A Single-Array-Based Method for Detecting Copy Number Variants Using Affymetrix High Density SNP Arrays and its Application to Breast Cancer

Libertas Academica
FREEDOM TO RESEARCH

Ming Li[1], Yalu Wen[2] and Wenjiang Fu[2,3]

[1]Division of Biostatistics, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, USA. [2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing MI, USA. [3]Department of Mathematics, University of Houston, Houston, TX, USA.

Supplementary Issue: Array Platform Modeling and Analysis (A)

ABSTRACT: Cumulative evidence has shown that structural variations, due to insertions, deletions, and inversions of DNA, may contribute considerably to the development of complex human diseases, such as breast cancer. High-throughput genotyping technologies, such as Affymetrix high density single-nucleotide polymorphism (SNP) arrays, have produced large amounts of genetic data for genome-wide SNP genotype calling and copy number estimation. Meanwhile, there is a great need for accurate and efficient statistical methods to detect copy number variants. In this article, we introduce a hidden-Markov-model (HMM)-based method, referred to as the PICR-CNV, for copy number inference. The proposed method first estimates copy number abundance for each single SNP on a single array based on the raw fluorescence values, and then standardizes the estimated copy number abundance to achieve equal footing among multiple arrays. This method requires no between-array normalization, and thus, maintains data integrity and independence of samples among individual subjects. In addition to our efforts to apply new statistical technology to raw fluorescence values, the HMM has been applied to the standardized copy number abundance in order to reduce experimental noise. Through simulations, we show our refined method is able to infer copy number variants accurately. Application of the proposed method to a breast cancer dataset helps to identify genomic regions significantly associated with the disease.

KEYWORDS: copy number variants, copy number standardization, hidden Markov model, Affymetrix high density SNP array, breast cancer

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Li et al. A Single-Array-Based Method for Detecting Copy Number Variants Using Affymetrix High Density SNP Arrays and its Application to Breast Cancer. *Cancer Informatics* 2014:13(S4) 95–103 doi: 10.4137/CIN.S15203.

RECEIVED: April 09, 2014. **RESUBMITTED:** June 03, 2015. **ACCEPTED FOR PUBLICATION:** June 04, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Technical Advance

FUNDING: Michigan State University College of Medicine Summer Oncology Research Scholarship, and the University of Arkansas for Medical Sciences College of Medicine Children's University Medical Group Fund Grant Program. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: fuw@epi.msu.edu

Introduction

Genome-wide association studies (GWAS) are useful for the discovery of genetic variants underlying complex human diseases, such as breast cancer and Type II diabetes.[1,2] These genetic association studies typically compare the allele/genotype frequency for each single-nucleotide polymorphism (SNP) between cases and controls. Large projects such as the HapMap and 1,000 Genomes have shown that, in addition to single-nucleotide sequence variations (SNVs), structural alterations, such as copy number variants (CNVs), also account for up to 7.3% of the genetic variation among humans and may be involved in the genetic susceptibility to diseases,[3–7] including cancers.[8,9]

CNVs were first identified in the early 2000s,[10,11] and have been found to exist pervasively in human genomes.[12,13] Two major platforms of DNA microarrays have been commonly used for copy number estimation, namely, Affymetrix high density SNP arrays and Illumina Bead arrays,[14,15] relying on the relative intensity, an indirect measurement of hybridization of fluorescently labeled DNA fragments to immobilized probes on the arrays. Sophisticated statistical models are required to accurately infer the actual copy number within samples. In the past few years, several methods have been proposed for copy number inference. For example, smoothing methods were used in early studies in the field,[16,17] which fit a smoothing curve for the intensities along the genomic region and use certain threshold to infer copy number levels. The smoothing methods have been shown to be effective in studies for detecting genomic region with copy number changes.[18] However, these methods suffer from two limitations, namely, difficulty in locating accurate boundaries and difficulty in significance testing for the alterations.[19] Another group of methods adopt certain change-point models for the underlying copy number levels.[20,21] A change-point model usually assumes that the SNPs come from segments that are uniformly distributed in human genome, and their underlying copy numbers are piecewise constants with a series of jump points. By maximizing the likelihood function, the parameters as well as the change points can be estimated for copy number inference. Such models are further extended by various formations of hidden Markov models (HMMs).[22–25] The HMM assumes that the observed intensities of SNPs are emitted by an underlying Markov chain. It usually explicitly specifies the distribution for the waiting time of copy number changes and the jumping probabilities between copy number states. These methods have emerged as promising tools for copy number inference.

Estimation of array intensity values is challenging due to presence of experimental noise, both within an array and among arrays of different samples. For example, it is commonly known that the level of ozone affects hybridization reactions, which can affect interpretation of the results.[26,27] In a recent study, we and others proposed a novel method to estimate copy number abundance on a single-array single-SNP basis, referred to as the probe intensity composition representation (PICR).[28] This method models the cross-hybridization between DNA sequences via their physical binding affinities. It has shown great potential for differentiating copy number signals from background noises. In this article, we propose to extend the PICR method with hidden Markov modeling for copy number inference, referred to as the PICR-CNV. The estimated copy number abundance at each SNP locus from PICR will first be standardized to achieve parity between multiple samples, to which an HMM will be further applied. Our method has two major advantages: 1) By estimating the CNV abundance through PICR, we expect reduction of background noise in intensity values,[28–30] and thus be able to boost the performance of HMM. 2) our method does not require between-samples array normalization, which maintains the data integrity and the independency of individual samples. The proposed method is compatible with Affymetrix high density SNP arrays for detection of CNVs.

Methods

This section is organized as follows: We first describe the design of Affymetrix 500K SNP array. Then we briefly review the estimation of copy number abundance for each single array by using a newly established RICR model.[28] We introduce the multi-array standardization of the copy number abundance to achieve equal footing between individuals. Finally, we explain the PICR-CNV by applying an HMM to integrate multiple SNPs for copy number inference.

Design of Affymetrix 500K SNP array. Oligonucleotide microarrays annotate each SNP using a set of 24 probes of 25-mer photolithographically synthesized immobilized nucleic acid sequences. The target sequences are labeled with 3'-fluorescent dye before hybridization to the array, and their abundance are often measured with the fluorescent intensity on the array after hybridization.[31–33] In a 500K SNP array, six quartets are adopted to interrogate a single dimorphic SNP site with its possible alleles commonly denoted as *A* and *B*. Each quartet consists of four types of probes that are 25 base pairs in length. These probes are designed either perfectly matching (PM) the target sequence or mismatching (MM) at a particular nucleotide site for each allele: perfect match *A*, mismatch *A*, perfect match *B*, and mismatch *B*, denoted, respectively, as *PA*, *MA*, *PB*, and *MB* for short. The probe sets are also designed to hybridize with either sense strands ($s = 1$) or antisense strands ($s = -1$). The quartets have different shifts (k) for the nucleotide on the probe sequence (k may take the values $-4, -3, -2, -1, 0, 1, 2, 3, 4$) from the center nucleotide of the probe sequence ($k = 0$ at position 13 of the 25 base pairs) (see Fig. 1A of Matsuzaki et al.[34] for detailed illustration.).

Estimation for copy number abundance by PICR. The PICR method takes into account the cross-hybridization between DNA sequences via a positional-dependent nearest neighbor (PDNN) model.[28] In PICR, the florescent intensity of a particular probe set is decomposed into four terms: the baseline intensity (b), the products of allelic copy numbers abundance (N_A, N_B) and the binding affinity between target and probe sequences with respect to different alleles (f_A, f_B), and a measurement error (ε) [Equation (1)]. The binding affinities (f_A, f_B) are inherently determined by the physical property of the DNA sequences. The allelic copy number abundance can then be estimated via a linear regression between the intensities and binding affinities. Each probe set may be perfectly matched or mismatched to either allele as described above (*PA*, *MA*, *PB*, *MB*).

$$\begin{aligned} &\vdots \\ I_{PA} &= b + N_A f_A^{PA} + N_B f_B^{PA} + \varepsilon_{PA} \\ I_{PB} &= b + N_A f_A^{PB} + N_B f_B^{PB} + \varepsilon_{PB} \\ I_{MA} &= b + N_A f_A^{MA} + N_B f_B^{MA} + \varepsilon_{MA} \\ I_{MB} &= b + N_A f_A^{MB} + N_B f_B^{MB} + \varepsilon_{MB} \\ &\vdots \end{aligned} \tag{1}$$

Multi-array equal footing by standardization. By using PICR, the allelic copy number abundance is estimated on a single-array single-SNP basis. Since all the raw fluorescence intensities are subject to experimental scales, which may vary among arrays, it is essential to achieve equal footing for multiple arrays before any further analysis. We propose to define a standardized copy number abundance (SCN) as

$$SCN_{i,j} = \frac{N_{i,j,A} + N_{i,j,B}}{se(N_{i,j,A} + N_{i,j,B})}; i = 1, 2, \ldots, N; j = 1, 2, \ldots, K \tag{2}$$

where $N_{i,j,A}$ ($N_{i,j,B}$) denotes the allelic copy number abundance for SNP j of subject i, and $se(N_{i,j,A} + N_{i,j,B})$ denotes its estimated standard deviation of $N_{i,j,A} + N_{i,j,B}$ via the linear regression model of Equation (1). Assuming the raw intensities are normally distributed among probe sets, these standardized copy numbers are expected to have identical distributions for $i = 1, \ldots, N$; $\forall j = 1, \ldots, K$, and hence, are expected to be on the same scale.

PICR-CNV: a hidden Markov model for copy number inference. *Modeling strategy and copy number states.* As illustrated by Equation (2), our objective is to detect total copy number changes among subjects. We assume that an interrogated locus covering an SNP may have five possible copy numbers states, with its total copy number ranging from 0 to 4 (Table 1). For simplicity, we also refer to the copy number at an interrogated locus as the copy number of the SNP locus in this article. Such copy number states are not observed directly, and hence, are latent. Following the same notation with existing methods,[22,24] the inference of these hidden states is based on two types of observations, log *R* ratios (*LRR*) and B allele frequencies (*BAF*), which can be calculated by the estimated allelic copy numbers abundance. We first estimate the standardized copy number abundance for the *j*th SNP of subject *i*, and define its *LRR* as

$$R_{i,j} = LRR_{i,j} = \log_2\left(\frac{SCN_{i,j}}{SCN_{j,ref}}\right);$$

$$\text{where } SCN_{j,ref} = \underset{i \in \{Control\}}{median}(SCN_{i,j})$$

The *SCN* estimates among controls are regarded as a reference level for each SNP locus. We further define the *BAF* as

$$B_{i,j} = BAF_{i,j} = \begin{cases} 0 & \theta_{i,j} \le a_j \\ (\theta_{i,j} - a_j)/(b_j - a_j) & a_j < \theta_{i,j} \le b_j; \\ 1 & b_j < \theta_{i,j} \end{cases}$$

where $\theta_{i,j} = \dfrac{\arctan(scn_{i,j,B} / scn_{i,j,A})}{\pi / 2}$ and a_j, b_j are the corresponding thresholds for accurate genotyping of SNP *j* with the PICR. Similar to a few previous studies, an HMM is adopted to integrate *LRR* and *BAF* for copy number inference.[22,24,25] Our method differs from the existing ones by using standardized copy number abundance to calculate corresponding *LRR* and *BAF* rather than the probe intensities.

Transition probability for the hidden copy number states. We assume that the copy number states at SNP loci follow a time-dependent continuous Markov process, with genomic position of SNPs as "time". The transition probability is dependent on the distance between SNPs. Let $z_{i,j}$ be the underlying copy number state for the *j*th SNP of subject *i*, and let $d_{j,j'}$ be the physical distance between SNP *j* and SNP *j′* on the chromosome based on reference genome. We define the transition probability between the copy number states of SNPs *j* and *j′* as

$$p_{s,s'}(d_{j,j'}) =$$

$$p(z_{i,j'} = s' | z_{i,j} = s) = \begin{cases} e^{-dj,j'/\lambda_s} & if \;\; s = s' \\ (1 - e^{-dj,j'/\lambda_s})\, p_{s,s'} & if \;\; s \ne s' \end{cases};$$

$$\text{where } 1 \le s,\ s' \le 5; \text{ and} \sum_{s' \ne s} p_{s,s'} = 1$$

Here, $p_{s,s'}$ $(d_{j,j'})$is the probability for a hidden state *s* at SNP *j* to stay at the same state at SNP *j′* over a distance of $d_{j,j'}$, which is modeled by an exponential distribution with parameter $1/\lambda_s$. Therefore, λ_s has the interpretation of the expected "time" (distance) for the copy number at a particular state *s*. The longer the distance, the less likely the copy number states will remain the same. Similar modeling strategies have been commonly adopted in previous studies.[22,24]

Table 1. Configuration of five possible copy number states.

STATE (*Z*)	COPY NUMBER	POSSIBLE GENOTYPES	EXPECTED *LRR*	EXPECTED *BAF*
1	0	– (Deletion)	log(0) =–∞	0
2	1	*A*; *B*	$\log_2(1/2) = -1$	0 1
3	2	*AA*; *AB*; *BB*	$\log_2(1) = 0$	0 0.5 1
4	3	*AAA*; *AAB*; *ABB*; *BBB*	$\log_2(3/2) = 0.585$	0 0.33 0.67 1
5	4	*AAAA*; *AAAB*; *AABB*; *ABBB*; *BBBB*	$\log_2(2) = 1$	0 0.25 0.5 0.75 1

Emission probability for the observations. Since the copy number states are not observed directly, a set of emission probabilities are used to model the distribution of the observed variables (*LRR* and *BAF*) given the copy number states at SNP loci. Similar to a few previous studies, we modeled *LRR* and *BAF* by mixture distributions.[22,24,25] Denote $z_{i,j}$, $R_{i,j}$, and $B_{i,j}$, as the underlying copy number state, *LRR*, and *BAF* for the *j*th SNP of subject *i*. We first assume that the *LRR* and *BAF* at a particular SNP locus are conditionally independent given its underlying copy number state, so that

$$p(R_{i,j}, B_{i,j} | z_{i,j}) = p(R_{i,j} | z_{i,j})\, p(B_{i,j} | z_{i,j})$$

Further, the emission probability of *LRR* is modeled with the mixture of a uniform distribution and a normal distribution as

$$p(R_{i,j} | z_{i,j} = s) = \frac{\pi_R}{R_M - R_m} + (1 - \pi_R) f(R_{i,j}, \mu_{R,s,} \sigma_{R,s});$$
$$1 \le i \le N; 1 \le j \le K; 1 \le s \le 5;$$

where $f(., \mu, \sigma)$ denotes the probability density function for a normal distribution with mean μ and variance σ^2. Here, we assume that the genotyping may fail with a small probability of π_R. Under such a case, *LRR* is observed as a background noise, which follows a uniform distribution between its possible minimum (R_m) and maximum values ($\mu_{R,s}$). Otherwise, it follows a normal distribution with a mean $\mu_{R,s}$ and variance ($\sigma^2_{R,s}$) with respect to its underlying copy number states. As illustrated by Table 1, the expected mean and the variance of *LRR* observations vary by the underlying copy number states. Similarly, the expected values of *BAF* also vary by the underlying copy number states and the underlying genotypes (Table 1). We model the emission probability of *BAF* at a particular SNP locus with the mixture of a uniform distribution and normal or truncated normal distributions:

$$p(B_{i,j}|z_{i,j}=s)$$

$$=\begin{cases}\pi_B+(1-\pi_B)\sum_{g=1}^{G_S} f(B_{i,j},\mu_{s,g},\sigma_{s,g}) & for\ 0<B_{i,j}<1\\ \pi_B+(1-\pi_B)\sum_{g=1}^{G_S} \psi_{s,g}\Phi(B_{i,j},\mu_{s,g},\sigma_{s,g}) & for\ B_{i,j}=0;\\ \pi_B+(1-\pi_B)\sum_{g=1}^{G_S} \psi_{s,g}(1-\Phi(B_{i,j},\mu_{s,g},\sigma_{s,g})) & for\ B_{i,j}=1\end{cases}$$

where $\Phi(.,\mu,\sigma)$ denotes the cumulative distribution function for a normal distribution with mean μ and variance σ^2; G_s denotes the total number of all possible genotypes at a SNP locus with copy number state *s*; and $\mu_{s,g}$ and $\sigma_{s,g}$ are the mean and standard deviation of *BAF* for a SNP locus with copy number state *s* and genotype *g* (Table 1). Further, $\psi_{s,g}$ denotes the prior probability of *BAF* for copy number state *s* and genotype *g*, which can be calculated by a binomial distribution based on the *B* allele frequency in the population (*bpf*).[22,24,25] For example, an SNP with genotype *AAB* has copy number 3 and an expected *BAF* of 1/3. The prior probability of the *BAF* can be calculated as

$$p(G_{i,j}=AAB\,|\,z_{i,j}=3)=\binom{3}{1}(bpf_j)^1(1-bpf_j)^2.$$

Parameter estimation and copy number inference. In practice, we assume $\pi_R=\pi_B=0.01$ as the empirical error rate for genotyping, and λ_s, $1\le s\le 5$, are predetermined to account for the size of copy number variants. The set of parameters that need to be estimated includes

$\Omega=\{\omega(s)=p(z=s)$ as starting probability; $s=1, 2, 3, 4, 5$
$P=(p_{s,s'})$ as transition probability; $1\le s,s'\le 5$
$\mu_{R,s}$; mean of *R*; $s=1, 2, 3, 4, 5$
$\sigma_{R,s}$; standard deviation of *R*; $s=1, 2, 3, 4, 5$
$\mu_{B,s,g}$; mean of *B*; $s=1, 2, 3, 4, 5$; $g=1,2.G_s$
$\sigma_{B,s,g}$; standard deviation of *B*; $s=1, 2, 3, 4, 5$; $g=1, 2.G_s\}$

The parameters in Ω are optimized by using a forward-backward algorithm, also known as the Baum–Welch algorithm.[35] After the parameter estimation, the inference of copy number states is carried out by the Viterbi algorithm.[36] The computational algorithms are commonly used in previous studies, and are not detailed here.

Results

Simulation study. In the simulation study, we simulated a segment of the genome with length of 10^6 base pairs. We first assumed 10*K* SNPs with their physical position uniformly distributed in the genome. Each SNP was simulated for its underlying copy number state, and the observed probe intensities were measured by *LRR* and *BAF*. PICR-CNV was then applied to infer the underlying copy number states. In the simulation, the expected lengths of the copy number states were set at $\lambda_3=50K$ for a normal copy number of two copies, and $\lambda_l=5K$; $l=1, 2, 4, 5$ for other copy number states. The transition probability between copy number states was set as

$$(p_{s,s'})=\begin{bmatrix}0 & 0.01 & 0.97 & 0.01 & 0.01\\ 0.01 & 0 & 0.97 & 0.01 & 0.01\\ 0.25 & 0.25 & 0 & 0.25 & 0.25\\ 0.01 & 0.01 & 0.97 & 0 & 0.01\\ 0.01 & 0.01 & 0.97 & 0.01 & 0\end{bmatrix}$$

The parameters for the emission probability of *LRR* were set as

State: 1 2 3 4 5
$\mu_{R,s}=(\log_2(1/10), \log_2(1/2), \log_2(1), \log_2(3/2), \log_2(2))$
$\sigma_{R,s}=(\log_2(1/10), \log_2(1/2), \log_2(1), \log_2(3/2), \log_2(2))$

The parameters for emission probability of *BAF* were set as

$$(\mu B,s,g)=\begin{bmatrix}0.5 & & & & \\ 0 & 1 & & & \\ 0 & 0.5 & 1 & & \\ 0 & 1/3 & 2/3 & & \\ 0 & 1/4 & 2/4 & 3/4 & 1\end{bmatrix}\text{and }(\sigma_{B,s,g})=\begin{bmatrix}0.25 & & & & \\ 0.05 & 0.05 & & & \\ 0.05 & 0.05 & 0.05 & & \\ 0.05 & 0.05 & 0.05 & 0.05 & \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05\end{bmatrix}\text{for}$$

$$\begin{array}{c}s=1\\ s=2\\ s=3\\ s=4\\ s=5\end{array}\begin{bmatrix}g=- & & & & \\ g=A & g=B & & & \\ g=AA & g=AB & g=BB & & \\ g=AAA & g=AAB & g=ABB & g=BBB & \\ g=AAAA & g=AAAB & g=AABB & g=ABBB & g=BBBB\end{bmatrix}$$

The observation of *B* was further truncated at 0 and 1.

We simulated 100 subjects by using the above model parameters. For each subject, the underlying copy number states and genotypes of 10K SNPs were first simulated in a sequential order according to the transition probabilities. The frequencies of allele *B* in the population followed a uniform distribution between [0.1, 0.9]. For each SNP, the observations of *LRR* (*R*) and *BAF* (*B*) were then simulated by using the emission probability according to its underlying copy number states and genotypes. Two subjects were randomly selected to estimate the parameters by using the Baum–Welch algorithm. The estimated parameters were then used to infer the underlying copy number states for all subjects by using the Viterbi algorithm. Owing to computational concerns, the convergence criteria were met when the summation of the absolute change of all parameters was less than 10^{-3}. We calculated the error rates for the inferred the copy number states of all SNPs in all subjects. Because the expected lengths of the copy number variants (λ_s) were predetermined and may have an impact on the performance of the inference, we also examined the error rates when they were incorrectly specified.

The simulation results are summarized in Table 2. It is seen that the proposed method was accurate for inferring the underlying copy number states when λ_s was correctly specified. The overall error rate for all SNPs is 1.34e–04. When λ_s was incorrectly specified, the error rate increased with the level of mis-specification. In our simulation, we found that the error rate was not seriously inflated with an up to 10-fold overspecification of λ_s. It was also noted that the error rate for SNPs with normal states of two copies decreased by the level of overspecification of λ_s. This was because the normal states of two copies had the largest expected length, and an SNP was more likely to be inferred as two copies when λ_s was large. On the other hand, the error rate for SNPs with normal state of two copies increased when λ_s was underspecified. Overall, the error rate was still properly controlled when λ_s was incorrectly specified.

Application to breast cancer data. We also applied the proposed method to study CNVs that are associated with breast cancer development, using a recent GWAS data among a genetically isolated population of Ashkenazi Jews (AJ),[37] in which all participants have their four grandparents of Jewish and of Eastern European ancestry. We are limiting our study to the inherited genetic variation, and potential somatic mutations are beyond the scope of our current study. The original study had three phases. The first phase included 249 breast cancer cases without *BRCA1* and *BRCA2* mutations, and 299 cancer-free AJ women as controls. The second phase was a replicate study using 343 candidate SNPs among 950 AJ cases and 979 AJ controls. The third phase was also a GWAS study that included 243 AJ cases and 187 controls. The participants from phase I and phase III were genotyped with Affymetrix 500K SNP array, while those from phase II were genotyped by Illumina GoldenGate assay. We focused our analysis on the phase I and phase III data. It is also worthwhile to note that samples from phase I were genotyped by using a combination of a commercial version and an early access version of Affymetrix 500K SNP arrays. This mismatch of arrays has imposed additional challenge to the application of existing methods that require between-array normalizations. However, since PICR is a single-array method and does not require multiple array algorithms, the application of PICR is straightforward as long as the raw florescent intensity values are valid.

We used phase III as an initial study for the analysis. The proposed method was first applied to 10 randomly selected controls for parameter training. The initial genotype calling was conducted by PICR, and all parameters in Ω were optimized and then used to infer the copy number states among all participants. We first examine the distribution of the sizes of identified CNVs (Fig. 1). The shape of this distribution is consistent with existing studies (Fig. 1 of Li et al.[38]). For each SNP locus, we further conducted a Kolmogorov–Smirnov (KS) test to compare the inferred copy numbers between cases and controls. The significant regions were selected if three consecutive SNPs showed significant copy number differences at a level of 1e–07. After the region was selected, a global *P*-value was further calculated by conducting a KS test using the average copy number of the SNPs within the region. The results are summarized in Table 3. The findings included 34 genomic regions

Table 2. Error rate for inference of copy number states with correctly and incorrectly specified expected length of copy number states.

	AVERAGE NO. OF SNP WITH COPY NUMBER STATE IN EACH SUBJECT					
HMM STATE	**1**	**2**	**3**	**4**	**5**	**TOTAL**
	557	163	8,875	185	220	10,000
λ used in HMM	**Error rates for copy number state inference**					
λ_{True}	5.92e–04	1.53e–04	2.37e–05	1.40e–04	1.32e–03	1.34e–04
$2\lambda_{True}{}^{a}$	3.97e–03	4.91e–04	1.69e–05	7.01e–04	4.46e–03	3.55e–04
$5\lambda_{True}$	4.18e–03	6.13e–4	1.80e–05	7.01e–04	4.51e–03	3.71e–04
$10\lambda_{True}$	4.38e–03	9.20e–04	1.80e–05	1.08e–03	4.87e–03	4.02e–04
$0.5\lambda_{True}$	9.69e–04	1.53e–04	3.27e–05	1.56e–04	1.32e–03	1.66e–04

Note: [a]means the model specified λ is 2 times greater than the true λ.

from 16 chromosomes. The region with the largest number of significant SNPs was 4q31.23. This region had 10 SNPs showing significant copy number difference between cases and controls. Besides region 4q31.23, two regions, 1p21.1 and 10q21.1, both have seven significant SNPs. Three regions have five SNPs with significant copy number differences, including 6q22.33, 6q27, and 11p12. These results indicate that copy number alterations on chromosome 4, 6, 1, and 11 may have a significant impact on the development of breast cancer.

We also applied the same procedure to the phase I data for replication. The results are also summarized in Table 3. Among the regions identified from phase III data, the copy number changes remained significant at five regions: 4q31.23, 6q13, 12q23.1, 13q14.3, and 2p21. These five regions contained 10, 5, 4, 4, and 3 SNPs, respectively.

Discussion and Conclusion

In this study, we have proposed an HMM-based method (PICR-CNV) for copy number inference. Through simulations, we have shown that the proposed method is highly accurate for copy number inference and robust against mis-specification of the predetermined model parameter. While it is not straightforward to evaluate the copy number inference with real data due to the unknown copy number status, we have evaluated the proposed standardization approach for genotyping accuracy. We applied PICR to 90 HapMap samples with Affymetrix Mapping 100K arrays, and found that the genotyping accuracies were improved by using standardized copy number abundance compared to using raw copy number abundance (99.70% vs 99.63%). Empirically, we also found that the standardized copy number abundance provided better genotype clustering than its alternative (Fig. 2). The proposed method was further illustrated with an application to breast cancer datasets. The analysis of breast cancer data also identified a few genomic regions that were significantly associated with breast cancer development. Most of these identified regions have been reported in the literature for potential involvement in breast cancer. One SNP in the region 4q31.23 has been recently reported to be significantly associated with breast cancer progression.[39] A gene *ARHGAP10-NR3C2*, which was located in the region, was also known to be related to carcinogenesis through structure alteration.[40] Possible copy number changes of the region were also observed from cancer cell line data.[41] Regions 1p21.1 and 10q21.1 have also been reported repeatedly for potential association with breast cancer. Chromosome arm 1p was suggested to contain multiple tumor suppressor genes.[42] Structure alterations of 1p21.1 have been observed from many studies.[42–45] Region 10q21.1 also has multiple candidate tumor suppressors, such as *ANX7* and *CDC2*.[46,47] Interestingly, for region 6q22.33, it was identified by the initial GWAS as a novel locus for breast cancer development.[37] Our analysis also confirmed this finding and also suggested that the copy number changes in the region may also play an important role.

The associations of the identified regions, including 4q31.23, 12q23.1, 13q14.3, and 2p21, were also replicated by using an independent dataset. The region of 4q31.23 was identified by phase III as the one with the largest number of significant SNPs. The long arm of chromosome 6 was reported to be frequently rearranged in human cancers.[48–50] The region of 6q13 was among the important regions that showed copy number alterations.[51,52] For region 12q23.1, a gene *SLC5A8* was identified by a previous study to be affected frequently by structure changes.[53,54] This gene was actively involved in the gene pathway for the development of primary human tumors.[55,56] The region 13q14.3 has been reported for copy number changes in various cancers, such as prostate cancer and breast cancer.[57–60] The structure changes of 2p12 was also suggested to be involved in cancer development.[61] While it is biologically plausible that the structure changes of these regions may play an important role in the development of breast cancer, additional studies are needed to further replicate the association and verify the biological functioning and mechanisms.

We are also aware that our method may have a few limitations. First, our copy number estimation method is based on the design of Affymetrix 500K SNP arrays. Further extension will be needed before applying it to Illumina platform or Affymetrix 6.0 arrays. Our current study is a secondary analysis of an existing GWAS dataset, extending previous genotype-based association study to copy-number-based association study. The

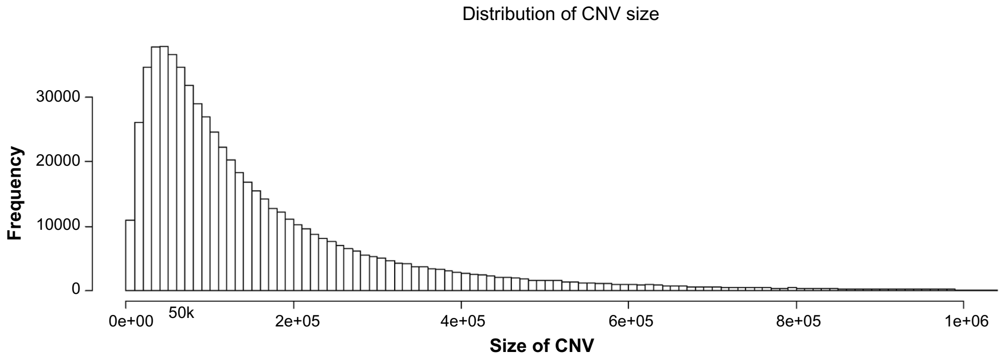

Figure 1. Distribution of the size of identified CNVs based on BRCA GWAS data.

Table 3. Regions showing significant copy number variation in phase III data and their replication in phase I data.

CHRO.	CYTOBAND	PHYSICAL LOCATION[a]	NO. OF SNPs	*P*-VALUE (PHASE III)[b]	*P*-VALUE (PHASE I)[c]
1	p21.1	102622376–102640646	7	2.62e–13	0.954
1	p12	120292824–120312909	3	7.62e–14	0.999
1	q22	154077091–154106555	3	2.453e–11	0.999
2	**p21**	45759616–45760637	3	1.106e–08	0.014
2	p12	81196767–81197522	3	7.232e–09	0.977
2	q21.1	131925407–131955270	3	4.872e–13	0.999
3	p14.3	57706175–57839689	3	1.228e–09	0.116
4	q26	117544365–117576957	3	4.577e–11	0.138
4	**q31.23**	148668320–148697327	10	9.43e–15	7.56e–05
4	q32.3	166885930–166957371	5	6.664e–11	0.189
5	q14.3	84350898–84398999	5	4.330e–14	0.720
5	q22.3	115145252–115178424	4	2.220e–16	0.893
6	**q13**	75247853–75311831	5	5.218e–15	0.034
6	q22.33	128476625–128533696	6	2.409e–13	0.806
6	q23.2	134651674–134672863	5	3.722e–10	0.999
6	q27	165234976–165247908	6	1.752e–09	0.996
7	q22.1	98318717–98361309	4	4.727e–11	0.103
7	q31.31	118754169–118754169	5	1e–17	0.524
8	q11.22	52786953–52796842	3	4.550e–10	0.840
8	q21.3	90963387–90964181	3	2.862e–08	0.772
8	q24.13	125649171–139914783	3	2.30e–08	0.973
8	**q24.3**	145891814–145948840	4	3.220e–15	7.96e–04
9	**p21.3**	22270796–22294230	5	6.249e–09	3.33e–03
10	q21.1	56853055–74432554	7	1.084e–09	0.998
11	p13	36306019–36366302	3	8.95e–11	0.223
11	p12	37905557–37916354	6	2.627e–09	0.968
11	q22.3	104741435–104806689	5	4.152e–14	0.999
12	**q23.1**	94977527–95052366	4	1.11e–16	1.07e–04
13	q13.3	34828145–34846106	4	8.975e–10	0.428
13	**q14.3**	51036156–51071687	4	5.268e–12	6.83e–09
13	q33.1	103334252–103344370	5	1.589e–09	0.964
14	q23.1	60136001–60140123	5	1.843e–12	0.996
18	p11.31	3597746–3635894	3	4.268e–10	0.417
X	q27.3	146596395–146646974	4	5.873e–14	0.086

Notes: [a]location based on Human Genome Assembly NCBI build 36.1. [b]Phase III included 243 cases and 187 controls. [c]Phase I included 249 cases and 299 controls.

Affymetrix SNPs array has been a major platform for SNP genotyping and copy number estimation. It was adopted by the Wellcome Trust Case Control Consortium (WTCCC) for intensive GWAS of 14,000 cases of seven common diseases and 3,000 shared controls.[62] Second, the current study only considered the total copy number changes at each locus. However, copy number changes may still occur without total number changes, such as balanced copy number with preferential loss of heterozygosity (LOH). Further extensions are needed to account for such copy number changes. Third, our method currently focuses on detecting the total copy number changes in an unrelated population. Detecting copy number status for related individuals or paternal (maternal) specific copy numbers is beyond the scope of current study.

Acknowledgments

We would like to thank Dr. Burt Gold from the National Cancer Institute and Drs. Vijai Joseph and Kenneth Offit from the Memorial Sloan Kettering Cancer Center, for helping us access the BRCA dataset and also for their critical comments that improved this manuscript. This work was supported in part by the Michigan State University College of Medicine

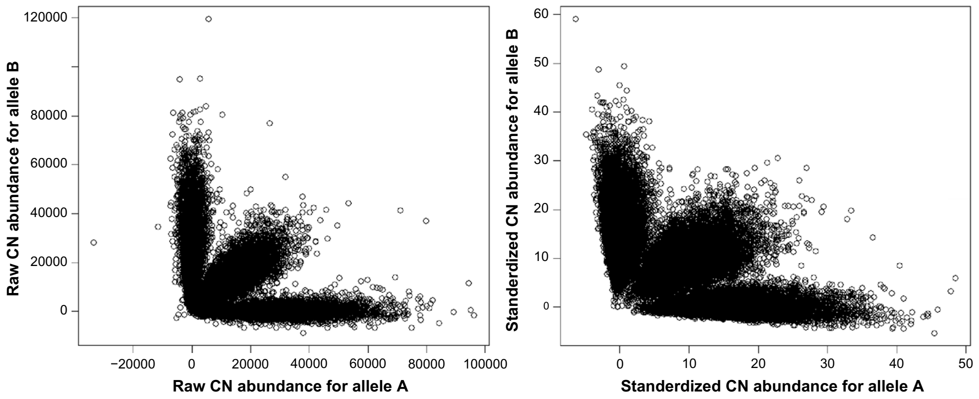

Figure 2. Raw and standardized copy number abundance for a randomly selected HapMap sample (NA12892).

Summer Oncology Research Scholarship and the University of Arkansas for Medical Sciences College of Medicine Children's University Medical Group Fund Grant Program.

Author Contributions

Conceived and designed the experiments: ML, WF. Analyzed the data: ML, YW, WF. Wrote the first draft of the manuscript: ML, WF. Contributed to the writing of the manuscript: YW. Agree with manuscript results and conclusions: ML, YW, WF. Jointly developed the structure and arguments for the paper: ML, WF. Made critical revisions and approved final version: YW, WF. All authors reviewed and approved of the final manuscript

REFERENCES

1. Easton DF, Pooley KA, Dunning AM, et al; SEARCH collaborators, AOCS Management Group. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007;447(7148):1087–93.
2. Zeggini E, Scott LJ, Saxena R, et al; Wellcome Trust Case Control Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008;40(5):638–45.
3. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11(6):446–50.
4. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–53.
5. Choy KW, Setlur SR, Lee C, Lau TK. The impact of human copy number variation on a new era of genetic testing. *BJOG*. 2010;117(4):391–8.
6. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.
7. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52–8.
8. Diskin SJ, Hou C, Glessner JT, et al. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*. 2009;459(7249):987–91.
9. Frank B, Bermejo JL, Hemminki K, et al. Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis*. 2007;28(7):1442–5.
10. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949–51.
11. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525–8.
12. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–54.
13. Perry GH, Ben-Dor A, Tsalenko A, et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet*. 2008;82(3):685–95.
14. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*. 2008;40(10): 1166–74.
15. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Res*. 2006;16(9):1136–48.
16. Pollack JR, Sørlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*. 2002;99(20):12963–8.
17. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*. 2004;20(18):3413–22.
18. Stankiewicz P, Park SS, Inoue K, Lupski JR. The evolutionary chromosome translocation 4;19 in gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Res*. 2001;11(7):1205–10.
19. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005;21(19):3763–70.
20. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*. 2007;23(6):657–63.
21. 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.
22. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an objective Bayes Hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013–25.
23. Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann Appl Stat*. 2008;2(2):687–713.
24. Wang K, Li M, Hadley D, 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.
25. Sun W, Wright FA, Tang Z, et al. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res*. 2009;37(16):5365–77.
26. Fare TL, Coffey EM, Dai H, et al. Effects of atmospheric ozone on microarray data quality. *Anal Chem*. 2003;75(17):4672–5.
27. Frueh FW. Impact of microarray data quality on genomic data submissions to the FDA. *Nat Biotechnol*. 2006;24(9):1105–7.
28. Wan L, Sun K, Ding Q, et al. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. *Nucleic Acids Res*. 2009;37(17):e117.
29. Li M, Wen Y, Lu Q, Fu WJ. An imputation approach for oligonucleotide microarrays. *PLoS One*. 2013;8(3):e58677.
30. Wen Y, Li M, Fu WJ. Catching the genomic wave in oligonucleotide single-nucleotide polymorphism arrays by modeling sequence binding. *J Comput Biol*. 2013;20(7):514–23.
31. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235): 467–70.
32. Brown CK, Madauss K, Lian W, Beck MR, Tolbert WD, Rodgers DW. Structure of neurolysin reveals a deep channel that limits substrate access. *Proc Natl Acad Sci U S A*. 2001;98(6):3127–32.
33. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. *Genome Res*. 2002;12(2):325–32.
34. Matsuzaki H, Loi H, Dong S, et al. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*. 2004;14(3):414–25.
35. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat*. 1970;41:164–71.

36. Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory*. 1967;It13(2):260.
37. Gold B, Kirchhoff T, Stefanov S, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A*. 2008;105(11):4340–5.
38. Li J, Yang T, Wang L, et al. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PLoS One*. 2009;4(11):e7958.
39. Azzato EM, Pharoah PD, Harrington P, et al. A genome-wide association study of prognosis in breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2010;19(4):1140–3.
40. Katoh M. Characterization of human ARHGAP10 gene in silico. *Int J Oncol*. 2004;25(4):1201–6.
41. Park JJ, Kang JK, Hong S, et al. Genome-wide combination profiling of copy number and methylation offers an approach for deciphering misregulation and development in cancer cells. *Gene*. 2008;407(1–2):139–47.
42. Aarts M, Dannenberg H, deLeeuw RJ, et al. Microarray-based CGH of sporadic and syndrome-related pheochromocytomas using a 0.1–0.2 Mb bacterial artificial chromosome array spanning chromosome arm 1p. *Genes Chromosomes Cancer*. 2006;45(1):83–93.
43. Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res*. 2006;8(1):R9.
44. Varma G, Varma R, Huang H, et al. Array comparative genomic hybridisation (aCGH) analysis of premenopausal breast cancers from a nuclear fallout area and matched cases from Western New York. *Br J Cancer*. 2005;93(6):699–708.
45. Bello MJ, de Campos JM, Vaquero J, Kusak ME, Sarasa JL, Rey JA. High-resolution analysis of chromosome arm 1p alterations in meningioma. *Cancer Genet Cytogenet*. 2000;120(1):30–6.
46. Srivastava M, Bubendorf L, Srikantan V, et al. ANX7, a candidate tumor suppressor gene for prostate cancer. *Proc Natl Acad Sci U S A*. 2001;98(8):4575–80.
47. Ohta T, Okamoto K, Isohashi F, et al. T-loop deletion of CDC2 from breast cancer tissues eliminates binding to cyclin B1 and cyclin-dependent kinase inhibitor p21. *Cancer Res*. 1998;58(6):1095–8.
48. Dutrillaux B, Gerbault-Seureau M, Zafrani B. Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. *Cancer Genet Cytogenet*. 1990;49(2):203–17.
49. Sheng ZM, Marchetti A, Buttitta F, et al. Multiple regions of chromosome 6q affected by loss of heterozygosity in primary human breast carcinomas. *Br J Cancer*. 1996;73(2):144–7.
50. Chappell SA, Walsh T, Walker RA, Shaw JA. Loss of heterozygosity at chromosome 6q in preinvasive and early invasive breast carcinomas. *Br J Cancer*. 1997;75(9):1324–9.
51. Rodriguez C, Causse A, Ursule E, Theillet C. At least five regions of imbalance on 6q in breast tumors, combining losses and gains. *Genes Chromosomes Cancer*. 2000;27(1):76–84.
52. Noviello C, Courjal F, Theillet C. Loss of heterozygosity on the long arm of chromosome 6 in breast cancer: possibly four regions of deletion. *Clin Cancer Res*. 1996;2(9):1601–6.
53. Hong C, Maunakea A, Jun P, et al. Shared epigenetic mechanisms in human and mouse gliomas inactivate expression of the growth suppressor SLC5A8. *Cancer Res*. 2005;65(9):3617–23.
54. Yamanaka S, Sunamura M, Furukawa T, et al. Chromosome 12, frequently deleted in human pancreatic cancer, may encode a tumor-suppressor gene that suppresses angiogenesis. *Lab Invest*. 2004;84(10):1339–51.
55. Gupta N, Martin PM, Prasad PD, Ganapathy V. SLC5A8 (SMCT1)-mediated transport of butyrate forms the basis for the tumor suppressive function of the transporter. *Life Sci*. 2006;78(21):2419–25.
56. Paroder V, Spencer SR, Paroder M, et al. Na(+)/monocarboxylate transport (SMCT) protein expression correlates with survival in colon cancer: molecular characterization of SMCT. *Proc Natl Acad Sci U S A*. 2006;103(19):7270–5.
57. Lerebours F, Bertheau P, Bieche I, et al. Evidence of chromosome regions and gene involvement in inflammatory breast cancer. *Int J Cancer*. 2002;102(6):618–22.
58. Bullrich F, Fujii H, Calin G, et al. Characterization of the 13q14 tumor suppressor locus in CLL: identification of ALT1, an alternative splice variant of the LEU2 gene. *Cancer Res*. 2001;61(18):6640–8.
59. Yin Z, Spitz MR, Babaian RJ, Strom SS, Troncoso P, Kagan J. Limiting the location of a putative human prostate cancer tumor suppressor gene at chromosome 13q14.3. *Oncogene*. 1999;18(52):7576–83.
60. Frank B, Klaes R, Burwinkel B. Familial cancer and ARLTS1. *N Engl J Med*. 2005;353(3):313–4.
61. Rippe V, Drieschner N, Meiboom M, et al. Identification of a gene rearranged by 2p21 aberrations in thyroid adenomas. *Oncogene*. 2003;22(38):6111–4.
62. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–78.