Hybridization and amplification rate correction for Affymetrix SNP arrays

Quan Wang, Peichao Peng, Minping Qian, Lin Wan and Minghua Deng

Abstract

Background: Copy number variation (CNV) is essential to understand the pathology of many complex diseases at the DNA level. Affymetrix SNP arrays, which are widely used for CNV studies, significantly depend on accurate copy number (CN) estimation. Nevertheless, CN estimation may be biased by several factors, including cross-hybridization and training sample batch, as well as genomic waves of intensities induced by sequence-dependent hybridization rate and amplification efficiency. Since many available algorithms only address one or two of the three factors, a high false discovery rate (FDR) often results when identifying CNV. Therefore, we have developed a new CNV detection pipeline which is based on hybridization and amplification rate correction (CNVhac).

Methods: CNVhac first estimates the allelic concentrations (ACs) of target sequences by using the sample independent parameters trained through physicochemical hybridization law. Then the raw CN is estimated by taking the ratio of AC to the corresponding average AC from a reference sample set for one specific site. Finally, a hidden Markov model (HMM) segmentation process is implemented to detect CNV regions.

Results: Based on public HapMap data, the results show that CNVhac effectively smoothes the genomic waves and facilitates more accurate raw CN estimates compared to other methods. Moreover, CNVhac alleviates, to a certain extent, the sample dependence of inference and makes CNV calling with appreciable low FDRs.

Conclusion: CNVhac is an effective approach to address the common difficulties in SNP array analysis, and the working principles of CNVhac can be easily extended to other platforms.

Keywords: SNP array, Copy number variation (CNV), Cross-hybridization, Genomic waves

Background

Copy number variations (CNVs) play an essential role in facilitating human diseases susceptibility [1,2] and have been shown to be one potential source of missing heritability of complex diseases [3]. Together with genome-wide association studies (GWAS), CNVs are predicted to be compelling in deciphering the pathology of human diseases [4]. SNP arrays have been widely used for CNV studies, and tremendous data have been generated [5-7]. Although high throughput sequencing technologies are emerging and have been applied to genetic variation (including CNV) studies, the cost of a sequencing-based approach is still higher than traditional SNP arrays, especially in library construction [8]. In addition, various studies have shown that the sequencing data are not sensitive to breakpoint detection [9-11]. Moreover, sequencing technologies have poor mutation detection capability when the sequencing coverage (read depth) is relatively low [12]. Thus, at their current stage of development, we believe that sequencing technologies are complementary, not substitute, tools of SNP arrays. Therefore, in this article, we aim to develop a new and more accurate CNV detection pipeline that avoids the common difficulties in SNP array analysis.

High quality CNV calls for accurate estimation of raw copy numbers and requires that statistical models be optimized [6]. Although many methods have been developed for CNV calling from array-based data [7,13-16], their accuracies are still far from satisfactory by the high incidence of false discovery rates (FDRs)
[5,17-19]. The high FDRs of these methods mainly arise from (1) cross-hybridization of probes [20], (2) genomic waves of intensities [21-23] and (3) sample dependence of outputs [24-26].

Cross-hybridization between probes and off-target sequences is a longstanding problem in microarray analysis [27-30]. Therefore, most previous methods have typically ignored cross-hybridization and focused on taking mean or median intensities of probes as the estimated raw CNs [15,31]. However, such estimated CNs hardly reflect the true allelic concentrations (ACs) of target sequences, and some studies [6,7,20] have shown that cross-hybridization, if not considered, can lead to large bias. To circumvent this problem, one prior investigation used PICR (probe intensity composite representation) to model the hybridization and cross-hybridization based on the underlying physicochemical principle of DNA/DNA duplex formation in array experiments, and then removed the effect of cross-hybridization and accurately estimated AC at a given SNP site through a statistical method [20]. Other similar models were also reported [28,32].

In addition to cross-hybridization, Maris et al. have stated that “whole-genome microarrays with large-insert clones designed to determine DNA copy number often show variation in hybridization intensity that is related to the genomic position of the clones.” [22] These ‘genomic waves’ have been observed in SNP arrays [21-23]. Genomic waves are shown to be correlated with GC-content [21,23] and may stem from the amplification of DNA fragments [33]. In the preprocessing of arrays, DNA samples are first digested with restriction enzymes, such as Nsp, and then ligated with adapters before amplification. However, owing to differences in amplification efficiencies of fragments, the PCR procedure can bring in artifacts which may give rise to genomic waves [33]. Presence of the waves will hamper detection of aberrations [23] and introduce hundreds of potentially confounding CNV artifacts that can obscure bona fide variants [33]. To solve this difficulty, a computational approach via fitting regression models with GC-content included as a predictor variable was proposed by [22], and this approach have improved the accuracy of CNV detection.

Finally, it has long been known that different sample batches can lead to inconsistent results, even if data are collected by the same lab [24-26]. Owing to this effect, statistical power in meta-analysis of multiple samples may be significantly reduced [34]. Almost all existing algorithms require multiple samples for training because of the numerous parameters, while different training sample batches can lead to different parameter estimation. The inconsistencies may be incurred by this sample-dependent parameter estimation. The effect has also been shown to be correlated with differences in batch sizes and the extent of homogeneity of samples in each batch. Hence, samples with high homogeneity are suggested to be placed into the same training batch [26]. Several other methods to adjust this batch effect have also been proposed, such as [25,35,36].

To the best of our knowledge, existing methods only address one or two of the three factors discussed above. In this study, we developed a novel CNV detection pipeline based on hybridization and amplification rate correction (CNVhac) to accurately detect CNVs for Affymetrix SNP array. In contrast to previous methods, CNVhac takes into account all three factors by proper modeling of cross-hybridization, smoothing genomic waves and alleviating sample batch dependence of parameter estimation, thus significantly improving the accuracy of CNV detection. Starting from dozens of basic constants concerning binding affinity, which can be well trained from one single array and are quite stable between arrays, CNVhac is able to get the binding affinity between all probes and sequences without suffering from sample batch dependence. Then CNVhac applies the PICR method [20] to address the effect of cross-hybridization. Finally, since we have found that the relative amplification efficiencies between different fragments are fairly stable from one array to another, a simple adjustment approach is proposed to smooth the genomic waves. Based on the accurate raw CN estimates, a hidden Markov model (HMM) is also proposed to detect breakpoints along the genome. The implementation of CNVhac with public datasets shows that our method does enhance the power of both raw CN estimation and CNV calling.

Methods

Dataset

Dataset I. ‘The International HapMap project’ [37] mapped 270 samples (30 YRI trios, 30 CEU trios, 45 CHB and 45 JPT individuals) to Affymetrix SNP 6.0 array to identify and catalog genetic similarities and variants in human beings. The raw SNP 6.0 dataset (http://www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx) is applied in this paper.

Dataset II. Conrad et al. recently used the ultra-high-resolution NimbleGen tiling arrays (42 M probes) to identify CNVs for HapMap samples [38]. The identified CNVs were then filtered by two other technologies (Agilent and Illumina). Finally, over 5000 regions that were cross-platform verified as CNV in at least one of the HapMap individuals of dataset I were selected [38] and referenced as benchmark in this article to assess the power of CNV calling in comparison with other algorithms. We have not performed any experimental research by ourselves, and both dataset I and II are
downloaded from public databases. Therefore, there is no ethical approval problem in this study.

**Estimation of raw CNs**
The problems usually confronted in the estimation of raw CNs are discussed in the background section. Array intensities not only rely on ACs of target sequences, but also probe binding affinities. Based on [20], we model hybridization and cross-hybridization with dozens of probe-independent parameters, which can be accurately estimated from single array and are consistent between arrays [39]. Another simple adjustment is proposed to calibrate the various amplification efficiencies.

**Modeling hybridization and cross-hybridization**
Considering one probe in a certain SNP probeset, we have the basic model [39,40]:

\[ I = I_s + I_{bg} + \epsilon, \]

where \( I_s \) and \( I_{bg} \) stand respectively for probe intensity, specific hybridization intensity caused by target sequences and background nonspecific binding intensity, and \( \epsilon \) is the measurement error. \( I_s \) has been further modeled by Langmuir-like adsorption principle, and Equation (1) can be rewritten as:

\[ I = I_s + I_{bg} + \epsilon = \frac{N}{1 + \exp(E)} + I_{bg} + \epsilon, \]

where \( N \) is AC of the target sequences, and \( E \) denotes specific binding free energy which can be modeled by position-dependent nearest-neighbor (PDNN) [39,40]:

\[ E = \sum_{i=1}^{24} \omega_i \lambda(b_i, b_{i+1}), \]

where \( \omega_i \) is a weight factor which is dependent on the position of consecutive bases along the oligonucleotides, \( b_i \) is the \( i \)-th nucleotide of probe sequence, and \( \lambda \) is the stacking energy of the pair of nearest-neighbors along the probe. With \( \lambda(b_i, b_{i+1}) \) and \( \omega \) known as basic constants which hardly change between arrays [39], \( N \) can be easily estimated by regression.

However, the model ignores cross-hybridization. There are two alleles (allele A and allele B) in the genome for a certain single polymorphic locus. For high sequence similarity, each allele has a high possibility of binding to the probe which is designed to interrogate the other allele. This cross-hybridization may bring bias when estimating the AC of target sequences (See [20] and Additional file 1). Therefore, we go one step further to improve the model by assuming that \( I_s \) follows an additive model of \( I_{A} \) and \( I_{B} \). Their meanings are clear: the contribution of allele A and B target sequences, respectively, to probe intensity. Both \( I_A \) and \( I_B \) can be modeled by Equation (2); thus our proposed model is

\[ I = \frac{N_A}{1 + \exp(E_A)} + \frac{N_B}{1 + \exp(E_B)} + I_{bg} + \epsilon, \]

where \( N_A \) and \( N_B \) are ACs for allele A and B, respectively, and \( E_A \) and \( E_B \) denote binding free energy. With quite a few probes in one probeset, the ordinary least squares (OLS) method yields unbiased estimates of \( N_A \) and \( N_B \). The summation of \( N_A \) and \( N_B \) gives the total concentration \( N \) (See [20] and Additional file 1). For the nonpolymorphic probe with only one allele, \( N \) can be straightforwardly obtained from Equation (2).

**Normalization between arrays**
In order to eliminate the systematic bias between arrays which may arise from the different library preparation conditions of the experimental process, we use the following transformation:

\[ N'_{mk} = N_{mk} \alpha_m, \]

where \( N_{mk} \) is the total concentration for array \( m \) at locus \( k \), and \( \alpha_m = 2/\text{median}(N_{mk}, k = 1, 2, \ldots, K) \) is the normalization factor for array \( m \) \((K = \) the total number of loci from one array).

** Calibration for amplification efficiency**
We have found that \( N'_{mk} \) are fairly stable from one array to another, except for CNV regions for one certain locus \( k \) (see Additional file 1); therefore, a simple adjustment approach is proposed to calibrate the various amplification efficiencies:

\[ \hat{N}_{mk} = N'_{mk} \gamma_k, \]

where \( \gamma_k = 2/\text{median}(N'_{mk}, m = 1, 2, \ldots, M) \) is the adjustment factor for each locus \( k \) \((M = \) the total number of reference samples). In order to estimate the adjustment factor \( \gamma_k \), a pool of reference samples is needed. In the case--control assay pattern, the control arrays are treated as the reference pool. In this article, the HapMap samples from dataset I are used to estimate \( \gamma_k \). CNVhac takes \( \hat{N}_{mk} \) as the estimated raw CN for locus \( k \) in array \( m \).

**CNV calling**
CNVhac implements a HMM-based algorithm to call CNVs. HMM methods have previously been successfully applied to other studies [13,41,42], and the main idea of our algorithm is similar to them. In our implementation of the HMM, the hidden state is the true CN \((0, 1, 2, 3 \text{ or } \geq 4)\) of each locus along the genome, and the observed state is our estimated raw CN \( \hat{N}_{mk} \). For each locus, the emission probabilities are estimated from a normal
distribution with true CN as mean. The transition probability of jumping out from normal state is presumed to be low, whereas jumping back to a normal CN or transitioning within the same state is relatively high. Furthermore, the distance between neighboring loci is correlated with transition probability [13]. Given the initial emission and transition probabilities, the Viterbi algorithm [43] is used to decode the hidden states. Then, the parameters can be updated iteratively until converging. A more detailed description of this method can be found in Additional file 1.

Results

The pipeline of CNVhac mainly consists of two major steps. The preprocessing step first estimates the raw CNs $N_{mk}$, and, second, the CNV calling step then searches for breakpoints through a HMM model. In this section, we compare CNVhac with two widely used raw CN estimation methods, CRMA_v2 (‘Copy-number estimation using Robust Multichip Analysis’ [6]) and cn.FARMS (‘factor analysis for robust microarray summarization’ [7]), to evaluate the accuracy of estimated raw CN $N_{mk}$. CRMA_v2 is an extension of CRMA [44] for estimating raw CNs for downstream analyses. cn.FARMS presents a probabilistic latent variable model for summarizing probes to obtain raw CN estimates. Both CRMA_v2 and cn.FARMS outperform other studies on raw CN estimation [6,7]. Meanwhile, to assess the performance of CNV calling, we compare CNVhac with another popular approach known as Birdsuite [13], which is asserted to be the best for CNV inference with Affymetrix SNP arrays [5]. Because Birdsuite does not estimate raw CNs, it is not considered in the comparison on raw CN estimation.

Raw CN estimation on HapMap CEU samples

We assess the performance of raw CN estimation from two aspects: the accuracy in classifying the sex of HapMap individuals and the amplitude of genomic waviness. Females have two copies of X chromosome, while males only one; therefore, the CN of X chromosome can naturally be used as the benchmark to evaluate the power of the raw CN estimates to differentiate between one or two copies. We collected the same 59 CEU parents in Dataset I to do this classification task as [7]. Children were excluded to avoid inherited biases. The sample of female founder NA12145 was also removed on the basis of its low true CN level [44]. All the loci in the pseudoautosomal regions (PAR1 and PAR2), segmental duplications (http://humanparalogy.gs.washington.edu/build36) and CNV regions [38] in chromosome X were excluded owing to CN contamination. Finally, 83121 polymorphic and nonpolymorphic loci were kept which gives 4904139 (=83121 × 59) single locus classification tasks. The receiver operating characteristic (ROC) curve is introduced to assess the performance of different methods. The horizontal axis of the ROC curve represents the false positive rate (the fraction of males classified as females), while the vertical axis stands for the true positive rate (the fraction of females classified as females). Figure 1 shows the ROC for CNVhac, CRMA_v2 and cn.FARMS, respectively. The areas under ROC curve (AUCs) of CNVhac, CRMA_v2 and cn.FARMS are 0.9684, 0.9603 and 0.9627, respectively. We see that CNVhac outperforms CRMA_v2 and cn.FARMS when distinguishing males from females based on the estimated raw CNs.

The better result of sex classification by CNVhac may be attributed to better control of genomic waviness. To assess the waviness, we investigated the estimated raw CNs of chromosome X used above. The three sets of raw CNs were separately scaled to the same median. For females, the median is set as 2 and for males 1. Figure 2 shows an example of dissimilar genomic wave patterns for one female CEU founder, NA06985. The fluctuation
of raw CNs is obvious in cn.FARMS, with somewhat less fluctuation in CRMA_v2. However, the waves are smoothed most effectively by CNVhac compared to the other methods. Figure 3 shows the density of raw CNs for female CEU founders and male founders, respectively. More precisely, we computed the variance of raw CNs. For females, the variances of cn.FARMS, CRMA_v2 and CNVhac are 0.2118, 0.1225 and 0.1112. For males, the variances are 0.2597, 0.0336 and 0.0289. For both females and males, CNVhac has the smallest variance (F test, all p-values are < 2e-16). This result implies that CNVhac can smooth the fluctuation through one simple, but effective, method.

**CNV calling on HapMap samples**

The cross-platform verified regions in dataset II are defined as true CNVs to assess the power of CNV detection for CNVhac and Birdsuite on the 269 samples from

---

**Figure 2** Genomic wave patterns on a segment of Chromosome X of one CEU female founder, NA06985, for (a) cn.FARMS, (b) CRMA_v2 and (c) CNVhac. CNVhac has the smallest amplitude of estimated raw CNs.

---

**Figure 3** Density of raw CNs estimated by different methods for (a) male CEU founders and (b) female CEU founders on chromosome X. Raw CNs are scaled to the same median (for males 1 and females 2). CNVhac shows significantly smaller variance than CRMA_v2 and cn.FARMS (F test, all p-values are < 2e-16).
dataset I (NA19012 is missing in the result of [38]). We filtered out those verified regions having fewer than 5 probes designed in Affymetrix SNP 6.0 array, resulting in 1381 verified regions for our evaluation. Each sample has a different number of CNVs annotated in the 1381 selected regions [38]. In total, we have 49662 true CNVs annotated in the 1381 regions across the 269 samples. We assessed the performance of each algorithm by calculating the ratio of the predicted CNVs, which are supported by true CNVs to all the predicted CNVs along the genome (precision), and the fraction of true CNVs, which are predicted by this algorithm (recall). The concordance principle for predicted and true CNVs is that when identifying rare CNVs [7]. Therefore, in the assessment of CNV calling power here, we removed those CNVs with a lower FDR.

Sample batch dependence of CNV calling
As described in the Background section, different parameters trained from different sample batches may cause an inconsistent inference. To evaluate the sample batch dependence of CNV calling of CNVhac, we compare it with Bird-suite. In CNVhac, estimating adjustment factor $\gamma_k$ is the only step requiring a batch of samples. In Section 3.2, all 270 HapMap samples were used to estimate $\gamma_k$. Here, we divided the 270 samples into 3 groups and then treated them as different pools of reference samples. Each group consisted of 90 samples. (The different choice of samples in each group can be found in Additional file 2). Adjustment factor $\gamma_k$ can be estimated within each group, respectively. With the different $\gamma_k$ raw CN estimates $N_{mk}$ change, as well as the CNV calling. For a specific sample $S_i$ three sets of CNV regions can be detected through different $\gamma_k$. We assess the sample batch dependence by computing the ratio of intersection to union. For Bird-suite, 3 groups were created by the same way. Next, sample $S_i$ was put to the other two groups which do not contain it. Hence, one can also obtain three sets of identified CNVs. We chose 6 individuals (2 CEU, 2 YRI, 1 JPT and 1 CHB) to call CNVs based on different groups. Table 1 displays the ratio of intersection to union, respectively, under default parameter setting. From this, we see that CNVhac shows significantly higher ratios than Bird-suite ($p$-value = 6.5e-3 by Wilcoxon rank-sum test). This indicates that CNVhac alleviates the sample batch dependence of CNV calling to a certain extent.

![Figure 4](http://www.biomedcentral.com/1755-8794/5/24)

**Figure 4** 1-precision versus recall curves for CNV detection on 269 HapMap samples. A curve that is located more toward the upper-left corner indicates better performance. Note: FDR is 1-precision. Compared to Bird-suite, CNVhac shows an appreciably lower FDR when calling CNVs.

Table 1 Results of CNV calling based on different training sample batches for CNVhac and Bird-suite

| Sample Batch | 1-precision (Ratio) | 1-Precision (U) | Recall (Ratio) | 1-Precision (U) |
|--------------|---------------------|----------------|---------------|----------------|
| Bird-suite   | G1                  | G2             | G3             | U              |
| G1           | 17                  | 19             | 21             | 14             | 0.64           | 15             | 17             | 18             | 15             | 17             | 0.88           |
| G2           | 22                  | 21             | 19             | 15             | 28             | 0.54           | 29             | 26             | 24             | 20             | 23             | 0.61           |
| G3           | 15                  | 20             | 10             | 23             | 0.43           | 16             | 20             | 20             | 15             | 21             | 0.71           |
| U            | 20                  | 21             | 21             | 14             | 15             | 0.56           | 21             | 21             | 18             | 16             | 23             | 0.73           |

†The number of predicted CNVs using group 1 for parameter training.

†The number of CNVs in intersection set of "G1", "G2" and "G3".

†The number of CNVs in union set of "G1", "G2" and "G3".

†The ratio of intersection to union.
Discussion

For years, the array-based technologies have been widely used for exploring CNV events. However, the inherent noise of microarray data may lead to high FDR when making inferences. In array experiments, hybridization is highly correlated with the sequence constitutions [27,28,30,32,39,40,46]. The binding affinities of probes can be subject to large variability by the various sequences. Most previous algorithms attempt to model the binding affinity through statistical or empirical methods [41,44], which need multiple samples for training parameters. However, such multiple samples may lead to another problem: sample dependence of outputs [26]. The various choices of training samples may result in different estimated parameters, leading, in turn, to incompatible results. All the algorithms which need multiple training samples have a possibility encountering this effect. Consequently, strategies based on single-array processing are preferred. Up to now, however, few single-array approaches have been presented. CRMA_v2 is a single-array preprocessing method for SNP array analysis. However, the raw CNs estimated by CRMA_v2 exhibit a wavy pattern, and thus may not be accurate enough for downstream CNV identification.

Motivated by addressing the cross-hybridization of probes, genomic waves of intensities and sample dependence of parameter estimation, we propose in this article a single-array preprocessing method, termed CNVhac, to estimate more accurate raw CNs. Based on the previous PICR method [20], we model the hybridization and cross-hybridization of probes through physicochemical law. Wan et al. have shown that the PICR model can address the cross-hybridization effect very well [20]. The genomic wave patterns of signal intensities are hypothesized to reflect the various amplification efficiencies of DNA fragments in the PCR process [33]. However, based on the diversity of sheared fragments and complicated PCR procedures, it is difficult to estimate the accurate amplification rate for each locus. Instead, we smooth the genomic waves by estimating an adjustment factor for each locus since we have found that the estimated CNs show a fairly stable pattern between loci (see Additional file 1). Compared to CRMA_v2 and cn.FARMS, this simple calibration method effectively reduces the amplitude of waviness. Note that the reduction of waviness is not simply a compression of variance in that CNVhac provides more accurate raw CN estimates which can well differentiate between one or two copies. Moreover, the number of parameters needed to estimate target concentration \( N_{mk} \) in CNVhac is much fewer than prior statistical models and can be estimated from one single array quite stably [39]. This property avoids the sample dependence of parameter estimation. Compared to one popular CNV detection method known as Birdsuite [5,13], CNVhac, indeed, alleviates the sample dependence of CNV calling more effectively. However, CNVhac needs a pool of reference samples to estimate \( \gamma_k \) for calibrating amplification efficiency. In the case–control assay pattern, the control samples are treated as the reference pool. While the dataset contains only case samples, anonymous normal samples, e.g., HapMap samples, can be used as the reference pool. Because of the different experimental conditions, the anonymous normal samples may bring sample-dependent bias for \( \gamma_k \). Actually, CNVhac cannot address this kind of sample dependence.

CNVs have attracted much attention in recent years because they are assumed to play a significant role in causing human disease [1,4]. Especially, some recent studies and reviews have shown that rare CNVs contribute much more to neuropsychiatric disorders than previously thought [2,47-51]. However, the mechanism underlying the influence of CNVs on human phenotypes is still not well understood. Furthermore, even a small fraction of false discoveries may introduce misunderstanding in the downstream association studies. Therefore, CNV calling methods are strongly desired to control the FDR [7]. On the basis of raw CN estimates with cross-hybridization and amplification rate correction, CNVhac can identify rare CNVs with a lower FDR compared to the powerful Birdsuite method. This result implies that CNVhac can accurately identify CNVs, especially rare CNVs, for downstream association studies.

Since CNVhac is a single-array based strategy, the running time could be reduced by executing CNVhac on multiple processors in parallel when analyzing a large set of samples. Also, since parameters are consistent between arrays, there is no need to reprocess the early data when new samples are hybridized.

Conclusion

Cross-hybridization and different amplification efficiencies of probes are the common difficulties in microarray analysis. Most studies attempt to solve the problem by training numerous model parameters from a large dataset, but this might incur inconsistent results. Moreover, the statistical power of this methodology may be significantly reduced when the training dataset is not big enough. In this article, we first addressed cross-hybridization problem through physico-chemical law and then proposed a simple adjustment for the various amplification rates. Our method, CNVhac, avoids complicated statistical models which need many samples for training. By comparing CNVhac with other methods, we have established that our simple process is effective and suitable for all Affymetrix SNP array types with similar
design standards. Finally, the working principle of CNVhac can be easily extended to other platforms, such as Illumina and Agilent arrays.

Endnotes

CNVhac: The algorithm is implemented in R and C++ and is available at http://www.math.pku.edu.cn/teachers/dengmh/CNVhac.

Additional files

Additional file 1: Supplementary Materials. It contains details of modeling hybridization, cross-hybridization and HMM, as well as one figure explaining stable total concentrations between arrays [S2].

Additional file 2: Constitution of different reference groups.

Abbreviations

CN: Copy number; CNV: Copy number variation; FDR: False discovery rate; AC: Allelic concentration; HMM: Hidden Markov Model; GWAS: Genome-wide association studies; PICR: Probe intensity composite representation; PDNN: Position-dependent nearest-neighbor; OLS: Ordinary least squares; CRMA: Copy-number estimation using Robust Multichip Analysis; cn.FARMS: Factor analysis for robust microarray summarization; ROC: Receiver operating characteristic; AUC: Area under ROC curve.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Linbo Wang and Yongjian Kang for helpful discussions.

Author details

1Center for Theoretical Biology, Peking University, Beijing 100871, People’s Republic of China. 2LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, People’s Republic of China. 3Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA, USA. 4National Center for Mathematics and Interdisciplinary Sciences, and the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People’s Republic of China. 5Center for Statistical Science, Peking University, Beijing 100871, People’s Republic of China.

Authors’ contributions

MPQ and MHD conceived the project. MPQ, LW and MHD proposed the main idea. QW and PCP developed the program. QW implemented the methods, analyzed the data, and wrote the manuscript. MPQ, LW and MHD finalized the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [No.31171262, No.11021463] and the National Key Basic Research Project of China [No.2009CB918503].

Received: 21 February 2012 Accepted: 12 June 2012 Published: 12 June 2012

References

1. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, et al. Genomewide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010, 464:713–720.

2. Grozeva D, Kitov G, Ivanov D, Jones IR, Jones L, Green EK, St Clair DM, Young AH, Ferrier N, Farmer AE, et al. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. Arch Gen Psychiatry 2010, 67:318–327.

3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature 2009, 461:747–753.

4. McCarroll SA. Extending genome-wide association studies to copy-number variation. Hum Mol Genet 2008, 17:R135–R142.

5. Zhang Q, Qian Y, Akula N, Aliley-Rodriguez N, Tang J, Genthon ES, Liu C. Accuracy of CNV Detection from GWAS Data. PLoS One 2011, 6:e14511.

6. Bengtsson H, Wrappati P, Speed TP: A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics 2009, 25:2149–2156.

7. Clevera DA, Mirteereck A, Mayr A, Klaambuger B, Tuefferd M, De Bondt A, Talloen W, Gohrmann H, Hochreiter S. cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate. Nucleic Acids Res 2011, 39:e79.

8. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 2009, 6:513–520.

9. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdian F, Kitzman JO, Baker C, Malig M, Mutili O, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 2009, 41:1061–1067.

10. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bhrin L, Sendur J, Eichler EE. Diversity of human copy number variation and multicopy genes. Science 2010, 330:941–946.

11. Alkan C, Cox BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011, 12:363–375.

12. Wang W, Wei Z, Lam TW, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep 2011, 1:55.

13. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Velitch T, Collins P, Dai Y, et al. Integrated genotyping and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008, 40:1253–1260.

14. Lin M, Wei LT, Sellers WR, Lieberfarb M, Wong WH, Li C. ChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004, 20:1233–1240.

15. Barnes C, Pignolo V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. A robust statistical method for case–control association testing with copy number variation. Nat Genet 2008, 40:1245–1252.

16. Pique-Regi R, Ortega A, As Carrasalde A. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. Biosinformatics 2009, 25:1223–1230.

17. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Methods 2007, 4:316–321.

18. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. Challenges and standards in integrating surveys of structural variation. Nat Genet 2007, 39:57–515.

19. Winchester L, Yao C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 2009, 8:353–366.

20. Wan L, Sun K, Ding Q, Cui Y, Li M, Wen Y, Elston RC, Qian M, Fu WJ. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Res 2009, 37:e117.

21. Marioni JC, Thorne NP, Dermitzakis ET, Stranger BE, Lynch AG, a et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 2007, 8:R228.

22. Diskin SJ, Li M, Hou C, Yang S, Gestrzner H, Makonavatar H, Bucan M, Malig M, Wang K. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008, 36:e126.

23. van de Wiel MA, Picard F, van Wieringen WN, Ylstra B. Preprocessing and downstream analysis of microarray DNA copy number profiles. Brief Bioinform 2010, 12(1):10–21. http://bib.oxfordjournals.org/content/12/1/10. short.

24. Lander ES. Arrays of hope. Nat Genet 1999, 21:3–4.

25. Johnson WE, Li C, Bafarovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007, 8:118–127.

26. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Xu J, Chen JJ, Han T, Caput J, et al. Assessing batch effects of genotype calling algorithm BRLMM for
High frequencies of de novo CNVs in 2011,
148
72
5
PICNIC: an algorithm to predict
17
36
2008,
Nucleic Acids Res
77
De novo CNVs in bipolar disorder: recurrent themes et al.
81
2003,
Origins and functional impact of
2009,
449
2004,
Hybridization and amplification rate
2010,
21
72
9
10
11
37
2007,
2008,
Multiple recurrent de
A second generation human

28. Held GA, Grinstein G, Tu Y: 
29. Hooyberghs J, Baiesi M, Ferrantini A, Carlon E: The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. Nucleic Acids Res 2009, 37:e53.
30. Hooyberghs J, Van Hummelen P, Carlon E: High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. Am J Hum Genet 2005, 77:709–726.
31. Ohno N, Suzuki S, Furusawa C, Agata T, Kashiwagi A, Shimizu H, Yono T: An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. Bioinformatics 2008, 24:1278–1285.
32. Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, Li HI, Qian H, Farinha P, Gascoyne RD, Marra MA: Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res 2008, 36:e80.
33. Pugh TJ, Delaney AD, Farnoud N, Varambally R, Ghosh D, Barrett T, Pandey A, Chinnaiyan AM: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 2004, 101:9309–9314.
34. Rabiner LR: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989, 77:257–286.
35. Rabiner LR: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989, 77:257–286.
36. Bengtsson H, Iztarz R, Carvalho B, Speed TP: Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics 2008, 24:795–767.
37. McCarron SA, Kuruvilla FG, Kam JM, Cavley S, Nemeroff J, Vissers L, Shapero MH, de Bakker P, Maller JB, Kirby A, et al: Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008, 40:1166–1174.
38. Mulders GC, Bakema GT, Carlon E: Inverse Langmuir method for oligonucleotide microarray analysis. BMC Bioinformatics 2009, 10:64.
39. Giriagian S, Eichert EE: De novo CNVs in bipolar disorder: recurrent themes or new directions? Neurogenet 2011, 72:885–887.
40. Kaminskiy EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, Moreno-De-Luca D, Moreno-De-Luca A, Mulle JG, Warren ST, et al: An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med 2011, 13:777–784.
41. Held GA, Grinstein G, Tu Y: Modeling of DNA microarray data by using physical properties of hybridization. Proc Natl Acad Sci U S A 2003, 100:7575–7580.
42. Held GA, Grinstein G, Tu Y: Relationship between gene expression and observed intensities in DNA microarrays—a modeling study. Nucleic Acids Res 2006, 34:e70.
43. Rabiner LR: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989, 77:257–286.
44. Bengtsson H, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Rider PM, Chasman DI, et al: Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 2009, 84:148–161.
45. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, McVean G, Abecasis G, Altshuler D, Brudno M, et al: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 2011, 70:863–885.
46. Malhotra D, McCarthy S, Michaelson JJ, Vacek V, Burdick RE, Yoon S, Cichon S, Corvin A, Gary S, Gershon ES, et al: High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron 2011, 72:951–963.
47. Malhotra D, Sebat J: CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. Cell 2012, 148:1223–1241.
48. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Rider PM, Chasman DI, et al: Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 2009, 84:148–161.
49. Sanders SJ, Erkan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-DeLuca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, et al: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 2011, 70:863–885.
50. Malhotra D, McCarthy S, Michaelson JJ, Vacek V, Burdick RE, Yoon S, Cichon S, Corvin A, Gary S, Gershon ES, et al: High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron 2011, 72:951–963.
51. Malhotra D, Sebat J: CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. Cell 2012, 148:1223–1241.
52. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Rider PM, Chasman DI, et al: Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 2009, 84:148–161.