Human Molecular Genetics, 2008, Vol. 17, Review Issue 2 ***R135–R142***
doi:10.1093/hmg/ddn282

Extending genome-wide association studies to copy-number variation

Steven A. McCarroll[1,2,3,*]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [2]Department of Molecular Biology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA and [3]Harvard Medical School, Boston, MA, USA

Received August 27, 2008; Revised and Accepted September 2, 2008

Appreciating the contribution of human genome copy-number variation (CNV) to clinical phenotypes is one of the compelling genetics challenges of the coming years. It is increasingly possible to pursue such investigations as an extension of genome-wide association studies (GWAS), enabled by innovations in the design and analysis of SNP (single nucleotide polymorphism) arrays and by progress in determining the genomic locations and population-genetic properties of the CNVs that segregate in the human population. Extensions of GWAS to CNV have already resulted in discoveries of both *de novo* and inherited CNV that are associated with risk of common disease. This review will discuss new approaches, recent findings and the analytical challenges involved in expanding GWAS to appreciate the contribution of CNV to human phenotypes.

INTRODUCTION

Genome-wide association studies (GWAS) have made hundreds of connections between clinical phenotypes and common sequence polymorphisms, each implicating a region of the human genome as playing a causal role in a disease. And yet for most common diseases, these discoveries collectively explain only a modest fraction (2–15%) of heritable variation in disease risk. One of the compelling challenges facing the next several years of human genetics is therefore to explain what accounts for the rest of heritable variation in phenotypes.

The human genome shows extensive copy-number variation (CNV), the presence of variable numbers of copies of large, multi-kilobase genomic regions in the genomes of different individuals (1–9). CNV could in principle account for a significant component of variation in disease risk. The overlap between copy-number variants (CNVs) and genes, the correlation of CNVs with gene expression levels (10), and the association of specific CNVs with clinical phenotypes (11–14) make it reasonable to hypothesize that CNV accounts for an appreciable component of human phenotypic variation. In the coming years, this hypothesis can be explored with the same genome-wide rigor and sensitivity with which common single nucleotide polymorphisms (SNPs) are now evaluated. GWAS are a good venue for such investigations: the large, well-curated patient cohorts collected for GWAS are well-suited to additional genetic analyses; the SNP arrays used to perform GWAS increasingly yield data that support CNV analysis in the same patients; and the SNP data from GWAS enable integrated analysis of SNPs and CNVs.

CNV POPULATION GENETICS

Appreciating the basic population-genetic properties of CNV is critical for designing whole-genome approaches for analyzing CNV in disease.

One central question has involved the extent to which CNVs are inherited versus arising as new mutations. Early studies of CNV in normal individuals referred to common CNVs as 'recurring' and frequently assumed independent mutational origins. However, subsequent studies, which typed a few CNVs in trios, indicated that these CNVs showed mendelian inheritance (6,7,15). More recently, a large-scale, high-resolution CNV study found that, when accurately typed, CNVs in normal individuals corresponded overwhelmingly to a model of mendelian inheritance of stable polymorphisms: $<1\%$ of the copy-number differences between any two individuals could not be explained by the simple inheritance of the same allele from a parent (16).

*To whom correspondence should be addressed at: Broad Institute of MIT and Harvard, 7 Cambridge Center, Room 6145, Cambridge, MA 02142, USA. Tel: +1 617 252 1902; Fax: +1 617 643 3293; Email: smccarro@broad.mit.edu

Another question involves the extent to which inherited CNV arises from common polymorphisms versus rare variants. Early surveys of CNV-containing regions (CNVRs), generally with techniques that had a resolution of fosmids and BACs (tens to hundreds of kilobases), frequently observed that overlapping CNV regions were affected in many individuals (1–3,5–7); the extent to which these resulted from the same polymorphism (versus a heterogeneous group of variants within the same large genomic region) was until recently unclear. Higher-resolution approaches now suggest that the great majority (90% or so) of such CNV regions are explained by copy-number polymorphisms (CNPs) in which the same sequence appears to be affected in each person, with some exceptional loci at which patterns are more complex (16,17). The relative contribution of rare and common variants to genetic variation can be measured as a fraction of the number of loci that differ in copy-number between any two unrelated individuals. In a recent analysis, >90% of the loci observed to differ in copy-number between pairs of individuals involved CNPs (those CNVs that segregate at an allele frequency >1%), and ~80% involved common CNPs (with minor allele frequency >5%) (16). This indicates that a large fraction of the copy-number differences between any two individuals arise from a limited set of common polymorphisms (16), analogous to an earlier observation that the largest component of human sequence variation (at fine scale) arises from common SNPs.

The CNV present in a study cohort will therefore consist of subsets of CNVs with different statistical properties and different propensities to affect heritable, familial and sporadic disease: common CNPs, rare CNVs and *de novo* copy-number mutations. The same can be said of fine-scale sequence variation, which includes common SNPs, rare sequence variants and an unknown number (estimated to be several dozen) of new sequence mutations in each person. For current GWAS on SNP arrays, though, there is a critical difference between fine-scale sequence variation and CNV: while current SNP array platforms ascertain only a pre-selected set of common sequence polymorphisms, the data from such platforms can in principle be used to identify common, rare, and *de novo* CNVs.

IMPROVEMENTS IN GENOTYPING PLATFORMS AND ANALYSIS METHODS

CNVs can perturb the collection of SNP data at a CNV locus by causing SNP intensity data to cluster poorly and to yield genotypes that appear to violate mendelian inheritance and Hardy–Weinberg equilibrium (5–7). For these reasons, the processes by which early commercial SNP arrays were designed—which involved evaluating potential SNP assays on screening arrays, then selecting high-performing assays to place on a commercial product—were later hypothesized to have the effect of excluding assays from many CNV loci. Comparison of a high-resolution map of segregating CNPs with the locations of SNPs on SNP arrays indicates that common CNPs (those CNPs that segregate at an allele frequency >5%) generally correspond to bald spots in the physical coverage of early SNP arrays, but that low-frequency CNPs and rare CNVs were covered at approximately the same density as the genome as a whole (16).

A newer generation of SNP arrays that include dedicated CNV content appear to have addressed this deficit. In one approach, we and collaborators at Affymetrix developed hybrid arrays consisting of a combination of SNP assays and 'copy-number' probes – non-polymorphic probes that are optimized for copy-number measurement, unconstrained by the locations of SNPs, and used to target regions of known and likely CNV (16). In another approach, developed by Illumina and DeCode Genetics, assays for SNPs within predicted and potential CNV regions were also added to genotyping arrays regardless of whether these SNP assays passed traditional QC criteria. Both approaches appear to have successfully gained physical access to the regions affected by common CNPs, yielding access to the majority of large- and intermediate-size (>5 kb), common CNPs (16). Although this represents great progress over earlier SNP arrays, the limitations of current SNP arrays should be kept in mind: current platforms have limited power to detect smaller CNVs (<20 kb), CNV in the genome's most duplication-rich corners (which may be hotspots for new mutation), and CNV in 'novel' regions of the human genome that are not part of the human reference sequence (9).

Progress has also been made in the development of algorithms for analyzing CNV. Although 'copy-number analysis' is frequently described as a single entity, an emerging approach is to treat common CNPs (which are present in all study cohorts) separately from the rare and novel CNVs that may be unique to each specific study cohort or patient. This approach represents a departure from earlier approaches for copy-number analysis, which treated all CNV analysis as a problem of *ab initio* discovery in each sample. Some new algorithms (18–20) treat 'CNP genotyping' as a distinct problem, defined not by *ab initio* discovery but rather by correct classification (or clustering) of each individual's copy-number state at each CNP locus. As the locations of segregating CNPs become known at ever-improving levels of precision—a process that is continuing with high-resolution arrays, complete resequencing of fosmids that contain CNV alleles (9), and analysis of whole-genome sequence (8) in many individuals—CNP genotyping can be supported by ever-better maps of the locations of common CNPs, and by the design of array platforms to target those CNPs.

GWAS, SPORADIC DISEASE AND *DE NOVO* CNV

Genomic disorders are sporadic disease occurrences caused by *de novo* structural mutations. The underlying mutations in many genomic disorders were identified over the past 20 years, with multiple loci identified as sites of recurring deletions and duplications that cause severe congenital and developmental phenotypes (21).

Even in common, generally heritable diseases of the type studied in GWAS, a subset of affected individuals might derive their affected status from a new mutation. This might be particularly true of diseases for which affected individuals have on average fewer children than unaffected individuals do—such as schizophrenia and severe forms of autism—since for such diseases to remain in the population, the pool of causal alleles would have to be replenished by recurring mutation.

The hypothesis that new structural mutations might contribute to the incidence of autism and schizophrenia was supported by findings that *de novo* copy-number mutations

Table 1. Extensions of GWAS to discover CNV–disease associations

Disease	Analysis approach	Locus	Type of CNV	Size (kb)	Frequency in population	Frequency in Cases	Effect size (OR)	References
Autism	Copy-number analysis of SNP array data	16p11.2	*De novo* deletion	593	1×10^{-4}	1%	100	(27)
Autism	Copy-number analysis of SNP array data	16p11.2	*De novo* duplication	593	3×10^{-4}	0.5%	16	(27)
Schizophrenia	Copy-number analysis of SNP array data	1q21.1	*De novo* deletion	1350	2×10^{-4}	0.3%	15	(25,26)
Schizophrenia	Copy-number analysis of SNP array data	15q13.3	*De novo* deletion	1580	2×10^{-4}	0.2%	12	(25,26)
Schizophrenia	Copy-number analysis of SNP array data	15q11.1	*De novo* deletion	470	0.2%	0.5%	2.7	(25)
Crohn's disease	SNP GWAS + SNP-CNP LD	*IRGM*	Inherited deletion polymorphism	20	7%	10%	1.5	(32)

Note that the *de novo* deletions and duplications above may also be inherited, though *de novo* mutation appeared to explain most or all of the cases in which inheritance could be evaluated. For the schizophrenia findings, frequency and effect size are estimated from the data in Ref. (25); for the autism findings, frequency and effect size are estimated from the replication cohorts in Ref. (27).

(regardless of their genomic location) are observed in a larger fraction of affected than unaffected individuals, particularly in the sporadic form of these diseases (22–24). These 'genomic burden' data implied that an unknown subset of the *de novo* CNVs identified were likely to be causal, and that identification of recurring hits at specific loci in larger cohorts might identify specific, causal loci.

Two large GWAS in schizophrenia recently identified novel schizophrenia genomic disorders involving recurring deletions at 1q21.1, 15q13, 3 and 15q11.1 (as well as confirming the known genomic disorder at the VCFS locus) (Table 1). One of the studies (25) used 2160 trios and 5558 parent–offspring pairs to identify *de novo* deletions as regions in which copy-number losses were observed in the offspring and excluded in both parents, or as regions that appeared to lack transmission of SNP alleles from parent to offspring. Researchers in the other study (26) searched for genomic regions in 3391 affected individuals in which deletions were present in multiple affected individuals but vanishingly rare in the unaffected population. The convergence of these studies on three of the same loci (1q21.1, 15q13.3, VCFS), together with data suggesting the deletions to be approximately as rare in the general population as the rate at which they arise as *de novo* events (25), indicates that the contribution of these deletions to schizophrenia is mostly (though perhaps not exclusively) through sporadic mutation.

In autism, copy-number analysis of GWAS data from 751 multiplex families led to the identification of a recurring microdeletion/duplication of a 493 kb segment at 16p11.2 that was detected in ~1% of autistic individuals; the deletion appears to be very highly penetrant for autism (27) (Table 1). Intriguingly, the study was able to uncover this recurring microdeletion/duplication syndrome despite a study design (trios with multiple affected offspring) that favored the discovery of inherited variants. In one family, this appears to be because the mutation was mosaic in the parental germline and transmitted to multiple offspring. In another family, the duplication event (which appears to be less penetrant than the deletion) was transmitted from a healthy parent. In other families, the mutation was *de novo* and present in only one of the affected offspring.

Although the heritability of common diseases has motivated their study in GWAS, the earlier discoveries largely involve extension of the class of non-inherited genomic disorders to include a subset of the patients with common, generally heritable diseases such as autism and schizophrenia. These discoveries leave unexplained the mysteries of (i) why these diseases are so heritable, and (ii) how much the bulk of human CNV—which is overwhelmingly inherited rather than *de novo*—contributes to disease. We next consider the emerging problem of designing genome-wide studies to appreciate the contribution of inherited CNV to clinical phenotypes.

TOWARD GWAS FOR INHERITED CNV

Inherited CNV presents different scientific opportunities and analytical challenges than *de novo* CNV (Table 2, Fig. 1):

Size of CNV events. The CNVs implicated in sporadic genomic disorders thus far have been 0.5–3 megabases in length (though it seems likely that additional genomic disorders due to smaller *de novo* CNVs have not yet been discovered). In contrast, the reservoir of inherited CNP appears, when analyzed at high resolution and with appropriate analysis methods, to contain only a few dozen segregating CNPs >100 kb (16).

Ascertainment on SNP arrays. Early SNP arrays had a severe design bias against including SNPs from the genomic segments affected by common CNPs, which made most common CNPs all but undetectable until the recent generation of SNP arrays.

Genotyping. As CNPs segregate at an appreciable frequency, with either allele potentially appearing in the homozygous state (and with the most common copy-number state often being greater or less than two), and because ~10% of CNPs appear to be multiallelic (16)—with three or more haplotypic copy-numbers segregating in the population—individuals can vary in copy-number across ranges (e.g. 0–2; or 2–4; or 0–4; or 2–8) that are not captured by simple description as a 'gain' or 'loss' relative to a 'normal' reference. Determining the disease association of common CNPs requires accurate resolution of all of the discrete copy-number levels that are present among individuals in a study cohort (14,28).

Table 2. Analyzing common, rare and *de novo* CNV in GWAS

	Common CNPs	Rare, inherited CNVs	Rare, *de novo* CNVs
Mechanism	Inherited	Inherited	New mutation
Component of disease burden explained	Heritable (familial)	Heritable (familial)	Sporadic
Types of disease	May be most relevant to common, late-onset diseases		May be most relevant to diseases of reduced fecundity and negative selection, such as mental disorders
Allelic state of patients	Because common, frequently homozygous and can give rise to three or more common CN levels in the population	Almost always heterozygous because variant is rare or a new mutation	
Suggested ascertainment strategy	Use of high-quality prior information about CNP locations; do not need to be discovered *ab initio* in each group of patients	*Ab initio* discovery using a stringent genome-wide significance threshold	
Suggested measurement strategy	Genotyping-like approaches to determine integer copy-number in each patient	Almost always single-copy gains and losses	
Suggested association analysis	Difference in allele frequency between Cases and Controls	Enrichment of a collection of rare variants in Cases or Controls	

Effect size. Relative to *de novo* mutation, which can arise despite intense negative selection and may therefore be highly or completely penetrant even for debilitating disease, one might expect the reservoir of inherited polymorphism to have more-modest effects on disease risk, particularly given the distribution of effect sizes that have been uncovered in GWAS. At the same time, the appreciable allele frequency of inherited CNPs should give GWAS ample statistical power for detecting such effects.

Confounds. As *de novo* mutations are independent mutational events, they can be analyzed in straightforward ways without fear of confounding by population structure and family relationships. Associations to inherited variation are confounded by population structure and cryptic relatedness and require additional analyses.

Genetic interpretation of a result. The large size of the CNVs involved in genomic disorders can make it difficult to identify the specific gene(s) relevant to the phenotype (Fig. 2A). Inherited CNPs are generally much smaller, such that in most cases, only one or a few genes will be implicated. However, the genetic interpretation of the disease association of inherited CNPs will prove complex for different reasons (Fig. 2B and C). As common CNPs are often in linkage disequilibrium (LD) with other polymorphisms (4,6,7,15,16), and because inherited, rare CNVs are often present on long, shared haplotypes (16), distinguishing the causal variant from other variants on the same haplotype will require integrated analysis of SNP and CNV data at the implicated locus.

COMMON, INHERITED CNPS

Approximately 80% of the copy-number differences between any two individuals appear to arise from common CNPs that segregate at an allele frequency $>5\%$, and $>90\%$ appear to arise from CNPs that segregate at an allele frequency $>1\%$ (16). Assessing the disease association of CNPs can therefore capture a large component of human CNV, and can utilize many of the association analytical frameworks that have already been developed for the SNPs in GWAS (Fig. 1B).

Three innovations increasingly make it possible to analyze common CNPs for association in GWAS. First, SNP arrays have been redesigned to eliminate much of the design bias against genomic segments affected by common CNPs, such that data are now collected from the majority of CNPs >5 kb (16,20). Secondly, high-resolution maps of the locations of CNPs increasingly make it possible to specify which probes interrogate each common CNP (16,20). Thirdly, new algorithms treat CNP genotyping as an explicit problem (18–20) and appear to yield more-accurate data at CNP loci than *ab initio* algorithms do (18).

This progress notwithstanding, genome-wide analysis of CNP–disease association is likely to be fraught with challenges and potential pitfalls in the short term. CNP-genotyping assays are much less mature than SNP genotyping assays: the high-resolution locations of common CNPs were often unknown at the time that array platforms were designed, with the result that CNP genotyping assays were not pre-screened or even pre-designed. As a result, CNP genotyping assays show a broad distribution of data quality: the data for many CNPs cluster into clear classes and are easily genotyped, while others are only partially resolved. SNP and CNP assays in which genotype classes are poorly resolved are prone to 'differential bias', in which differences in the origin and handling of samples (typically confounded by Case/Control status) give rise to patterns in the data that result in false association with phenotype (19,29). Such bias appears to be pervasive in published CNV data sets (19); one recent study proposes an approach for addressing differential bias by integrating genotyping with association testing (19). Regardless of the approach used, it will be critical to both

- examine the genome-wide distribution of an association test statistic for evidence of inflation, and
- examine the quality of the raw data underlying any putative CNP–disease association, to determine the extent to which genotyping results are supported by clear, unambiguous categories in the underlying data.

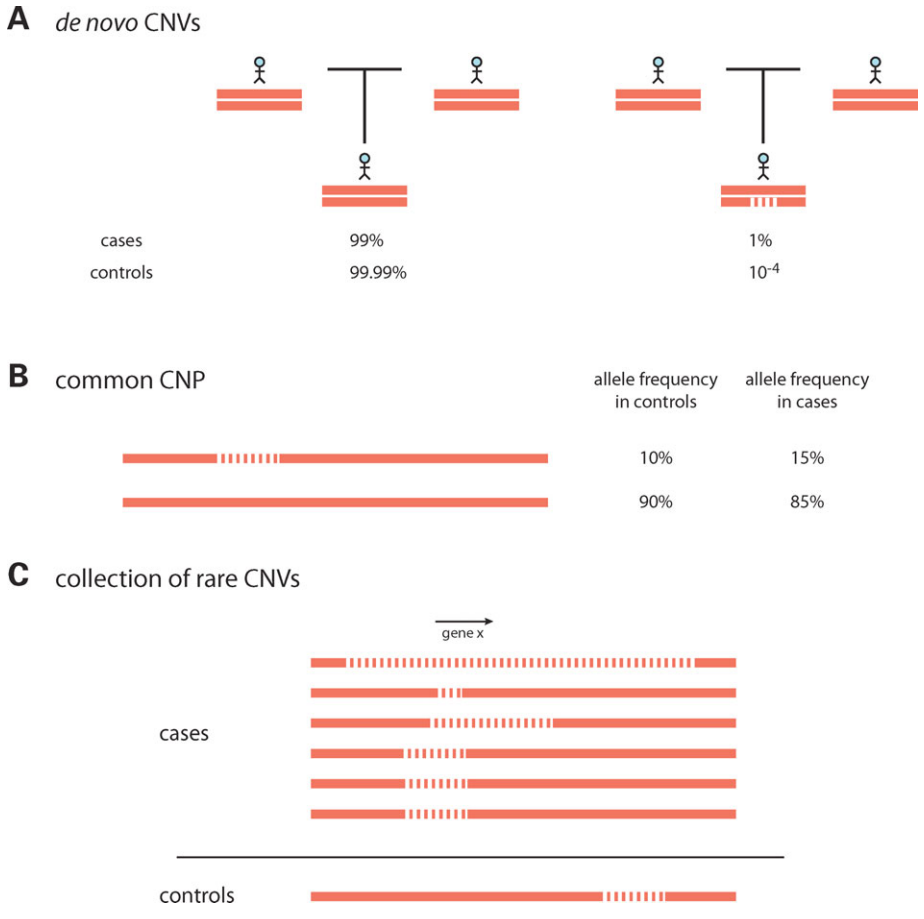

Figure 1. Scenarios of association for *de novo* CNVs, common CNPs, and collections of individually rare CNVs. Vertical stripes on each haplotype indicate the location of a CNV. (**A**) The occurrence of *de novo* CNV at a specific locus may be vanishingly rare in the general population but occur at an appreciable frequency among those affected with a disease. (**B**) For common CNPs, one allele or genotype may be significantly more common among affected than unaffected individuals, similar to GWAS association analysis for SNPs. (**C**) Association analysis of rare CNVs may in some cases involve analyzing sets of individually rare CNVs that are grouped into collections using clear, *a priori* criteria, such as impact on the same protein-coding gene.

Many CNPs segregate with different allele frequencies in different populations, a phenomenon that (as measured by F_{st}) appears to resemble the allele-frequency differentiation of SNPs and is therefore likely to represent drift in allele frequencies in reproductively isolated populations (16). It will therefore be critical to evaluate each study cohort for population structure, an analysis that is extremely powerful when informed by genome-wide SNP genotypes (30) and that may also be possible (in a more-limited form) using CNV data (7).

A lively debate has surrounded the extent to which CNPs are in LD with SNPs (4–7,15,31). The largest empirical analysis, based on integer genotypes for hundreds of common CNPs (those CNPs that segregate at allele frequencies of $\geq$5%), found that common CNPs were almost as well-tagged as SNPs of the same frequency (16). One implication of this result is that the disease association of many (though by no means all) common CNPs could also be assessed by combining GWAS SNP data with a map of the LD relationships between SNPs and CNPs. Such an approach might be particularly useful for data from first-generation GWAS platforms from which the genomic regions affected by common CNPs were substantially excluded.

This principle was recently used to identify the association with Crohn's disease of a common, 20 kb deletion polymorphism immediately upstream of *IRGM* (32) (Table 1). The deletion polymorphism is in perfect LD with SNPs at *IRGM* that were previously found (33,34) to be associated with Crohn's; the deletion was also directly associated with Crohn's disease in an independent patient cohort (32). Supporting the possibility that this upstream deletion is functionally relevant, the deletion and reference haplotypes of *IRGM* are expressed in different cell types (32).

Given the strong LD between SNPs and many CNPs, interpreting a disease–CNP association will frequently require analysis of the disease association of surrounding sequence polymorphisms (Fig. 2B). At *IRGM*, the strong LD between the deletion and surrounding SNPs means that no single variant has been determined to be the causal variant based on genetic association alone (32); the definitive identification of a single causal variant will require additional work and possibly functional studies that consider each variant in isolation.

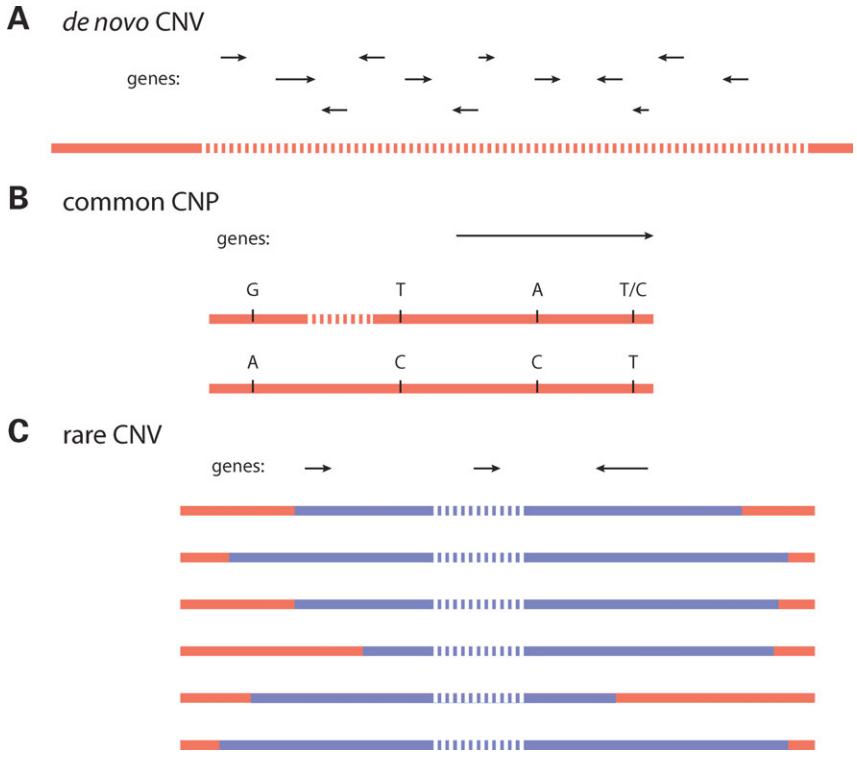

Figure 2. Challenges in interpreting the disease association of *de novo* CNVs, common CNPs, and rare CNVs. (**A**) The *de novo* CNVs identified in genomic disorders to date have generally been 0.5–3 megabases long and spanned many genes, making it difficult to implicate a specific gene in the etiology of the disease. (**B**) Common CNPs are frequently in strong LD with surrounding genetic polymorphisms, making it difficult to identify a single causal variant. Integrated association analysis of SNP and CNP genotypes may be critical. (**C**) Rare CNVs that are shared by unrelated individuals are frequently present on a long shared haplotype (16) (blue); the possibility that some other feature of this haplotype might be the causal feature needs to be evaluated in an integrated analysis of CNV and SNP haplotypes.

RARE, INHERITED CNVS

Investigation of the disease association of rare CNVs is a leading edge in the next frontier of genetic research, which involves analyzing the disease association of collections of rare variants. The sizes of many of the cohorts now collected for GWAS would in principle enable detection of the disease association of low-frequency variants, particularly if such variants have fairly penetrant effects or can be evaluated in reasonable sets (defined *a priori* by clear, plausible criteria) to increase statistical power (Fig. 2C). Although association analysis of rare sequence variants (35,36) requires extensive resequencing, SNP array platforms should allow ascertainment of a considerable fraction of the rare CNVs that are present in a disease cohort. Thus, GWAS can increasingly become studies of both common and rare CNVs as well as common SNPs.

When a rare CNV is detected across the same genomic segment in apparently unrelated individuals, it is usually present on a shared SNP haplotype (frequently quite long), indicating recent shared ancestry at the locus (16). This finding should inform how the putative disease association of a rare CNV is interpreted: the conclusion that the association arises from the CNV—and not from some other feature of a long, shared haplotype—should not be taken for granted. Instead, this should be considered a hypothesis to be explored in an integrated analysis of the SNP and CNV data: the extent of the shared SNP haplotype around the CNV can be documented, and the entire associated haplotype evaluated (Fig. 2C).

In studies of rare CNVs (as indeed in studies of rare sequence variants) it will be important to be vigilant about the potentially confounding effects of non-uniform sensitivity, differential bias, population structure and cryptic relatedness (Box 1).

DIRECTIONS

Over the coming years, a promising hypothesis—that CNV influences disease risk broadly in the population and across disease types—will finally receive an ample, well-powered test. CNV analysis in large cohorts will also offer an early look at the extent to which rare variants shape risk of common disease; such inquiries may set early precedents for subsequent efforts to study rare variants through large-scale sequencing. GWAS will be a powerful venue for such investigations, particularly by enabling the integrated analysis of SNPs, haplotypes and CNVs. Such efforts will help elucidate the molecular etiology of common disease, and will begin to shape our understanding of how multiple forms of genetic variation—fine-scale and large-scale, inherited and *de novo*, common and rare—act in concert to influence human phenotypes.

Box 1. Potential confounds in the association analysis of rare CNVs in disease

Sensitivity to detect and differential bias. Detection of CNVs from array data is sensitive to the detailed noise properties of each hybridization, which in turn arise from sample and hybridization quality. (A survey of the supplementary data sets underlying many CNV-discovery approaches indicates that the number of CNVs detected per sample can vary 4-fold from sample to sample.) As the DNA from affected and unaffected individuals in GWAS often originates at different clinical sites, is extracted at different times, or is analyzed in different experimental plates or batches, 'differential bias' (29)—the confounding of detection sensitivity with Case/Control status—is pervasive. The problem is particularly severe in analyses of the 'genomic burden' of CNVs across the genome, since such analyses aggregate the effects of confounds at thousands of loci. It is therefore important to carefully dissect the sample- and batch-specific influences on detection sensitivity, and to develop carefully controlled analyses. Algorithmic approaches for this are an urgent need in the field.

Population structure. Ancestry may be an unrecognized confound in association studies of rare variants. Populations with African ancestry appear to have far more rare CNVs and low-frequency CNPs than non-African populations do (16) (as indeed they also harbor more rare sequence variants). Careful analysis of population structure in a GWAS cohort is therefore essential. Genome-wide SNP data enable powerful analyses of population structure, both across the genome (30) and at each individual locus (37).

Relatedness. When GWAS cohorts have been analyzed for cryptic relatedness using genome-wide SNP genotypes, such analyses have frequently found that some DNA samples are cryptically related to each other – cousins, siblings, or even the same individual ascertained at different medical centers. For obvious reasons, relatives are far more likely than the general population to share rare sequence variants and rare CNVs. It will therefore be critical to exclude the possibility that association of a rare CNV with phenotype arises in part from cryptic relationships among the affected individuals who share the rare variant.

FUNDING

The author would like to acknowledge support from a Lilly Life Sciences Research Fellowship.

ACKNOWLEDGEMENTS

The author would like to thank Joshua Korn, Jennifer Stone, Doug Levinson, Mark Daly and David Altshuler for thoughtful conversations and comments on this manuscript.

Conflict of Interest statement. None declared.

REFERENCES

1. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
2. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
3. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
4. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.*, **38**, 82–85.
5. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
6. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J. *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
7. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
8. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
9. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
10. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.
11. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.
12. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E. *et al.* (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439**, 851–855.
13. Fanciulli, M., Norsworthy, P.J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J.M., Gough, S.C., de Smith, A., Blakemore, A.I. *et al.* (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.
14. Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M. *et al.* (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.*, **40**, 23–25.
15. Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M. *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, **79**, 275–290.
16. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A. *et al.* (2008) Integrated detection and population genetic analysis of SNPs

and copy number variation. *Nat. Genet.*, published online 7 September 2008, doi:10.1038/ng.238.
17. Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, **82**, 685–695.
18. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Nizzari, M., Gabriel, S.B., Purcell, S. *et al.* (2008) Integrated genotype calling and assocaition analysis of SNPs, common copy number polymorphisms, and rare CNVs. *Nat. Genet.*, published online 7 September 2008, doi:10.1038/ng.237.
19. Barnes, C. and Hurles, M.E. (2008) A robust statistical method for case–control association testing with copy number variation. *Nat. Genet.*, published online 7 September 2008, doi:10.1038/ng.206.
20. Cooper, G.M., Zerr, T.R., Kidd, J.M., Eichler, E.E. and Nickerson, D.A. (2008) Assessment of CNV detection via genome-wide SNP genotyping. *Nat. Genet.*, published online 7 September 2008, doi:10.1038/ng.236.
21. Lupski, J.R. (2007) Genomic rearrangements and sporadic disease. *Nat. Genet.*, **39**, S43–S47.
22. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–449.
23. Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A. and Karayiorgou, M. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.*, **40**, 880–885.
24. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A. *et al.* (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539–543.
25. Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E. *et al.* (2008) Large recurrent microdeletions associated with schizophrenia. *Nature*, published online 30 July 2008, PMID: 18668039.
26. Stone, J.L., O'Donovan, M.C., Gurling, H., Kirov, G.K., Blackwood, D.H., Corvin, A., Craddock, N.J., Gill, M., Hultman, C.M., Lichtenstein, P. *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, published online 30 July 2008, PMID: 18668038.
27. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T. *et al.* (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.*, **358**, 667–675.
28. McCarroll, S.A. (2008) Copy-number analysis goes more than skin deep. *Nat. Genet.*, **40**, 5–6.
29. Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E. *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case–control association study. *Nat. Genet.*, **37**, 1243–1246.
30. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
31. McCarroll, S.A. and Altshuler, D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.
32. McCarroll, S.A., Huett, A.S., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H. *et al.* (2008) Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Gen.*, published online 24 August 2008, doi:10.1038/ng.215, PMID: 18724369.
33. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
34. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G., Nimmo, E.R., Cummings, F.R., Soars, D. *et al.* (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.*, **39**, 830–832.
35. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. and Hobbs, H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.
36. Cohen, J.C., Boerwinkle, E., Mosley, T.H. and Hobbs, H.H., Jr (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.*, **354**, 1264–1272.
37. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D. *et al.* (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, **74**, 979–1000.