BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 742–743

GRR: graphical representation of relationship errors

Gonçalo R. Abecasis, Stacey S. Cherny, W. O. C. Cookson and Lon R. Cardon

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7RZ, UK

Received on February 1, 2001; revised on March 27, 2001; accepted on March 28, 2001

ABSTRACT

Summary: A graphical tool for verifying assumed relationships between individuals in genetic studies is described. GRR can detect many common errors using genotypes from many markers.

Availability: GRR is available at http://bioinformatics.well.ox.ac.uk/GRR.

Contact: goncalo@well.ox.ac.uk; lon@well.ox.ac.uk

Many large scale linkage and association studies have been conducted and their popularity is increasing. Simple, efficient, quality control procedures are essential to the successful completion of these studies. A common problem in genetic studies is the misspecification of relationships between DNA samples (Ott, 1991). Misspecification of relationships can lead to inaccurate or biased results and it is therefore important to verify all assumed relationships.

The effects of relationship misspecification are varied. In studies using family data, problems such as non-paternity and the mislabeling of monozygotic (MZ) twins as non-twin full sibs, as well as sample mix-ups can lead to mistaken inferences about allele sharing. For example, MZ twins will always share more alleles than other sibling pairs while, on average, half-siblings will share fewer alleles than full-siblings. In larger pedigrees, the potential for relationship misspecification is greater and the detection of these problems is even harder.

In studies using samples of unrelated individuals, such as association and pharmacogenetic applications, the presence of related individuals can lead to a misleading inference about statistical significance. For example, although it may be interesting to find that individuals with resistance to a certain drug share a certain genotype, the finding is less striking if some of the individuals are related.

The correct genetic relationship between any two individuals defines an expected pattern of allele sharing between them. The details of this pattern can be complex, and will depend on the exact type of relationship, marker characteristics, population history and inbreeding. Statistics for verifying relationships through patterns of allele sharing have been proposed, with various degrees of sophistication, computing time requirements and assumptions (Boehnke and Cox, 1997; Goring and Ott, 1997; Broman and Weber, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000). Here we describe a simple, general approach for verifying that individuals with the same specified relationship have similar patterns of allele sharing. Unlike other approaches, our method does not require specification of allele frequencies or any other population parameter. It is expected to be robust to a small level of random errors in the data and applicable to large inbred samples. In addition to relationship misspecification, our method can detect some other problems such as sample duplications and switches.

The method is defined as follows: first, classify each pair of individuals according to their assumed relationship (such as sib-pairs, parent–offspring pairs, unrelated individuals, etc.). Second, calculate the mean (μ_{ij}) and variance (σ_{ij}) of identical-by-state allele sharing over a number of polymorphic loci for each pair of individuals, i and j. If the sample is homogeneous, we expect each group to display a characteristic pattern of allele sharing. For example, sib-pairs will be expected to share more alleles on average than unrelated individuals, while parent–offspring pairs (which share at least one chromosome) are expected to show less variability in allele sharing than sib-pairs (which may share zero, one or two chromosomes). A convenient way to identify individuals with patterns of allele sharing inconsistent with their specified relationship is to colour code and plot these mean—variance statistics (Figure 1).

The figure presents typical results for a genome scan in a non-inbred sample. Several distinct clusters are present: unrelated individuals have the lowest average sharing and high variance (coloured in blue); half-siblings have higher sharing on average (coloured in green) and full-siblings have even higher sharing (coloured in red); parent–offspring pairs have a similar degree of allele sharing to sib-pairs but with lower variance (coloured in yellow). All

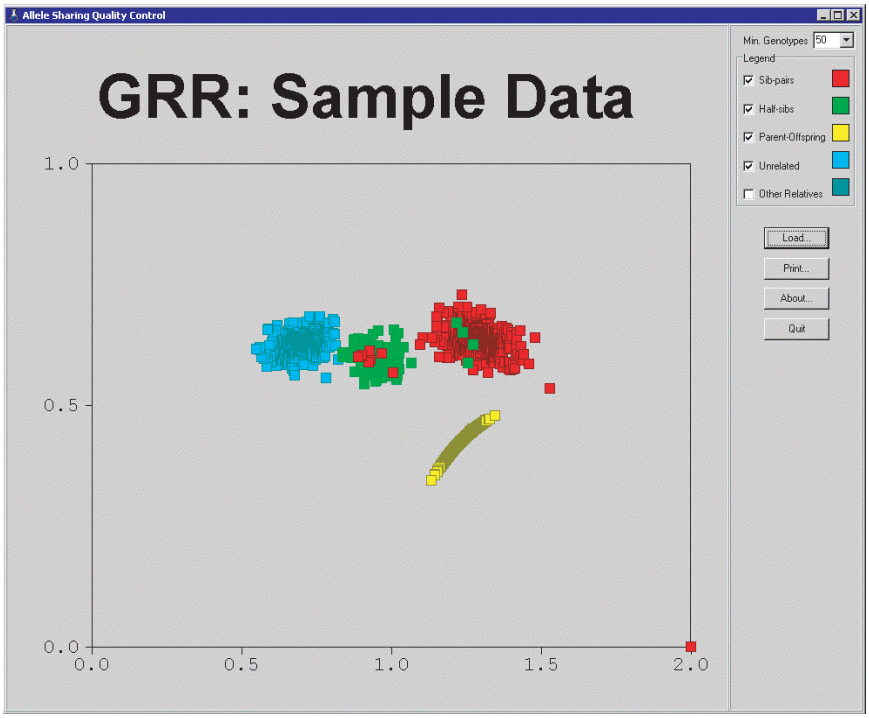

Fig. 1. Sample screen shot. Features described in text.

other relative pairs are grouped together and not displayed by default. Note that some sibling and full-sibling pairs have been misclassified and appear in other clusters. A single sib-pair displays maximum average sharing (bottom right corner) and corresponds to a pair of identical twins.

To ensure that outlier points are easily identifiable, GRR implements an outlier rating scheme and places likely outliers on top of less interesting points. This scheme is implemented by calculating the mean and variance of each allele-sharing statistic within each relationship group. Then each individual's scores are standardized to obtain $Z_{\mu_{ij}}$ and $Z_{\sigma_{ij}}$ and assigned the outlier scores $R_{ij} = \max(|Z_{\mu_{ij}}|, |Z_{\sigma_{ij}}|)$ so that points with higher scores are layered on top of lower rated points. Alternative schemes for layering data points, such as the Mahalanobis (1936) distance, can be selected by the user.

GRR recognizes standard genetic formats for genotype and family structure data, including linkage and QTDT format files (Ott, 1991; Abecasis *et al.*, 2000). Interactive features allow the user to select individual families and inspect statistics for any pair of individuals by clicking the appropriate plot area.

This approach is simple to implement and can be incorporated into many genetic analysis databases and quality control protocols. The method performs efficiently in genome scan linkage panels, although as few as 50 unlinked markers may be sufficient to verify first-degree relationships in family samples or to verify that no close relatives or gross stratification are present in samples of unrelated individuals.

ACKNOWLEDGEMENTS

This research was supported by the Wellcome Trust and by grant EY-12562 from the National Institutes of Health, USA.

REFERENCES

Abecasis,G.R., Cardon,L.R. and Cookson,W.O.C. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.

Boehnke,M. and Cox,N.J. (1997) Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.*, **61**, 423–429.

Broman,K.W. and Weber,J.L. (1998) Estimation of pairwise relationships in the presence of genotyping errors. *Am. J. Hum. Genet.*, **63**, 1563–1564.

Epstein,M.P., Duren,W.L. and Boehnke,M. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.*, **67**, 1219–1231.

Goring,H.H. and Ott,J. (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur. J. Hum. Genet.*, **5**, 69–77.

Mahalanobis,P.C. (1936) On the generalized distance in statistics. *Proc. Natl Inst. Sci. India*, **2**, 49.

McPeek,M.S. and Sun,L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.*, **66**, 1076–1094.

Ott,J. (1991) *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.