Statistical Challenges and Opportunities in Copy Number Variant Association Studies

Patrick Breheny1,2*, Yinglei Li2 and Richard Charnigo1,2
1Department of Biostatistics, University of Kentucky, 725 Rose Street, Lexington KY, USA
2Department of Statistics, University of Kentucky, 725 Rose Street, Lexington KY, USA

According to classical genetics, humans have two copies of each region of DNA. During the past decade, however, a large body of research has emerged demonstrating that this is something of an oversimplification [1-3]. Even phenotypically normal individuals have many stretches of their genome in which more or fewer than two copies are found – these stretches have been estimated to constitute roughly 5% of the entire genome [4]. Such genetic variations are referred to as Copy Number Variants (CNVs).

In the wake of the Human Genome Project, a tremendous effort has been spent on understanding the genetic basis of human variation and disease through Genome-Wide Association (GWA) studies. The vast majority of this effort has focused on one-base-pair differences between individuals, termed Single Nucleotide Polymorphisms (SNPs). As the research on copy-number variation demonstrates, however, SNPs represent only one type of genetic variation.

One effort to quantify the relative contributions of SNPs and CNVs on gene expression estimated that SNPs were responsible for 84% of the explainable variation, while CNVs were responsible for 17%, with only 1% resulting from overlapping effects [5]. This 17% represents a sizable degree of genetic variation that has been understudied. Furthermore, some have argued that CNVs are more likely, a priori, to play a role in common diseases because, given that they represent a more subtle, quantitative genetic variation, they are less likely to have been selected out of the population by evolutionary pressures [6]. Indeed, CNVs have been linked to a number of diseases such as Crohn’s disease, psoriasis, and autism [7-9].

Fortunately, CNV information can be mined from existing data collected by GWA studies, thereby avoiding the considerable costs of carrying out new studies. One of the limiting factors—perhaps the limiting factor—in carrying out genome-wide CNV association studies, however, is a challenge of analyzing the data. While methods to determine the locations in which an individual has gained or lost copies of genetic material are fairly well-developed [10,11], methods for integrating these CNV calls into an association study are “still in [their] infancy” [12]. Relative to that of CNVs, genome-wide analysis of SNPs is straightforward: at every genetic marker, each individual is genotyped (AA, AB, or BB) and an association test is carried out. An adjustment for multiple comparisons then preserves the overall type I error rate.

A similar analytic strategy does not readily apply to CNV association studies, for two primary reasons. First, the uncertainty in a CNV call is much greater than that in a SNP call. For each type of calling, the goal is to classify a sample into one of three groups (AA/AB/BB for a SNP, gain/loss/neutral for a CNV) based on probe intensity measurements. However, for a SNP, one obtains a two-dimensional measure (intensities for both the A and B probes); for CNVs, one obtains only a one-dimensional measure (total intensity). Consequently, there is a much greater separation between classes for SNPs, and more extensive misclassification in CNV genotyping.

The second reason is that, unlike SNPs, CNVs span multiple markers and introduce an added complexity: that of estimating the boundaries of the CNV. These two features of CNV data are illustrated in the left panel of figure 1, which comes from an analysis of real data described in Breheny et al. [13]. As noted earlier, CNV calling is based on a one-dimensional measure of intensity; the gray region was determined to have a loss of copy, leading to lower intensity throughout that region. As figure 1 indicates, there is no clear separation between intensities originating from the white (neutral) and gray (loss) regions. Furthermore, the precise boundaries of the CNV are not obvious.

Each of these two features of CNV data complicates association testing. Ignoring misclassification error may considerably diminish the power of a test [14], while the imprecise estimation of boundaries makes it difficult to determine whether two partially overlapping CNVs represent the same genetic variation. Reasonable decision rules for two overlapping CNVs may be proposed; however, even with as few as three CNVs, patterns may arise for which there is no unambiguous resolution. For example, consider a scenario with three CNVs: A, B, and C. Suppose A has 50% overlap with B, B has 50% overlap with C, but A and C have no overlap. How many association tests should one carry out? A variety of ad-hoc rules have been proposed to address this scenario, but one can easily imagine how intractable the problem may become with, say, 25 partially overlapping CNVs. Dealing with partial overlap is both burdensome in practice and likely to be statistically inefficient.

To avoid these complications, one may avoid CNV calling altogether and carry out marker-level testing (as opposed to the previous approach, which we refer to as variant-level testing). Marker-level tests can be simple, such as carrying out t-tests of CNV intensities between cases and controls, or more complex, involving mixture models to incorporate uncertainty in copy number [15]. However, due to the noise in intensity measurements, single-marker tests tend to have low power and require very large sample sizes (n>4,000 for the study in Barnes et al. [15]).

An intriguing possibility is to supplement the power of single-marker tests by pooling results from neighboring tests. This idea is illustrated in the right panel of figure 1. If a CNV spanning multiple marker is present and associated with the phenotypic outcome of interest, this will induce marker-level associations spanning the genomic region covered by the CNV. Although no single p-value in

*Corresponding author: Patrick Breheny, Department of Biostatistics, University of Kentucky, 725 Rose Street, Lexington KY, USA, E-mail: patrick.breheny@uky.edu

Received November 17, 2012; Accepted November 19, 2012; Published November 26, 2012


Copyright: © 2012 Breheny P, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
There are interesting possibilities for improving variant-level tests as well, based on the idea of joint CNV calling. Rather than calling CNVs separately for each sample, several authors [17-19] have recently proposed methods for jointly calling common CNVs across multiple samples, potentially eliminating the partial overlap issue discussed earlier. These methods are still new, however, and the feasibility of extending them to CNV association studies on the genome-wide level has not yet been investigated.

Our focus in this editorial has been on analytical approaches for incorporating information across markers and across samples when performing association tests. We do not wish to downplay other important statistical issues in CNV association studies, such as normalization of the data, proper experimental design, and controlling for confounding factors. Rather, we hope to have highlighted some interesting features of CNV data, limitations of existing approaches, possible avenues for improvement, and open statistical questions surrounding this important area of scientific inquiry.

References
