Association tests and software for copy number variant data

Vincent Plagnol

JDRF/WT Diabetes and Inflammation Laboratory, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0XY, UK
Correspondence to: Tel: +44 (0)1223 762107; E-mail: vincent.plagnol@cimr.cam.ac.uk

Date received (in revised form): 7th October 2008

Abstract

Recent studies have suggested that copy number variation (CNV) significantly contributes to genetic predisposition to several common disorders. These findings, combined with the imperfect tagging of CNVs by single nucleotide polymorphisms (SNPs), have motivated the development of association studies directly targeting CNVs. Several assays, including comparative genomic hybridisation arrays, SNP genotyping arrays, or DNA quantification through real-time polymerase chain reaction analysis, allow direct assessment of CNV status in cohorts sufficiently large to provide adequate statistical power for association studies. When analysing data provided by these assays, association tests for CNV data are not fundamentally different from SNP-based association tests. The main difference arises when the quality of the CNV assay is not sufficient to convert unequivocally the raw measurement into discrete calls — a common issue, given the technological limitations of current CNV assays. When this is the case, association tests are more appropriately based on the raw continuous measurement provided by the CNV assay, instead of potentially inaccurate discrete calls, thus motivating the development of new statistical methods. Here, the programs available for CNV association testing for case control or family data are reviewed, using either discrete calls or raw continuous data.

Keywords: *calling algorithm*, *association studies*, *CNV*, *genetics*, *software*

Introduction

The use of genome-wide association studies (GWAS) has successfully linked genetic variants with susceptibility to a wide range of common polygenic diseases.[1] Such GWAS, however, have almost exclusively focused on single nucleotide polymorphisms (SNPs). Additional studies targeted to specific copy number variant loci,[2,3] as well as the alternative approach consisting of using SNPs to tag copy number variations (CNVs),[4] have suggested that CNVs may contribute significantly to genetic predisposition to several common diseases. The tagging of CNVs using SNPs is often imperfect, and therefore motivates the development of association studies directly targeting CNVs.

Importantly, association tests directly targeting CNVs rely on the development of maps of common CNV polymorphism in the human genome.[5–7] The process of discovering common CNVs is related but distinct from the association testing procedure. Software reviewed here assume that the user is considering known CNV regions, with properly mapped boundaries. Programs designed to identify previously unknown CNVs (eg Wang *et al.*[8]) are not discussed here. Moreover, association tests outlined in this review typically consider the common variant/small effect situation. Different approaches must be used when dealing with rare but highly penetrant CNVs.

Various assays can estimate CNV status directly — in particular, real-time polymerase chain reaction (rtPCR) analysis,[2,3] comparative genomic hybridisation (CGH) arrays[5,9] and SNP genotyping arrays.[10] The throughput of these assays is sufficient to analyse the large cohorts required to detect subtle effects previously reported by SNP-based association studies. These CNV assays typically generate a one-dimensional continuous measure per DNA sample and per CNV probe. Provided that the quality of the CNV assay is sufficiently clear to convert these raw continuous measures into discrete calls, CNV-based association testing is analogous to the SNP situation, with a few minor differences, which are outlined in this review. When uncertainty in the calling makes this discretisation difficult, however, a different statistical approach is necessary. CNV association testing must then be based on the raw continuous data instead of potentially inaccurate discrete calls. Here, these alternative approaches are outlined, both for case control and family-based association studies.

Discrete or continuous genotypes

Even though the underlying CNV state is a discrete integer, CNV assays typically only provide a continuous measure for each CNV and individual, and this continuous trait is a surrogate for the actual, discrete CNV state. Therefore, when testing for association at a CNV locus, the first choice is to decide whether to analyse the CNV as a continuous or a discrete measurement.

In an ideal scenario, the distribution of this surrogate measure can be properly separated in discrete clusters, which can then be linked to discrete numbers of DNA copies. In this case, it is appropriate to discretise the data and base the association test on these discrete calls. Unfortunately, and because of technical difficulties associated with the assessment of CNV status, this continuous measure cannot always unequivocally be converted into discrete numbers of DNA copies. Erroneous calls can inflate the rate of false-positive associations and limit the statistical power to detect true associations.[11] Therefore, when discretisation is difficult, it is preferable to base the association test on the raw continuous measurement.[12,13]

Association testing using CNV status as a discrete trait

If the CNV status is summarised using a discrete measure, the association test is analogous to the SNP situation, and traditional statistical software such as R (http://www.r-project.org), or more specialised tools such as PLINK[14], are appropriate; however, additional factors complicate the analysis of CNVs compared with SNPs.

First, association tests require assumptions on the model linking CNV status to disease risk. For example, a traditional Cochran-Armitage test[15] assumes that, on the log-scale, the risk is proportional to the number of copies. Such assumptions may not be appropriate for some CNVs; for example, for some diseases only extreme numbers of copies may have a causal role in disease aetiology. Especially for multi-allelic CNVs, for which a wide range of models can be investigated, the choice of model must be driven by prior belief on disease aetiology. In the absence of prior knowledge, it is advisable to constrain the analysis to a small number of simple models to avoid misleading *p*-values generated by multiple testing.

Secondly, CNV assays do not always provide information about the exact number of copies, but often only on the 'relative' numbers. For example, one cannot always distinguish CNVs with zero, one or two copies from CNVs with two, three or four copies.

Thirdly, the phase is often unknown and only the total number of copies can be measured; for example, most CNV genotyping assays cannot distinguish two haplotypes with one copy on each from one haplotype with two copies combined with a complete deletion. Phasing CNVs is difficult, but if sufficient marker density is available it can be achieved using software such as fastPHASE.[16]

Lastly, it is common when using SNP genotyping arrays to assess CNV status to obtain information jointly on SNP genotypes and DNA copy number. Rather than a single number, each individual is therefore summarised by a pair; the first element indicates the number of DNA copies and the second element indicates the associated SNP genotype. This added information can potentially provide additional power when testing for association, and a recent study[17] has proposed a new approach to integrating SNP and CNV information in a unique association testing procedure. This joint SNP/CNV association is implemented as part of the analysis software PLINK.[14]

Combining information from multiple probes within the same CNV

In contrast to SNPs, CNVs can extend over large genomic regions. CNV genotyping arrays, in particular high-density CGH arrays,[9] typically use multiple probes to interrogate a single CNV. This is particularly useful when we expect variable probe performance or, more generally, when measurements are noisy and a combination of probes can provide more accurate measurements. Some association testing procedures require the combination of information from all probes into a single measurement for each individual,[11] while others deal directly with multi-marker data.[18]

Several procedures can be considered to combine data across multiple probes. When the genomic location of the CNV is well defined, and there is strong evidence that all probes are located within the copy number variable region, simply averaging the probe intensity across all probes for each individual is sensible. In many situations, however, the boundaries of the CNV are not well defined, and some probes may lie outside of the region of interest. When this is the case, it is advantageous to identify and down-weight, or even simply remove, these non-informative probes to lower the measurement noise. This down-weighting procedure can be done using a principal component analysis[11] that should down-weight the non-informative probes provided that the signal generated by the probes in the CNV region is sufficiently strong.

This down-weighting procedure is also relevant when dealing with complex CNV regions, where several distinct CNVs may be overlapping. A principal component analysis can help separate a set of probes in genetically relevant discrete groups. Distinct association tests can then be carried out separately for each group of probes.

Association testing using CNV status as a continuous trait

When the uncertainty in the calling procedure is too large to call the data confidently, it is advisable to base the association test on the raw, continuous data summary.[12,13] In the context of family-based association studies, PBAT implements a broad class of association tests and has been extended to deal with raw measurements from CNV assays.[18] In its simplest form, this test compares the CNV measure for an affected offspring with the average CNV measure of both parents in parent–offspring trios. A consistently elevated number in affected offspring indicates an association between high copy number and elevated disease risk. A commercial version of the PBAT statistical tools, including a graphical user interface, is now part of the Golden Helix genetic association software (http://www.goldenhelix.com).

Case-control association studies, however, require a different analytical approach. A previous study[19] has highlighted the potential for different DNA sourcing, handling, extraction or storage to affect the measures obtained from genotyping assays, potentially creating biases between sample collections that do not reflect actual genotype differences. Therefore, even when cases and controls are analysed using the same genotyping assay and at the same time, differences in the distribution of the raw CNV measures cannot always be interpreted directly as actual genotype differences.

To circumvent these issues, Barnes *et al.*[11] proposed a joint calling/association testing approach implemented in the R package CNVtools. This approach provides calls by clustering the data using a Gaussian mixture, but the association test statistic is designed to account for the uncertainty in these calls. In addition, the calling procedure implements a hierarchical clustering model that lets the parameters of the Gaussian mixture vary across cohorts in order to account for potential biases correlated with the origin and handling of DNA samples.[19] This method is, in fact, general and can be used to combine data generated by different CNV genotyping assays at the same CNV locus.

A drawback of the clustering approach implemented by CNVtools is the fact that the clustering procedure relies on the data quality being sufficient to fit a Gaussian mixture with a limited number of components. For example, when analysing highly polymorphic short tandem repeats, it is unreasonable to expect that the CNV assay can properly separate samples that differ by one or very few repeats. It makes little sense, in this situation, to cluster the CNV measurement, and CNVtools is not an appropriate tool for that type of data. The PBAT testing procedure,[18] however, is still usable, provided that one is analysing family data.

Conclusion

While a wide range of options is available for analysing the CNV status as a discrete trait, software tools to analyse directly raw continuous measurements from CNV assays are currently limited to PBAT[18] for family and CNVtools for case control association studies. Both PBAT and CNVtools share a frequentist statistical approach; however, a Bayesian approach to this problem appears well suited to average over the uncertainty in the calling procedure efficiently, which is the main hurdle for CNV association tests. As GWAS targeting CNVs become more common, we expect that several Bayesian approaches will be developed to complement currently available software.

Acknowledgment

Vincent Plagnol is a Juvenile Diabetes Research Foundation postdoctoral fellow.

References

1. Wellcome Trust Case Control Consortium (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature* Vol. 447, pp. 661–678.
2. Aitman, T.J., Dong, R., Vyse, T.J. *et al.* (2006), 'Copy number polymorphism in FCGR3 predisposes to glomerulonephritis in rats and humans', *Nature* Vol. 439, pp. 851–855.
3. Gonzalez, E., Kulkarni, H., Bolivar, H. *et al.* (2005), 'The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility', *Science* Vol. 307, pp. 1434–1440.
4. McCarroll, S.A., Huett, A., Kuballa, P. *et al.* (2008), 'Deletion polymorphism upstream of irgm associated with altered irgm expression and Crohn's disease', *Nat. Genet.* Vol. 40, pp. 1107–1112.
5. Redon, R., Ishikawa, S., Fitch, K.R. *et al.* (2006), 'Global variation in copy number in the human genome', *Nature* Vol. 444, pp. 444–454.
6. Kidd, J.M., Cooper, G.M., Donahue, W.F. *et al.* (2008), 'Mapping and sequencing of structural variation from eight human genomes', *Nature* Vol. 453, pp. 56–64.
7. Conrad, D.F., Andrews, D.T., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2005), 'A high-resolution survey of deletion polymorphism in the human genome', *Nat. Genet.* Vol. 38, pp. 75–81.
8. Wang, K., Li, M., Hadley, D. *et al.* (2007), 'Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data', *Genome Res.* Vol. 17, pp. 1665–1674.
9. Perry, G.H., Ben-Dor, A., Tsalenko, A. *et al.* (2008), 'The fine-scale and complex architecture of human copy-number variation', *Am. J. Hum. Genet.* Vol. 82, pp. 685–695.
10. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. and Nickerson, D.A. (2008), 'Systematic assessment of copy number variant detection via genome-wide SNP genotyping', *Nat. Genet.* Vol. 40, pp. 1199–1203.
11. Barnes, C., Plagnol, V., Fitzgerald, T. *et al.* (2008), 'A robust statistical method for case-control association testing with copy number variation', *Nat. Genet.* Vol. 40, pp. 1245–1252.
12. McCarroll, S.A. and Altshuler, D.M. (2007), 'Copy-number variation and association studies of human disease', *Nat. Genet.* Vol. 39 (7 Suppl.), pp. S37–S42.
13. Stranger, B.E., Forrest, M.S., Dunning, M. *et al.* (2007), 'Relative impact of nucleotide and copy number variation on gene expression phenotypes', *Science* Vol. 315, pp. 848–853.
14. Purcell, S., Neale, B., Todd-Brown, K. *et al.* (2007), 'Plink: A tool set for whole-genome association and population-based linkage analyses', *Am. J. Hum. Genet.* Vol. 81, pp. 559–575.
15. Armitage, P. (1955), 'Tests for linear trends in proportions and frequencies', *Biometrics* Vol. 11, pp. 375–386.
16. Scheet, P. and Stephens, M. (2006), 'A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase', *Am. J. Hum. Genet.* Vol. 78, pp. 629–644.
17. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A. *et al.* (2008), 'Integrated genotype calling and association analysis of SNPS, common copy number polymorphisms and rare CNVs', *Nat. Genet.* Vol. 40, pp. 1253–1260.
18. Ionita-Laza, I., Perry, G.H., Raby, B.A. *et al.* (2008), 'On the analysis of copy-number variations in genome-wide association studies: A translation of the family-based association test', *Genet. Epidemiol.* Vol. 32, pp. 273–284.
19. Clayton, D.G., Walker, M.M., Smyth, D.J. *et al.* (2005), 'Population structure, differential bias and genomic control in a large-scale, case-control association study', *Nat. Genet.* Vol. 37, pp. 1243–1246.