BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 2 2000
Pages 182–183

GOLD—Graphical Overview of Linkage Disequilibrium

G. R. Abecasis* and W. O. C. Cookson

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

Received on July 26, 1999; revised and accepted on September 17, 1999

Abstract

Summary: *We describe a software package that provides a graphical summary of linkage disequilibrium in human genetic data. It allows for the analysis of family data and is well suited to the analysis of dense genetic maps.*
Availability: *http://www.well.ox.ac.uk/asthma/GOLD*
Contact: *goncalo@well.ox.ac.uk*

Precise estimates of the location of complex disease genes should permit their identification through positional cloning, even when understanding of the underlying biochemical pathways is limited (Collins, 1992). Public and private genome projects are investing a great deal of effort in the identification of polymorphic sites in the human population. These efforts are cataloguing increasing numbers of single-nucleotide polymorphisms (SNPs) which are well suited to automated high-throughput analysis. A dense genetic map of the human genome should be provided by SNPs in the near future.

Traditional linkage analysis, based on allele sharing between relatives, identifies broad chromosomal regions that are likely to contain disease genes. However, the resolution of these methods is limited by the number of recombination events in typical pedigrees and impractical for positional cloning efforts in complex disease. Fine-mapping within the broad regions identified by allele-sharing methods is a major challenge. Gene mapping strategies based on linkage disequilibrium are expected to have much greater resolution, and should be able to capitalize on dense SNP maps as they become available (Risch and Merikangas, 1996).

As ancestral haplotypes propagate through a population, their physical length is reduced by recombination events. Recombination events between markers separated by very short distances are very rare. Individuals inheriting a disease mutation from a common, but possibly distant, ancestor are expected to share a region of the ancestral haplotype in which the mutation originated. Markers within this shared haplotype are non-randomly associated with the disease and each other, and are said to be in linkage disequilibrium. Association studies using family based controls can distinguish between linkage disequilibrium and other sources of association, such as population admixture.

While markers in linkage disequilibrium are expected to be tightly linked, the precise extent of shared haplotypes depends on a stochastic process which involves a number of factors including migration, selection, mutation and genomic patterns of recombination. It is expected that patterns of linkage disequilibrium will vary not only between populations, but also between loci in the same population. Information on the patterns of linkage disequilibrium within the populations and loci to be studied is essential for organizing gene-mapping efforts (Kruglyak, 1999). The extent of linkage disequilibrium in a particular locus and population dictate the choice of SNP markers to be genotyped, the resolution that might be achieved in the study and the sample size required for analysis.

Traditionally, linkage disequilibrium is described by pair-wise measures, such as Lewontin's standardized disequilibrium coefficient D′ (Lewontin and Kojima, 1960). The number of pair-wise statistics to be estimated rises exponentially with the number of markers so that their interpretation is cumbersome for dense maps. As a consequence, it is exceedingly difficult to draw meaningful conclusions from tabular summaries of disequilibrium coefficients. We have implemented a novel graphical summary of these pair-wise measures in the Graphical Overview of Linkage Disequilibrium (GOLD) package which provides an easy to interpret graphical representation of the patterns of disequilibrium in a region and their relationship to the underlying genetic or physical map. These graphical summaries are especially well-suited for comparative analysis amongst loci or populations, and are scalable to moderate numbers of markers.

Using input in the form of founder haplotype estimates, GOLD calculates pair-wise disequilibrium measures and provides a graphical overview of disequilibrium patterns. The Simwalk2 (ftp://watson.hgen.pitt.edu/pub/simwalk2) program allows haplotype estimation in general pedigrees,

*To whom correspondence should be addressed.

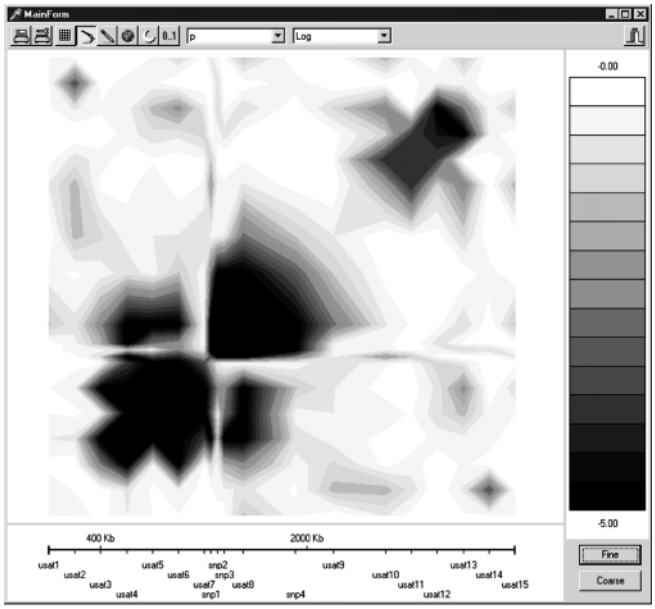

Fig. 1. Typical graphical output. In this case Log(p) is plotted for a set of markers, p measures the significance of observed association. A color version is available online.

and is used to determine founder haplotypes from an arbitrary data-set (Sobel and Lange, 1996). This family-based inference of haplotypes should be robust even in the presence of population admixture. For situations where family data is not available, a limited interface to Estimate Haplotype (EH) (ftp://linkage.rockefeller.edu/software/eh) is provided. The current version estimates some commonly used genetic measures of linkage disequilibrium (such as D, D′ and Δ^2), as well as conventional measures of association (Cramer's V, uncertainty coefficient U and the contingency table chi-squared and significance levels). These statistics are output for inspection in a traditional contingency table. Definitions of these statistics are provided in our website, but additional user-defined measures of association [such as higher order measures; Weir (1996)] may be plotted.

Finally, GOLD summarizes these statistics in a color plot. The horizontal and vertical axes are scaled according to physical (or genetic) distances between markers. For each marker pair m_i and m_j, the pair-wise disequilibrium statistics are color coded (bright red and dark blue are opposite ends of the scale) and plotted atposition $[x, y] = [\text{pos}(m_i), \text{pos}(m_j)]$, where pos denotes the marker location on a user defined physical map. Optionally, an interpolation algorithm can be used to complete the plot, and provide an indication of the rate of change of disequilibrium. Figure 1 illustrates typical graphical output. Interactive options allow the user to specify a disequilibrium measure to be plotted and to select an arbitrary physical or genetic map. In this example, note that although marker spacing is approximately even, the patterns of disequilibrium in the region are irregular. Command-line tools of GOLD were implemented in C++, while the graphical interface was implemented in Delphi Pascal and is compatible with Microsoft Windows.

The GOLD program differs from other tools for the analysis of linkage disequilibrium (such as Arlequin; http://anthropologie.unige.ch/arlequin/) because it provides a distinct graphical representation of disequilibrium patterns. Its automated interface to Simwalk2 makes analysis of disequilibrium patterns in family data convenient and easy to interpret and should facilitate analysis of linkage disequilibrium as the number of publicly available SNPs increases from 1000s to 100 000s (Masood, 1999).

Acknowledgements

We would like to thank Lon Cardon and Eric Sobel for helpful discussions on an earlier version of this manuscript. G.R.A. holds a Wellcome Trust Prize Studentship. W.O.C.C. is a Wellcome Trust Senior Clinical Research Fellow.

References

Collins,F.S. (1992) Positional cloning: let's not call it reverse anymore. *Nat. Genet.*, **1**, 3–6.

Kruglyak,L. (1999) Prospects for whole genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.

Lewontin,R.C. and Kojima,K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 450–472.

Masood,E. (1999) Consortium plans free map of human genome. *Nature*, **398**, 545–546.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Sobel,E. and Lange,K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.

Weir,B.S. (1996) *Genetic Data Analysis II*. Sinauer, Massachusetts.