Bayesian EM Algorithm for Scoring Polymorphic Deletions From SNP Data and Application to a Common CNV on 8q24

Sebastian Zöllner,[1–4]* Gang Su,[3] William C. L. Stewart,[1,4] Yi Chen,[1] Melvin G McInnis[2] and Margit Burmeister[2–5]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Psychiatry, University of Michigan, Ann Arbor, Michigan*
[3]*Bioinformatics Program, University of Michigan, Ann Arbor, Michigan*
[4]*Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan*
[5]*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan*

Copy number variations (CNVs) in the human genome provide exciting candidates for functional polymorphisms. Hence, we now assess association between CNV carrier status and diseases status by evaluating the signal intensity of SNP genotyping assays. Here, we present a novel statistical method designed to perform such inference and apply this method to a known CNV in a bipolar disorder linkage region. Using Bayesian computations we calculate the posterior probability for carrier status of a CNV in each individual of a sample by jointly analyzing genotype information and hybridization intensity. We model the signal intensity as a mixture of normal distributions, allowing for locus-specific and allele-specific distributions. Using an expectation maximization algorithm we estimate the parameters of these distributions and use these estimates for inferring carrier status of each individual and for the boundaries of the CNV. We applied the method to a sample of 3,512 individuals to a previously described common deletion on 8q24, a region consistently showing linkage to bipolar disorder, and unambiguously inferred 172 heterozygous and 1 homozygous deletion carrier. We observed no significant association between bipolar disorder and carrier status.

We carefully assessed the validity of the inferred carrier status and observed no indication of errors. Furthermore, the algorithm precisely identifies the boundaries of the CNV. Finally, we assessed the power of this algorithm to detect shorter CNVs by sub-sampling from the SNPs covered by this deletion, demonstrating that our EM algorithm produces precise estimates of carrier status. *Genet. Epidemiol*. 33:357–368, 2009. © 2008 Wiley-Liss, Inc.

Key words: copy number variation; association mapping; EM; deletion; 8q24

Contract grant sponsor: NIMH; Contract grant numbers: R01MH70775; HL084729-02.

*Correspondence to: Sebastian Zöllner, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: szoellne@umich.edu

Received 9 October 2008; Revised 9 October 2008; Accepted 21 October 2008
Published online 9 December 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20391

INTRODUCTION

After the human genome was sequenced it became clear that large segments of the human sequence exists in differing copy number [Redon et al., 2006]. Such copy number variations (CNVs) often include genes and open reading frames and hence are compelling candidates for risk variants for complex traits. Recent studies demonstrated association between a common deletion and reduced risk of HIV infection [Gonzalez et al., 2005], a novel neurodegenerative disorder caused by a deletion on 17q2 [Sharp et al., 2006] and an association between de novo deletions in several regions of the genome and autism [Sebat et al., 2007; Weiss et al., 2008] and schizophrenia [Walsh et al., 2008]. Presently, several efforts are underway to catalog all common CNVs in humans [Eichler et al., 2007], thus a map of most common CNVs may soon be available.

The next challenge will be to evaluate if common CNVs affect disease risk for common complex disorders. To this end, CNVs need to be typed in large samples of cases and controls. Several experimental methods exist to assess copy number, such as competitive genetic hybridization. However, it is efficient to use SNP genotype information generated for association mapping for the additional purpose of testing for CNVs as multiple genome-wide association studies, typing 300.000–500.000 SNPs in hundreds or thousands of cases and controls have been carried out. Thus, there is interest how to use the SNP genotype data to infer carrier status of known CNVs, and to test those CNVs for association with the studied disease phenotype and QTLs. Using genotyping technology, it is not possible to observe a duplicated or hemizygous region in the genome directly; hemizygous regions will be interpreted as homozygous sequence and duplicated regions may easily stump modern genotyping algorithms and appear as failed genotypes.

At least three methods to infer carrier status from genotyping data are conceivable: (1) tagging the alleles of each CNV with adjacent SNPs; (2) identifying segregating deletions based on non-Mendelian inheritance errors

(NMIs) in family data; or (3) assessing the carrier status from the signal intensity of the genotyping reaction. The first method is dependent on linkage disequilibrium (LD) between the CNV and the surrounding SNPs. Such LD will be high for CNVs that are the result of one single past event. Common CNVs in unique regions of the genome are often in strong LD to neighboring SNPs [McCarrol et al., 2006; Hinds et al., 2006]. However the mean r^2 between CNVs and flanking markers is significantly lower than the mean r^2 between pairs of SNPs [Redon et al., 2006] and CNVs in duplication-rich regions often show little LD to flanking markers [Locke et al., 2006]. Likely, CNVs in segmental duplications have a higher mutation rate [Sharp et al., 2005] and thus less LD, while CNVs outside of those regions are the result of single mutation events [Locke et al., 2006]. Therefore, it is not possible to tag every CNV using SNPs.

The second method, using NMIs to identify segregating deletions, has proven to be a powerful approach to localize deletions [Conrad et al., 2006; McCarrol et al., 2006]. If the deletion is transmitted from a hemizygous parent to a hemizygous offspring, both will appear to be homozygous for the allele carried on the other chromosome. Hence, if the chromosome transmitted by the other parent is different from the allele carried by the hemizygous parent, the offspring will appear to be homozygous for an allele not carried by one parent and will be counted as an NMI. However, only families where a deletion is actually transmitted will produce an NMI that can be detected. Thus, many carriers in the parental generation will not be identified, making it challenging to apply methods commonly used for family data to test for association between a phenotype and a deletion inferred from NMIs, as such tests rely on transmission distortion [Spielman and Ewens, 1996; Horvath et al., 2001]. Kohler and Cutler [2007] have recently developed a method to overcome this limitation, combining NMIs, deviations from Hardy-Weinberg Equilibrium (HWE), and frequency of missing data. However it is not clear how robust this method is to population genetic effects affecting HWE such as inbreeding and population substructure. Further, this method is contingent on the availability of family data.

Hence, it is often advantageous to infer CNV carrier status directly from the genotyping intensity signal. The two most commonly used high throughput genotype platforms generate a signal for a genotype whose intensity depends on the number of alleles present. However, interpreting this signal is challenging because the inference of CNV status is confounded with the genotype calling based on the same signal. Furthermore it is not obvious how to model the distribution of hybridization signals across multiple markers. Often such data are analyzed with somewhat ad hoc methods [Weiss et al., 2008]. Statistically more rigorous methods jointly model uncertainty about the location of the CNV and the carrier status of each individual in the sample [Wagenstaller et al., 2007]. Hidden Markov methods are the most commonly used tool; Komura et al. [2006] extended the SW-ARRAY algorithm [Price et al., 2005] to infer CNVs from data generated from an Affymetrix 500 K chip. Colella et al. [2007] proposed an objective Bayes Hidden Markov Model to infer location and carrier status of CNVs from Illumina Bead array data. PenCNV [Wang et al., 2007] extends this model to include information of related individuals. Furthermore, several methods that have been designed for CGH array [e.g. Fridlyand et al., 2004; Henrichsen et al., 2008] can be extended to genotyping data. As such methods model the uncertainty of the location of the CNV, they have relatively high error rates when calling CNVs; PenCNV has an error rate of 25% for CNVs of any length [Wang et al., 2007] and 8% for CNVs encompassing 10 SNPs or more [Wang et al., 2007; Jakobsson et al., 2008]. Such imprecise estimates of carrier status reduce the power of a test for association between a deletion and a phenotype.

However, once the location of a CNV is known, it may be more efficient to infer only the carrier status of the individuals in the sample. For the following reasons, we can assume, that most candidate regions are known. For association studies of complex disorders, we expect two types of risk affecting CNVs, either common, transmitted CNVs of moderate effect, or novel mutations. As complex diseases rarely show strong linkage peaks, transmitted CNVs with larger effects are unlikely. Common CNVs likely were observed before and their approximate location is recorded in one of the existing CNV databases. On the other hand, CNVs with high rates of recurring mutation are usually located in segmental duplications [Sharp et al., 2006]. These regions have also been well characterized and can be specifically targeted. Hence, we can reduce noise in the data by focusing on the regions most likely to contain CNVs of interest.

Here, we present a novel EM algorithm to infer the carrier status in unrelated individuals based on both SNP genotype calls and hybridization intensity data, assuming that the general location of the CNV is known, even though the specific boundaries may be unspecified. We model the signal intensity as a mixture of normal distributions allowing for locus-specific and allele-specific hybridization signals, similar to approach by Marioni et al. [2007] applied to CGH data. We combine information across markers and use genotype information to identify obligate non-carriers and use these known non-carriers to enhance the estimates of signal distribution and the precise boundary of the CNV. Thus, we generate the posterior probability for the carrier status of each individual in the sample. This estimate can then be used to test for association, e.g. in a logistic regression. We implemented this algorithm in the freely available program CNVEM.

We applied this method to a sample of 3,512 individuals from 737 families typed for 1,536 SNPs on 8q24 [Zandi et al., 2008], a region that has repeatedly shown linkage to bipolar disorder [Avramopoulos et al., 2004; McInnis et al., 2003; McQueen et al., 2005]. D8S272 marks a common 192 kb deletion localized in this region [Yu et al., 2002], called Variation_0337 in the database of genomic variations [Iafrate et al., 2004]. Twelve successfully called SNPs are located in the region covered by this deletion. We demonstrated the segregation of this deletion in our sample by considering NMIs of covered SNPs. Applying our EM-algorithm, we inferred the carrier status for each individual. After performing several quality-control checks on the inferred carrying status and validating a subset of the inferred deletion carriers by PCR, we found no indication of any error in the inferred carrier status. Using LAMP [Li et al., 2005, 2006], a family-based test for association, we and found no evidence for association between the deletion carrier status and bipolar disorder.

The 8q24 dataset is well suited to explore the properties of CNVEM and other CNV-typing algorithms. By generating and analyzing subsamples from that dataset, we showed that EM-methods provide precise estimates of carrier status for CNV with minor allele frequency above 1%, spanning five or more SNPs. Furthermore, we demonstrated the capacity of CNVEM to fine-map the borders of a CNV.

METHODS

Here, we describe a Bayesian approach for predicting carriers and boundaries of a CNV from the allelic hybridization intensity data of L SNPs sampled in n individuals. For ease of exposition, we limit our initial description to the simplest possible case wherein $m>0$ individuals carry a deletion, $n-m>0$ individuals do not, heterozygous genotypes are observed without error, and the deletion boundaries are constant across carriers (in the sense that the deletion is assumed to cover the L SNPs in every carrier). Our method is, however, easily extended to more complicated settings where any of the following may be present: carriers of duplications, genotype errors, or SNPs that are outside the span of the CNV, and we present extensions for the latter two in subsequent sections. For computational convenience, we ignore the unlikely possibility that individuals could carry multiple copies of a duplication, or two copies of a deletion. The latter is confounded with missing data at consecutive markers, and these individuals (if any), are often detected based on this pattern.

PREDICTING CARRIERS WITHOUT GENOTYPING ERROR

Let $G_{ij} \in \{AA, A-, AB, B-, BB\}$ be the true genotype of individual i at SNP j, for $i = 1, \ldots, n$, and $j = 1, \ldots, L$ with "$-$" denoting a deletion. Furthermore, let $D_{ij} \in \{\emptyset, A, AB, B\}$ be the observed alleles of genotype G_{ij} as generated by genotyping algorithms such as BRLMM [Rabbee and Speed, 2006] and DM [Di and Cawley, 2005]. For a pair of indices in the set $\Gamma \equiv \{(i,j) : D_{ij} \in \{A, B\}\}$, let H_{ij} be the hybridization intensity of allele D_{ij}. Hence, the observed data are $\mathbf{H} \equiv \{H_{ij} : (i,j) \in \Gamma\}$ and $\mathbf{D} \equiv \{D_{ij}\}$. Now, define $\mathbf{C} \equiv (C_1, \ldots, C_n)$ where $C_i \in \{0, 1\}$ is the carrier status of individual i. Note that G_{ij} is a deterministic function of C_i and D_{ij}, and that we condition on $\mathbf{D} = \mathbf{d}$ throughout. The goal then, is to predict carriers and non-carriers in the sample based on the posterior distribution of $\mathbf{C}$.

Under the assumption that H_{ij} is conditionally independent of all other variables given G_{ij}, and that C_i depends only on variables specific to individual i, the posterior distribution of $\mathbf{C}$ is

$$\begin{aligned}\Pr(\mathbf{C}|\text{Data}) &= \Pr(\mathbf{C}|\mathbf{H}, \mathbf{D}) \\ &\propto \Pr(\mathbf{C}, \mathbf{H}|\mathbf{D}) \\ &= \Pr(\mathbf{H}|\mathbf{C}, \mathbf{D}) \Pr(\mathbf{C}|\mathbf{D}) \\ &= \prod_{\Gamma} \Pr(H_{ij}|G_{ij}) \prod_{i} \Pr(C_i|\mathbf{D}_{i\cdot}). \qquad (1)\end{aligned}$$

In the absence of any information about the conditional distribution of $\mathbf{C}$ given $\mathbf{D}$, one may adopt a uniform prior (as we do here)

$$\Pr(C_i = 1|\mathbf{D}_{i\cdot}) = \begin{cases} \psi & \text{if } D_{ij} \in \{A, B\} \ \ \forall j, \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

by setting $\psi = 0.5$. Alternatively, one could consider incorporating information about the frequency of the deletion from public databases like DGV (Database of Genomic Variants, http://projects.tcag.ca/variation/) into the prior as well. Also, note that the prior given in (2) implicitly assumes that any individual who is heterozygous at any SNP is an obligate non-carrier. We will relax this assumption later, when we consider the possibility that we observed heterozygous genotypes in hemizygous individuals due to genotyping errors.

For each $(i,j) \in \Gamma$, we model the conditional distribution of H_{ij} given $G_{ij} = g$ as a normal random variate with mean $\mu(g)$, and variance $\sigma^2(g)$. Thus, there are eight L unknown parameters in the model, and to find the maximum likelihood (ML) estimates of these parameters, we use the expectation-maximization (EM) [Dempster et al., 1977] algorithm. For example, if we define p_i as the posterior probability that individual i is a carrier, then the parameter updates for the AA genotype at SNP j are

$$\mu_j(AA) = \frac{1}{\sum_{i:D_{ij}=A}(1-p_i)} \sum_{i:D_{ij}=A} (1-p_i) H_{ij}, \qquad (3)$$

$$\sigma_j^2(AA) = \frac{1}{\sum_{i:D_{ij}=A}(1-p_i)} \times \sum_{i:D_{ij}=A} (1-p_i)(H_{ij} - \mu_j(AA))^2. \qquad (4)$$

In the case where g is $A-$, $B-$, or BB, parameter updates are defined by analogous equations. For a complete derivation of all of the EM updates, see Appendix B. Hence, the ML estimates for the AA and BB genotypes are influenced by both potential carriers and obligate non-carriers, due to the hybridization intensities of the latter at observed homozygous genotypes. Given the ML estimates $(\mu(g), \sigma^2(g))$ for all $g \in \{AA, A-, BB, B-\}$, the posterior probability of being a carrier is easily computed from (1).

PREDICTING CARRIERS WITH GENOTYPING ERROR

In the presence of genotyping error, heterozygous genotypes may be observed in carriers. Let $K_i \leq L$ denote the number of observed heterozygous genotypes in individual i. To account for genotyping error, we assume that a hemizygous genotype is observed as a heterozygous genotype with probability ε, and that the conditional distribution of K_i given $C_i = 1$ is $\text{Bin}(L, \varepsilon)$.

Furthermore, when ε is small (e.g. 0.01),

$$\begin{aligned}&\Pr_\varepsilon(C_i = 1|K_i = 0) \\ &\quad\approx \Pr(C_i = 1|D_{ij} \in \{A, B\} \forall j) \equiv \psi;\end{aligned}$$

and, for deletions that span two or more SNPs, $\Pr_\varepsilon(C_i = 1|K_i > 1)$ is generally very close to zero. As a result,

$$\Pr_\varepsilon(C_i = 1|K_i = k) = \begin{cases} \psi & \text{if } k = 0, \\ \dfrac{\Pr_\varepsilon(K_i = 1|C_i = 1)\Pr(C_i = 1)}{\Pr_\varepsilon(K_i = 1)} & \text{if } k = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $\Pr_\varepsilon(K_i = k)$ for $k = 0, 1$ is estimated from the observed genotypes in the data, and $\Pr(C_i = 1)$ is estimated as $\psi \Pr_\varepsilon(K_i = 0)[\Pr_\varepsilon(K_i = 0|C_i = 1)]^{-1}$. To compute $\Pr_\varepsilon(\mathbf{C}|\mathbf{H}, \mathbf{D})$ in the presence of genotyping error, we continue to ignore the observed heterozygous genotypes in the calculation of $\Pr(H_{ij}|G_{ij})$, but we replace $\Pr(C_i = 1|D_{i\cdot})$ in (1) with $\Pr_\varepsilon(C_i|K_i = k)$ in (5). Thereafter, carrier status prediction proceeds as before, only this time, $\Pr_\varepsilon(\mathbf{C}|\mathbf{H}, \mathbf{D})$ is used instead of $\Pr(\mathbf{C}|\mathbf{H}, \mathbf{D})$. Note that this model may also identify individuals hemizygous for deletions that differ in length from the deletion assessed by the algorithm.

ESTIMATING DELETION BOUNDARIES

To estimate the boundaries of a deletion, we consider the L SNPs in physical order and we suppose now that some of the L SNPs may not be spanned by the deletion. Therefore, let α and ω denote the left-most and right-most SNPs spanned by the deletion. We estimate the deletion boundaries by finding the pair (α, ω) that maximize

$$P_{\alpha,\omega} = \prod_{j=1}^{\alpha-1} \Pr(\mathbf{H}_{\cdot j}|\mathbf{D}_{\cdot j}) \prod_{j=\alpha}^{\omega} \underset{\varepsilon}{\Pr}(\mathbf{H}_{\cdot j}|\mathbf{C}^*, \mathbf{D}_{\cdot j}) \times \prod_{j=\omega+1}^{L} \Pr(\mathbf{H}_{\cdot j}|\mathbf{D}_{\cdot j}),$$

with $\mathbf{H}_{\cdot j} \equiv (H_{1j}, \ldots, H_{nj})$, $\mathbf{D}_{\cdot j}$ defined similarly, $\mathbf{C}^* \equiv \operatorname{argmax} \Pr_\varepsilon(\mathbf{C}|\mathbf{H}, \mathbf{D})$ relative to the SNPs spanned by the deletion, and the hybridization intensities of obligate non-carriers are not modeled at heterozygous genotypes. To compute the outer products in the expression above, we estimate $\mu_j(AA)$, $\mu_j(BB)$ and $\sigma_j^2(AA)$, $\sigma_j^2(BB)$ using the sample mean and variance of the observed hybridization intensities for each homozygous genotype, since these distributions are assumed to be univariate normal. Then, the EM algorithm and $\Pr_\varepsilon(\mathbf{C}|\mathbf{H}, \mathbf{D})$ are used to estimate the parameters needed to compute the middle product. This model assumes that the CNV has the same boundaries in all carriers. This is a likely scenario if the CNV is the result of a single mutation event. However, present data on CNVs is insufficient to assess the heterogeneity of boundaries of CNVs caused by recurring mutations.

TESTING FOR ASSOCIATION

After calculating the posterior probability of being a carrier, p_i, for each $i = 1, \ldots, n$, several methods can be used to test for association between carrier status and disease. If all calls of carrier status are unambiguous (all $p_i \geq 0.99$ or $p_i \leq 0.01$), we can directly consider the inferred carrier status as the true carrier status. Ambiguity ($0.01 < p_i < 0.99$) can be resolved by selecting less stringent thresholds. While the latter approach ignores some information, an alternative test which does not dichotomize the continuous p is also possible. Specifically, we can calculate the expected number of carriers in cases as $\sum_{i \in \{\text{cases}\}} p_i$, and the same quantity in controls $\sum_{i \in \{\text{controls}\}} p_i$. Then, the two could be compared with a χ^2 test with one degree of freedom. Similarly, one could also consider the logistic regression of p_i onto disease status, in which case covariates are easily included as well.

DATA AND MATERIALS

The interval between Mb123.0 and Mb131.1 on chromosomal region 8q24 has twice met criteria for genome-wide significance [Lander and Kruglyak, 1995] for linkage with bipolar disorder [Cichon et al., 2001; McInnis et al., 2003; Avramopoulos et al., 2004]. Moreover, McQueen et al. [2005] recently pooled the primary genotype data from 11 BP genome-wide linkage scans (including individuals from our study) and reported two regions, 6q21 and 8q24, achieving genome-wide significance. As reported before [Zandi et al., 2008], we typed 1,536 SNPs across the region, 1,458 of those SNPs passed quality control filters. The sample consisted of 3,512 subjects from 737 multiplex families in which 1,954 subjects were affected (1,546 bipolar I disorder, 314 bipolar II disorder, and 94 subjects with schizo-affective disorder, bipolar type). The families were collected by the NIMH bipolar initiative [Nurnberger et al., 1997], and a sample collected by our group in the Mood Disorders Research Program at Johns Hopkins University [McInnis et al., 2003]. Genotyping was performed by the Center for Inherited Disease Research (CIDR) using an Illumina BeadLab system with Golden Gate chemistries. We performed single-locus tests with FBAT [Horvath et al., 2001] and Geno-PDT [Martin et al., 2003], and multi-locus test using HBAT [Horvath et al., 2004] and multi-locus Geno-PDT. None of the 1,458 SNPs showed strong evidence for association to bipolar disorder, the most significant P-value after data cleaning was 2.82×10^{-4} [Zandi et al., 2008].

However, the common deletion Variation_0337, marked by D8S272, as described by Yu et al. [2002], located between 137.7 and 137.9 Mb is a potential risk allele. Yu et al. [2002] observed the deletion in each of six population samples, one sample consisting of families with autism cases, a sample of individuals with learning disabilities, a sample of Alzheimer patients, two control samples and a sample of CEPH founders with allele frequencies ranging from 0.013 to 0.104, averaging 0.038. No association of the deletion with autism, learning disability, or Alzheimer was reported. The deletion does not contain any refseq genes, however several mRNAs and ESTS are mapped to this region (See the UCSC Genome Browser, http://genome.ucsc.edu/.). It covers 14 SNPs in our dataset: rs2613825, rs305276, rs305312, rs10505666, rs305279, rs2582431, rs2610077, rs10505665, rs305274, rs2613841, rs2613837, rs7825584, rs2649120, rs2681672. Two SNPs did not pass Illumina's default QC criteria and were not called (rs2649120 and rs2613825), the other 12 SNPs produced genotype calls.

VALIDATING THE PRESENCE OF A DELETION

To assess whether Variation_0337 segregated in our sample we examined NMIs in the SNPs covered by the deletion [Conrad et al., 2006; McCarrol et al., 2006]. While we expect consecutive NMIs to be highly specific for the presence of deletions, using NMIs to infer a deletion is not necessarily a very sensitive method; especially in sibpairs the probability to detect an NMI is zero. Furthermore, deletions need to be transmitted to generate an NMI, deletions that only occur in the parental generation cannot be inferred by NMIs. We counted a total of 269 NMIs in the 12 sites covering the deletion, compared to 393 NMIs over all other 1,446 SNPs, indicating a significant excess of NMIs ($p<10^{-100}$) in sites covered by the deletion, a clear indicator that the deletion is segregating in the sample. To assess the number of deletion carriers, we considered all nuclear families with at least one NMI, assuming that the NMI was caused by a segregating deletion, and identified the pair of individuals that are obligate carriers under such a model. All individuals implicated to carry a deletion by at least two NMIs were considered certain carriers; we counted 76 such individuals. Furthermore, we counted 18 individuals that were implicated by an NMI at exactly one of the 12 markers. Finally, one individual had missing genotypes for all 12 SNPs, showing low signal intensities of all 12 genotyping reactions, but not for other SNPs. We verified by PCR that this individual was homozygous for the deletion (data not shown).

RESAMPLED DATA

To obtain more general information about the performance of CNVEM, we generated datasets by sub-sampling k of the 12 markers without replacement, maintaining the patterns of LD. We analyzed the generated dataset using the EM algorithm to calculate for each individual i the probability of carrying the deletion $\tilde{p}_i$. Assuming the carrier status p_i obtained from the full dataset is the true carrier status, we compared $\tilde{p}_i$ to p_i by calculating the error E from

$$E=\frac{1}{z}\sum_{i=1}^{n}|p_i-\tilde{p}_i|,$$

where z is the true number of hemizygous individuals in the sample. Thus, E can roughly be considered the number of individuals falsely assigned to be hemizygous per true hemizygous individual. We generated 100,000 subsamples of size $k\in\{1,\ldots,12\}$ markers and calculated the mean and standard deviation of the error term E.

To assess the impact of sample size and CNV-frequency on the precision of our algorithm, we sampled with replacement n individuals from our sample of 4,001 sets of genotypes (including Illumina control individuals and duplicate individuals) and applied the EM algorithm, while maintaining the same proportion of hemizygous individuals as in the original sample (4.9%). For each subsample, we used CNVEM to infer the expected carrier status of each individual; then we compared the inferred carrier status with the inferred status based on the full dataset, calculating E. We thus generated 100,000 subsamples and summarized the distribution of E by calculating its mean and standard deviation. We repeated this analysis using hemizygous frequencies of 1, 2, 3, 10, and 20%.

PCR ANALYSIS OF SELECTED SAMPLES

Based on the junction sequence and primers F0 and R0 of [Yu et al., 2002], PCR reactions were performed with the following three primers: F0: gatcaagggatgatgagtatctc, F01: ggctgagtgaaaggaatgtg, and R0: gtgtagtggagccactatgctc. In the presence of the deletion, primer F01 is deleted, and F0 and R0 result in a fragment of 180 bp, in the absence of the deletion, primers F0 and R0 are more than 100 kb apart and thus will not give a fragment, but F01 and R0 together will result in a 249 bp fragment. Annealing temperature was 55°C, extension 72°C. Fragments were separated by agarose gel electrophoresis in the presence of ethidium bromide (data not shown).

RESULTS

We applied CNVEM to estimate the carrier status of each individual in the sample using the normalized signal intensity of the Illumina Golden Gate assay as the signal strength (see Illumina white papers at https://icom.illumina.com.icom/software.ilmn). We included all duplicated individuals and CIDR control individuals in the analysis; thus the sample consisted of 4,001 individual in total [Zandi et al., 2008]. After running five steps of the algorithm, the likelihood of the estimates converged to a local maximum. We ran the process for 30 steps to ensure convergence, running time was approximately 2 sec. To ensure that the process converged to the global maximum, we repeated the analysis starting the EM-algorithm from multiple random starting points; we always generated the identical result.

PREDICTING DELETION CARRIERS

We applied CNVEM to the individuals in the 8q24 dataset, treating the individuals as unrelated and 997 individuals in the 8q24 dataset are homozygous for all 12 SNPs and thus potential carriers of the deletion; 775 individuals are heterozygous for a single SNP and thus potential carriers with a genotyping error. After running the algorithm without modeling genotyping error, we observed posterior probabilities $p_i=1$, or $p_i=0$ for all individuals i, thus we could assign carrier status unequivocally, inferring 172 hemizygous carriers. After re-running the analysis while allowing for genotyping error, we observed one individual with one heterozygous genotype and a carry probability $p_i=0.83$ while carry probabilities for all other individuals were >0.99 or <0.01, almost identical to the analysis without genotyping error. However, this individual was not experimentally confirmed being hemizygous (see below). As the status of this individual is unclear, we removed him from further analysis. We thus inferred 172 hemizygous carriers in addition to the one homozygous carrier for a deletion frequency of 2.5% and carrier frequency of 4.9% in the 3,512 individuals of the 8q24 sample. Among the 885 pedigree founders, 40 hemizygous carriers were inferred (deletion frequency 2.3%). We could not reject HWE in either the entire sample ($p=0.42$) or the founders ($p=0.46$). A deletion frequency of 2.5% is consistent with previously reported frequencies of 0.013–0.104 [Yu et al., 2002].

To evaluate the validity of these inferences, we compared the set of individuals inferred to carry the deletion by the EM-algorithm with the set of individuals inferred by NMI-errors. All 74 individuals implicated by at least two NMIs and 16 of the 18 individuals implicated by one NMI were also inferred to be deletion carriers by the

EM algorithm. The remaining two individuals are a parent-offspring pair; both are heterozygous for multiple SNPs in the deletion. This suggests that the NMI in this family is caused by genotyping error, rather than by a deletion that had not been inferred by the EM. Hence, all deletions inferred by NMIs were also called by the EM algorithm, proving a high sensitivity of the algorithm. The algorithm, however, was able to call an additional 82 carriers not implicated by NMIs. Using a binomial approximation suggests a 95% confidence interval for the false-negative rate of $[0, 0.033]$.

As a second control of the inferred carrier status, we tested the deletion for Mendelian segregation. As our EM-analysis does not account for family information, we used the segregation of the deletion as independent control for the inferred carrier status. We assessed if offspring inferred to be hemizygous also had at least one parent inferred to be hemizygous. Failure to see such a pattern would have two possible explanations, either a recurring deletion event had occurred or the inferred carrier status was false in either parent or the offspring. In fact, each offspring carrying the deletion also had at least one parent carrying the deletion. Thus, our algorithm had a high power to identify carriers of the deletion. Furthermore, this result also indicated no evidence for recurring deletion events in this region in the 5,346 transmission events covered by our dataset. This is concordant with the observation that CNVs outside of segmental duplications are the result of rare events [Locke et al., 2006].

Finally, we used the junction sequences and two primers published by Yu et al. [2002] to develop a three primer PCR reaction, which results in amplified fragments of 180 bp in the presence and of 249 bp in the absence of the deletion, and both fragments in heterozygote samples. Thirteen samples (one predicted homozygote for the deletion, nine predicted heterozygotes for the deletion, two predicted non-deleted samples and the putative genotyping error) were PCR amplified. The individual identified as putative genotyping error only amplified the 249 bp fragment, indicating that this individual carries no deletion with the same 3′ breakpoint as D8S272. For all other individuals the fragment sizes observed were exactly as predicted in all cases (data not shown).

In summary, using a model without genotyping error, we unambiguously inferred the carrier status for all individuals in the sample and did not see any indications of erroneous assignments among the 172 hemizygous carriers, nor among the non-carriers. Including genotyping error in the model added ambiguity for only one individual. We consider the inferred carrier status for all other individuals to be highly reliable. This allows testing for association between the carrier status and bipolar disorder (BP). Furthermore, we can generate subsamples with known carrier status for each individual and thus assess the performance of CNVEM for smaller and less common CNVs.

TESTING FOR ASSOCIATION

We used LAMP [Li et al., 2005, 2006], a ML method that jointly tests association and linkage in families, to asses the evidence for association between D8S272 and bipolar disorder (BP); we did not reject the hypothesis of no association ($p = 0.42$). Thus, the D8S272 deletion has no major effect on the risk of BP and is unlikely to explain the linkage signal to 8q24.

ESTIMATION OF CNV BORDERS

To assess the ability of the program to correctly infer the boundaries of a deletion, we reanalyzed the 8q24 data, assuming that the boundaries of the deletion are not known precisely. We selected a set of 20 consecutive SNPs consisting of the 12 markers covered by the deletion and an additional 4 markers on each side of the CNV region. Thus, the deletion covers markers 5 through 16. To estimate these start- and endpoints, we calculated the posterior probability for every pair of start- and endpoints as described in the "Methods" section. We performed one calculation using the model assuming no genotyping error. The configuration starting at SNP 5, ending at SNP 16 had the highest posterior probability (Table I). Furthermore all of the top 10 most likely border configurations had less than 2% miscalls of carrier status. Markers 17 and 18 have heterozygosities below 0.05 and thus provide little information to exclude non-carriers. Thus, including markers 17 and 18 in the putative CNV does not substantially reduce the posterior probability. For a CNV with unknown boundaries such an observation would indicate some uncertainty about the terminal boundary of the CNV.

When inferring the borders of the CNV using a model with genotyping error, we observe that the configuration starting at SNP 5, ending at SNP 17 had the highest posterior probability. However, this posterior probability is close to the probability of the 5–16 configuration and the same carriers are inferred for both sets of boundaries (Table I). Further when modeling genotyping error, the 10 most probable border configurations have more similar posterior probabilities than the 10 most probable border configurations in a model without genotyping error.

TABLE I. Normalized log-likelihood of CNV-boundaries

First	Last			First	Last		
SNP	SNP	LPP	Error	SNP	SNP	LPP	Error
5	16	61,896	0	5	17	61,913	0
5	18	61,613	0.017	5	16	61,908	0
5	17	61,609	0.017	5	18	61,625	0.018
6	16	60,788	0	5	19	61,433	0.030
5	15	60,778	0	4	18	61,424	0.034
6	18	60,517	0.017	4	17	61,420	0.034
6	17	60,513	0.018	4	16	61,414	0.034
7	16	59,839	0	3	18	61,231	0.046
6	15	59,673	0	3	17	61,228	0.046
7	18	59,590	0.017	3	16	61,221	0.046

We normalized the log-likelihoods for each pair of boundaries by subtracting the log-likelihood of the data without a segregating CNV and ranked them by the result. The left half of the table displays the 10 most likely boundary pairs and their unnormalized log posterior probability using a model with no genotyping error (LPP), and the error of the called carrier status using these boundaries; the right side of the table provides the 10 most likely boundary pairs, their unnormalized log-likelihood, and the error using a model with genotyping error. CNV, copy number variations.

Hence, modeling genotyping error reduces the ability to fine map the borders of a CNV.

RESAMPLED DATA

We explored the properties of the algorithm by randomly generating subsets of the original dataset, varying the sample size and number of markers. We compared inference results based on subsets of the data to inferences from the entire dataset using the error statistic *E*, which is normalized for the total numbers of CNV carriers in the population. Varying the number of SNPs covered by the CNV between 1 and 12, we observed that datasets of four or less SNPs generate on average an error statistic >0.25 (Table II), roughly equivalent of one false call for every four true carriers in the sample. Furthermore, some posterior probabilities were near 0.5, further indicating that the data were insufficient to assign carrier status. For CNVs covered by 5 or 6 markers, we observed an acceptable error statistic of 0.05 and 0.02; for CNVs covered by 8 or more markers, the error was negligible.

Heterozygosity of the covered SNPs and the LD between the covered SNPs has a large influence on the resolution of the algorithm. Markers with low minor allele frequency are less likely to be heterozygous in non-carriers and therefore less likely to exclude these individuals as carriers. Furthermore, markers in strong LD will not independently exclude individuals as CNV-carriers. In each simulated dataset, we summarized heterozygosity and LD by counting the number of individuals that are homozygous for all markers covering the CNV. Then we calculated the correlation coefficient (*r*) between this summary statistic and *E* for each number of markers. Values of *r* ranged from 0.4 to 0.9, showing a strong effect of marker selection on the large standard deviation of *E* (Table II). Hence, seven to eight randomly selected SNPs are necessary to reliably assign carrier status; if the SNPs are selected to have high heterozygosity and little mutual LD (e.g. tag SNPs) four to six SNPs covering the CNV are sufficient.

To study the impact of sample size and deletion frequency on the precision of the estimate, we generated sub-samples of 100, 200, 400, 600, 800, 1,000, 1,500, 2,000, 2,500, and 3,000 individuals from the NIMH data, setting the frequency of hemizygous individuals to 1, 2, 3, 4.9% (as observed in the NIMH sample), 10, and 20%. Based on the generated dataset, we inferred the carrier status of individuals in that subsample by applying the algorithm on all 12 SNPs covered by the deletion (Fig. 1) and calculating the error of the inferred carry probabilities relative to the true number of hemizygous individuals. The frequency of deletion carriers has a large effect on the precision of the estimate. Deletions that occur in only 1% of individuals are very difficult to infer, even for samples of 3,000 individuals (Fig. 1A). On the other hand deletions occurring in 10 or 20% of families can be inferred with almost no error even in samples of 100 individuals (Fig. 1B). For deletions with a frequency between 2 and 3%, the precision of the inference is strongly dependent on the sample size, for 2% (3%) frequency, more than 1,500 (500) individuals are required to reduce the error below 0.05. For a deletion with the frequency of Variation_0337 (4.9%), samples of 200 individuals are sufficient to observe only seven erroneous

TABLE II. Error of CNV carrier status inference dependent on the number of SNPs covering the CNV

	1	2	3	4	5	6	7	8	9	10	11	12
Mean *E*	8.745	3.977	1.364	0.257	0.054	0.019	0.010	0.007	0.005	0.003	0.001	0.000
Standard deviation	2.250	2.015	1.312	0.507	0.139	0.041	0.011	0.008	0.006	0.005	0.002	0.000

The first line displays the number of SNPs that was randomly sampled to generate a dataset. The second and third line shows the mean error statistic and the standard deviation of such datasets over 100.000 simulated datasets. CNV, copy number variations.

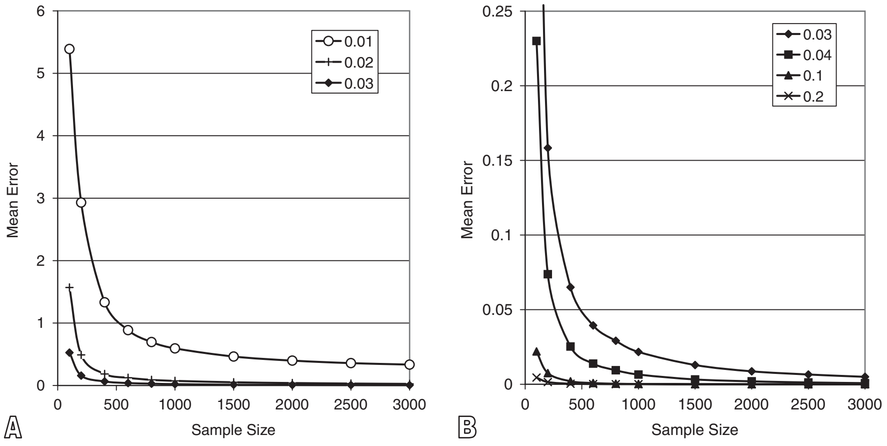

Fig. 1. Impact of CNV frequency and sample size on the precision of the inference procedure. The mean error (vertical axis) is shown for sample sizes displayed along the vertical axis. Results are presented for sample frequencies of hemizygous individuals of 1, 2, and 3 (panel A) and 3, 4.4, 10, and 20% (panel B).

inferences per 100 true hemizygous individuals. Thus, samples of 200 individuals assessed for all 12 SNPs have approximately the same error as samples of 4,000 individuals assessed for only five SNPs.

DISCUSSION

CNVs are a major part of human genetic variation [Iafrate et al., 2004; Conrad et al., 2006; Redon et al., 2006; Jakobsson et al., 2008]. The most cost-effective way of assessing their carrier status in a genome-wide association study is using signal intensities of the genotyping reaction. We describe a novel EM-algorithm to analyze such signal information together with the SNP genotypes for large samples. We use genotype information to identify obligate non-carriers of the CNV (i.e. individuals heterozygotes for at least one SNP) and hybridization signal intensities to infer carrier status in all other individuals. By including such obligate non-carriers in the analysis we improve estimates of signal intensity distributions. Our method does not consider family information and can thus be applied both to related and unrelated individuals. It is computationally efficient and can infer carrier status for >10.000 individuals in seconds. The method was implemented the method in the program CNVEM available from the authors at http://http://www.sph.umich.edu/csg/zollner.

We applied CNVEM to infer carrier status of the Variation_0337 deletion in the 8q24 dataset in the NIMH sample, unequivocally inferring 172 hemizygous carriers and one homozygous carrier. We verified the carrier status using the family information and observed no contradiction between inferred carrier status, family relatedness, and SNP genotype. Furthermore, we used PCR to verify a subset of inferred carriers experimentally, again observing no error in the inferred carrier status. When including genotyping error in our model, we identified one putative additional carrier. Using PCR, we could not detect Variation_0337 in this individual. Hence, the individual either carries a deletion with different boundaries or the inference is erroneous. We tested the inferred deletion status for association with bipolar disorder using LAMP [Li et al., 2005, 2006], a family-based test for association. We observed no significant result.

We further analyzed the properties of CNVEM with data from the 8q24 dataset. Resampling SNPs from this dataset provided important information on the limits of CNVEM and related algorithms. Our algorithm is precise even for small sample sizes of 200 individuals at the hemizygous frequency of 4.4% of Variation_0337. Large samples of $>1,500$ individuals are required for precise inference only for less common CNVs with frequencies below 3%. However, if CNVs contribute to the risk of a disease, we would expect them to be more common in samples of affected individuals, even if their frequency in the general population is low.

On the other hand, the number of SNPs covered by the CNV has a larger effect on the error of the inference procedure; CNVs covered by fewer than four SNPs were inferred with low reliability. For CNVs covered between four and six SNPs, the resolution was uneven, particularly depending on the LD between the covered SNPs. Only for CNVs covered by 7 or more markers did the algorithm perform well, having error rates of 1%, regardless of the specific subset of SNPs selected. Two reasons may explain this high level of precision:

We focus on specific region rather than the entire genome. At least 90% of all SNPs will lie outside of CNV regions. Even within CNV region, at least 90% of individuals will have the baseline copy number. Hence, evaluating the entire genome requires isolating 1% signal from data that consists mostly of noise. Therefore, the quality of calling CNV carrier status can be improved by concentrating on known CNV regions. While the precise borders of CNVs in databases are generally unknown, we have shown that our algorithm can overcome such uncertainty and finemap the borders of a CNV.

Furthermore, by focussing on smaller regions, we are able to apply detailed models of hybridization intensity in CNVEM, as we assume known CNV location. The distribution of hybridization signals across markers is highly heterogeneous, even after the signal intensities have been normalized by Illumina's normalization algorithm. Methods such as QuantiSNP [Colella et al., 2007] and PenCNV [Wang et al., 2007] normalize the hybridization intensity at each SNP using the signal intensity of a canonical genotyping cluster. Such methods are appropriate when the entire genome is scanned for signals of copy number variation; however they reduce the fit of the modeling of the hybridization intensity [Wagenstaller et al., 2007]. More detail about the models of hybridization intensity are given in Appendix A.

With the current 500,000–1,000,000 SNP panels, the median density is about one SNP every 1.5–3.2 kb, thus our algorithm allows calling most deletions that are larger than about 10–20 kb. In the set of CNVs described by Kidd et al. [2008] the median length of CNVs discovered by SNP arrays was 33.4 kb, thus our method can detect most such variation present in the human genome [Cooper et al., 2008]. Furthermore, both Illumina and Affymetrix have recently developed chips that hybridize additional probes for CNV-detection, potentially decreasing the average size of deletions identifiable with CNVEM.

However, it is not clear how the properties of the algorithm depend on the genotyping platform and SNP calling algorithm used to generate genotypes and hybridization intensities. Illumina technology generally generates fewer genotyping errors than Affymetrix technology and large numbers of genotyping errors are likely to reduce the precision of CNVEM.

The algorithm in CNVEM can be applied to duplications as well as deletions. However, for inferring duplications, the resolution of the algorithm is lower for two reasons: The relative signal difference between an individual carrying two copies and an individual carrying three copies is smaller than the relative difference between an individual carrying two copies and an individual carrying one copy. Furthermore, it is more challenging to identify obligate non-carriers from genotype calls. Finally, in the presence of a segregating duplication, more genotypes will be impossible to call, requiring a more detailed modeling of markers with missing genotype calls.

A further avenue of improving carrier status calls is analyzing repeat measurements of the genotyping signal as occur within families. During the inference, we disregard all the family information in the data and treat individuals as unrelated. This allows us to test the inferred carrier status by assessing NMIs of the deleted allele, as described earlier. However calling CNV carrier status would be more powerful if we combined the evidence of a segregating CNV across generations. As chromosomes are transmitted from parent to offspring, they are genotyped several times in different individuals. Modeling the transmission can thus combine the signal across individuals and increase the resolution of our method. Furthermore, in such a model we can include NMIs and departure from HWE to directly call carriers [Kohler and Cutler, 2007].

In summary the present algorithm provides a simple and powerful tool to assess CNV carrier status from genotype signal that can be applied to score all common large deletions in the human genome with high accuracy. CNVEM, the implementation of this program is computationally efficient even for large samples thus making it possible to extend genome-wide association studies to these polymorphisms.

ACKNOWLEDGMENTS

We thank Don Conrad, Michael Boehnke, and Michele Gornick for helpful discussions. This work was supported by NIMH grant R01MH70775 to M.G.M. and HL084729-02 to S.Z.

REFERENCES

Avramopoulos D, Willour VL, Zandi PP, Huo Y, MacKinnon DF, Potash JB, DePaulo Jr JR, McInnis MG. 2004. Linkage of bipolar affective disorder on chromosome 8q24: follow-up and parametric analysis. Mol Psychiatry 9:191–196.

Cichon S, Schmidt-Wolf G, Schumacher J, Muller DJ, Hurter M, Schulze TG, Albus M, Borrmann-Hassenbach M, Franzek E, Lanczik M, Fritze J, Kreiner R, Weigelt B, Minges J, Lichtermann D, Lerer B, Kanyas K, Strauch K, Windemuth C, Baur MP, Wienker TF, Maier W, Rietschel M, Propping P, Nothen MM. 2001. A possible susceptibility locus for bipolar affective disorder in chromosomal region 10q25–q26. Mol Psychiatry 6:342–349.

Colella S, Yau C, Taylor J, Mirza G, Butler H, Clouston P, Bassett A, Seller A, Holmes C, 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.

Conrad D, Andrews T, Carter N, Hurles M, Pritchard J. 2006. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40:1199–1203.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38.

Di X, Cawley S. 2005. Alternative base calling method for resequencing microarrays. Conf Proc IEEE Eng Med Biol Soc 3:2809–2812.

Eichler E, Nickerson D, Altshuler D, Bowcock A, Brooks L, Carter N, Church D, Felsenfeld A, Guyer M, Lee C, Lupski J, Mullikin J, Pritchard J, Sebat J, Sherry S, Smith D, Valle D, Waterston R. 2007. Completing the map of human genetic variation. Nature 447:161–165.

Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. 2004. Hidden Markov models approach to the analysis of array CGH. J Multivar Anal 90:132–153.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs R, Freedman B, Quinones M, Bamshad M, Murthy K, Rovin B, Bradley W, Clark R, Anderson S, O'Connell R, Agan B, Ahuja S, Bologna R, Sen L, Dolan M, Ahuja S. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440.

Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Ruedi M, Kaessmann H, Reymond A. 2008. Segmental copy number variation shapes tissue transcriptomes, submitted.

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85.

Horvath S, Xu X, Laird NM. 2001. The family based association test method: strategies for studying general genotype–phenotype associations. Eur J Hum Genet 9:301–306.

Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. 2004. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61–69.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36:949–951.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64.

Kohler JR, Cutler DJ. 2007. Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. Am J Hum Genet 81:684–699.

Komura D, Shen F, Ishikawa S, Fitch K, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles M, Lee C, Scherer S, Jones K, Shapero M, Huang J, Aburatani H. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 16:1575–1584.

Lander E, Kruglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247.

Li M, Boehnke M, Abecasis G. 2006. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 78:778–792.

Li M, Boehnke M, Abecasis GR. 2005. Joint modeling of linkage and association: identifying snps responsible for a linkage signal. Am J Hum Genet 76:934–949.

Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, Eichler EE. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79:275–290.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavaré S, Hurles ME. 2007. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 8:R228.

Martin ER, Bass MP, Gilbert JR, Pericak-Vance MA, Hauser ER. 2003. Genotype-based association test for general pedigrees: the genotype-pdt. Genet Epidemiol 25:203–213.

McCarrol SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM. 2006. Common deletion polymorphisms in the human genome. Nat Genet 38:86–92.

McInnis MG, Lan TH, Willour VL, McMahon FJ, Simpson SG, Addington AM, MacKinnon DF, Potash JB, Mahoney AT, Chellis J, Huo Y, Swift-Scanlan T, Chen H, Koskela R, Stine OC, Jamison KR, Holmans P, Folstein S, Ranade K, Friddle C, Botstein D, Marr T, Beaty TH, Zandi P, DePaulo JR. 2003. Genome-wide scan of bipolar disorder in 65 pedigrees: supportive evidence for linkage at 8q24, 18q22, 4q32, 2p12, and 13q12. Mol Psychiatry 8:288–298.

McQueen MB, Devlin B, Faraone SV, Nimgaonkar VL, Sklar P, Smoller JW, Jamra RA, Albus M, Bacanu SA, Baron M, Barrett TB, Berrettini W, Blacker D, Byerley W, Coryell SCW, Craddock N, Daly MJ, DePaulo JR, Edenberg HJ, Foroud T, Gill M, Gilliam TC, Hamshere M, Jones I, Jones L, Juo S-H, Kelsoe JR, Lambert D, Lange C, Lerer B, Liu J, Maier W, MacKinnon JD, McInnis MG, McMahon FJ, Murphy DL, Nothen MM, Nurnberger Jr JI, Pato CN, Pato MT, Potash JB, Propping P, Pulver AE, Rice JP, Rietschel M, Scheftner W, Schumacher J, Segurado R, Steen KV, Xie W, Zandi PP, Laird NM. 2005. Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. Am J Hum Genet 77:582–595.

Nurnberger JI, DePaulo JR, Gershon ES, Reich T, Blehar MC, Edenberg HJ, Foroud T, Miller M, Bowman E, Mayeda A, Rau NL, Smiley C, Conneally PM, McMahon F, Meyers D, Simpson S, McInnis M, Stine OC, Detera-Wadleigh S, Goldin L, Guroff J, Maxwell E, Kazuba D, Gejman PV, Badner J, Sanders A, Rice J, Bierut L, Goate A. 1997. Genomic survey of bipolar illness in the NIMH genetics initiative pedigrees: a preliminary report. Am J Med Genet 74:227–237.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. 2006. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Res 16:1136–1148.

Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, Smith L, Greenfield A, Tiganescu A, Buckle V, Ventress N, Ayyub H, Salhan A, Pedraza-Diaz S, Broxholme J, Ragoussis J, Higgs DR, Flint J, Knight SJ. 2005. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. Nucleic Acids Res 33:3455–3464.

Rabbee N, Speed TP. 2006. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12.

Redon R, Ishikawa S, Fitch K, Feuk L, Perry G, Andrews T, Fiegler H, Shapero M, Carson A, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global variation in copy number in the human genome. Nature 23:444–454.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YJH, Spence S, Lee A, Puura K, Lehtimaki T, Ledbetter D, Gregersen P, Bregman J, Sutcliffe J, Jobanputra V, Chung W, Warburton D, King M, Skuse D, Geschwind D, Gilliam T, Ye K, Wigler M. 2007. Strong association of de novo copy number mutations with autism. Science 316:445–449.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38:1038–1042.

Spielman RS, Ewens WJ. 1996. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989.

Wagenstaller J, Spranger S, Lorenz-Depiereux B, Kazmierczak B, Nathrath M, Wahl D, Heye B, Glaser D, Liebscher V, Meitinger T, Strom, T. 2007. Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. Am J Hum Genet 81:768–779.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King M-C, Sebat J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, 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.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Autism Consortium MJD. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358:667–675.

Yu CE, Dawson G, Munson J, D'Souza I, Osterling J, Estes A, Leutenegger AL, Flodman P, Smith M, Raskind W, Spence MA, McMahon W, Wijsman EM, Schellenberg GD. 2002. Presence of large deletions in kindreds with autism. Am J Hum Genet 71:100–115.

Zandi PP, Zöllner S, Avramopoulos D, Willour VL, Chen Y, Qin ZS, Burmeister M, Miao K, Gopalakrishnan S, McEachin R, Potash JB, DePaulo Jr JR, McInnis MG. 2008. Family-based SNP association study on 8q24 in bipolar disorder. Am J Med Genet B Neuropsychatr Genet 147:612–618.

APPENDIX A

DISTRIBUTION OF NORMALIZED SIGNAL INTENSITIES

How to best model the normalized signal intensity of a genotyping reaction is an open question. A priori, it is not obvious how such distributions vary between markers and alleles after an appropriate normalization has been performed. In the algorithm presented here, we carefully model the signal distribution of each marker and each allele. Other methods instead normalize the hybridization intensities into the summary statistic Log R ratio (LRR), the logarithm base 2 of the observed total signal intensity divided by the signal intensity of a canonical genotyping cluster for that SNP [Peiffer et al., 2006; Wang et al., 2007]. This statistic is modeled to be independent of genotype and identically distributed across markers [Colella et al., 2007; Wang et al., 2007].

To assess whether more careful modeling improves the calling of carrier status, we used the 8q24 dataset to estimate properties of the signal distribution for all genotypes. Based on the inferred carrier status, we calculated the mean and standard deviation of the hybridization intensity signal for each SNP (Fig. 2). We assessed the signal distribution of the major allele in individuals that were inferred to be homozygous or hemizygous for the major allele; we estimated the signal distribution for the minor allele in individuals that were inferred to be homozygous or hemizygous for the minor allele and we estimated the signal intensity for both alleles in heterozygous individuals. Figure 2 reveals that for all markers the signal intensity of hemizygous individuals is lower than the signal intensity of homozygous individuals, indicating that each marker provides some information about the deletion carrier status. However, individual markers do not provide sufficient resolutions to reliably infer deletion carriers. Thus, combining the information across multiple markers is required.

Moreover, the patterns of hybridization intensity are highly heterogeneous between markers and between alleles. Our results indicate large differences between the mean signal intensity at individual SNPs. In extreme cases, this leads to homozygous markers at some loci having a weaker signal than hemizygous markers at an other locus. For example, the signal intensity for homozygote at marker 6 is lower than the mean signal intensity of a hemizygous individual at marker 12.

Even within the same marker the signal distribution between the A and B allele can be markedly different. In markers 5 and 8, the signal intensity of the homozygous for the A allele is lower than the signal intensity or the hemizygous of the B allele. Clearly is not possible to use a single distribution to model the genotyping signal across all genotypes and all loci. Furthermore, the heterozygous individuals is consistently lower than the signal intensity of the hemizygous individuals 2. Hence, heterozygous individuals cannot be used to precisely estimate the signal distribution in hemizygous individuals.

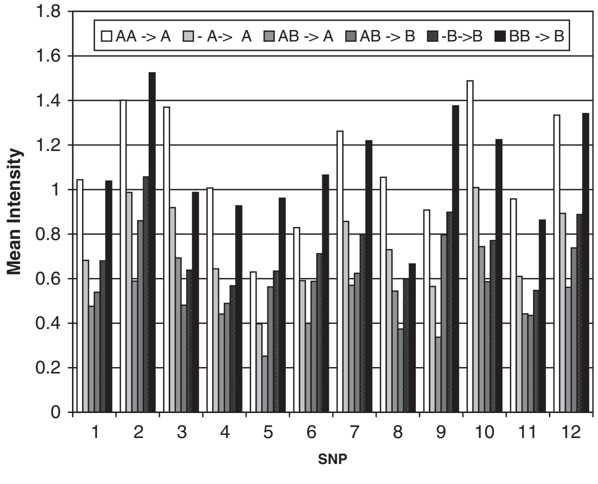

Fig. 2. Estimated mean signal intensities. For each of the 12 SNPs covered by the deletion, this figure displays the mean intensity of the normalized hybridization signal dependent on the genotype as the height of a bar. For each SNP along the horizontal axis, the first three bars indicate the mean intensity of the hybridization signal of allele *A* for *AA* homozygotes (*AA*→*A*), A-hemizygotes (-*A*→*A*) and for *AB* heterozygotes (*AB*→*A*). The next three bars show the mean intensity of the hybridization signal of the *B* allele for *AB* heterozygotes (*AB*→ *B*), *B*-hemizygotes (-*B*→*B*) and *BB* homozygotes (*BB*→*B*).

For comparison, we also calculated the mean LRR for each genotype at each marker 3, observing much less heterogeneity across markers than for genotyping signal intensity. For almost all markers, average LRR of a hemizygous individual is significantly lower than the average LRR of a heterozygous or homozygous individual. However, while models assume that LRR is independent of genotype, the average LRR between genotypes is different for different genotypes. While the expected LRR of a locus with two copies is 0, the mean LRR is consistently >0 in homozygous individuals and consistently <0 in heterozygous individuals. Furthermore, the difference between the average mean LRR of an AA homozygote is significantly different from the average LRR of a BB homozygote (two-sided *t*-test 1df, $\alpha = 0.05$) for all markers. This heterogeneity extends across markers, 60% of all pairs of markers have significantly different mean (*t*-test 1 df $\alpha = 0.05$). Particularly for markers with low minor allele frequency, the mean LRR at markers homozygote for the minor allele differ strongly from the mean LRR observed at other loci (see Fig. 3).

Nevertheless, estimate carrier status based on one mixture distribution of LRR requires estimating fewer parameters (four parameters total) than using CNVEM (eight parameters per SNP). To assess if using the more complicated model of CNVEM improves the fit to the data, we re-analyzed the 8q24 dataset using LRR rather than signal intensity. We apply the CNVEM algorithm to calculate the maximal likelihood, modeling LRR as sampled from either the two copy or the one copy distribution. To compare the fit of the 4 parameter model with the fit of the 96 parameter model, we used a likelihood ratio test. The log-likelihood under the 4 parameter model was 24,981, the log-likelihood under

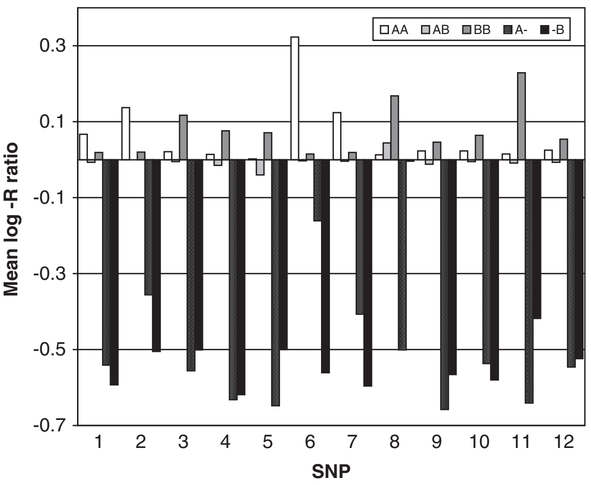

Fig. 3. Estimated mean log R ratios (LRR) of signal intensities. For each of the 12 SNPs covered by the deletion, this figure displays the mean LRR of the normalized hybridization signal dependent on the genotype as the height of a bar. For each SNP along the horizontal axis, the bars indicate the mean LRR intensity of the hybridization signal for *AA* homozygotes (*AA*), *AB* heterozygotes (*AB*), *BB* homozygotes (*BB*), *B*-hemizygotes (-*B*) and *A*-hemizygotes (-*A*).

the 96 parameter model was 38,196. Modeling the signal intensity of each marker individually thus resulted in a significantly better fit to the data ($p<10^{-100}$). Notably, using only LRR to estimate carrier status also results in several ambiguous calls with final probability of carrying the CNV neither zero or one.

APPENDIX B

EM UPDATING EQUATIONS

Here, we derive the EM updating equations used to compute the ML estimates of $(\mu,\theta)\equiv((\mu_j(g),\sigma_j^2(g)):g\in(AA,A-,BB,B-))$, $j=1,\ldots,L$. Given current estimates $(\mu^{(k)},\theta^{(k)})$, the EM updates are defined by the following recursion:

$$(\mu^{(k+1)},\theta^{(k+1)}) \equiv \operatorname{argmax} \mathbf{E}[\log \Pr(\mathbf{H},\mathbf{C}|\mathbf{D},\mu,\theta)|\mathbf{H},\mathbf{D})],$$

where the expectation is indexed by $(\mu^{(k)},\theta^{(k)})$, the maximization occurs of (μ,θ), **H** are the observed hybridization intensities, **D** are the distinct alleles, and **C** are the carrier status indicator variables (as describe in "Methods"). Now,

$$\begin{aligned}&\mathbf{E}[\log \Pr(\mathbf{H},\mathbf{C}|\mathbf{D},\mu,\theta)|\mathbf{H},\mathbf{D})]\\&\quad=\mathbf{E}\left[\sum_{\Gamma}\log \Pr(H_{ij}|C_i,D_{ij})+\log \Pr(C_i|\mathbf{D}_{i\cdot})|\mathbf{H},\mathbf{D}\right]\\&\quad=\sum_{\Gamma}\left[\sum_{C_i}(\log \Pr(H_{ij}|C_i,D_{ij})+\log \Pr(C_i|\mathbf{D}_{i\cdot})\rho_{ij}\right],\end{aligned}$$

where $\Gamma\equiv\{(i,j):D_{ij}\in\{A,B\}\}$, $\rho_{ij}\equiv\Pr(C_i\mathbf{H},\mathbf{D},\mu^{(k)},\theta^{(k)})$.

To maximize the preceding expression over (μ,θ), we must compute its gradient, set that to zero, and then solve those equations for μ and θ. Since the prior distribution of C_i given $\mathbf{D}_{i\cdot}$ is independent of (μ,θ), the gradient, denoted by ∇ is

$$\begin{aligned}&\nabla\mathbf{E}[\log \Pr(\mathbf{H},\mathbf{C}|\mathbf{D},\mu,\theta)\mathbf{H},\mathbf{D})]\\&\quad=\sum_{\Gamma}\left[\sum_{C_i}\nabla\log \Pr(H_{ij}C_i,D_{ij})\rho_{ij}\right].\end{aligned}$$

Given $G_{ij}(C_i,D_{ij})=g$, and that $H_{ij}\sim N(\mu_j(g),\theta_j(g))$, the gradient is easily managed. Hence, after equating the gradient to zero, the resulting equations for (μ,θ) are

$$0=\sum_{\{i:D_{ij}\sim g\}}(H_{ij}-\mu_j(g))\rho_{ij}(C_i\sim g),$$

$$0=\sum_{\{i:D_{ij}\sim g\}}\left[\frac{(H_{ij}-\mu_j(g))^2}{\theta_j}-1\right]\rho_{ij}(C_i\sim g),$$

for all $g\in\{AA,A-,BB,B-\}$ and for all $j=1,\ldots,L$, where $x\sim y$ indicates that x is consistent with y. By definition, the solutions to these equations are

$$\mu_j^{(k+1)}(g)=\frac{\sum_{\{i:D_{ij}\sim g\}}\rho_{ij}(C_i\sim g)H_{ij}}{\sum_{\{i:D_{ij}\sim g\}}\rho_{ij}(C_i\sim g)},$$

$$\theta_j^{(k+1)}(g)=\frac{\sum_{\{i:D_{ij}\sim g\}}[(H_{ij}-\mu_j(g))^2\rho_{ij}(C_i\sim g)]}{\sum_{\{i:D_{ij}\sim g\}}\rho_{ij}(C_i\sim g)}$$

$$\forall g \text{ and } j.$$

Note that equations (3) and (4) in "Methods" are just special cases of the general equations given here, with $g=AA$. It is well known that convergence of the EM algorithm only applies to local maxima of $L^{(c)}(\mu,\theta;\mathbf{H})$, where the superscript c indicates that the likelihood is conditional on **D**. Therefore, to increase the chance of finding the global maximum, we repeat the algorithm multiple times with different starting points each time. Then, we select the parameter estimates with the highest likelihood. In our analyses, the likelihoods appeared to be unimodal.