VIEWPOINT

In search of meaningful copy number variations

BY **ALAN PACKER**

24 OCTOBER 2008

Image courtesy: Nature Genetics
Closing in: New tools such as Canary may help find copy number variations that are valid and reproducible.

In the past few months, researchers have published dozens of reports linking single-nucleotide polymorphisms (SNPs) with susceptibility to a range of common diseases.

For neuropsychiatric diseases, however, the number of such variants robustly associated with greater risk has been quite small, including **limited success reported for autism spectrum disorders**.

Fortunately, there has been progress in other areas, particularly in the search for rare copy number variants (CNV). These deletions or duplications of genomic segments, typically greater than 1 kilobase in size, appear to be overrepresented in individuals with autism.

The excitement surrounding these findings has been tempered somewhat by the difficulty of making a statistically sound argument for any particular CNV in disease risk.

Four papers recently published in *Nature Genetics* provide a much needed experimental and statistical framework for CNV association studies. They report no new discoveries about the genetics of disease, but enable future discoveries by putting the entire field on solid footing.

Lessons learned:

To understand why these papers are vital to the search for CNVs that predispose to autism and other disorders, it?s important to look back at the record of SNP-based association studies, which until recently has been quite poor. The pre-2007 literature is chock full of studies associating common SNPs with complex diseases such as diabetes, cancer, coronary heart disease, asthma and others, very few of which held up in subsequent attempts to replicate them.

There are several reasons for this poor performance. In retrospect, the most significant problem was simply that these studies were 'underpowered?. In other words, investigators genotyped too few SNPs in too few individuals to detect the associations that we now know to be real ? namely, associations conferring a very modest increase (or decrease) in risk.

This problem was overcome with (i) the development of a haplotype map (HapMap) that identified several hundred thousand SNPs covering the entire human genome (ii) increasingly inexpensive SNP genotyping technology and (iii) international collaborations that generated sample sizes of tens of thousands of individuals.

Two other confounders also contributed to the false positive results. The first is genotyping error, which can skew results. The second is insufficient statistical rigor: if hundreds of thousands of SNPs are tested for association, some false positive results are guaranteed by chance.

Boosting the signal:

CNV-based association studies are no less prone to these difficulties, and in some respects may require even more care to avoid results that cannot be replicated.

The four *Nature Genetics* papers tackle many of these issues. Two of the papers report the work of a team led by David Altshuler; the first[1] details the development of a new chip (produced in collaboration with chip-manufacturer Affymetrix) that simultaneously genotypes more than 900,000 SNPs and CNVs at more than 1.8 million genomic locations.

This chip allows the most comprehensive interrogation of variation in the human genome to date. Most important, Altshuler and colleagues generate a new high-resolution map of human CNVs showing that most are much smaller and encompass a smaller fraction of the genome than previously reported, a finding that will undoubtedly improve the accuracy of CNV-association studies.

The second paper by this team addresses the issues of CNV genotyping[2]. Although the genotyping of SNPs can be plagued by error, in principle the calling of a particular nucleotide is straightforward: a particular variant is either there or it?s not.

CNVs present a new challenge, in that there can be multiple alleles segregating in the population (0, 1, 2, 3 copies at a particular locus, and so on). The number of copies can be roughly quantified, but current methods often do not provide the sort of accurate integer measurement that can allow an investigator to conclude with confidence the number of copies that are actually present.

Further problems are introduced by the fact that CNVs have typically been discovered and genotyped by the same method. Errors that may be acceptable in the generation of an initial survey of variants are almost certain to produce flawed results when included in association studies.

The authors developed a new method (Canary), which uses multiple probes to interrogate the same CNV segment, thus generating highly correlated and reproducible measurements from which one can infer a copy-number genotype, as well as a score reflecting the confidence of each call.

Validation of these calls was determined by comparing them with those generated by an independent method, and averaged an accuracy of more than 97 percent. Canary is part of a package of new tools (Birdsuite) that integrates these CNV calls with SNP genotypes, allowing for a more complete and accurate assessment of genomic variation.

False positives:

A complementary approach is outlined by Gregory Cooper and colleagues in yet another paper[3]. These researchers developed a novel genotyping algorithm that accurately infers many CNV genotypes by carrying out genome-wide SNP genotyping. Because SNPs at distinct locations in the genome are frequently inherited with nearby CNVs, they show that CNV genotypes can be inferred by genotyping SNPs with as few as two probes.

Finally, in the fourth paper, Matthew Hurles and colleagues present a statistical framework for making robust associations between CNVs and phenotypes[4]. They recognize that, even aided by the new tools mentioned above, CNV genotyping will be less than perfect.

Intuitively, this will be particularly problematic when there are differences in genotyping quality between cases and controls in a test of association, which the authors show often leads to a large number of false positive results.

In a series of simulations the researchers assessed six different statistical approaches to CNV-based association, and hit upon one that is relatively robust to the sort of ‘noisy? data that is likely to emerge from genotyping efforts. Their methods are packaged in a software package called CNVtools that seems likely to be widely used.

Eventually, one can imagine genotyping being replaced completely by genomic sequencing, which is now a mature technology that would eliminate many of the pitfalls associated with current approaches for assessing variation. Until then, the methods outlined in these four papers should enable disease associations to be made more rapidly and with higher confidence.

References:

1. McCarroll S. *et al. Nat. Genet.* **40**, 1166-1174 (2008) **PubMed ?**
2. Korn J.M. *et al. Nat. Genet.* **40**, 1253-1260 (2008) **PubMed ?**
3. Cooper G.M. *et al. Nat. Genet.* **40**, 1199-1203 (2008) **PubMed ?**
4. Barnes C. *et al. Nat. Genet.* **40**, 1245-1252 (2008) **PubMed ?**