Original Paper

Human Heredity

Hum Hered 2011;71:141–147
DOI: 10.1159/000324683

Received: July 19, 2010
Accepted after revision: January 25, 2011
Published online: July 20, 2011

# Copy Number Variation Accuracy in Genome-Wide Association Studies

Peng Lin[a] Sarah M. Hartz[a] Jen-Chyong Wang[a] Robert F. Krueger[b]
Tatiana M. Foroud[c] Howard J. Edenberg[c] John I. Nurnberger, Jr[c]
Andrew I. Brooks[d] Jay A. Tischfield[d] Laura Almasy[e] Bradley T. Webb[f]
Victor M. Hesselbrock[g] Bernice Porjesz[h] Alison M. Goate[a] Laura J. Bierut[a]
John P. Rice[a] COGA Collaborators, COGEND Collaborators, GENEVA Investigators

[a]Department of Psychiatry, Washington University, St. Louis, Mo., [b]Department of Psychology, University of Minnesota, Minneapolis, Minn., [c]School of Medicine, Indiana University, Indianapolis, Ind., [d]Department of Genetics, Rutgers University, Piscataway, N.J., [e]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Tex., [f]Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, Va., [g]Department of Psychiatry, University of Connecticut Health Center, Farmington, Conn., and [h]The Henri Begleiter Neurodynamics Laboratory, State University of New York Downstate Medical Center, Brooklyn, N.Y., USA

**Key Words**
Accuracy · Copy number variations · False positives · Genome-wide association studies

**Abstract**

***Background/Aim:*** Copy number variations (CNVs) are a major source of alterations among individuals and are a potential risk factor in many diseases. Numerous diseases have been linked to deletions and duplications of these chromosomal segments. Data from genome-wide association studies and other microarrays may be used to identify CNVs by several different computer programs, but the reliability of the results has been questioned. ***Methods:*** To help researchers reduce the number of false-positive CNVs that need to be followed up with laboratory testing, we evaluated the relative performance of CNVPartition, PennCNV and QuantiSNP, and developed a statistical method for estimating sensitivity and positive predictive values of CNV calls and tested it on 96 duplicate samples in our dataset. ***Results:*** We found that the positive predictive rate increases with the number of probes in the CNV and the size of the CNV, with the highest positive predicted rates in CNVs of at least 500 kb and at least 100 probes. Our analysis also indicates that identifying CNVs reported by multiple programs can greatly improve the reproducibility rate and the positive predicted rate. ***Conclusion:*** Our methods can be used by investigators to identify CNVs in genome-wide data with greater reliability.

## Introduction

Copy number variations (CNVs) are duplications or deletions of a particular segment of an individual's genome. Over the past 10 years, evidence has accumulated that CNVs play an important role in disease [1–7]. It is hypothesized that a CNV changes the expression level of genes in or near those regions, leading to various phenotypes as well as diseases [8]. Therefore, CNVs constitute a major source of interindividual variation that could contribute to common disorders and complex traits [9].

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2011 S. Karger AG, Basel

Accessible online at:
www.karger.com/hhe

John P. Rice
Department of Psychiatry, Washington University School of Medicine
660 S. Euclid Ave
St. Louis, MO 63110 (USA)
Tel. +1 314 286 2572, E-Mail john@zork.wustl.edu

The advent of genome-wide association studies (GWASs) has led to the possibility of discovering CNVs across the genome. So far, many CNV detection programs have been developed for this purpose, including CNVPartition, PennCNV, and QuantiSNP.

However, despite the obvious scientific importance of understanding the role that CNVs play in human disease, there is some controversy regarding the use of GWAS data to detect CNVs. First, a recent study suggested that disease-related CNVs detected from GWAS data are well tagged by SNPs, and, therefore, CNVs do not add further information [10]. Second, there is evidence that different methods for identifying CNVs from GWAS data report different results, even when applied to the same array data [11].

To address the first controversy, although many common CNVs that are well typed in a microarray can be tagged by SNPs [10], there are at least three reasons why testing the association between a trait and CNVs remains important. First, CNVs may be the true causative variant of the trait, and will therefore show a stronger association than a SNP tag. For example, the copy number of the salivary amylase gene (*AMY1*) is positively correlated with the salivary amylase protein level [8]. Second, the number of common CNV loci is limited, and therefore, the typical GWAS significance level of $p < 5 \times 10^{-8}$ is overly conservative [12]. After adjusting for multiple tests in GWAS, SNP-tagging-associated CNVs are unlikely to be statistically significant at this stringent threshold, although they would be significant in a setting where only CNVs were tested. Third, de novo CNVs are not well tagged by SNPs. In addition, tagging a recurrent CNV by multiple SNPs demands heavy computation. Thus, despite the potential for some CNVs to be tagged by SNPs, many researchers continue to look for CNVs in GWAS data [13].

The second controversy with localizing CNVs is the imprecision of estimation. Methodologies for measurement of CNVs in GWAS microarrays continue to evolve, leading to the varied results mentioned above. Currently, most methods that make use of SNP microarray data to detect CNVs depend on log R ratio and B-allele frequency from microarray data. One simple and straightforward method draws log R ratio and B-allele frequency as the y-axis and chromosome position as the x-axis. When a deletion or duplication occurs, the pattern of log R ratio and B-allele frequency will change accordingly [14]. However, this method requires extremely high data quality and necessitates that investigators spot pattern changes. Subsequently, more sophisticated methods of identifying CNVs have attempted to adjust undesirable microarray artifacts, such as genomic waves [15], and build a mathematical model to detect CNVs from those data. Numerous programs have been written for this purpose. The most widely used are CNVPartition (http://www.illumina.com/software/illumina_connect.ilmn), PennCNV [16] and QuantiSNP [17]. Although all 3 programs use standard statistics from the observed data to estimate the location of CNVs, they use different iterative mathematical methods. CNVPartition uses a likelihood-based algorithm, PennCNV implements a hidden Markov model, and QuantiSNP uses an objective Bayes hidden-Markov model. A detailed comparison of these different algorithms can be found in the study by Dellinger et al. [11]. These 3 programs have often helped to find putative disease-related CNVs [18–22]. Moreover, several recent studies have used SNP microarray data to study the characteristics of CNVs [14, 23]. However, there is evidence that the varied algorithms identify different CNVs even with the same data, questioning the reliability of using these programs to detect CNVs [11].

Although laboratory confirmation is necessary to validate CNVs derived from SNP array platforms [2, 12, 19–22], it is not economically feasible to validate all CNVs in a genome-wide scale, especially for the purpose of estimating measurement accuracy. Here, using duplicates in a GWAS sample, we develop an algorithm to better evaluate the accuracy of CNVs predicted by several CNV calling algorithms for GWAS data. Whether a CNV that is called the first time can be confirmed the second time is restricted by sensitivity and specificity. This gives some insight about CNV calling accuracy to investigators wishing to evaluate CNVs found in SNP microarray data that might be associated with disease.

## Methods

### *Data and Quality Control*

The dataset was collected as part of the Study of Addiction: Genetics and Environment (SAGE) [24]. SAGE is part of the Gene Environment Association Studies (GENEVA) project (http://genevastudy.org/) [13]. All participants in SAGE provided written informed consent for genetic studies and agreed to share their DNA and phenotypic information for research purposes. The institutional review boards at all data collection sites granted approval for the use of the data. In this study, all samples were deidentified and only subjects who consented to health research were included.

Samples were genotyped on the Illumina Human 1M array at the Center for Inherited Disease Research at the Johns Hopkins University. The Illumina 1M array has a total of 1,072,820 probes, of which 23,812 are 'intensity-only' probes. Data cleaning procedures included using HapMap controls, detection of gender misannotation and chromosomal anomalies, cryptic relatedness, population structure, batch effects, and Mendelian and duplication error detection [24, 25]. In this study, 107 study subjects were genotyped in duplicate on the Illumina 1M array. These subjects were selected randomly from the study sample for the purpose of assessing genotyping accuracy. The mean of the SNP calling discordance rate between the duplicates was 0.02%. These duplicates were further compared against each other to determine the accuracy of CNV calling.

*CNV Calling*

We used 3 common programs to call CNVs: CNVPartition, PennCNV, and QuantiSNP. We also implemented a procedure to adjust genomic waves when we called CNVs by PennCNV and QuantiSNP [15]. Both PennCNV and QuantiSNP report data quality control measures. In order to pass the quality control, subjects and their replication need to be considered as good quality by both PennCNV and QuantiSNP. After quality control, 96 subjects and their replications passed these filters. CNVPartition does not provide any quality control information for individual subjects. We also removed all CNV calls with log Bayes factor <10, which is recommended by QuantiSNP (see online suppl. materials for more details, www.karger.com/doi/10.1159/000324683).

Each program also reports a confidence score based on different mathematical models. The confidence score is a positive number representing the likelihood that there is a CNV at that region, with a higher number representing a greater probability of a CNV in that region. The confidence scores for the 3 programs are calculated differently and are on different scales. CNVPartition uses a likelihood-based method to compute the confidence score (http://www.illumina.com/software/illumina_connect.ilmn). QuantiSNP computes a Bayes factor by comparing the evidence of the region containing deletions or duplications to that having 2 copies, and reports the log Bayes factor as the confidence [17]. PennCNV reports an experimental confidence score that is not well documented [16]. These confidence scores allow users to filter out CNV regions that are likely to be false positives. Due to variability in the confidence score distributed among the 3 programs, we converted the confidence scores within each program into percentiles and used them as covariates for modeling.

*Comparative Statistics*

These CNV calls are then compared against each other among duplicate samples. Concordance is defined as the percentage of regions that have been consistent in the existence or absence of CNVs between duplicate samples. However, this measure is misleading, because a large percentage agreement is the chance agreement of negatives.

In addition to the concordance rate, we present the reproducibility rate. We define a CNV as being reproduced when the percentage of overlap of these 2 CNVs is >30% of the region which the 2 CNVs cover. The reproducibility rate is defined as the percentage of CNVs that can be reproduced at time point 2 among CNVs that are discovered at time point 1.

*Statistical Modeling*

Whether a CNV discovered at the 1st time point can be confirmed at the 2nd time point is restricted by sensitivity and specificity. In turn, this information can be used to estimate sensitivity and specificity. Using a model derived in previous work [26–28], we calculated CNV sensitivity and a positive predicted rate with logistic regression parameters derived from CNV characteristics. All CNVs called by any program or >1 program were used to fit the model. We also added the consistency rate – the number of programs reporting a CNV at a particular locus – as a covariate. The mathematical model allows us to estimate the cumulative probability of being true for a set of CNVs with similar characteristics, and thus avoids the issue of testing whether a particular CNV is true or not.

Based on this model, we estimated the probability that an observed CNV is a true positive, and further the sensitivity for different methods. Duplications and deletions were modeled separately. The percentile of confidence scores from CNVPartition, PennCNV and QuantiSNP, as well as the consistency rate, were all significant for duplications or deletions, and thus were included in our model (see online suppl. materials for more details).

*Model Validation*

Based on our model, we were able to calculate the positive predicted rate for each CNV. We grouped CNVs with similar positive predicted rates together and compared the positive predicted rate of each group against the proportion of CNVs from that group that can be reproduced. We reported a CNV as reproduced in duplicate if the CNV detected by the 2 independent genotyping methods shares >30% of the total coverage. We were able to obtain agreement between theoretical positive predicted rate and experimental reproducibility in duplicates (online suppl. fig. S1).

We also randomly selected 90% of replicate pairs, and randomly assigned status as discovery or replication, and then we calculated the positive predicted rate for ‘any of 3 programs’. We repeated the process 100 times. The positive predicted rate was stable across many repeats (online suppl. fig. S2), indicating that our result is not subject to serious random fluctuations.

## Results

We tested the concordance rate of CNV calls from each program in duplicate samples. The concordance rates for the 3 programs range from 98.0 to 99.3% (table 1). However, concordance rate is not a good indicator of CNV calling reliability, because the concordance rate also includes the agreement of the absence of CNVs. Similar to SNPs with very low minor allele frequencies [29], a large portion of agreement is due to the chance agreement of negatives. Because of this, we believe that the reproducibility rate is a more appropriate measure for CNV calling reliability. We reported a CNV as reproduced in the duplicate if the CNV detected by the 2 independent genotyping methods shares >30% of the total coverage. The reproducibility among deletions ranged from 59 to 62%, and the reproducibility among duplications ranged from 43 to 57% (table 1). This highlights the variation between methods and the low reliability of all 3 methods.

**Table 1.** Concordance and reproducibility rates for CNVPartition, PennCNV and QuantiSNP

| | CNVPartition | | PennCNV | | QuantiSNP | |
|---|---|---|---|---|---|---|
| | concordance | reproducibility[a] | concordance | reproducibility[b] | concordance | reproducibility[c] |
| Duplication | 99.2% | 54% | 98.0% | 41% | 98.9% | 48% |
| Deletion | 99.3% | 61% | 98.6% | 62% | 98.9% | 63% |

[a] Reproduced by CNVPartition. [b] Reproduced by PennCNV. [c] Reproduced by QuantiSNP.

**Table 2.** Positive predicted rate $R_+$, sensitivity p' and total number of CNVs for different CNV calling methods

| Method | Duplication | | | | Deletion | | | |
|---|---|---|---|---|---|---|---|---|
| | p' | $R_+$ | reproducibility rate[a] | total CNVs, n | p' | $R_+$ | reproducibility rate[a] | total CNVs, n |
| CNVPartition | 0.77 | 0.69 | 0.63 | 849 | 0.75 | 0.78 | 0.77 | 2,227 |
| PennCNV | 0.92 | 0.46 | 0.41 | 2,001 | 0.94 | 0.65 | 0.64 | 2,348 |
| QuantiSNP | 0.83 | 0.58 | 0.55 | 1,177 | 0.91 | 0.69 | 0.68 | 4,171 |
| Any of 3 programs | 0.94 | 0.43 | 0.40 | 2,199 | 0.96 | 0.59 | 0.59 | 5,767 |
| Any 2 of 3 programs | 0.82 | 0.61 | 0.56 | 1,169 | 0.88 | 0.76 | 0.75 | 3,565 |
| All 3 programs | 0.75 | 0.79 | 0.73 | 642 | 0.72 | 0.89 | 0.85 | 1,816 |

[a] Reproduced by any one of the 3 programs (CNVPartition, PennCNV, or QuantiSNP).

We then estimated the reproducibility rate, the positive predicted rate and the sensitivity for each CNV calling method (table 2). As expected, deletions have higher reproducibility rates, higher positive predictive rates and better sensitivity. For both duplications and deletions, the method that requires CNVs to be reported by all 3 programs has the highest reproducibility rate and the highest positive predicted rate.

False CNV calling may be caused by intensity variation (noise) from the microarray. A short CNV segment with few probes is particularly vulnerable to noise. Because of this, we estimated both the reproducibility rate and the positive predicted rate $R_+$ within four subcategories for each method based upon the number of probes contained within the CNV (table 3). Some of these subcategories are often used in the literature as thresholds for quality controls [16]. Not surprisingly, a higher positive predicted rate $R_+$ was seen when there were more probes in a single CNV. We also tested the relationship between the size of CNV segments and positive predicted rate $R_+$ (online suppl. table S1). As expected, the result was similar to table 3, because a larger CNV segment typically contains more probes.

The primary purpose of this study was to determine the reliability of CNVs found in microarrays, such as in GWAS. We found that if a CNV is reported by all 3 programs, it has the highest positive predicted rate. Moreover, in a microarray, probes are not always evenly spaced. We hypothesized that the combination of the number of probes and the size would boost the positive predicted rate. We tested this hypothesis using both the number of probes and the size as filters. The results suggest that a minimum of 10 probes and 10-kb pairs are necessary to reach a positive predicted rate >80% (table 4).

## Discussion

Data from GWASs can be used to estimate locations of CNVs and their potential effects on disease. There is disturbing evidence that calling CNVs from SNP microarray data is not reliable [11]. For this reason, investigators are interested in quantifying the reliability. To our knowledge, this is the first study that compares CNV calls from a considerable number of duplicate samples.

**Table 3.** The positive predicted rate $R_+$ within subcategories defined by the number of probes (from <10 to ≥100)

| | Duplication | | | | Deletion | | | |
|---|---|---|---|---|---|---|---|---|
| | <10 | 10–50 | 50–100 | ≥ 100 | <10 | 10–50 | 50–100 | ≥ 100 |
| CNVPartition | 0.54 (150) | 0.69 (486) | 0.77 (143) | 0.88 (84) | 0.70 (1,176) | 0.85 (974) | 0.87 (82) | 0.88 (54) |
| PennCNV | 0.29 (1,009) | 0.56 (930) | 0.88 (117) | 0.95 (39) | 0.58 (3,433) | 0.81 (1,334) | 0.95 (74) | 0.99 (45) |
| QuantiSNP | 0.42 (228) | 0.57 (757) | 0.70 (163) | 0.84 (77) | 0.63 (2,411) | 0.74 (1,644) | 0.74 (178) | 0.89 (81) |
| All 3 programs | 0.65 (122) | 0.77 (461) | 0.87 (157) | 0.91 (83) | 0.84 (993) | 0.92 (1,013) | 0.96 (101) | 0.98 (67) |

The total number of CNVs is indicated in parentheses. In calculating the total number of CNVs, we included a CNV for a certain category if the report from any one of the specified programs satisfied this category.

**Table 4.** The positive predicted rate $R_+$ for the '3 of 3 method'

| Probes | Duplication size | | | | | Deletion size | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | <1 kb | 1–10 kb | 10–100 kb | 100–500 kb | ≥ 500 kb | <1 kb | 1–10 kb | 10–100 kb | 100–500 kb | ≥ 500 kb |
| <10 | 0.58 (3) | 0.68 (39) | 0.64 (81) | 0.62 (6) | – | 0.80 (47) | 0.83 (595) | 0.84 (426) | 0.76 (9) | – |
| 10–50 | – | 0.72 (20) | 0.79 (250) | 0.74 (239) | 0.72 (20) | – | 0.91 (144) | 0.93 (731) | 0.91 (247) | 0.86 (24) |
| 50–100 | – | – | 0.89 (19) | 0.91 (73) | 0.82 (74) | – | – | 0.95 (39) | 0.98 (45) | 0.96 (24) |
| >100 | – | – | 0.79 (2) | 0.96 (29) | 0.90 (56) | – | – | 0.99 (3) | 0.99 (32) | 0.97 (37) |

The total number of CNVs is indicated in parentheses. In calculating the total number of CNVs, we included a CNV for a certain category if the report from any one of the specified programs satisfied this category.

Although experimental validation is necessary for CNV association studies, it is both demanding and costly and should be limited to regions most likely to contain true CNVs associated with disease. In this study, we introduced a convenient way to identify potential false-positive CNVs on a genome-wide scale, using an estimated positive predicted rate for CNV callings. Our results confirmed that combining CNVs from different programs is one way to improve the positive predicted rate.

In this study, we found that 10 probes and 10 kb in size maximize CNV calling quality. We also discovered that deletions are much easier to detect than duplications. The reason is that when calling genotypes from the microarray, 1 deletion represents a 50% decrease in signal intensity, rather than the 33% increase caused by 1 duplication. In addition, B-allele frequencies – a reported measure from microarrays – of those SNPs at a particular deletion region usually take the value of 0 or 100%, leading to a distinctive pattern that is relatively easy to spot.

Different methods for estimating the locations of CNVs use different mathematical models. Both PennCNV and QuantiSNP use hidden Markov models [16, 17], while CNVPartition estimates model parameters using bivariate Gaussian distributions. Each method has its own strengths, but all also have relatively high frequencies of false-positive CNVs. The '3 of 3' method, however, minimizes false positives.

When 3 different programs call the same CNV, different boundaries may be reported, leading to a quandary on how to categorize this particular CNV. To resolve this, we included all CNVs for one category if a CNV reported by any program satisfies the category. Therefore, the total number of CNVs for '3 of 3 programs' may be higher than the total number of CNVs reported by each program alone.

Moreover, the reproducibility in our paper is defined either as being reproduced by itself or being reproduced by any of the 3 algorithms. The exact definition is indicated in tables 1 and 2. The reason for this is to address both self-reproducibility and across-the-spectrum reproducibility. In table 1, we adopted 'being reproduced by itself' as the criterion in order to show self-reproducibility. That is because self-reproducibility is a good indicator of reliability when the truth is not known, and also a good point to start with. The fact that a program cannot even reproduce its result is surely a good sign of poor reliability. In table 2, we want to compare the reproducibility among the 3 algorithms and the 3 combinational methods, therefore, a consistent criterion, which is across-the-spectrum reproducibility for this table, is needed in order to make the comparison fair and meaningful.

The sensitivity here is restricted to CNVs that can be detected by a microarray. In our data from 96 subjects and their replications, we identified 2,348 potential regions across the genome for deletions and 851 potential regions for duplications. For any particular potential region, at least 1 of these 96 subjects had a duplication or deletion in this region. Among these regions, the true base rate *k* is 0.016 for deletions and 0.012 for duplications (see online suppl. materials). We restricted our study only to these potential regions. Some CNVs in the genome may be located at particular regions where no probes or very few probes exist. Those CNVs can never be detected by microarray technology, and therefore are excluded from the estimation of sensitivity. The sensitivity here may be better understood as the sensitivity adjusted by the total number of those potential CNV regions. Therefore, the sensitivity reported by our study should not be directly compared to other studies [11, 30].

Based on our model parameters, investigators can estimate the probability that an estimated CNV is true. Interested researchers can estimate the positive predicted rate for their own data if confidence scores and some other information can be provided. Finally, it is important to emphasize that there are benefits to be gained from utilizing multiple CNV calling approaches and then comparing the results between them. This can maximize the sensitivity for discovery, maximize the positive predicted rate for verification, or balance the sensitivity and the positive predicted rate to a desired point. As GWASs move forward from SNPs to CNVs, investigators can better identify CNVs associated with human disease using multiple estimation programs and calculating the positive predictive rates of observed CNVs.

## Acknowledgments

Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI; U01 HG004422). SAGE is one of the GWASs funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for the collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (R01 DA013423, and R01 DA019963). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract 'High throughput genotyping for studying the genetic contributions to human disease' (HHSN268200782096C).

## References

1 Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C: Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 2006;38:1372–1374.

2 Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK: The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434–1440.

3 Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT: Copy number polymorphism in FCGR3 predisposes to glomerulonephritis in rats and humans. Nature 2006;439:851–855.

4 Padiath QS, Saigoh K, Schiffmann R, Asahara H, Yamada T, Koeppen A, Hogan K, Ptacek LJ, Fu YH: Lamin B1 duplications cause autosomal dominant leukodystrophy. Nat Genet 2006;38:1114–1123.

5 Lee JA, Lupski JR: Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 2006;52:103–121.

6 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 YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M: Strong association of de novo copy number mutations with autism. Science 2007;316:445–449.

7 Fan YS, Jayakar P, Zhu H, Barbouth D, Sacharow S, Morales A, Carver V, Benke P, Mundy P, Elsas LJ: Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. Hum Mutat 2007;28:1124–1132.

8 Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC: Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007;39:1256–1260.

9 Beckmann JS, Estivill X, Antonarakis SE: Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 2007;8:639–646.

10 Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, et al.: Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713–720.

11 Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ: Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 2010; 38:e105.

12 Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME: Origins and functional impact of copy number variation in the human genome. Nature 2010;464:704–712.

13 Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF, Feenstra B, Feingold E, Fornage M, Haiman CA, Harris EL, Hayes MG, Heit JA, Hu FB, Kang JH, Laurie CC, Ling H, Manolio TA, Marazita ML, Mathias RA, Mirel DB, Paschall J, Pasquale LR, Pugh EW, Rice JP, Udren J, van Dam RM, Wang X, Wiggs JL, Williams K, u K: The Gene, Environment Association Studies Consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 2010;34:364–372.

14 Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE: Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 2009;84:148–161.

15 Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008;36:e126.

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

17 Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J: QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35:2013–2025.

18 Wain LV, Pedroso I, Landers JE, Breen G, Shaw CE, Leigh PN, Brown RH, Tobin MD, Al-Chalabi A: The role of copy number variation in susceptibility to amyotrophic lateral sclerosis: genome-wide association study and comparison with published loci. PLoS One 2009;4:e8175.

19 Breitling LP, Dahmen N, Mittelstrass K, Illig T, Rujescu D, Raum E, Winterer G, Brenner H: Smoking cessation and variations in nicotinic acetylcholine receptor subunits alpha-5, alpha-3, and beta-4 genes. Biol Psychiatry 2009;65:691–695.

20 Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI, Alvarez Retuerto AI, Imielinski M, Hadley D, Bradfield JP, Kim C, Gidaya NB, Lindquist I, Hutman T, Sigman M, Kustanovich V, Lajonchere CM, Singleton A, Kim J, Wassink TH, McMahon WM, Owley T, Sweeney JA, Coon H, Nurnberger JI, Li M, Cantor RM, Minshew NJ, Sutcliffe JS, Cook EH, Dawson G, Buxbaum JD, Grant SF, Schellenberg GD, Geschwind DH, Hakonarson H: Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. PLoS Genet 2009;5:e1000536.

21 Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM: Copy number variation at 1q21.1 associated with neuroblastoma. Nature 2009;459:987–991.

22 Bowden W, Skorupski J, Kovanci E, Rajkovic A: Detection of novel copy number variants in uterine leiomyomas using high-resolution SNP arrays. Mol Hum Reprod 2009;15:563–568.

23 Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Deng H, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Recker RR, Cheng J, Deng HW: Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS One 2009;4:e7958.

24 Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S, Hinrichs AL, Almasy L, Breslau N, Culverhouse RC, Dick DM, Edenberg HJ, Foroud T, Grucza RA, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Krueger RF, Kuperman S, Lynskey M, Mann K, Neuman RJ, Nothen MM, Nurnberger JI Jr, Porjesz B, Ridinger M, Saccone NL, Saccone SF, Schuckit MA, Tischfield JA, Wang JC, Rietschel M, Goate AM, Rice JP: A genome-wide association study of alcohol dependence. Proc Natl Acad Sci USA 2010;107: 5082–5087.

25 Laurie C, Bierut L, Bhangale T, Boehm F, Caporaso N, Doheny K, Gabriel S, Harris E, Hu F, Jacobs K, Kraft P, Landi M, Manolio T, McHugh C, Mirel D, Pugh E, Rice J, Weir BS, The GENEVA Investigators: Genotype data cleaning for whole-genome association studies, in preparation.

26 Rice JP, Endicott J, Knesevich MA, Rochberg N: The estimation of diagnostic sensitivity using stability data: an application to major depressive disorder. J Psychiatr Res 1987;21: 337–345.

27 Rice JP, McDonald-Scott P, Endicott J, Coryell W, Grove WM, Keller MB, Altis D: The stability of diagnosis with an application to bipolar II disorder. Psychiatry Res 1986;19: 285–296.

28 Rice JP, Rochberg N, Endicott J, Lavori PW, Miller C: Stability of psychiatric diagnoses. An application to the affective disorders. Arch Gen Psychiatry 1992;49:824–830.

29 Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, Goate AM, Bierut LJ, Rice JP: A new statistic to evaluate imputation reliability. PLoS One 2010;5:e9697.

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