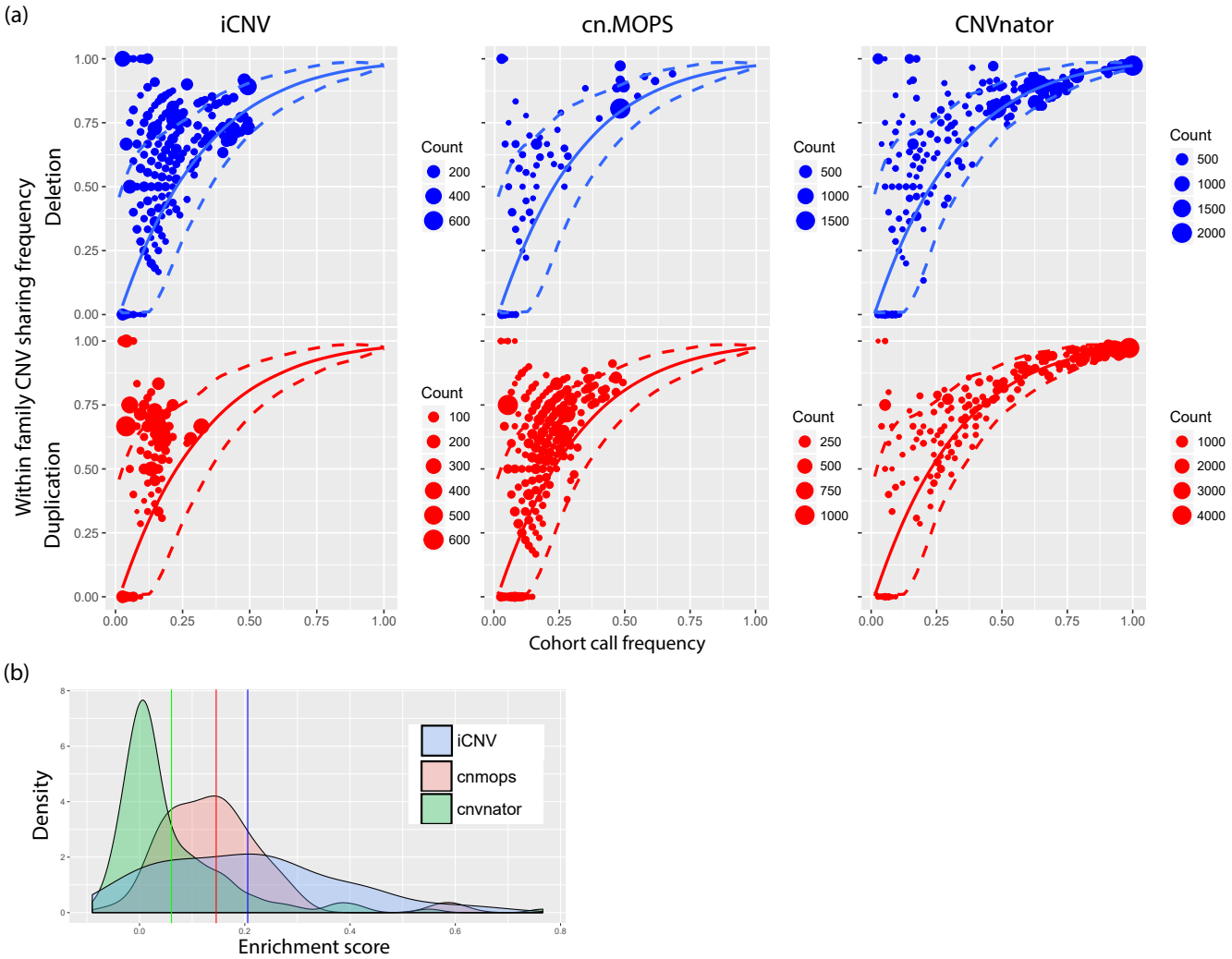

Figure 6. (a) CNV detection assessment of iCNV, cn.MOPS, CNVnator by real dataset with pedigree information in WGS only scheme. Plot shows the enrichment of CNVs (deletion or duplication) with in a pedigree. The solid line and the dash lines are the null distribution mean and 95% confidence interval, respectively, calculated by permutation. Size of the dots shows number of CNV events counts. (b) Enrichment score distribution plot between iCNV, cn.MOPS and CNVnator. Larger number indicates higher enrichment. Solid line is the distribution mean for each method.

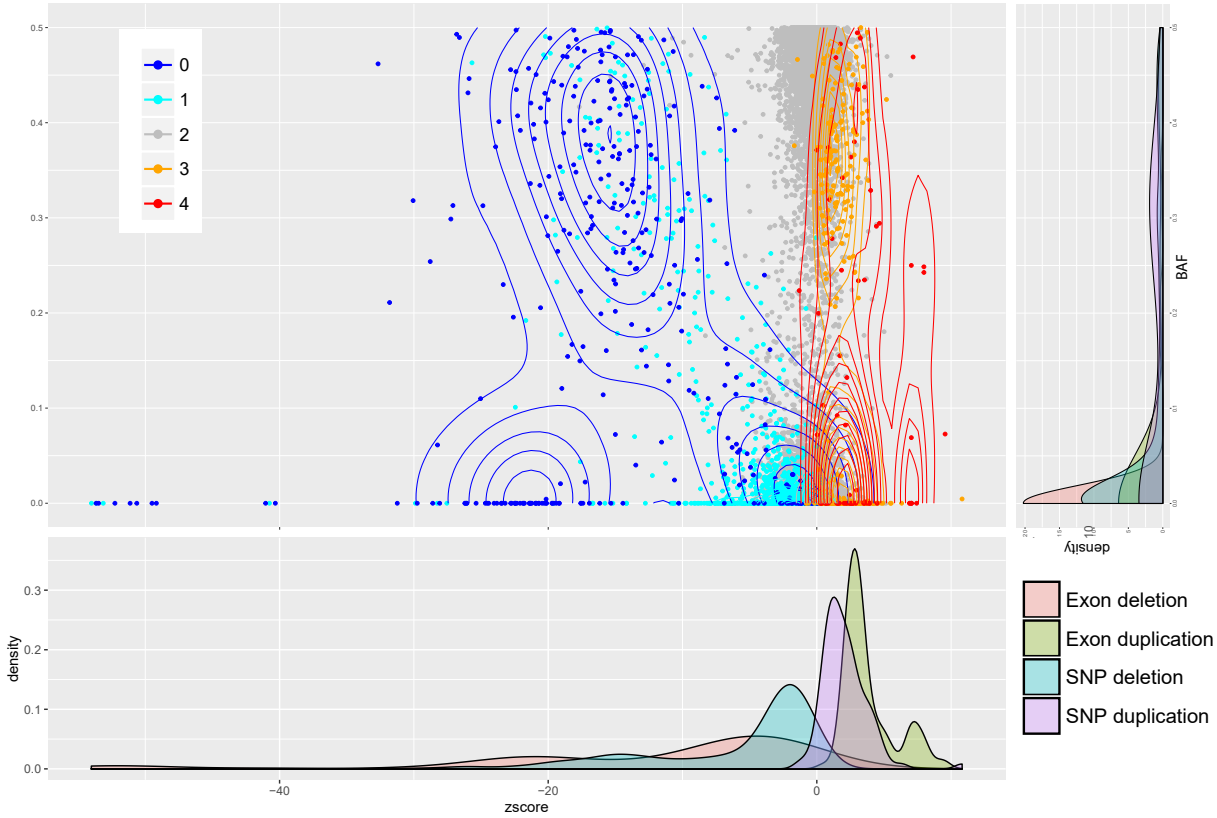

Figure 7. Z-score and BAFs distribution of CNV genotype inference. Contour plot shows distribution density. Each margin respectively shows density distribution of Z-score and BAF in each CNV genotype category.