Published online 2 October 2008

Nucleic Acids Research, 2008, Vol. 36, No. 21 ***e138***
doi:10.1093/nar/gkn641

Modeling genetic inheritance of copy number variations

Kai Wang[1,2,*], Zhen Chen[3], Mahlet G. Tadesse[4], Joseph Glessner[2], Struan F. A. Grant[2], Hakon Hakonarson[2], Maja Bucan[1] and Mingyao Li[3]

[1]Department of Genetics, University of Pennsylvania, [2]Center for Applied Genomics and Division of Human Genetics, The Children's Hospital of Philadelphia, [3]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104 and [4]Department of Mathematics, Georgetown University, Washington, DC 20057, USA

Received July 25, 2008; Revised September 5, 2008; Accepted September 16, 2008

ABSTRACT

Copy number variations (CNVs) are being used as genetic markers or functional candidates in gene-mapping studies. However, unlike single nucleotide polymorphism or microsatellite genotyping techniques, most CNV detection methods are limited to detecting total copy numbers, rather than copy number in each of the two homologous chromosomes. To address this issue, we developed a statistical framework for intensity-based CNV detection platforms using family data. Our algorithm identifies CNVs for a family simultaneously, thus avoiding the generation of calls with Mendelian inconsistency while maintaining the ability to detect *de novo* CNVs. Applications to simulated data and real data indicate that our method significantly improves both call rates and accuracy of boundary inference, compared to existing approaches. We further illustrate the use of Mendelian inheritance to infer SNP allele compositions in each of the two homologous chromosomes in CNV regions using real data. Finally, we applied our method to a set of families genotyped using both the Illumina HumanHap550 and Affymetrix genome-wide 5.0 arrays to demonstrate its performance on both inherited and *de novo* CNVs. In conclusion, our method produces accurate CNV calls, gives probabilistic estimates of CNV transmission and builds a solid foundation for the development of linkage and association tests utilizing CNVs.

INTRODUCTION

A central strategy in the genetic study of human diseases is to identify genomic DNA variations related to clinical phenotypes. Human genomic variation exists in many forms, including single nucleotide polymorphisms (SNPs), simple repeat elements, microsatellites and structural variations such as copy number variations (CNVs) (1). A CNV is defined as a chromosomal segment, at least 1 kb in length, whose copy number varies in comparison with a reference genome (2). A significant fraction of CNVs are likely to have functional consequences, due to gene dosage alteration, disruption of genes, positional effects or the uncovering of deleterious alleles (3,4). Thus, comprehensive identification and cataloging of CNVs will greatly benefit the genetic and functional analysis of human genome variation.

Multiple techniques have been developed to detect deletions or duplications in the human genome and other mammalian genomes (5), and many of them depend on analyzing patterns of signal intensities across the genome. Traditionally, large chromosome rearrangements have been detected by array-comparative genomic hybridization (CGH) techniques that analyze the fluorescence signal intensities of clones (6–9). Another comparable platform for CNV detection is whole genome oligonucleotide arrays. Since design of the arrays does not depend on SNPs, such technology can achieve complete genome coverage with higher precision for boundary inference of CNVs. Due to recent increased popularity of genome-wide association studies, high-density SNP genotyping arrays have been commonly used for CNV detection and analysis. With such arrays, signal intensity is measured for each allele of a given SNP, and analysis of signal intensities across all SNPs in the genome is used to infer CNVs (10,11). More recently, to improve the coverage of SNP arrays for CNV analysis, manufacturers of SNP genotyping arrays, such as Affymetrix and Illumina, have incorporated nonpolymorphic (NP) markers into their SNP genotyping arrays, especially in known CNV regions.

*To whom correspondence should be addressed. Tel: 267 426 2378; Fax: 267 426 0363; Email: wangk@chop.edu
Correspondence may also be addressed to Mingyao Li. Tel: 215 746 3916; Fax: 215 573 4865; Email: mingyao@mail.med.upenn.edu

Although traditionally 'losses' and 'gains' have been used to describe the major classes of CNVs, CNVs in a diploid genome are indeed chromosome-specific events. That is, CNVs can exist in any of the two homologous chromosomes, such as being deleted on one chromosome but duplicated on the other. Knowing chromosome-specific copy number is important to the development of linkage and association tests for CNVs. However, those commonly used CNV detection techniques mentioned above all depend on signal intensity measures, and are therefore unable to infer copy number in each homologous chromosome. The efficient utilization of family information can potentially help circumvent this issue. Furthermore, since most CNVs follow Mendelian inheritance (8), the use of family information can improve the sensitivity and specificity of CNV detection (12). In fact, family-based designs are now commonly used in genome-wide association studies, making it highly desirable to develop methods to infer chromosome-specific copy numbers. For example, in a recent CNV study on autism spectrum disorders, 751 families have been genotyped by the Affymetrix genome-wide 5.0 Human SNP arrays (13); in our ongoing study, 943 autism families were genotyped using the Illumina HumanHap550 SNP arrays (14). Other family-based genome-wide association studies include the Framingham heart study (15), a multiple sclerosis study (16) and type I diabetes studies (17,18).

To use family information in analysis of CNVs, Kosta *et al.* (19) developed an approach to infer chromosome-specific copy numbers for nuclear families after the total copy numbers are obtained from quantitative PCR. In our previous CNV analysis (12), we incorporated family information in a two-step procedure in which family members were first used independently to generate CNV calls, and then combined together to post-validate calls obtained in the first step by incorporating family relationships. Although this approach has been shown to significantly increase the sensitivity and specificity of CNV detection, the family information is not optimally used. Moreover, if the CNV boundary is inferred incorrectly in the first step, it cannot be corrected in the second step. More recently, Marioni *et al.* (20) discussed similar issues extensively for array CGH data, and proposed that copy numbers can be inferred on each chromosome, using HapMap family data as examples.

Efficient utilization of family information in CNV detection requires incorporation of the family relationships when modeling the joint probability distribution of signal intensities for family members. Similar to traditional multipoint linkage analysis with families, such a modeling procedure requires consideration of two levels of dependency—the dependency of signal intensities both between adjacent markers for each family member and at the same marker between family members. The first level of dependency can be modeled by a hidden Markov chain, in which the degree of dependency is determined by transition probabilities of the hidden copy number states, whereas the second level of dependency is determined by Mendelian inheritance. However, unlike the analysis of SNPs or microsatellites, family-based CNV studies are limited by the technical platforms, which can only give intensity estimates of the total copy number of a diploid genome. The analysis of CNVs in families is further complicated by the occurrence of *de novo* events, which occur as germline, somatic or cell line-induced chromosome aberrations in offspring that were not inherited from either parent.

To address these complications, we describe a unified statistical framework developed to jointly model the signal intensities for a parents–offspring trio. We demonstrate that our model is computationally feasible and can be used to analyze trios in a more efficient manner than existing methods, which do not consider family relationships or use family relationships separately (12). By computer simulations and analysis of experimentally validated CNVs on real data, we demonstrate its superior performance in increasing call rates and in identifying the exact boundaries of CNVs. In addition, by analyzing a set of families genotyped using both the Illumina and Affymetrix SNP arrays, we further show the applicability of our method on different technical platforms and in detecting both inherited and *de novo* CNVs. Although CNV detection only concerns the total copy number, our model gives probabilistic estimates of chromosome-specific copy numbers, which can be used for the future development of linkage and association tests that require chromosome-specific copy number information.

METHODS

Overview of the hidden Markov model framework

The hidden Markov model (HMM) is a statistical technique that models data generated from an underlying Markov process. The HMM assumes that the distribution of an observed data point depends on an unobserved (hidden) state, where the elements of the hidden states follow a Markov process. Since CNV detection typically involves aggregating information from multiple consecutive SNPs, HMM provides a natural framework for modeling dependence structures between copy numbers at nearby markers. Figure 1 shows a schematic representation of our proposed model for joint CNV distribution in a parents–offspring trio. The model consists of a chain for the copy number states of the father, a chain for the copy number states of the mother, a chain for the *de novo* event status of the offspring and these three chains are independent of each other. Although the offspring copy number at each marker is dependent on the copy number at the previous marker, it is also determined by six other elements: the copy number states of parents and the *de novo* status at both the current marker and the previous marker (dashed lines in Figure 1). Below we will describe how to explicitly model the joint CNV distribution of a parents–offspring trio through likelihood calculation of the signal intensities.

Signal intensities for Illumina SNP arrays

To illustrate our method, we focus on data generated from the Illumina SNP arrays and the Affymetrix arrays with both SNP and NP markers. Illumina SNP arrays produce two measures on signal intensities at each SNP—log *R* ratio (LRR) and B allele frequency (BAF), and these

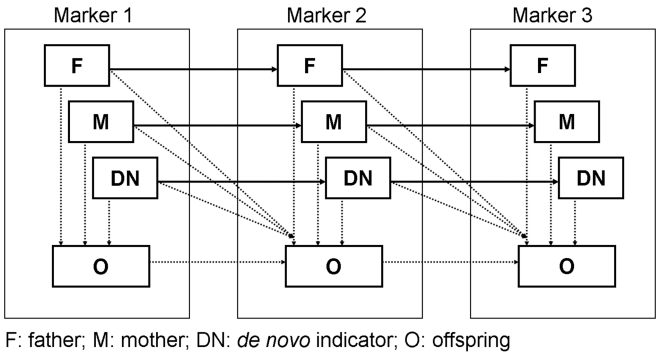

Figure 1. Illustration of the hidden Markov model framework for modeling genetic inheritance of CNVs in parents–offspring trios. F, M and O represent copy number states of the father, mother and offspring, respectively, and DN is an indicator variable for *de novo* event status of the offspring.

two measures were originally proposed by Illumina for copy number inference (10). To obtain LRR and BAF, for each SNP, the raw signal intensities are subject to a normalization procedure, which produces the X- and Y-values, representing normalized signal intensity for the A and B alleles, respectively. Two measures, $R = X + Y$, and $\theta = \arctan(Y/X)/(\pi/2)$, are then calculated for each SNP. As a normalized measure of total signal intensity, LRR is then calculated as $\log_2(R_{\text{observed}}/R_{\text{expected}})$, where R_{expected} is computed from linear interpolation of the canonical genotype clusters. The BAF is a normalized measure of the relative signal intensity ratio of the B and A alleles. Let θ_g, $g \in \{\text{AA, AB, BB}\}$ denote the mean θ value for genotype cluster g obtained from a set of reference samples. The corresponding BAFs are defined as 0.0, 0.5 and 1.0, respectively. Then, for a subject with θ_{subject}, the BAF is defined through linear interpolation among the three clusters:

$$\text{BAF} = \begin{cases} 0, & \text{if } \theta_{\text{subject}} < \theta_{\text{AA}} \\ 0.5(\theta_{\text{subject}} - \theta_{\text{AA}})/(\theta_{\text{AB}} - \theta_{\text{AA}}), & \text{if } \theta_{\text{AA}} \le \theta_{\text{subject}} < \theta_{\text{AB}} \\ 0.5 + 0.5(\theta_{\text{subject}} - \theta_{\text{AB}})/(\theta_{\text{BB}} - \theta_{\text{AB}}), & \text{if } \theta_{\text{AB}} \le \theta_{\text{subject}} < \theta_{\text{BB}} \\ 1, & \text{if } \theta_{\text{subject}} \ge \theta_{\text{BB}} \end{cases} \quad \mathbf{1}$$

Signal intensities for Affymetrix SNP arrays

The Affymetrix genome-wide 5.0 and 6.0 SNP arrays contain approximately equal numbers of SNP markers and NP markers to improve the genome coverage. We followed a similar procedure as used by the Illumina platform to derive the LRR and BAF values for SNP markers, and the LRR values for NP markers. We used the Affymetrix Power Tools (http://www.affymetrix.com/support/developer/powertools/changelog/index.html) to perform data normalization, signal extraction and genotype calling from raw CEL files generated in genotyping experiments. For each SNP marker, we then relied on the allele-specific signal intensities for the AA, AB and BB genotypes on all genotyped samples to construct three canonical genotype clusters. Since each NP marker has only one reference cluster, we set the center value of the cluster as the median of the signal intensities for all genotyped samples. Once the canonical genotype clusters are constructed, we can then transform the signal intensity values for each SNP into R, LRR, θ and BAF values. The method described below uses both LRR and BAF values, but for NP markers, the BAF information is ignored in the likelihood calculation.

Likelihood of signal intensities for a parents–offspring trio

Assume a parents–offspring trio is genotyped at T consecutive SNPs. For SNP j ($1 \le j \le T$), let $r_j = (r_{j,f}, r_{j,m}, r_{j,o})$ denote the triplet of LRRs of the father, the mother and the offspring, $b_j = (b_{j,f}, b_{j,m}, b_{j,o})$ denote the corresponding BAFs, $z_j = (z_{j,f}, z_{j,m}, z_{j,o})$ denote the underlying hidden copy number states, and DN_j (1: *de novo* event; 0: inherited from parents) denote the *de novo* event status of the offspring. The observed signal intensities for the trio can be represented by $r = (r_1,\ldots,r_T)$, $b = (b_1,\ldots,b_T)$, and the hidden copy number states can be represented by $z = (z_1,\ldots,z_T)$. Let λ denote all parameters in the HMM (including means and standard deviations in the emission probabilities of signal intensities, initial probabilities of copy number states and transition probabilities). The likelihood of the signal intensities for the trio is

$$\begin{aligned} &P(r_1,\ldots,r_T,b_1,\ldots,b_T|\lambda) \\ &= \sum_{z_1}\cdots\sum_{z_T}\sum_{\text{DN}_1}\cdots\sum_{\text{DN}_T}\Big\{P(r_1,\ldots,r_T|z_1,\ldots,z_T,\lambda) \\ &\quad \times P(b_1,\ldots,b_T|z_1,\ldots,z_T,\lambda) \\ &\quad \times P(z_1,\ldots,z_T|\text{DN}_1,\ldots,\text{DN}_T,\lambda) \\ &\quad \times P(\text{DN}_1,\ldots,\text{DN}_T|\lambda)\Big\} \\ &= \sum_{z_1}\cdots\sum_{z_T}\sum_{\text{DN}_1}\cdots\sum_{\text{DN}_T}\Big\{P(r_1|z_1,\lambda)P(b_1|z_1,\lambda) \\ &\quad P(z_1|\text{DN}_1,\lambda)P(\text{DN}_1|\lambda) \\ &\quad \times \prod_{j=2}^{T} P(r_j|z_j,\lambda)P(b_j|z_j,\lambda)P(z_j|z_{j-1},\text{DN}_j,\text{DN}_{j-1},\lambda) \\ &\quad P(\text{DN}_j|\text{DN}_{j-1},\lambda)\Big\}. \end{aligned} \quad \mathbf{2}$$

Figure 1 provides a schematic representation of the dependence structure specified in Equation (2). This equation requires a few simplifying but reasonable assumptions, including the conditional independence of LRR and BAF values at each marker (supported by empirical data), the conditional independence of LRR/BAF values between adjacent markers, as well as the conditional independence of BAF values and the *de novo* event status at each marker. For the starting SNP, its contribution to the likelihood is the product of the emission probability of the signal intensities, the initial probability of copy

number states, and the initial probability of the *de novo* event status. Based on empirical data from HapMap, we estimate that $\varepsilon = P(\mathrm{DN}_1 = 1|\lambda) = 1.5e-6$. For the remaining SNPs ($2 \le j \le T$), the contribution of each SNP to the likelihood is the product of the emission probability, the transition probability of copy number states and the transition probability of the *de novo* event status. The challenge of the HMM lies in the inference of the hidden copy number states of each marker and the *de novo* event status of the offspring, given the observed signal intensities. Below, we describe elements needed in the HMM calculation. We note that the calculation in Equation (2) can be easily extended to nuclear families with multiple offspring, in which each additional offspring requires a variable specifying copy number state and a variable indicating *de novo* status for each marker.

Hidden copy number states. We adopt a five-state definition of hidden copy number states (Table 1) to reflect possible copy number changes, including double-copy deletion (zero copies), single-copy deletion (one copy), normal state (two copies), single-copy duplication (three copies) and double or more copy duplication (four or more copies). A copy number of more than four is usually indistinguishable from four copies in patterns of signal intensity, so we combine this rare scenario with four copies.

Emission probabilities of signal intensity. Given the copy number states of the father, the mother and the offspring, their signal intensities are independent, thus for marker j, $P(r_j|z_j,\lambda) = \prod_{k\in\{f,m,o\}} P(r_{j,k}|z_{j,k},\lambda)$. We propose to model the emission probability of the LRRs as a normal distribution based on empirical observations, $P(r_{j,k}|z_{j,k},\lambda) = \phi(r_{j,k};\mu_{z_{j,k}},\sigma_{z_{j,k}})$, where $\phi(r_{j,k};\mu_{z_{j,k}},\sigma_{z_{j,k}})$ is the normal density function with unknown mean $\mu_{z_{j,k}}$ and SD $\sigma_{z_{j,k}}$.

The emission probability of BAF is slightly more complicated than the LRR. For the zero-copy state, we used a mixture of normal with mean 0.5 and unknown SD, and a point mass at 0 or 1 to model the distribution of BAF. For each state other than the zero-copy state, there are multiple possible genotypes with distinct patterns of BAF. Let $C(z_{j,k})$ denote the total number of genotypes (Table 1) for state $z_{j,k}$ of individual k at SNP j. For each genotype that is consistent with the copy number state, let g denote the number of copies of allele B. Let $p_{j,B}$ be the population frequency of allele B at marker j, which can be estimated from a set of reference samples such as the HapMap. Then the emission probability of BAF can be modeled as a mixture of distributions, $P(b_{j,k}|z_{j,k},\lambda) = \sum_g P(b_{j,k}|g,z_{j,k},\lambda) P(g|z_{j,k},\lambda)$, where

$$P(g|z_{j,k},\lambda) = \binom{C(z_{j,k})}{g} p_{j,B}^{g}(1-p_{j,B})^{C(z_{j,k})-g} \quad \mathbf{3}$$

is the probability of genotype g, and

$$P(b_{j,k}|g,z_{j,k},\lambda) = \begin{cases} \phi(b_{j,k};\mu_{\mathrm{BAF},z_{j,k},g},\sigma_{\mathrm{BAF},z_{j,k},g}), & \text{if } 0 < g < C(z_{j,k}) \\ I_{\{b_{j,k}=0\}}M_0 + I_{\{0<b_{j,k}<1\}} \times \phi(b_{k,j};\mu_{\mathrm{BAF},z_{j,k},g},\sigma_{\mathrm{BAF},z_{j,k},g}), & \text{if } g = 0 \\ I_{\{b_{j,k}=1\}}M_1 + I_{\{0<b_{j,k}<1\}} \times \phi(b_{k,j};\mu_{\mathrm{BAF},z_{j,k},g},\sigma_{\mathrm{BAF},z_{j,k},g}), & \text{if } g = C(z_{j,k}) \end{cases} \quad \mathbf{4}$$

The use of truncated normal distribution is due to the truncation in BAF calculation. The point mass probabilities are set as $M_0 = M_1 = 0.5$.

Initial probability of copy number states. For the first marker, the initial probability of the copy number states for the trio is

$$\begin{aligned} P(z_1|\mathrm{DN}_1,\lambda) &= P(z_{1,f}|\lambda)P(z_{1,m}|\lambda)P(z_{1,o}|z_{1,f},z_{1,m},\mathrm{DN}_1,\lambda) \\ &= \pi_{z_{1,f}}\pi_{z_{1,m}}P(z_{1,o}|z_{1,f},z_{1,m},\mathrm{DN}_1,\lambda) \end{aligned} \quad \mathbf{5}$$

where the first two terms are the initial probabilities of copy number states for the father and mother, respectively, and the third term is the conditional probability of copy number state of the offspring given the parental copy number states and *de novo* event status of the offspring. If $\mathrm{DN}_1 = 1$, then a *de novo* event occurs. Assuming the offspring is equally likely to take one of the five copy number states, then $P(z_{1,o}|z_{1,f},z_{1,m},\mathrm{DN}_1 = 1,\lambda) = 1/5$. We note that this is a simplified assumption for computational convenience, since in reality the probability of some *de novo* events (such as when duplicating two additional copies) requires more dramatic changes in genomic contents than others (such as when duplicating one copy). If $\mathrm{DN}_1 = 0$, then the offspring's copy number is determined by the parental copy numbers through Mendelian inheritance.

To model Mendelian inheritance of CNVs, it is necessary to specify models for chromosome-specific copy numbers. Given the total copy number, there might be multiple compatible chromosome-specific copy number configurations. A detailed illustration is given in Suppplementary Figure 1. Due to combinatorial complexities, the likelihood function should explicitly incorporate and appropriately weigh different configurations of chromosome-specific copy numbers. To model the probability distribution of chromosome-specific copy number at a single marker, here we propose a single-parameter model with the parameter a, which specifies the probability of the less likely chromosome-specific copy number configuration (Table 2). Once the parental chromosome-specific copy numbers are known, the probability distribution of the offspring's chromosome-specific copy numbers can then

Table 1. Description of the five copy number states

Total copy number	Genotypes	Description (autosomal markers)
0	Null	Deletion of two copies
1	A, B	Deletion of one copy
2	AA, AB, BB	Normal state
3	AAA, AAB, ABB, BBB	Duplication of one copy
4	AAAA, AAAB, AABB, ABBB, BBBB	Duplication of two or more copies

be obtained following Mendel's first law (Supplementary Tables 1–3).

Transition probabilities of copy number states. For a parents–offspring trio, the transition probability of their copy number states from SNP $j-1$ to SNP j is

$$\begin{aligned} &P(z_j|z_{j-1},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda) \\ &= P(z_{j,f},z_{j,m},z_{j,o}|z_{j-1,f},z_{j-1,m},z_{j-1,o},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda) \\ &= P(z_{j,o}|z_{j,f},z_{j,m},z_{j-1,f},z_{j-1,m},z_{j-1,o},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda) \\ &\quad P(z_{j,f}|z_{j-1,f},\lambda)P(z_{j,m}|z_{j-1,m},\lambda) \end{aligned} \quad \mathbf{6}$$

The transition probability describes the probability of having a copy number state change between two adjacent SNPs. Intuitively, the copy number state is unlikely to change for SNPs that are nearby but is more likely to change for SNPs that are far apart. To appropriately model such spatial dependency, we use the following model to characterize the transition probability for the parents ($k = f$ or m),

$$P(z_{j,k}=l|z_{j-1,k}=h,\lambda)=\begin{cases} 1-\sum_{l\neq h}\gamma_{h,l}(1-e^{-d_j/D}) & \text{if } l=h \\ \gamma_{h,l}(1-e^{-d_j/D}) & \text{if } l\neq h \end{cases} \quad \mathbf{7}$$

where d_j is the physical distance between SNPs $j-1$ and j, and D is a constant that is set as 100 kb. The values of γ's are treated as unknown parameters.

Table 2. Probabilistic model specifying chromosome-specific copy numbers at a single marker, given the total copy number

Chromosome	Total copy number	Chromosome-specific copy numbers	Probability
Autosome	0	0/0	1
	1	0/1	1
	2	1/1	$1-a$
		0/2	a
	3	1/2	$1-a$
		0/3	a
	4	2/2	0.5
		1/3	0.5
Male chromosome X	0	0	1
	1	1	1
	2	2	1
	3	3	1
	4	4	1
Female chromosome X	0	0/0	1
	1	0/1	1
	2	1/1	$1-a$
		0/2	a
	3	1/2	$1-a$
		0/3	a
	4	2/2	0.5
		1/3	0.5

In the last column, a is the probability of the less likely chromosome-specific copy number configuration, and is simplified to be same for different copy numbers. Our simplified model only considers those combinatorial configurations listed in the table, while other extremely rare combinations are treated as having probability of zero in the modeling procedure.

For the offspring, we need to calculate $P(z_{j,o}|z_{j,f},z_{j,m},z_{j-1,f},z_{j-1,m},z_{j-1,o},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)$. We note that

$$\begin{aligned} &P(z_{j,o}|z_{j,f},z_{j,m},z_{j-1,f},z_{j-1,m},z_{j-1,o},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda) \\ &= \frac{P(z_{j,o},z_{j,f},z_{j,m},z_{j-1,o},z_{j-1,f},z_{j-1,m}|\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)}{P(z_{j,f},z_{j,m},z_{j-1,f},z_{j-1,m},z_{j-1,o}|\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)} \\ &= \frac{P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j-1,m},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)}{\sum_{z_{j,o}} P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j,m-1},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)} \end{aligned} \quad \mathbf{8}$$

Thus, we need to calculate $P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j-1,m},\mathrm{DN}_j,\mathrm{DN}_{j-1},\lambda)$. This probability can be classified into four categories: *de novo* at both SNPs, *de novo* at only one SNP and inherited at both SNPs.

When both SNPs are *de novo*, the parental copy numbers become irrelevant, implying that we can assume the offspring's copy number states follow a hidden Markov chain that is independent of the parents. Under this assumption,

$$\begin{aligned} &P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j-1,m},\mathrm{DN}_j=1,\mathrm{DN}_{j-1}=1,\lambda) \\ &= P(z_{j,o},z_{j-1,o}|\mathrm{DN}_j=1,\mathrm{DN}_{j-1}=1,\lambda) \\ &= P(z_{j,o}|z_{j-1,o},\mathrm{DN}_j=1,\mathrm{DN}_{j-1}=1,\lambda) \\ &\quad P(z_{j-1,o}|\mathrm{DN}_{j-1}=1,\lambda), \end{aligned} \quad \mathbf{9}$$

where $P(z_{j-1,o}|\mathrm{DN}_{j-1}=1,\lambda)=1/5$, and $P(z_{j,o}|z_{j-1,o},\mathrm{DN}_j=1,\mathrm{DN}_{j-1}=1,\lambda)$ can be calculated based on the transition probability as described earlier for the parents.

When SNP $j-1$ is *de novo* and SNP j is inherited, then a CNV breakpoint occurs between markers $j-1$ and j, thus it is reasonable to assume that the copy number states of the offspring at these two markers are independent. Under this assumption,

$$\begin{aligned} &P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j-1,m},\mathrm{DN}_j=0,\mathrm{DN}_{j-1}=1,\lambda) \\ &= P(z_{j,o}|z_{j,f},z_{j,m},\mathrm{DN}_j=0,\lambda)P(z_{j-1,o}|\mathrm{DN}_{j-1}=1,\lambda) \\ &= \frac{1}{5}P(z_{j,o}|z_{j,f},z_{j,m},\mathrm{DN}_j=0,\lambda) \end{aligned} \quad \mathbf{10}$$

where $P(z_{j,o}|z_{j,f},z_{j,m},\mathrm{DN}_j=0,\lambda)$ can be calculated based on Mendelian inheritance as specified in Supplementary Tables 1–3. The conditional probability when SNP $j-1$ is inherited and SNP j is *de novo* can be calculated in a similar fashion.

When both SNPs $j-1$ and j are inherited, the probability $P(z_{j,o},z_{j-1,o}|z_{j,f},z_{j-1,f},z_{j,m},z_{j-1,m},\mathrm{DN}_j=0,\mathrm{DN}_{j-1}=0,\lambda)$ can be calculated based on Mendelian inheritance. Given the high density of SNPs on Illumina's whole-genome SNP genotyping arrays, it is reasonable to assume that there is no recombination between SNPs $j-1$ and j, suggesting that we can treat these two SNPs as a single unit when calculating the Mendelian inheritance probabilities. To model Mendelian inheritance, we need to specify models for chromosome-specific copy numbers given the total copy numbers at two adjacent SNPs. Following a similar derivation of the chromosome-specific copy number model for a single SNP, here we propose a single-parameter model with parameter, b, which

specifies the probability of the less likely chromosome-specific copy number configuration in which copy number changes occur at both chromosomes (Table 3). Such an assumption is reasonable since it is unlikely that copy number changes occur on both chromosomes unless the CNV is common. Due to the high-dimensionality ($25 \times 25 \times 25$) of the table for two-marker copy number inheritance, we do not provide it in the manuscript, but it is available in the source code of our software.

Transition probabilities of de novo event statuses for the offspring. The majority of the CNVs in the offspring are inherited from the parents, but a small fraction of the offspring's CNVs may occur due to meiotic recombination, mitotic recombination, or cell line-induced chromosome rearrangements. The transition probability of *de novo* event status describes the probability of the offspring's CNV changing from inherited to *de novo* or vice versa. Clearly, the transition probability is dependent on distance between two adjacent markers since markers that are close to each other are more likely to be located in the same inherited or *de novo* region. To model such spatial dependency, we adopt the same transition probability model that was previously described for copy number states but with different transition parameters,

$$P(\mathrm{DN}_j = l | \mathrm{DN}_{j-1} = h, \lambda) = \begin{cases} 1 - \sum_{l \neq h} \delta_{h,l}(1 - e^{-d_j/D}) & \text{if } l = h \\ \delta_{h,l}(1 - e^{-d_j/D}) & \text{if } l \neq h \end{cases}, \quad \mathbf{11}$$

where the values of δs are treated as unknown parameters.

Parameter estimation and CNV calling. Inference on the hidden copy number states requires estimation of all unknown parameters, including μs and σs for the signal intensity, initial probabilities for copy number states π, the transition probability matrix $\Gamma = (\gamma_{h,l})$ for copy number states, the transition probability matrix for the *de novo* event status $\Delta = (\delta_{h,l})$ and a and b, the parameters in the single- and two-marker chromosome-specific copy number models. It is computationally challenging to estimate these parameters given the high dimension of the data. Moreover, a single sample may not carry sufficient information for estimating all model parameters. However, assuming the samples are homogeneous, then we can select a set of training samples with large CNV regions through visually examining patterns of LRRs and BAFs to estimate the corresponding μs and σs for regions with different numbers of copies. In our analysis, we fixed the values of a and b at 0.0009. Evaluations with different values of a and b suggest that the results are robust to misspecification of their values (data not shown). The initial probabilities π, the *de novo* rate ε and the Δ matrix are estimated from previously published HapMap CNV results (12). To estimate the parameters in the transition matrix Γ, we used the Baum–Welch algorithm (21) to maximize the likelihood in Equation (2). Given a set of HMM parameters and the signal intensity data from a trio, we then used the Viterbi algorithm (22) to infer the most likely path (state sequences for all SNPs along each chromosome) for each of the individuals in the trio simultaneously. A CNV is called from the most likely state sequence, whenever a stretch of states that is different from the normal state is observed.

Table 3. Probabilistic model specifying the relative probability of CNV haplo-genotypes, given the total copy numbers at two adjacent markers

Copy numbers at two adjacent markers	Chromosome-specific copy numbers	Probability
0 0	0 0 || 0 0	1
0 1	0 0 || 0 1	1
0 2	0 0 || 0 2	*b*
0 2	0 1 || 0 1	1 − *b*
0 3	0 0 || 0 3	*b*
0 3	0 1 || 0 2	1 − *b*
0 4	0 1 || 0 3	0.5
0 4	0 2 || 0 2	0.5
1 1	0 0 || 1 1	1
1 2	0 0 || 1 2	*b*
1 2	0 1 || 1 1	1 − *b*
1 3	0 0 || 1 3	*b*
1 3	0 2 || 1 1	1 − *b*
1 4	0 3 || 1 1	0.5
1 4	0 2 || 1 2	0.5
2 2	1 1 || 1 1	1 − *b*
2 2	0 0 || 2 2	*b*
2 3	1 1 || 1 2	1 − *b*
2 3	0 0 || 2 3	*b*
2 4	1 1 || 1 3	0.5 − 0.5*b*
2 4	1 2 || 1 2	0.5 − 0.5*b*
2 4	0 2 || 2 2	*b*
3 3	0 0 || 3 3	*b*
3 3	1 1 || 2 2	1 − *b*
3 4	1 1 || 2 3	0.5
3 4	1 2 || 2 2	0.5
4 4	2 2 || 2 2	0.5
4 4	1 1 || 3 3	0.5

In the last column, *b* is the probability of the less likely chromosome-specific copy number configuration in which copy number change occurs at both of the two homologous chromosomes. Our simplified model only considers those combinatorial configurations listed in the table, while other extremely rare combinations are treated as having probability of zero in the modeling procedure.

Availability

All the CNV calling algorithms have been implemented in the latest version of PennCNV, which is freely and publicly available at http://www.openbioinformatics.org/penncnv/. The Affymetrix LRR/BAF data transformation programs, which were used in this study for the Affymetrix genome-wide 5.0 arrays, were also made available as a beta version. A set of standard HMM models are provided for commonly used arrays; however, like commercial software such as Partek and GoldenHelix, users have the freedom to tweak all HMM parameters, the CNV inheritance models, as well as the population frequency of B allele parameters, which are suitable for custom-made arrays.

RESULTS

We have developed a joint-calling algorithm for CNV detection in parent–offspring trios, using a hidden

Markov framework that simultaneously models family relationship and signal intensities. This CNV calling algorithm differs substantially from the previously described family-based CNV calling algorithm (12) in that, first, copy number estimates are given with respect to a parent–offspring trio simultaneously in one step instead of two steps (Figure 1), and second, it gives probabilistic estimates of chromosome-specific copy numbers. Therefore, we compared the performance of the proposed joint-calling algorithm with existing algorithms that either do not incorporate family relationship or use them separately (12). We first performed simulation studies and then analyzed experimentally validated CNVs from multiple families in a real dataset genotyped using the Illumina SNP arrays. Furthermore, we used several concrete examples from real data to demonstrate how chromosome-specific copy numbers and SNP allele composition within CNVs can be inferred from family data. Finally, to demonstrate the versatility of the proposed method, we tested it on inherited and *de novo* CNVs from a set of families genotyped using both the Illumina HumanHap550 and the Affymetrix genome-wide 5.0 SNP arrays.

Comparative analysis of CNV detection on simulated data

To evaluate the performance of the proposed joint-calling algorithm under various scenarios of CNV inheritance, we performed computer simulations. We generated signal intensity data, as represented by LRR and BAF values, for 27,742 SNPs on chromosome 11 for the HumanHap550 SNP array, based on allele frequency and CNV size distribution from the empirical data obtained from the HapMap CEU samples (12). We tested a total of eight different inheritance scenarios of parent–offspring CNV combinations (Figure 2); for each scenario, we called CNVs using (i) the individual-calling algorithm that treats family members as if they were unrelated, (ii) the posterior-calling algorithm as described before (12) and (iii) the joint-calling algorithm as proposed in this paper. A total of 1000 data sets are simulated for each scenario, and there are either 1000 or 2000 true CNVs for each scenario depending on whether the

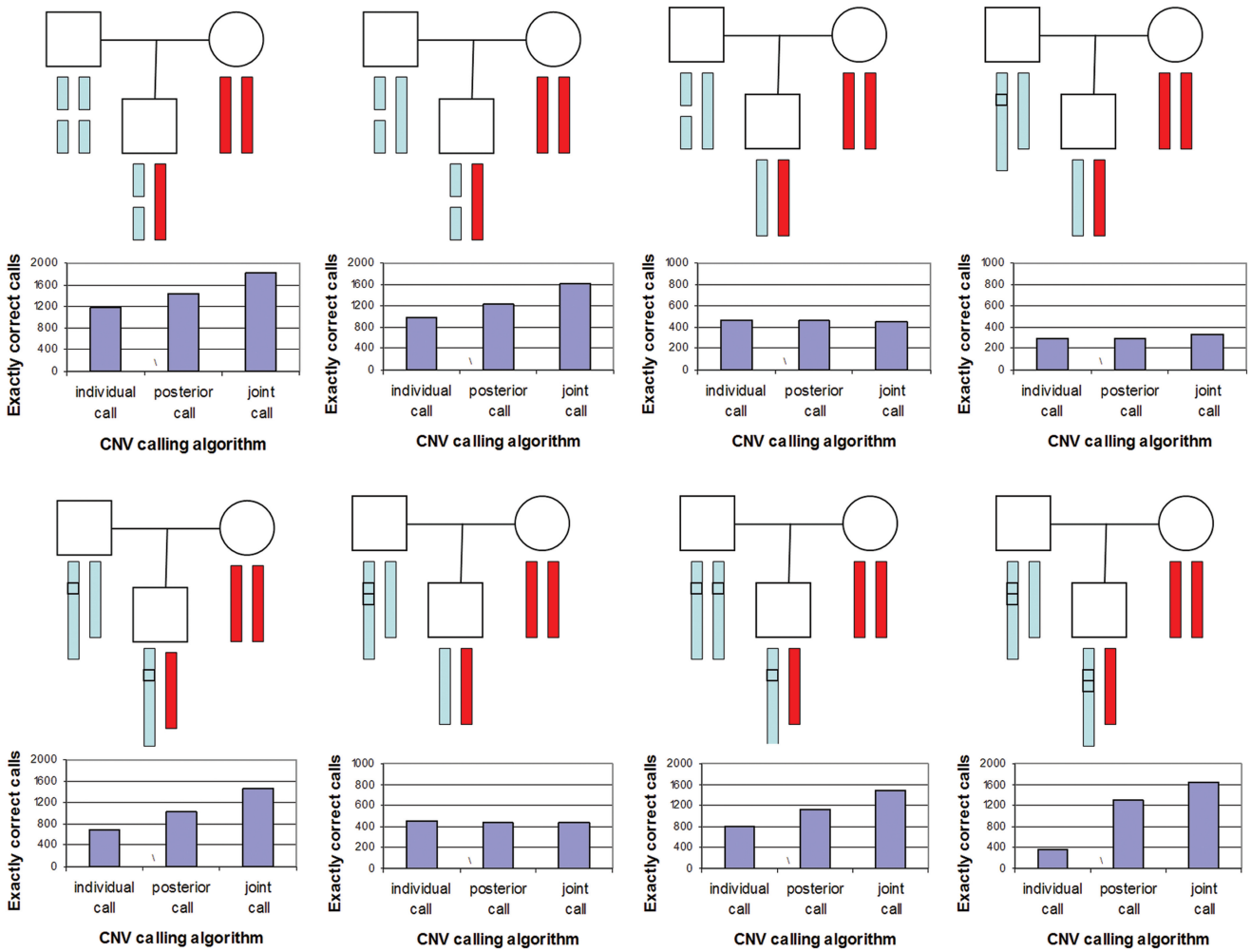

Figure 2. Comparative analysis of three CNV calling algorithms on simulated data. For each of the eight scenarios, 1000 trio data sets were simulated and analyzed. We evaluated whether each calling algorithm can identify 'exactly correct' CNV calls (calls with the exact CNV boundaries and the exact copy number as true CNVs). The joint-calling algorithm has the overall best performance, especially for inherited duplications.

CNVs are transmitted to the offspring. Given that *de novo* CNVs are rare (23–25), we did not consider them in the simulations; instead, we evaluated *de novo* CNVs in real data analysis as shown in a later section.

For each simulation scenario, the number of 'exactly correct' CNV calls (CNV calls with identical copy number and identical boundaries as the true CNVs) is shown in Figure 2. We can see that the three calling algorithms have similar performance for scenarios 3, 4 and 6, but for the other scenarios, the joint-calling algorithm yielded a substantially larger number of 'exactly correct' calls. Another important criterion of comparing different CNV calling algorithms is the number of false positive and false negative calls. Here, we refer to a CNV call as false positive if the call does not overlap with the true CNV, and we refer to a CNV as false negative if it is not detected by the CNV calling algorithm. Supplementary Figure 2 shows the numbers of false positive and false negative calls for the simulated data. We observed that when the offspring's CNV is inherited, the performance of the joint-calling algorithm far exceeds the other two algorithms, especially for duplication CNVs. For example, for scenario 7, where the father has duplication on both chromosomes and the offspring has duplication on only one chromosome, the number of false negative calls for the individual-calling algorithm is 646; it drops down to 404 and 162, respectively, for the posterior-calling algorithm and the joint-calling algorithm. Our results suggest that efficient utilization of family information can significantly improve the call rates as well as the accuracy of CNV boundary inference.

Comparative analysis of CNV detection on real data

To test the performance of the joint-calling algorithm in real data, we examined 10 families (34 subjects) from the Autism Genetic Resource Exchange (AGRE) Consortium. All study samples were genotyped using the Illumina HumanHap550 SNP array (14). To compare the performance of the three calling algorithms, we focused on the 4p16.1 deletions between *WDR1* and *ZNF518B*, which spans only four SNPs, making the CNV detection especially difficult. We designed PCR-walking experiments to validate the CNVs and mapped the approximate breakpoints. We then selected a pair of primers to infer the true copy numbers of all subjects by PCR amplification of the genomic segment encompassing CNV breakpoints. Finally, we re-sequenced the short PCR product to confirm that the breakpoints are identical among unrelated families. Since the true copy number for all subjects are known experimentally (Supplementary Figure 3), we compared the CNV calls for three algorithms with the true copy numbers. Collapsing all families together, there are a total of 30 true CNVs. For the individual-based calling algorithm, only 15 CNVs were detected, implicating a relatively high false negative rate. In contrast, both the posterior calling algorithm and the joint-calling algorithm are capable of detecting all 30 CNVs in all families. However, for one family, the posterior calling algorithm identified a CNV call with only three SNPs, resulting in a slight discordance in boundary inference. The joint-calling algorithm, on the other hand, completely recovered all true CNVs, and all the CNV calls have the correct boundaries with four SNPs. This comparative analysis on real data corroborate our analysis on simulated data, and confirms that the joint-calling algorithm improves accuracy in boundary inference and leads to decreased false negative rate.

Inference of chromosome-specific copy numbers from family data

Family information can be used to infer CNV genotypes, that is, chromosome-specific copy numbers on each of the two homologous chromosomes. To illustrate this point by a concrete example, we show in Figure 3 an AGRE family in which all family members carry a ~130 kb duplication on 22q11.21, which encompasses the *PRODH* and *DGCR6* gene. The results from the individual-calling, the posterior-calling and the joint-calling algorithms are concordant for this family, revealing that the first child has four copies of this CNV region, yet the father, the mother and the sibling in this family carry three copies. As shown in Table 2, when the total copy number is 3, the corresponding chromosome-specific copy numbers can be either 1/2 or 0/3, and when the total copy number is 4, the corresponding chromosome-specific copy numbers can be either 1/3 or 2/2. Despite such uncertainty, when the family relationship is considered, we can infer confidently that the chromosome-specific copy numbers for the father, the mother, the first child and the second child must be 1/2, 1/2, 2/2 and 1/2, respectively, and that the first child inherits the duplicated chromosome from both parents. If only one child is available in this family, we can still infer the most likely chromosome-specific copy number combinations in the parents–offspring trio, albeit with less confidence. This example is merely an illustration of how additional family information can be used to increase confidence of chromosome-specific copy number estimates, compared to the 'prior distribution' in Table 2.

Inference of chromosome-specific SNP genotypes in CNVs from family data

For CNV calls generated on SNP genotyping arrays, we can also use the SNP genotypes within the CNV to infer the SNP allele composition for each of the two homologous chromosomes, that is, chromosome-specific SNP genotypes. Unlike the 'called SNP genotypes' given by a genotype calling software, which comprises three types of allele compositions (AA, AB and BB), the 'real SNP genotypes' within a CNV can be jointly inferred from the BAF values and the total copy numbers (for example, A, BB, ABB and AABB, in Table 1). Knowing the SNP allele composition within the inherited CNV is important for the development of linkage and association tests on CNVs for disease phenotypes. To further illustrate this, we used a large segregating pedigree in AGRE as an example for such analysis: there are six offspring in this family, and five of them are affected by autism spectrum disorders (including four with strict autism diagnosis and one with broad spectrum diagnosis). Figure 4 displays the chromosome-specific SNP genotypes on the first 10 SNPs in the 10q11.22 duplication CNV region for each

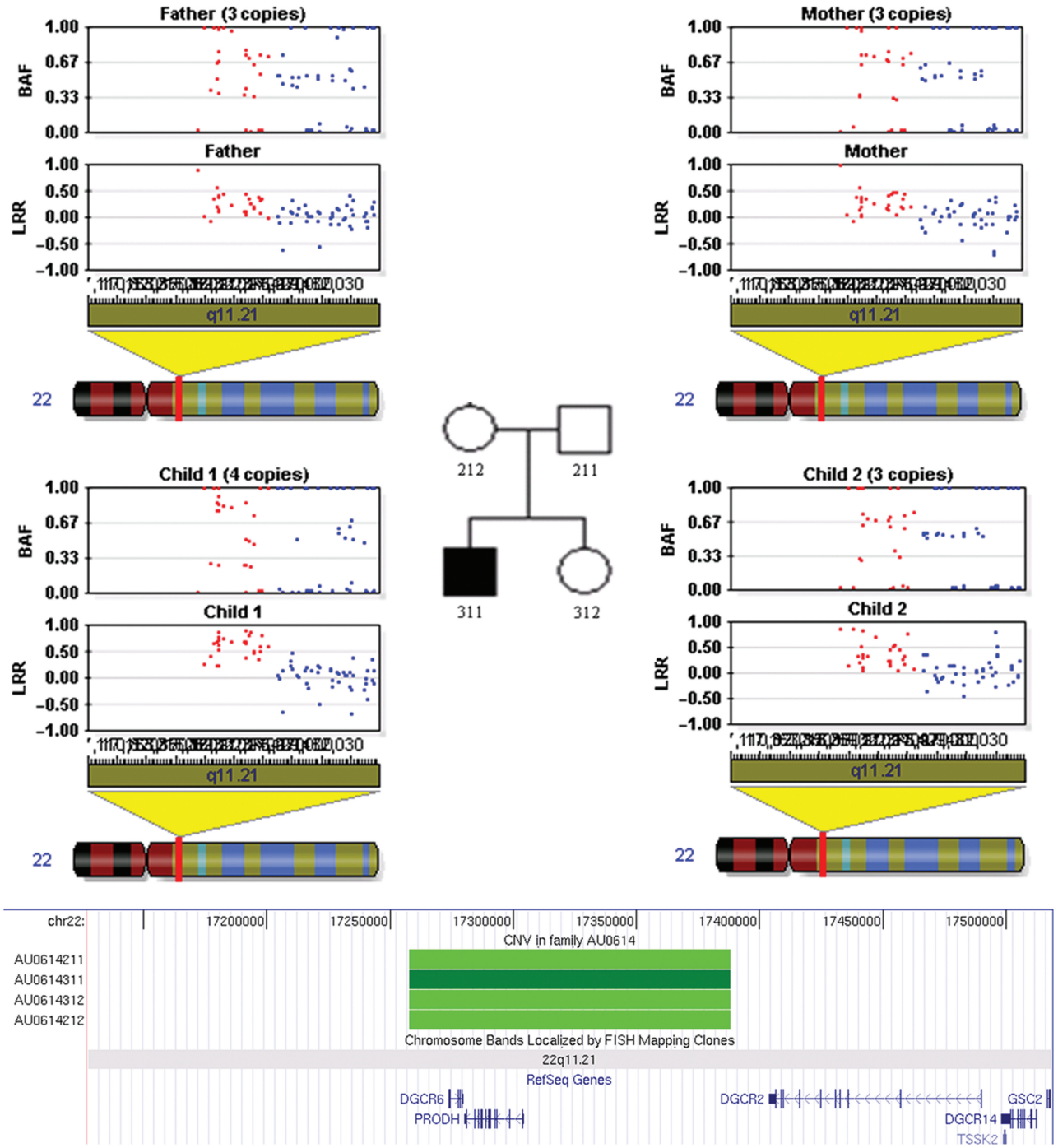

Figure 3. Illustration of the signal intensity patterns (LRR and BAF values in upper panel) at a CNV region on 22q11.21 in four members in an AGRE family. This CNV region encompasses the *DCGR6* and *PRODH* gene, as shown in the genome browser (26) shot (lower panel), where the CNV region for each individual is represented by a bar in the browser track (green = three copies, dark green = four copies). With the family information, we can infer that the first child inherits duplications from both the father and the mother, resulting in having four copies of the chromosome region.

of the individuals. By examining SNP genotypes, we can disambiguate the four parental CNV haplotypes with clear SNP allele composition (Figure 4, Supplementary Table 4).

Furthermore, Supplementary Figure 4 and Supplementary Table 5 show another example of chromosome-specific SNP genotypes in this pedigree at a duplication CNV on 6q27, which co-segregates with autism in this family. However, unlike the 10q11.22 duplication, since the mother is homozygous without copy number change in the 6q27 region, the transmission patterns of the two maternal chromosomes cannot be discriminated. In addition, we also analyzed the 22q11.21 duplication in the family presented in Figure 3: the use of family relationship

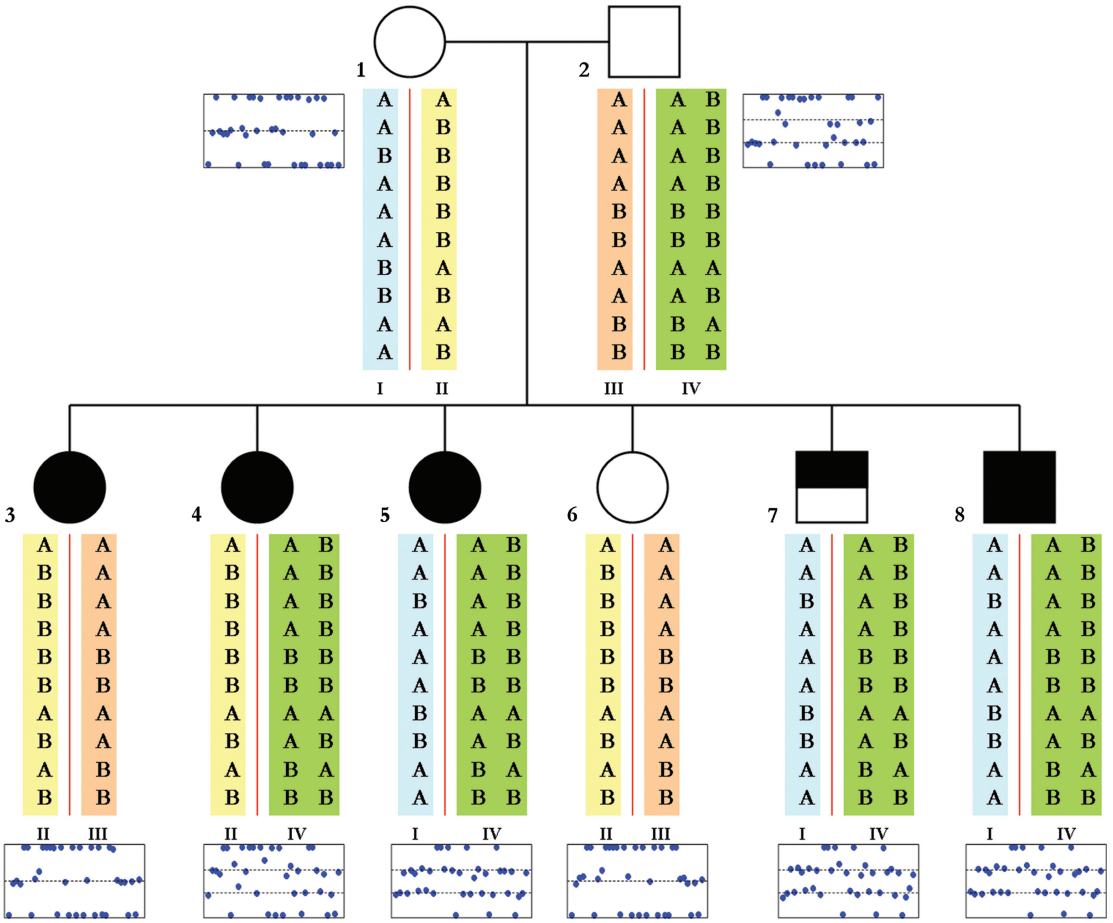

Figure 4. Illustration of a duplication CNV on 10q11.22 that exists in the father and is transmitted to four offspring. The CNV calls are made on six trios separately by the joint-calling algorithm. For each individual, the BAF values for all SNPs within the CNV and the chromosome-specific SNP genotypes (for the first 10 SNPs) are displayed, and the SNP genotypes for the entire region are listed at Supplementary Table 4. The four different parental CNV haplotypes are marked by different colors and denoted by I through IV beneath the genotypes. Combining information from total copy number and the SNP genotypes, we can infer the SNP allele compositions within each homologous chromosome confidently for each offspring.

allows the identification of the SNP allele composition and parental origin for each of the two duplicated homologous chromosomes in the first child, who carries four copies of this region (Supplementary Figure 5 and Supplementary Table 6). All these examples suggest the importance of examining family relationship and incorporating SNP genotypes into the analysis of CNVs. Efficient utilization of such information can provide valuable insights into studying the biological aspects of CNVs, including their evolutionary history as well as their genetic transmission patterns.

Detection of inherited and *de novo* CNVs from Illumina and Affymetrix SNP arrays

Although our algorithm was originally developed for Illumina data, the algorithm is general enough and can be readily applied to data generated from other technical platforms. To demonstrate such utility, we analyzed a set of AGRE families genotyped with both the Illumina HumanHap550 SNP arrays by us and the Affymetrix genome-wide 5.0 SNP arrays by others (13). All these families contain at least one family member with experimentally validated 16p11.2 deletion or duplication, including three inherited CNVs from the father in family AU0029 and five *de novo* CNVs in the offspring in four other families. This CNV region is flanked by two ~146 kb segmental duplications which share 99.6% sequence identity to each other and are 593 kb apart (Figure 5). For the Illumina HumanHap550 array, the CNV is covered by 47 SNPs with 530 kb in length. The Affymetrix genome-wide 5.0 Human SNP array contains 82 markers between segmental duplications; however, it also contains three additional markers within segmental duplications without unique genomic location, therefore we removed the three markers from our analysis. The exactly correct CNV calls from the Affymetrix array should contain 82 markers (28 SNP markers, 54 NP markers) with 569 kb in length. These families provide an ideal basis for comparison of different CNV calling algorithms and different technical platforms.

We compared the performance of the individual-calling, the posterior-calling and the joint-calling algorithms (Figure 5). All three algorithms gave correct CNV calls in all individuals carrying the CNV, indicating the high sensitivity and specificity of these algorithms in

			Illumina HumanHap550 SNP Array						Affymetrix Genome-wide 5.0 Human SNP Array					
			Individual call		Posterior call		Joint call		Individual call		Posterior call		Joint call	
Type	ID	Inheritance	#SNP	Length	#SNP	Length	#SNP	Length	#marker	Length	#marker	Length	#marker	Length
Dup	AU002903	inherited	43	483,152	43	483,152	46	509,622	68	481,604	68	481,604	**82**	**569,091**
Dup	AU002904	inherited	46	509,622	46	509,622	46	509,622	**82**	**569,091**	**82**	**569,091**	**82**	**569,091**
Dup	AU002905	inherited	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	**82**	**569,091**	**82**	**569,091**	**82**	**569,091**
Del	AU0154302	de novo	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	75	499,802	75	499,802	73	495,864
Del	AU0154303	de novo	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	80	499,802	80	499,802	75	499,802
Del	AU029803	de novo	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	68	481,604	68	481,604	68	481,604
Del	AU041905	de novo	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	**82**	**569,091**	**82**	**569,091**	**82**	**569,091**
Del	AU0938301	de novo	**47**	**530,466**	**47**	**530,466**	**47**	**530,466**	77	519,942	77	519,942	**82**	**569,091**

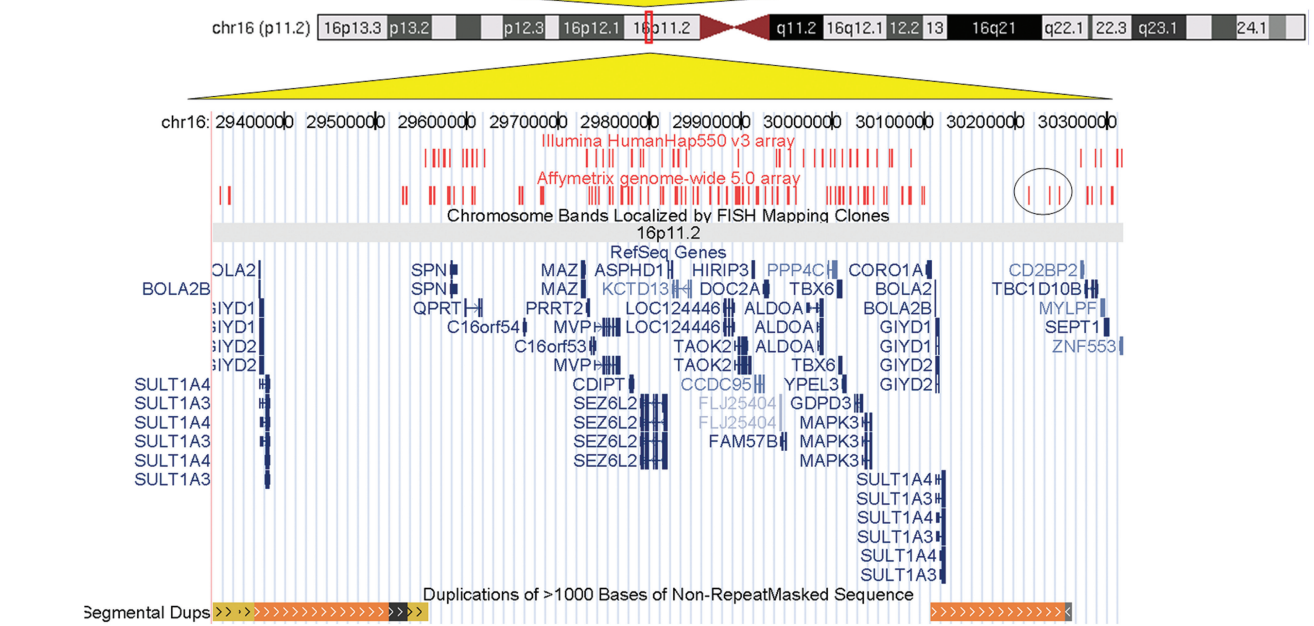

Figure 5. Comparison of three CNV calling algorithms in identifying the exact boundaries of the 16p11.2 CNV in offspring of a set of families genotyped by both Illumina HumanHap550 SNP array and Affymetrix genome-wide 5.0 Human SNP array. The CNV calls with exact boundaries were marked by bold font in the table in upper panel. The CNV region is displayed within UCSC genome browser (26), with two tracks representing marker coverage in two different arrays, as well as the RefSeq Genes track showing genes within the CNV. The three Affymetrix CN markers located within segmental duplication regions are marked by a circle and are removed from analysis. Two ~146 kb flanking segmental duplications are shown as dark orange bars in the Segmental Dups track. The joint-calling algorithm makes more exactly correct CNV calls than the other two calling algorithms.

detecting large-sized inherited or *de novo* CNVs. However, the algorithms differ in their ability to detect the exact CNV boundaries, which is especially obvious for the Affymetrix array due to its higher levels of background noise in signal intensity data. This example illustrates the ability of the joint-calling algorithm in detecting both inherited and *de novo* CNVs with accurate boundary prediction, and its broad applicability to arrays that incorporate NP markers.

DISCUSSION

We have developed a formal statistical framework to model the genetic inheritance of CNVs, via a HMM that simultaneously considers family relationship and signal intensities for parent–offspring trios. Our method considers the trio as a unit and calls their CNVs simultaneously, thus avoiding the generation of calls that are Mendelian inconsistent while maintaining the ability to detect *de novo* events. Moreover, our method allows the probabilistic estimation of chromosome-specific copy numbers, which can be used in subsequent CNV analysis. By extensive simulations and analysis of real family data, we showed that when the offspring's CNVs are inherited from the parents, the proposed method improves the call rates and the accuracy of boundary inference over existing methods. Although we present the algorithm for parent–offspring trios only, our method can be extended to analysis of nuclear families with multiple offspring, and we demonstrated the utility of using information from additional family members in Figure 3. Altogether, we hope that our method and software (http://www.openbioinformatics.org/penncnv) will be of great value to genome-wide CNV studies using family data.

Although we described our CNV calling algorithm for data generated from Illumina HumanHap550 and Affymetrix genome-wide 5.0 SNP arrays, we note that data derived from Illumina Human1M and Affymetrix genome-wide 6.0 SNP arrays are similar in nature and can therefore be analyzed directly with the proposed

algorithm. Moreover, our algorithm can be extended to other platforms, such as array-CGH experiments, or oligonucleotide tiling arrays. For these non-SNP arrays, since no allele frequency information can be inferred, only the signal intensities at each marker contribute to likelihood calculation. We note that due to the lower precision of array-CGH experiments, one might consider using 'loss' and 'gain', rather than the exact copy number, in the model. In such cases, the number of hidden copy number states reduces to three, and the various CNV inheritance tables need to be adjusted accordingly by combining copy numbers zero and one into a single 'loss' state, and copy numbers three and four into a single 'gain' state.

Compared to a previously published posterior-calling algorithm (12), there are several distinct advantages of the proposed joint-calling algorithm. First, instead of using family information separately, the joint-calling algorithm jointly models the family information with signal intensities and thus uses data in the most efficient manner. Second, if the CNV boundary is inferred incorrectly in the first step in the posterior-calling algorithm, then it cannot be corrected in the second step; however, as evidenced by our simulation results and analysis of the AGRE families, for inherited CNVs, the joint-calling algorithm is more likely to infer the correct boundary. Another unique feature of the joint-calling algorithm is the ability to give probabilistic estimate of chromosome-specific copy numbers, which is only feasible when family information is simultaneously modeled with signal intensities.

Despite the distinct advantages of the joint-calling algorithm, we recognize that it is computationally intensive and requires more assumptions than the posterior-calling algorithm. First, in the joint-calling algorithm, the family relationship needs to be jointly modeled with the signal intensities, thus requires $5 \times 5 \times 5 \times 2$ (five states for each individual and two *de novo* states for the offspring) states in the HMM for each marker in the genome; however, the original formula for the posterior-calling algorithm needs only six HMM states multiplied by three individuals in a trio. Second, due to the increased number of hidden states, the joint-calling algorithm requires more memory than the posterior-calling algorithm, which may be a problem for future ultra high-density oligonucleotide arrays with dozens of millions of markers. However, these problems can be solved by analyzing chromosome segments sequentially and then combining results together. Third, we note that the joint-calling algorithm makes more assumptions (such as the parameters used in Tables 2 and 3, as well as the *de novo* indicator transition rate) than the posterior-calling algorithm. The inherent complexity of the model dictates that some parameters must be estimated directly from empirical data rather than inferred from maximum likelihood. However, we note that the accuracy of these parameters only affects rare scenarios, and has little effects on the overall CNV calls. For example, increasing the transition rate of DN indicator from 'inherited' to '*de novo*' 10-fold only makes the detection of *de novo* event in the child less sensitive, but has virtually no effect on the detection of inherited CNVs in the trio or non-transmitted CNVs in the parents, which comprise the majority of CNVs in a family (data not shown).

For CNVs identified from high-density SNP genotyping data, another important piece of information is the corresponding SNP genotypes for markers within the CNVs; for example, SNP genotypes can be used to characterize parental origin of *de novo* events (12). In addition, SNP genotypes can help interpret inherited CNVs and extract more biological information, including probabilistic models for chromosome-specific copy numbers. We demonstrated the utilization of SNP genotype information on an AGRE family in which we can infer with certainty on the chromosome-specific copy numbers and the corresponding chromosome-specific SNP genotypes. Results from such analysis can be used to evaluate the preferential transmission pattern in a transmission disequilibrium test framework. In addition, information on the parental origin of a CNV will be particularly important for analysis of allelic imbalance and can help interpret gene expression differences from two homologous chromosomes.

In summary, we have developed a statistical framework to model the genetic inheritance of CNVs for parents–offspring trios. The likelihood calculation makes it easily extendable to nuclear families with multiple offspring. We believe that the application, adaptation and extension of our model in future studies will greatly facilitate the development of CNV detection algorithms for data generated from various technical platforms, and will foster the development of powerful and efficient linkage and association tests utilizing CNVs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Edmund Weisberg in the Center for Clinical Epidemiology and Biostatistics at the University of Pennsylvania for editing assistance. We thank two anonymous reviewers for their insightful suggestions on realistic data simulations and presentation of concrete real data examples. We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium (members of the consortium listed in Appendix 1) and the participating AGRE families. We are most grateful to the Children's Hospital of Philadelphia and the Broad Institute for providing us with access to the genotype data from the Illumina and Affymetrix platforms, respectively. National Institute of Mental Health (grant 1U24MH081810 to Clara M. Lajonchere PI partially); National Institutes of Health (grant R01-MH604687).

FUNDING

NARSAD Distinguished Investigator Award (to M.B.); University Research Foundation grant, McCabe Pilot Award from the University of Pennsylvania (to M.L.); National Institute of Health (grant R01HG004517),

an Institute Development Award to the Center for Applied Genomics from the Children's Hospital of Philadelphia (to H.H., S.F.A.G., J.G.). Funding for open access charge: National Institute of Health (grant R01HG004517 to M.L.).

Conflict of interest statement. None declared.

REFERENCES

1. Eichler,E.E., Nickerson,D.A., Altshuler,D., Bowcock,A.M., Brooks,L.D., Carter,N.P., Church,D.M., Felsenfeld,A., Guyer,M., Lee,C. *et al.* (2007) Completing the map of human genetic variation. *Nature*, **447**, 161–165.
2. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
3. Beckmann,J.S., Estivill,X. and Antonarakis,S.E. (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.*, **8**, 639–646.
4. Estivill,X. and Armengol,L. (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.*, **3**, 1787–1799.
5. Carter,N. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
6. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
7. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
8. Locke,D.P., Sharp,A.J., McCarroll,S.A., McGrath,S.D., Newman,T.L., Cheng,Z., Schwartz,S., Albertson,D.G., Pinkel,D., Altshuler,D.M. *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, **79**, 275–290.
9. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
10. Peiffer,D.A., Le,J.M., Steemers,F.J., Chang,W., Jenniges,T., Garcia,F., Haden,K., Li,J., Shaw,C.A., Belmont,J. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.
11. Komura,D., Shen,F., Ishikawa,S., Fitch,K.R., Chen,W., Zhang,J., Liu,G., Ihara,S., Nakamura,H., Hurles,M.E. *et al.* (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, **16**, 1575–1584.
12. Wang,K., Li,M., Hadley,D., Liu,R., Glessner,J., Grant,S.F.A., 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.
13. Weiss,L.A., Shen,Y., Korn,J.M., Arking,D.E., Miller,D.T., Fossdal,R., Saemundsen,E., Stefansson,H., Ferreira,M.A., Green,T. *et al.* (2008) Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N. Engl. J. Med.*, **358**, 667–675.
14. Bucan,M., Abrahams,B., Wang,K., Glessner,J., Herman,E.I., Sonnenblick,L.I., Alvarez Reteuro,A.I., Imielinski,M., Hadley,D., Bradfield,J.P. *et al.* (2008) Genome-wide analysis of exonic deletions identify novel autism susceptibility genes. *Submitted.*
15. Cupples,L.A., Arruda,H.T., Benjamin,E.J., D'Agostino,R.B. Sr, Demissie,S., DeStefano,A.L., Dupuis,J., Falls,K.M., Fox,C.S., Gottlieb,D.J. *et al.* (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med. Genet.*, **8** (Suppl. 1), S1.
16. Hafler,D.A., Compston,A., Sawcer,S., Lander,E.S., Daly,M.J., De Jager,P.L., de Bakker,P.I., Gabriel,S.B., Mirel,D.B., Ivinson,A.J. *et al.* (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.*, **357**, 851–862.
17. Todd,J.A., Walker,N.M., Cooper,J.D., Smyth,D.J., Downes,K., Plagnol,V., Bailey,R., Nejentsev,S., Field,S.F., Payne,F. *et al.* (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.*, **39**, 857–864.
18. Hakonarson,H., Grant,S.F., Bradfield,J.P., Marchand,L., Kim,C.E., Glessner,J.T., Grabs,R., Casalunovo,T., Taback,S.P., Frackelton,E.C. *et al.* (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature*, **448**, 591–594.
19. Kosta,K., Sabroe,I., Goke,J., Nibbs,R.J., Tsanakas,J., Whyte,M.K. and Teare,M.D. (2007) A Bayesian approach to copy-number-polymorphism analysis in nuclear pedigrees. *Am. J. Hum. Genet.*, **81**, 808–812.
20. Marioni,J.C., White,M., Tavare,S. and Lynch,A.G. (2008) Hidden copy number variation in the HapMap population. *Proc. Natl Acad. Sci. USA*, **105**, 10067–10072.
21. Baum,L.E., Petrie,T., Soules,G. and Weiss,N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.
22. Viterbi,A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.
23. Marshall,C.R., Noor,A., Vincent,J.B., Lionel,A.C., Feuk,L., Skaug,J., Shago,M., Moessner,R., Pinto,D., Ren,Y. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 477–488.
24. Sebat,J., Lakshmi,B., Malhotra,D., Troge,J., Lese-Martin,C., Walsh,T., Yamrom,B., Yoon,S., Krasnitz,A., Kendall,J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–449.
25. Xu,B., Roos,J.L., Levy,S., van Rensburg,E.J., Gogos,J.A. and Karayiorgou,M. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.*, **40**, 880–885.
26. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.

APPENDIX 1: THE AGRE CONSORTIUM

Daniel Geschwind, MD, PhD, UCLA, Los Angeles, CA; Maja Bucan, PhD, University of Pennsylvania, Philadelphia, PA; W. Ted Brown, MD, PhD, FACMG, N.Y.S. Institute for Basic Research in Developmental Disabilities, Staten Island, NY; Rita M. Cantor, PhD, UCLA School of Medicine, Los Angeles, CA; John N. Constantino, MD, Washington University School of Medicine, St Louis, MO; T.Conrad Gilliam, PhD, University of Chicago, Chicago, IL; Martha Herbert, MD, PhD, Harvard Medical School, Boston, MA; Clara Lajonchere, PhD, Autism Speaks, Los Angeles, CA; David H. Ledbetter, PhD, Emory University, Atlanta, GA; Christa Lese-Martin, PhD, Emory University, Atlanta, GA; Janet Miller, J.D., PhD, Autism Speaks, Los Angeles, CA; Stanley F. Nelson, MD, UCLA School of Medicine, Los Angeles, CA; Gerard D. Schellenberg, PhD, University of Washington, Seattle, WA; Carol A. Samango-Sprouse, Ed.D, George Washington University, Washington, DC; Sarah Spence, MD, PhD, UCLA, Los Angeles, CA; Matthew State, MD, PhD, Yale University, New Haven, CT; Rudolph E. Tanzi, PhD, Massachusetts General Hospital, Boston, MA.