BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 13 2011, pages 1873–1875
doi:10.1093/bioinformatics/btr264

Genetics and population analysis

Advance Access publication May 5, 2011

famCNV: copy number variant association for quantitative traits in families

Hariklia Eleftherohorinou[1,†], Johanna C. Andersson-Assarsson[2,3,†], Robin G. Walters[1,3], Julia S. El-Sayed Moustafa[3], Lachlan Coin[1], Peter Jacobson[2], Lena M. S. Carlsson[2], Alexandra I. F. Blakemore[3], Philippe Froguel[3], Andrew J. Walley[3] and Mario Falchi[3,*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St Mary's Hospital, London, UK, [2]Department of Molecular and Clinical Medicine, The Sahlgrenska Academy, Gothenburg University, and Sahlgrenska Center for Cardiovascular and Metabolic Research, SE-413 07 Gothenburg, Sweden and [3]Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK

Associate editor: Jeffrey Barrett

ABSTRACT

Summary: A program package to enable genome-wide association of copy number variants (CNVs) with quantitative phenotypes in families of arbitrary size and complexity. Intensity signals that act as proxies for the number of copies are modeled in a variance component framework and association with traits is assessed through formal likelihood testing.

Availability and implementation: The Java package is made available at www.imperial.ac.uk/medicine/people/m.falchi/.

Contact: m.falchi@imperial.ac.uk

Received on November 23, 2010; revised on April 9, 2011; accepted on April 12, 2011

Copy number variants (CNVs) have recently attracted increasing interest as a source of phenotypic variation in humans, with both rare and common CNVs being associated with different complex diseases (e.g. Diskin *et al.*, 2009; Walters *et al.*, 2010) and gene expression levels (Henrichsen *et al.*, 2009). The availability of high-coverage SNP arrays and the development of methods for making CNV calls from SNP data are now being exploited to evaluate genome-wide association with CNVs (e.g. Glessner *et al.*, 2009). At present, FBAT (Ionita-Laza *et al.*, 2008) is the only other available algorithm that performs association test of normalized signal intensity measurement variations, reflecting CNVs in family-based datasets. The family-based association test implemented in FBAT uses only the within-family variation. Although that is one of the possible approaches to control for population stratification, disregarding the between-family variation might lead to decreased statistical power (Aulchenko *et al.*, 2007). famCNV uses information from both within- and between-family variations, while population stratification issues can be detected and corrected using standard methods. Although qualitative CNV calls from existing algorithms can be incorporated as multiallele loci in a family-based association test, direct testing of the underlying quantitative CNV measurement has been shown to be more robust in identifying genuine associations in many cases (McCarroll and Altshuler, 2007; Stranger *et al.*, 2007). Moreover, quantitative trait designs are less likely to be affected by differential errors that can arise in case-control studies. We have implemented a Java program, famCNV, that uses intensity signals such as the $\log_2$ ratio from array comparative genomic hybridization (aCGH) or the log ratio of observed to expected signal intensity (LRR) from Illumina genotyping arrays. The quantitative trait of interest is analysed in a variance component framework to model the resemblance among relatives and to remove possible bias from familiality (Abecasis *et al.*, 2000). A linear mixed model is fitted to data to partition the total phenotypic variance into a polygenic component σ_a^2 determined by additive genetic effects and a residual variance σ_e^2 determined by the unique individual environment, including measurement errors. The variance covariance matrix Ω for the subjects in each family is $\Omega = 2\Phi\sigma_a^2 + I\sigma_e^2$, where 2Φ is the matrix of the expected proportion of alleles that are identical-by-descent among the family members used to correct for the sample structure and I is an identity matrix.

Variance component analysis allows simultaneous modeling of the trait mean as $\mu = \gamma + A\beta$, where γ is the trait mean, A is a matrix of covariate effects and β the corresponding regression coefficients. The test is evaluated by comparing the likelihood of the full model, where signal intensities are included in the mean model and the likelihood of the null model where the effect of the signal intensities is constrained to zero. Twice the difference between the logarithms of the full and the null likelihoods asymptotically follows a chi-square distribution with one degree of freedom and can be used to evaluate the difference in goodness-of-fit between the two models (Amos, 1994).

When using data from genotype arrays, the B Allele frequency (BAF—representing the proportion of the total signal deriving from the B allele) can also be included in both the full and null models, to control for misleading differences in an LRR distribution between genotypes that may have arisen due to errors during the array normalization process (Peiffer *et al.*, 2006).

The identified loci, either *via* the reported *P*-value or *via* meeting a threshold for the false discovery rate (FDR; Storey and Tibshirani, 2003), are expected to reflect probes that are both genuinely copy

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

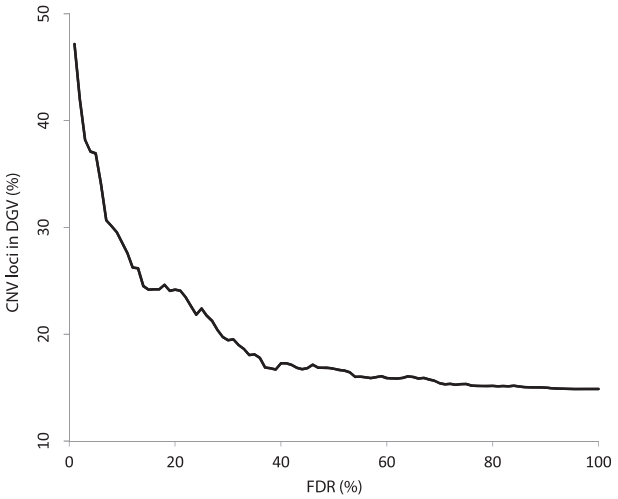

Fig. 1. Overlap of CNVs identified by famCNV with CNV loci catalogued in the Database of Genomic Variants at different FDR levels. At an FDR of 1%, 47% of QT-associated probes lay within reported CNVs. In contrast, only 15% of non-associated probes (FDR > 80%) lay within DGV CNVs.

number-variable and whose copy number is associated with the quantitative trait under examination. To evaluate this approach, we used famCNV to identify CNVs that were associated in *cis* with variation in gene expression. We applied our approach to a cohort of 154 nuclear families, including 732 subjects for which both genotyping and gene expression data were available (Carlsson *et al*., 2009). For all siblings, gene expression levels were measured in subcutaneous adipose tissue using Affymetrix U133Plus2.0 microarrays and DNA from peripheral blood was genotyped using the Illumina 610K-Quad array. As quantitative traits, we selected 28 580 transcripts with known position in the genome and 275 348 SNPs for which intensity data were available. For each of these transcripts, our preliminary analysis focussed on SNP markers lying within the transcript plus 10 kb upstream and downstream. The median (first–third quartile) number of markers tested per transcript was 10 (6–21).

Using this dataset, 189 probes were identified as putative CNV loci associated with transcript level of the neighbouring gene (161 different transcripts), at FDR of 10%. We performed an in-silico experiment to determine whether these regions have been independently identified as copy number-variable in other studies. The probe locations were checked against CNVs and indels catalogued in the Database of Genomic Variants (DGV, version 9). We selected from the database CNVs and indels detected through aCGH and sequencing studies, and determined the number of our association signals that were included in the DGV-reported boundaries. Figure 1 shows the results for CNVs identified by famCNV at different FDR levels, demonstrating that the proportion lying within reported CNVs rises with an increasingly stringent significance threshold.

As a second level of validation of the presence of common CNVs at the predicted loci, we examined the overlap between the CNV regions identified by famCNV in the family study and the calls made by an alternative CNV prediction algorithm, cnvHap (Coin *et al*., 2010), in an independent dataset of Caucasian ancestry. For this, we used a cohort of 902 adult French subjects genotyped using the Illumina Human CNV370-duo array (Meyre *et al*., 2009). No phenotypic data were considered by cnvHap, as this analysis was aimed solely at confirmation of the presence of common structural variants.

For this comparison, we first limited the CNV calls of cnvHap to those that spanned at least two SNPs and were present in at least five samples. We found that the regions identified by famCNV at 5% FDR were enriched 6.7-fold ($P < 0.01$) for probes that intersect with cnvHap CNV calls, when compared to the frequency observed in 100 random datasets of probes not found to show association with transcript levels by famCNV (FDR > 80%). Similarly significant results were obtained after applying a wide range of alternative CNV-calling criteria to cnvHap.

The package is entirely written in the Java language and is therefore executable on all supported operating systems. The variance component algorithm follows the QTDT implementation (Abecasis *et al*., 2000). Files can be imported into famCNV in the PED and MAP file formats as used by the PLINK package (Purcell *et al*., 2007). For analysis of gene expression data, transcript locations can be provided to allow the restriction of the analyses within a user-defined window around each transcript. To help distinguish between putative rare and common CNVs, famCNV also calculates the Gini coefficient (Gini, 1921) to determine whether the positive contribution to the final chi-square statistics has been determined by a uniformly distributed effect or by a small number of strongly associated families.

The program, documentation and example files are available at http://www.imperial.ac.uk/medicine/people/m.falchi/.

ACKNOWLEDGEMENTS

The authors thank Leonardo Bottolo, Adam J. De Smith and the staff of the Imperial College High-Performance Computing Service for their advice and support.

Funding: This study was funded by Grant no. 079534/z/06/z from the Wellcome Trust. The SOS Sibpair study is supported by the Swedish Research Council (K2010-55X-11285-13), the Swedish Foundation for Strategic Research to Sahlgrenska Center for Cardiovascular and Metabolic Research, the Swedish Diabetes foundation and the Swedish federal government under the LUA/ALF agreement.

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al*. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.

Amos,C.I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.*, **54**, 535–543.

Aulchenko,Y.S. *et al*. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–578.

Carlsson,L.M. *et al*. (2009) ALK7 expression is specific for adipose tissue, reduced in obesity and correlates to factors implicated in metabolic disease. *Biochem. Biophys. Res. Commun.*, **382**, 309–314.

Coin,L.J. *et al*. (2010) cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. *Nat. Methods*, **7**, 541–546.

Diskin,S.J. *et al*. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, **459**, 987–991.

Gini,C. (1921) Measurement of inequality of incomes. *Econ. J.*, **31**, 124–126.

Glessner,J.T. *et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, **459**, 569–573.

Henrichsen,C.N. *et al.* (2009) Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.*, **41**, 424–429.

Ionita-Laza,I. *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.*, **32**, 273–284.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.

Meyre,D. *et al.* (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.*, **41**, 157–159.

Peiffer,D.A. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Stranger,B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Walters,R.G. *et al.* (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, **463**, 671–675.