Extending genome-wide association studies to copy-number variation

Steven A. McCarroll

1Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, 2Department of Molecular Biology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA and 3Harvard 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 \textit{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).
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.

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.

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).

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
The “genomic burden” data implied that an unknown subset of the sporadic form of these diseases (22–24). These ‘genomic fraction of affected than unaffected individuals, particularly in (regardless of their genomic location) are observed in a larger 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 micro-deletion/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

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

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.
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 ≥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.
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.
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


