Statistica Sinica **23** (2013), 1463-1477
doi:http://dx.doi.org/10.5705/ss.2012.071s

TESTING ASSOCIATIONS OF COPY NUMBER VARIATIONS IN GENOME-WIDE ASSOCIATION STUDIES

Jianxin Shi and Peng Li

National Cancer Institute

Abstract: Copy number variations (CNVs) are a major source of genetic variation in humans. In large-scale genome-wide association studies (GWAS), CNVs have been detected from the intensity data generated by SNP genotyping arrays and then tested for association. This strategy lacks statistical power for detecting associations with short CNVs. In this article, we propose methods for testing the association for each probe, based on a Hidden Markov Model that leverages information from nearby probes in the same CNV region. Our methods do not require specifying CNV regions, are convenient for genome-wide scan data, and work for both population-based and family-based studies. Through simulation studies, we found that loss of efficiency due to CNV calling uncertainty was very small even for short CNVs covering as few as four probes in case-control studies. The efficiency loss was larger for short CNVs in family studies. We applied our methods to a large family-based GWAS of autism in 831 trios, and identified a genomic region on chromosome 17q22 harboring deletions that may contribute to the disease risk. Our methods are computationally efficient, requiring only two hours to analyze the genome-wide intensity data of all trios using a single Linux core.

Key words and phrases: Copy number variation, family-based study, genome-wide association study, Hidden Markov Model, TDT.

1. Introduction

Copy number variations (CNVs) are pervasive in the human genome and have been reported to be associated with many complex diseases in genome-wide association studies (GWAS) based on SNP arrays. Unlike GWAS SNP analysis, where no causality can be inferred, detected rare CNVs are very likely to be causal. Identifying disease-causing rare CNVs helps to elucidate the etiology of complex diseases and may ultimately contribute to molecular diagnosis and treatment.

CNVs are associated with a disease if cases are more likely to carry a CNV in case-control studies (Figure 1A and 1B) or if the CNV carried by parents is more likely to be transmitted to affected offspring in family-based studies (Figure 1C). Most CNV analyses in GWAS have been based on a two-stage strategy: (1) detect CNVs for each subject using CNV detection packages (see e.g., Olshen et al.

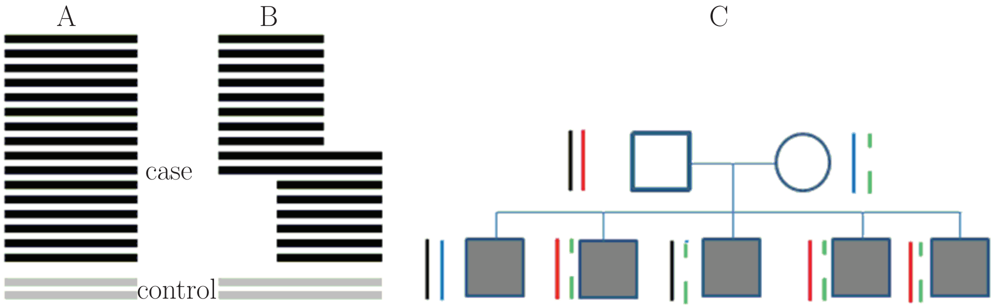

Figure 1. CNVs are associated with the disease in case-control studies and family studies. A: CNVs have the same boundary in case-control studies. B. CNVs overlap but do not have the same boundary. C: The mother carries a deletion. If the deletion is associated with disease status, it is more likely to be transmitted to the affected offsprings.

(2004), Wang et al. (2007), Colella et al. (2007), Korn et al. (2008) Wang et al. (2009), Zhang et al (2010), Jeng, Cai, and Li (2010), Shi and Li (2012)) and (2) perform probe-wise or segment-wise association tests to identify candidate susceptibility CNV regions. This strategy has proven successful for detecting long CNVs associated with diseases. However, the majority of germline CNVs are short and covered by only a few probes in current genotyping platforms. Hence, associations of short CNVs with diseases cannot be detected with good power.

A second strategy is to directly test the association without calling CNVs from the intensity data. For example, Ionita-Laza et al. (2008) proposed a method for nuclear families that tested the association with each probe using its Log R Ratio (LRR) as a CNV surrogate. Eleftherohorinou et al. (2011) proposed a similar strategy for detecting CNV associations for quantitative traits in nuclear families. Barnes et al. (2008) developed a likelihood framework to test associations for case-control studies in given CNV regions. These methods work well only for known CNV regions but may miss rare, undocumented CNVs, which are of primary interest in GWAS. Also, they may lose power when the CNVs do not completely overlap. In addition, these methods may further lose power because they do not utilize the B-Allele Frequency (BAF).

In this article, we develop methods for detecting CNV associations in both case-control and family-based studies. Our methods integrate intensity information from probes in a CNV region using a Hidden Markov Model (HMM). Our methods do not require the specification of CNV regions, and therefore are convenient to use in GWAS. We demonstrate through extensive simulations that, under realistic parameter settings in case-control studies, our methods have high relative efficiency even for short CNVs covering as few as four probes while maintaining expected type-I error rates. Relative efficiency is lower in family-based studies than case-control studies. Type-I error rates are robust to misspecification of model parameters in both case-control studies and family-based studies. Finally, we applied our methods to a GWAS of autism based on nuclear families and identified one promising genomic regions harboring deletions that may contribute to the risk of autism.

2. Methods

2.1. Characteristics of the intensity data for CNV inferences

CNVs are inferred from the raw intensity measurements obtained from the genotyping array. Each DNA sample is preprocessed and then genotyped at M SNP probes. For each probe, we denote the two alleles as A and B. Each probe has two measurements derived from the raw intensity data: the Log R Ratio (LRR) and the B Allele Frequency (BAF). The LRR measures the total intensity of the fluorescence used to label the probe in the assay; LRR is then an approximation of the total amount of DNA. The BAF measures the proportion of the DNA attributable to the B allele. For example, for a probe with copy number (CN) 3, the four possible genotypes AAA, AAB, ABB, and BBB have BAFs close to 0, 1/3, 2/3 and 1, respectively. BAFs are truncated to [0,1]. The distributions of LRRs and BAFs are illustrated in Figure 2. For convenience, we use $\mathbf{\Omega}$ to represent all parameters characterizing these distributions. In this paper, we assume that $\mathbf{\Omega}$ is known.

2.2. Testing association with CNVs in case-control studies

Consider a set of N subjects genotyped at M markers on one chromosome. For the ith subject, let $Y_i \in \{0,1\}$ denote the binary disease status. Let X_{it} denote the LRR, B_{it} the BAF and C_{it} the CN at the tth probe. Here, $C_{it} \in \{0,1,2,3,4\} : C_{it} = 2$ suggests copy normal; $C_{it} = 0, 1$ denote deletions; $C_{it} = 3, 4$ denote duplications. Let $O_{it} = (X_{it}, B_{it})$ and $\mathbf{O}_i = (O_{i1}, \ldots, O_{iM})$. Hence, $\mathbf{O}_i$ denotes all observed information for the subject. Given C_{it}, X_{it} and B_{it} are independent,

$$P\{O_{it} = (X_{it}, B_{it})|C_{it}\} = P\{X_{it}|C_{it}\}P\{B_{it}|C_{it}\}. \tag{2.1}$$

In Appendix A, we describe the computation of $P\{X_{it}|C_{it}\}$ and $P\{B_{it}|C_{it}\}$.

We develop a score statistic to test if the CNV status at probe t is associated with the disease status. We assume that

$$\frac{P(Y_i = 1|C_{it} = c)}{P(Y_i = 0|C_{it} = c)} = e^{\alpha+\beta S(c)}. \tag{2.2}$$

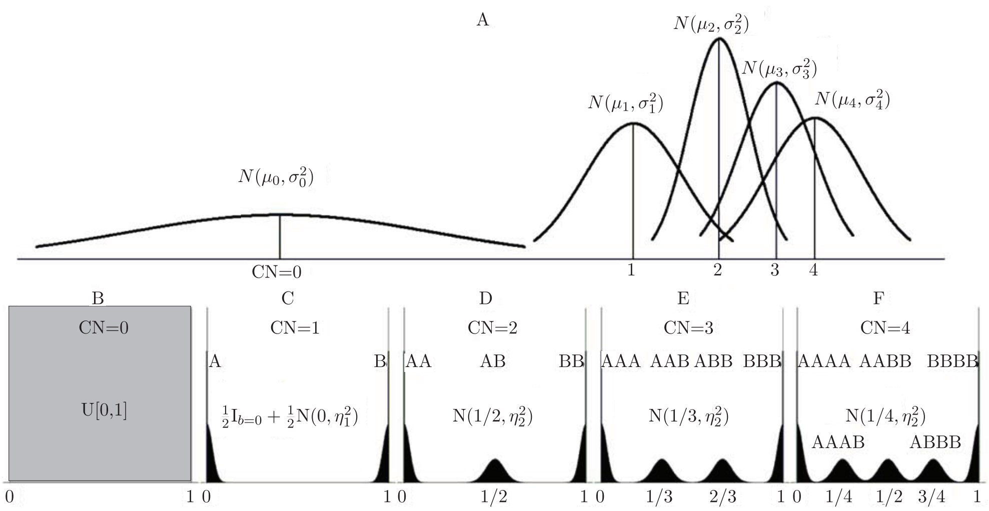

Figure 2. A: The distribution of LRRs for probes with different copy numbers. Estimated from the intensity data of long, experimentally validated CNVs in Autism Genetic Resource Exchange (www.agre.org) Consortium. Note that the distributions of CN1, CN2, and CN3 at each probe are not well separated. Figures B-F illustrate the distribution of BAFs for different copy numbers and genotypes.

Here $S(c)$ is the score assigned to each CNV status c, which makes our framework flexible enough to test different CNV effects. We take $p_{tc} = P\{C_{it} = c|Y_i = 0\}$ to be the CNV frequencies in the control population. Then

$$P\{C_{it} = c|Y_i = 1\} = \frac{e^{\beta S(c)}p_{tc}}{\sum_{c'} e^{\beta S(c')}p_{tc'}}. \tag{2.3}$$

The retrospective likelihood is $P\{\mathbf{O}_i|Y_i\} = \sum_c P\{\mathbf{O}_i|C_{it} = c, Y_i\}P\{C_{it} = c|Y_i\}$. We assume that the intensity data $\mathbf{O}_i$ and Y_i are independent given C_{it}, a reasonable assumption if the cases and controls are genotyped using the same genotyping platform and balanced in plates. Otherwise, parameters $\mathbf{\Omega}$ should be estimated separately. Under this assumption,

$$P\{\mathbf{O}_i|Y_i\} = \sum_c P\{\mathbf{O}_i|C_{it} = c\}P\{C_{it} = c|Y_i\}. \tag{2.4}$$

We define

$$T_{itc} = P\{\mathbf{O}_i|C_{it} = c\}. \tag{2.5}$$

This probability effectively captures all available information in the nearby probes for inferring C_{it}. In Section 2.3., we compute T_{itc} based on a HMM.

Let $P_t = (p_{t0}, \ldots, p_{t4})$. Take $l(\beta, P_t) = \sum_{i=1}^{N} \log P\{\mathbf{O}_i|Y_i\}$ as the retrospective log likelihood function summarizing all subjects. Combining (2.3), (2.4), and (2.5) gives:

$$l(\beta, P_t) = \sum_{i=1}^{N} \Big\{ \log \sum_{c=0}^{4} e^{\beta Y_i S(c)} p_{tc} T_{itc} - \log \sum_{c=0}^{4} e^{\beta Y_i S(c)} p_{tc} \Big\}. \tag{2.6}$$

Interests center on testing $H_0 : \beta = 0$, that the copy number score $S(c)$ at probe t is not associated with the disease status. In Appendix B, we derive a score statistic Z_t. The nuisance parameters P_t are estimated using the EM algorithm under H_0.

We define Δ to be the set of probes with estimated CNV (deletion, duplication or both) frequencies between 0.01 and 0.05. We only test the association for probes in Δ, because rarer CNVs are not powered and more common CNVs are tagged by common SNPs. Let $Z_{max} = \max_{t \in \Delta} |Z_t|$. The genome-wide significance is approximated by permutations. In practice, restricting analysis to Δ excludes 90% to 95% probes and hence substantially reduces the multiple testing burden. The high correlation among the score statistics in the same CNV region further reduces multiple testing burden, which will be evaluated using simulations.

Comment: The framework can be applied to other genotyping platforms, e.g. array-comparative genomic hybridization (aCGH) with only non-polymorphic probes and genotyping platforms with both polymorphic and non-polymorphic probes. Because non-polymorphic probes do not have BAFs, (2.1) reduces to $P\{X_{it}|C_{it}\}$.

2.3. Compute T_{itc} based on a Hidden Markov Model

We choose to use a HMM to integrate the intensity of probes in the neighborhood to maximize the power. This is particularly convenient because of no requirement for specifying CNV regions. The settings of HMM are similar to most HMMs designed for CNV analysis (e.g., Wang et al. (2007), Colella et al. (2007)).

At probe t, the hidden state is C_{it}. The emission probability $P\{O_{it}|C_{it}\}$ is computed in (2.1). The transition probability

$$a_{t,kj} = P\{C_{i,t+1} = j | C_{i,t} = k\} = \begin{cases} A_{kj}(1 - e^{-\lambda d_t}) \text{ if } k \neq j; \\ 1 - \sum_{l \neq k} a_{t,kl} \;\; \text{if } k = j, \end{cases}$$

is assumed to depend on the physical distance d_t, a constant λ and the baseline transition probability matrix A. Both λ and A can be estimated based on Hapmap samples genotyped using the same genotyping platform because the

CNV status of these samples have been accurately inferred by deep sequencing and experimentally validated.

Note that $T_{itc} = P\{\mathbf{O}_i|C_{it} = c\}$ can be written as $T_{itc} = P\{O_{i1}, \ldots, O_{i,t-1}| C_{it} = c\}P\{O_{it}|C_{it} = c\}P\{O_{i,t+1}, \ldots, O_{i,M}|C_{it} = c\}$. Let $T^+_{itc} = P\{O_{i,t+1}, \ldots, O_{i,M}|C_{it} = c\}$ and $T^-_{itc} = P\{O_{i1}, \ldots, O_{i,t-1}|C_{it} = c\}$. Then, T^+_{itc} can be computed efficiently using the backward algorithm in HMM by the induction rule

$$T^+_{itc} = \sum_{j=0}^{4} a_{t,cj} P\{O_{i,t+1}|C_{i,t+1} = j\}T^+_{i,t+1,j}. \tag{2.7}$$

The initial condition is $T_{iMc} = 1$ for all c. Similarly, T^-_{itc} can be computed by reversing the index of the Markov chain. We only need to compute T_{itc} once.

2.4. Testing associations of CNVs in nuclear families

Consider a genetic study with N complete nuclear families genotyped at M SNP probes. Family i has n_i affected offsprings. For the jth subject in the ith family, let $\mathbf{O}_{ij}$ denote the intensity at all probes and C_{ijt} be the copy number at probe t. Here, $j \in \{f, m\}$ indexes parents and $j \in \{1, \ldots, n_i\}$ indexes offsprings. If the CNV status is known for each subject, the associations can be tested using the transmission disequilibrium test (TDT; Spielman, McGinnis, and Ewens (1993)) or FBAT (Laird, Horvath, and Xu (2000)).

Ideally, one shall make full use of family information to jointly call CNVs (Wang et al. (2008)) and then perform TDT or FBAT. However, the sensitivity of detecting a CNV transmitted from a parent to an offspring is higher than a CNV not transmitted. This unequal sensitivity creates a substantial bias in TDT, particularly for short CNVs with great calling uncertainty (simulation results not shown). A second option is to develop a statistic based on a likelihood function conditioning on the phenotype of the offsprings and the intensity data of the parents, following the argument of Schaid (1996), and Laird, Horvath, and Xu (2000). However, the score statistic based on this strategy is again biased under H_0 (simulations results not shown) due to the fact that the transmitted CNVs have higher posterior probabilities than untransmitted CNVs.

Hence, we discarded family information when inferring CNV status and derived a FBAT-type statistic (Laird, Horvath, and Xu (2000); Laird and Lange (2006)) $V = \sum_{i=1}^{N}\sum_{j=1}^{n_i} v_{ijt}$ with

$$v_{ijt} = E[S(C_{ijt})|\mathbf{O}_{ij}] - E[S(C_{ijt})|\mathbf{O}_{im}, \mathbf{O}_{if}]. \tag{2.8}$$

Here $\mathbf{O}_{im}$ and $\mathbf{O}_{if}$ are the intensities of parents. The variance of V is empirically estimated by $Var(V) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} v^2_{ijt}$. The normalized statistic $V/Var(V)^{1/2} \sim N(0, 1)$ under H_0.

Table 1. Parameters characterizing the distribution of LRRs and BAFs.

	Mean of LRRs $(\mu_0,\mu_1,\mu_2,\mu_3,\mu_4)$	SD of LRRs $(\sigma_0,\sigma_1,\sigma_2,\sigma_3,\sigma_4)$	SD of BAFs (η_1,η_2)
Ω^a	(-3,-0.45,0,0.30,0.5)	(1,0.26,0.16,0.19,0.22)	(0.02,0.05)
Ω^b	(-3,-0.36,0,0.25,0.5)	(1,0.26,0.16,0.19,0.22)	(0.02,0.05)
Ω^c	(-3,-0.54,0,0.36,0.5)	(1,0.26,0.16,0.19,0.22)	(0.02,0.05)

[a] estimated from the long, experimentally validated CNVs.

[b] a set of parameters with weaker signal-to-noise ratio.

[c] a set of parameters with stronger signal-to-noise ratio.

To compute the conditional expectations in (2.8), we need to compute the conditional distributions

$$P\{C_{ijt}|\mathbf{O}_{ij}\} \propto P\{\mathbf{O}_{ij}|C_{ijt}\}P\{C_{ijt}\},$$

$$P\{C_{ijt}|\mathbf{O}_{im},\mathbf{O}_{if}\} = \sum_{C_{ift},C_{imt}} P\{C_{ijt}|C_{ift},C_{imt}\}P\{C_{ift}|\mathbf{O}_{if}\}P\{C_{imt}|\mathbf{O}_{im}\},$$

where C_{ift}, C_{imt} are the parental CNV statuses. Again, $P\{\mathbf{O}_{ij}|C_{ijt}\}$ is computed using HMM. The population CNV frequencies $P\{C_{ijt}\}$ are computed based on the parents using the EM algorithm under H_0. Also, $P\{C_{ijt}|C_{ift},C_{imt}\}$ is determined by the Mendel's rule.

We will show in simulations that this test has correct type-I error rates and reasonably high relative efficiency in realistic parameters settings. One can extend this approach to incomplete nuclear families, quantitative traits or pedigrees with phenotyped parents (Laird and Lange (2006)).

3. Results

3.1. Simulations for case-control studies

We performed simulations using autosomal SNPs that were present on both the Illumina HumanHap550 SNP array and the Hapmap II SNP list. For computational efficiency, we randomly chose segments covering 30 SNPs and put simulated CNVs in the middle of the segments. Because the whole chromosome can be naturally partitioned into independent segments with short CNVs, chromosome-wide or genome-wide characteristics of the test statistics can be easily obtained from segment-based simulations. For each subject, we first simulated CNV status for each SNP and then simulated LRR and BAF using parameters $\mathbf{\Omega}$ (Table 1), that were estimated based on the intensity data from experimentally validated long CNVs. Each simulation had 1,000 cases and 1,000 controls. CNV frequencies were 0.025 for either deletions or duplications in the controls.

First, we performed 100,000 simulations to evaluate the probe-wise type-I error rates. Because we used a prespecified set of parameters $\mathbf{\Omega}$ when testing

Table 2. Probe-wise type-I error rates of the score statistic for case-control studies, estimated based 100,000 simulations, α=0.001. Each simulation has 1,000 cases and 1,000 controls. Population frequency of CNV=0.025. The score statistics use parameter Ω as default. Intensity data are simulated using Ω, Ω_1, and Ω_2 to investigate whether type-I error rates are robust to Ω. CN1 represents deletions with one copy. CN3 represents duplications with three copies.

	Ω	Ω	Ω_1	Ω_1	Ω_2	Ω_2
#probes	CN1	CN3	CN1	CN3	CN1	CN3
4	0.0008	0.0006	0.0008	0.0092	0.0008	0.0006
5	0.0008	0.0008	0.0008	0.0096	0.0008	0.0008
6	0.0008	0.0009	0.0009	0.0097	0.0008	0.0009
8	0.0009	0.0009	0.0008	0.0093	0.0009	0.0008

associations, we investigated whether the type-I error rates were inflated if $\mathbf{\Omega}$ was misspecified. To achieve this goal, we simulated more (less) noisy intensity data using $\mathbf{\Omega}_1(\mathbf{\Omega}_2)$, and ran our algorithm using $\mathbf{\Omega}$ as default. See Table 1 for the specification of $\mathbf{\Omega}, \mathbf{\Omega}_1$, and $\mathbf{\Omega}_2$. The results in (Table 2) suggest that the statistic has the expected type-I error rates when parameters are misspecified.

We then investigated the relative efficiency (RE) of the statistic. Let Z_0 be the statistic assuming known CNV status. Here $RE = (E(Z_t)/E(Z_0))^2$ evaluated at H_1, where Z_t is the score statistic at probe t. Thus, $1 - RE$ measures the loss of the effective sample size due to CNV calling uncertainty. Figure 3A and 3B are the results for duplications and deletions, respectively. RE is higher for longer CNVs. Surprisingly, Z_t achieves a very high RE even for CNVs covering as few as four probes, suggesting that power loss due to CNV calling uncertainty is minimal.

Next, we performed simulations to approximate the region-wide p-value $P_0(\max_t |Z_t| > 3.29)$. We found that $cor(Z_t, Z_{t+1}) \approx 0.99$ in CNV regions, suggesting that the Bonferroni correction was conservative. In fact, even for CNVs with eight probes, the region-wide p-value was only slightly larger than the probe-wise p-value (Table 3).

Finally, we compared the performance of our approach with CNVTools (Barnes et al. (2008)) , the most widely-used algorithm for detecting CNV associations in case-control GWAS. To use CNVTools, one has to specify a region that may harbor recurrent CNVs. Once a genomic region is specified, CNVTools tests the CNV association by performing principal component analysis (PCA) on the LRR of all probes in the region fitting a latent variable model to the PCA scores under H_0 and H_1, and deriving a likelihood ratio statistic. We performed simulations under two realistic situations: subjects had either CNVs with identical boundaries (Figure 1A) or overlapping CNVs but with

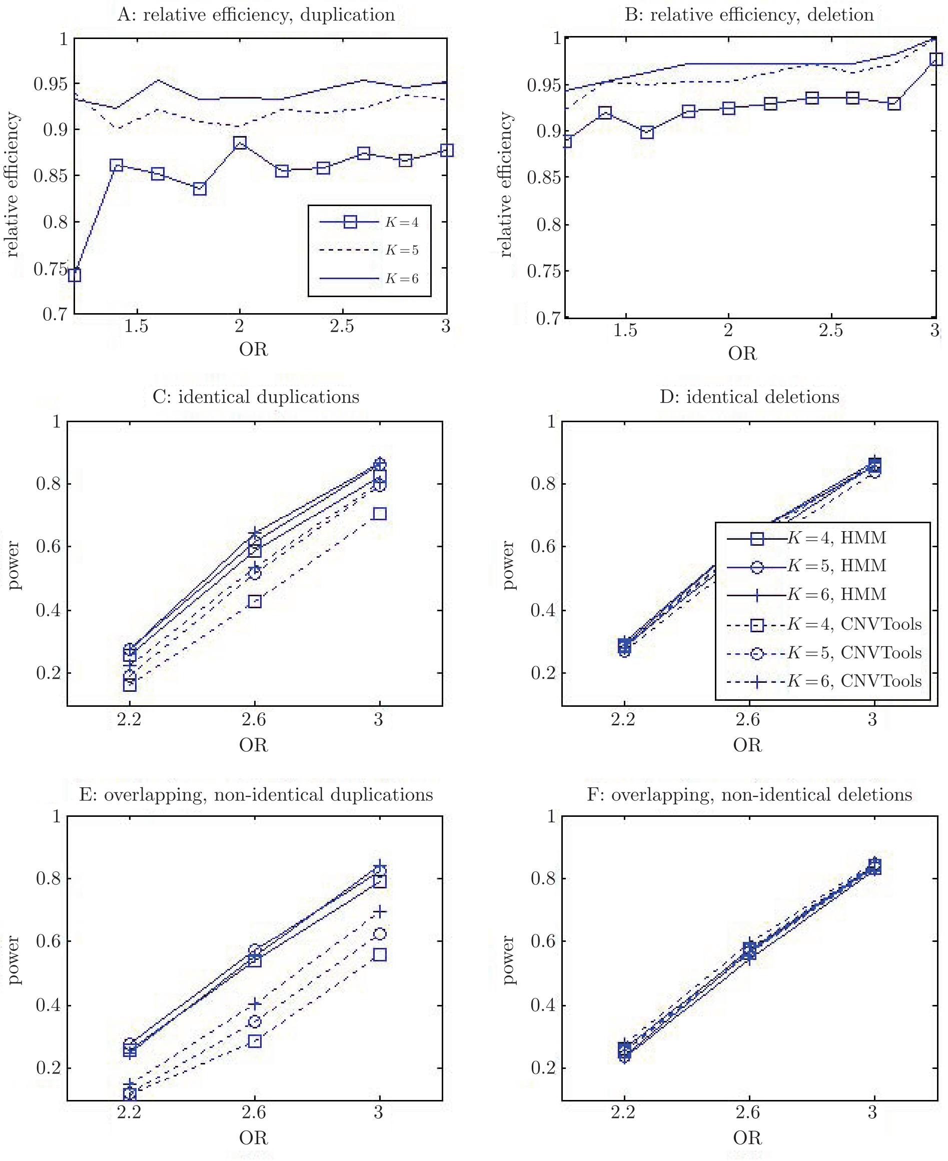

Figure 3. Shown in panels A and B are the relative efficiency of the score test in case-control studies for CNVs covering K probes. CNV frequency is 2.5% in controls. Figures C and D compare the power of our method based on HMM and CNVTools for detecting associations of CNVs with identical boundaries (see Figure 1A). Results are based on 1,000 simulations, $\alpha = 0.0001$. Figures E and F are the results of power comparison for overlapping CNVs with different boundaries (see Figure 1B).

Table 3. Probe-wise type-I error rate $P_0\{|Z_t| > 3.29\}$ and region-wise type-I error rate $P_0\{\max_t |Z_t| > 3.29\}$ estimated based 100,000 simulations of datasets with 1,000 cases and 1,000 controls. Population frequency of CNV=0.025.

	$P_0\{\lvert Z_t\rvert > 3.29\}$		$P_0\{\max_t \lvert Z_t\rvert > 3.29\}$	
#probes	CN1	CN3	CN1	CN3
4	0.0008	0.0006	0.0010	0.0009
5	0.0008	0.0008	0.0012	0.0012
6	0.0008	0.0009	0.0013	0.0014
8	0.0009	0.0009	0.0015	0.0014

variable boundaries (Figure 1B). For our score statistics, an association was detected if $\max_t Z_t > U_0$, where U_0 was chosen to have region-wide $p = 0.0001$ ($P_0(\max_t |Z_t| > U_0) = 0.0001$) based on 100,000 simulations. CNVTools detected an association if its $p < 0.0001$. We estimated the power using 1,000 simulations for each setting. We found that CNVTools did not converge in about 15%-20% of simulations because the PCA scores were not sufficiently separated to fit a latent variable model. Thus, the power with CNVTools was based on the simulations successfully analyzed by CNVTools. The simulation results are in Figure 3. Both methods have similar performance for detecting deletion associations. However, our method is more powerful when detecting the associations with short duplications. This is expected because our method used BAF information, which is particularly informative for inferring duplications, while that CNVTools cannot.

3.2. Simulations for family-based studies

For each nuclear family, we simulated the CNV status for four parental haplotypes and then randomly transmitted a haplotype to each offspring. Given CNV status, we simulated LRRs and BAFs according to $\mathbf{\Omega}$. We performed 10,000 simulations to evaluate the type-I error rates. For each simulation, there were 1,000 affected offspring in total. Our family-based statistic controlled the type-I error rates in all scenarios and was robust to misspecification of $\mathbf{\Omega}$. Results are reported in Supplementary Table 1. Figures 4A−D show the RE for CNVs covering different numbers of probes in families with one or two offspring. The RE is sufficiently high for detecting deletions although it is slightly lower than that in case-control studies. However, the RE is low for detecting short duplications.

3.3. Application to autism genome-wide association data

We applied our family-based statistic to a GWAS of autism spectrum disorders (ASDs) performed by the Autism Genome Project (AGP) Consortium. The

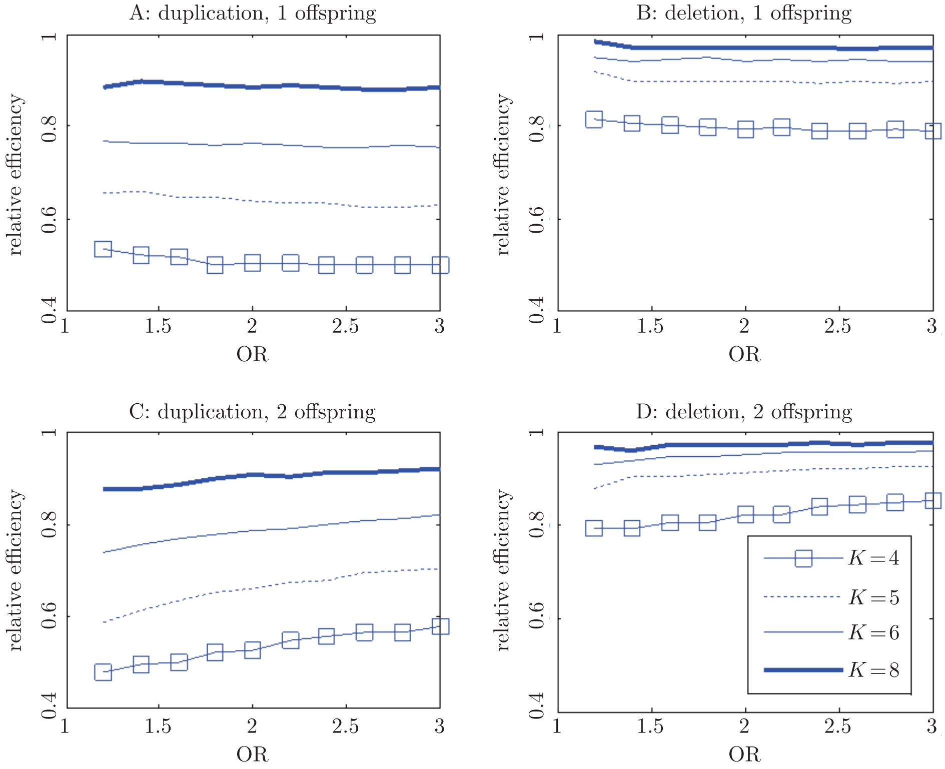

Figure 4. Shown in figures A and B are the relative efficiencies for studies based on trios. Each study has 1,000 trios. Population CNV frequency is 2.5%. Figure C and D show the relative efficiencies for nuclear families with two affected offsprings.

study has about 1,400 trios, each of which has an affected offspring. All subjects were genotyped using Illumina 1M genotyping arrays with 1,068,909 probes. We obtained the intensity data from dbgap (`http://www.ncbi.nlm.nih.gov`) and extracted LRRs and BAFs using GenomeStudio. Subjects were removed if their genome-wide LRR standard deviations were larger than 0.30. The LRRs were adjusted for GC content. For each subject, the LRRs were normalized to have median zero and standard deviation 0.16. After quality control, we analyzed 831 complete trios.

We implemented our algorithms using C++. We first partitioned the whole genome into short segments, each of which had 2,000 probes. After genome partition and data normalization, analyzing the association for all segments took less than two hours using a single core in a Linux server. We only analyzed probes with estimated frequencies of CNVs (including both deletions and duplication) between 0.01 and 0.05. In total, 105,000 probes were tested for associa-

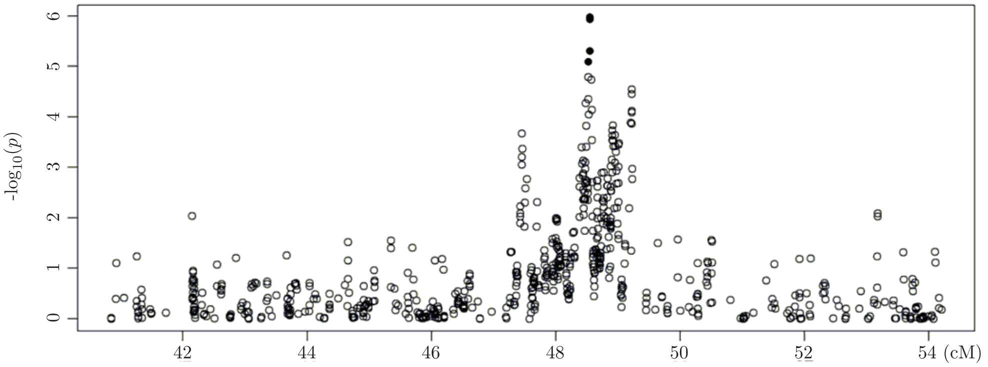

Figure 5. Shown are the p-values for probes on chromosome 17q22. Within this region, the estimated frequencies of deletions range from 0.01 to 0.05.

tion. Singled-sided p-values were calculated. The Manhattan plot is presented in Supplementary Figure 1. One region located between 48cM and 49 cM on chromosome 17q22 showed strong associations with multiple consecutive probes (Figure 5). The strongest p-value in this region was 1.06×10^{-6}, close to the threshold $5 \times 10^{-7} = 0.05/105000$ based on the Bonferroni correction. Because our statistics are highly correlated in the same CNV regions, the Bonferroni correction is very conservative. Lab validation and further replication studies in independent samples are recommended to investigate this region.

4. Discussion

Detecting CNV associations from GWAS is statistically challenging but scientifically important. We have developed methods for both case-control and family-based studies. The features of our methods include integrating both LRRs and BAFs, efficiently integrating information from nearby probes, ery fast computation and without the need to specify CNV regions. Simulation studies demonstrated that our approach had a better power to detect association with short duplications compared to CNVTools in case-control studies. Our methods may prove useful for analyzing existing GWAS of complex diseases.

Perhaps surprising, the power loss in case-control studies due to calling uncertainty was very small even for short CNVs. However, the power loss was larger in family-based studies, and the loss could be substantial for short duplications because of greater calling uncertainty. Hence, it would be valuable to develop more powerful methods for testing CNVs in family-based studies by appropriately integrating family information while maintaining correct type-I error rates. Another limitation is that we have assumed homogenous data quality for the purpose of theoretical investigation. With data, one has to correct artifacts due to GC content, batch effect, and potential differential genotyping bias.

For case-controls studies, we used the retrospective likelihood framework. Compared with the prospective likelihood methods, this method is expected to be more powerful when assuming Hardy-Weinberg equilibrium (HWE)(see e.g., Epstein and Satten (2003) Chatterjee et al. (2009)). However, type-I error rates may be inflated if HWE fails. It would be interesting to investigate the power gain and the robustness of type-I error rates for CNV association testing assuming HWE.

Acknowledgement

The authors are supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute at the National Institutes of Health. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

Appendix

Appendix A

For probe t, given C_{it}, we assume that the LRR follows a mixture distribution

$$P\{X_{it}|C_{it}=c\} \sim (1-\pi)N(\mu_c,\sigma_c^2)+\pi U[-R_0,R_0], \tag{A.1}$$

where π represents the level of the background noise. Let genotype G_{it} be the number of the B allele with value ranging from 0 to C_{it}. The conditional distribution $P(B_{it}=b|C_{it}=c,G_{it}=g)$ is given in Figure 2. Furthermore, we assume that the B allele has a frequency f_t in the population. Hence, assuming HWE for the SNP, we have

$$P(B_{it}=b|C_{it}=c)=\sum_{g=0}^{c}\frac{c!}{(g!(c-g)!)}f_t^g(1-f_t)^{c-g}P(B_{it}=b|C_{it}=c,G_{it}=g). \tag{A.2}$$

Appendix B

Based on (2.6), we have

$$\frac{\partial l}{\partial \beta}\Big|_{\beta=0}=\sum_{i=1}^{N}\Big(\frac{\sum_{c=0}^{4}p_{tc}T_{itc}Y_iS(c)}{\sum_{c=0}^{4}p_{tc}T_{itc}}-\sum_{c=0}^{4}p_{tc}Y_iS(c)\Big). \tag{B.1}$$

The nuisance parameters $P_t = (p_{t0}, \ldots, p_{t4})$ are estimated using the EM algorithm under H_0. When $\beta = 0$, the likelihood (2.6) reduces to $\sum_{i=1}^{N} \log \sum_{c=0}^{4} p_{tc} T_{itc}$. Let $\hat{p}_{tc}^{j}$ be the estimate at step j. Then, the updating rule is

$$\hat{p}_{tc}^{j+1} = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{p}_{tc}^{j} T_{itc}}{\sum_{c=0}^{4} \hat{p}_{tc}^{j} T_{itc}}.$$

Because $\sum_{c=0}^{4} p_{tc} = 1$, we reparameterize $(p_{t0}, \ldots, p_{t4})$ as $p_{t0} = 1/(1 + \sum_{j=0}^{4} e^{u_j})$ and $p_{tk} = e^{u_k}/(1 + \sum_{j=0}^{4} e^{u_j})$ for $k = 1, \ldots, 4$. Here, $u_1, \ldots, u_4 \in (-\infty, \infty)$. Given $\hat{P}_t$, u_j is determined. We then compute the empirical Fisher's information matrix $H(\beta, u_1, \ldots, u_4)$ evaluated at $\beta = 0$ and $\hat{P}_t$. The score statistic is then

$$Z_t = \frac{\partial l}{\partial \beta}\Big|_{\beta=0} / [H^{-1}]_{11}^{\frac{1}{2}}.$$

References

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D. and Hurles, M. E. (2008). A robust statistical method for case-control association testing with copy number variation. *Nature Genet. 40*, 1245-1252.

Chatterjee, N., Chen, Y. H., Luo, S., and Carroll, R. J. (2009). Analysis of case-control association studies: SNPs, imputation and haplotypes. *Statist. Sci.* **24**, 489-502.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Basset, A. S., Seller, A., Holmes, C. C. and Ragoussis, J. (2007). QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013-2025.

Eleftherohorinou, H., Andersson-Assarsson, J. C, Walters, R. G., Moustafa, J. S, Coin, L., Jacobson, P., Carlsson, L. M. S, Blakemore, A., Froguel, P., Walley, A. J., and Falchi, M. (2011). famCNV: copy number variant association for quantitative traits in families. *Bioinformatics* **27**, 873-875.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *Amer. J. Hum. Genet.* **73**, 1316-1329.

Ionita-Laza, I., Perry, G. H., Raby, B. A., Klanderman, B., Lee, C., Laird, N. M., Weiss, S. T., and Lange, C. (2008). On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epi.* **32**, 273-284.

Jeng, X. J., Cai, T. and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105**, 1156-1166.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J. and Altshuler, D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genet.* **40**, 1253-1260.

Laird, N. M., Horvath, S. and Xu, X. (2000). Implementing a unified approach to family based tests of association. *Genet. Epi.* **19**(Suppl 1), S36-S42.

Laird, N. M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genet.* **7**, 385-394.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572.

Schaid, D. J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genet. Epi.* **13**, 423-449.

Shi, J. X. and Li, P. (2012). An integrative segmentation method for detecting germline copy number variations in SNP arrays. *Genet. Epi.* **36**, 373-383.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506-516.

Wang, H., Veldink, J. H., Blauw, H., Van Den Berg L. H., Ophoff, R. A., and Sabatti, C. (2009). Markov models for inferring copy number variations from genotype data on Illumina platforms. *Hum. Hered.* **68**, 1-22.

Wang, K., Chen, Z., Tadesse, M.G., Glessner, J., Grant, S.F., Hakonarson, H., Bucan, M., and Li, M. (2008). Modeling genetic inheritance of copy number variations. *Nucleic Acids Res.* **36**, e138.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665-1674.

Zhang, Z. Y., Lange, K., Ophoff, R., and Sabatti, C. (2010). Reconstructing DNA copy number by penalized estimation and imputation. *Ann. Appl. Stat.* **4**, 1749-1773.

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, RM8040, Rockville, MD20852, USA.

E-mail: Jianxin.Shi@nih.gov

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, RM8040, Rockville, MD20852, USA.

E-mail: lip4@mail.nih.gov

(Received March 2012; accepted April 2013)