SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal

Emily L. Webb, Gabrielle S. Sellick and Richard S. Houlston*

Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK

Received on March 4, 2005; revised on April 8, 2005; accepted on April 11, 2005
Advance Access publication April 19, 2005

ABSTRACT
Summary: SNPLINK is a Perl script that performs full genome linkage analysis of high-density single nucleotide polymorphism (SNP) marker sets. The presence of linkage disequilibrium (LD) between closely spaced SNP markers can falsely inflate linkage statistics. SNPLINK removes LD from the marker sets in an automated fashion before carrying out linkage analysis. SNPLINK can compute both parametric and non-parametric statistics, utilizing the freely available Allegro and Merlin software. Graphical outputs of whole genome multipoint linkage statistics are provided allowing comparison of results before and after the removal of LD.

Availability: SNPLINK is freely available for non-commercial research institutions. For full details see www.icr.ac.uk/cancgen/molgen/MolPopGen_Bioinformatics.htm

Contact: richard.houlston@icr.ac.uk

The recent availability of dense single nucleotide polymorphism (SNP) maps coupled with technological and cost developments in performing highly paralleled genotyping, now make it practical to use these markers for genomewide linkage searches (Matise et al., 2003). The advantages of conducting searches using SNPs over conventional microsatellite scans based on markers spaced at 10 Mb intervals have been well documented (Evans and Cardon, 2004). Specifically, SNPs are far more abundant than microsatellites throughout the genome and therefore yield a higher information content, giving superior power to detect linkage when typed at a high density (Evans and Cardon, 2004). Furthermore, the denseness of SNP marker maps means that there is potential for improved localization of disease genes (Kruglyak, 1997; John et al., 2004).

A number of analytical platforms capable of simultaneously scoring large numbers of SNPs (such as the Affymetrix GeneChip® Human Mapping 10K Array) have recently become available to mainstream genetic researchers. The performance of these SNP arrays, both in terms of efficiency and precision, indicates that they are likely to become the dominant technology for performing genomewide linkage searches (Sellick et al., 2004).

The current linkage software packages that are in popular use, such as Genehunter (Kruglyak et al., 1996), Merlin (Abecasis et al., 2002) and Allegro (Gudbjartsson et al., 2000) generate multipoint statistics under the assumption of linkage equilibrium between markers. While this would nearly always be the case for more widely, evenly spaced microsatellite markers, linkage disequilibrium (LD) is negatively correlated with increasing distance (Dawson et al., 2002; Ke et al., 2004), and has a high probability of existence between tightly spaced markers. In the Affymetrix GeneChip® Human Mapping 10K Array, SNPs are often found to be in clusters with distances as small as 1 × 10^{-6} Mb between consecutive markers.

The very property of SNP maps that affords the advantages detailed above therefore leads to the potential problem of LD between markers affecting linkage statistics (Evans and Cardon, 2004; Schaid, 2002). This is not an issue if the genotypes of all family members are included in the analysis. However, if there is missing data, for example, if parental genotypes are unknown, existing linkage software programs estimate the probability distribution of unknown genotypes by assigning equal probabilities to all possible inheritance vectors. Therefore, if markers are in LD (and are thus more likely to be inherited together) certain haplotypes may occur more frequently than expected under linkage equilibrium, thereby falsely inflating linkage statistics. While this type of bias can be reduced by including the genotypes of additional unaffected pedigree members in the analysis (Huang et al., 2004), it cannot be entirely eliminated and may even increase as the number of consecutive markers in LD increases.

Since the use of dense SNP marker sets in linkage analysis is a very recent development, no general approach has been proposed to deal with the existence of LD. Ideally, programs should incorporate LD between markers into the likelihood calculations so that expected haplotype frequencies are correctly estimated. In the absence of such software, other approaches are required to address the issue in the interim.

To this end, we have developed a Perl Script, SNPLINK that interfaces with the linkage software Merlin and Allegro, to undertake automated analyses of genomewide linkage scans, but unlike other programs it addresses the issue of LD. SNPLINK accepts genotype data in standard LINKAGE format as generated by the GeneChip® DNA Analysis Software (GDAS v3.0.2) from Affymetrix or by the pedigree database program ProgenyLab 6.0 (Progeny Software Inc, IN). Non-Mendelian errors are identified and removed using the Merlin option error. SNPLINK then implements full-genome multipoint linkage analysis using Allegro for parametric, and Merlin for non-parametric analyses. Both Allegro and Merlin are capable of analysing large numbers of markers (tested to a maximum of 945 SNPs) whereas other linkage programs, such as Genehunter (version 2.1), are restricted to fewer markers than are available.

*To whom correspondence should be addressed.
we suggest that the choice of disequilibrium measure to be used and the cut off point for the presence of LD should be determined after preliminary assessment of the data. Specifically, we recommend that the cut point should be chosen as the value at which the minimum of the density of the LD measure is attained.

In conclusion, the advantages of using SNPs for linkage analysis are clearly apparent, but care must be taken to avoid false linkage results due to LD between markers. SNPLINK provides a user-friendly, fully automated program for the systematic removal of LD between SNP markers and performs full genomewide linkage analysis both before and after LD is removed. SNPLINK has been extensively tested on real data and successfully used to perform several linkage scans using the Affymetrix 10K array.

ACKNOWLEDGEMENTS

This work was supported by grants from Cancer Research UK. G.S.S. is in receipt of a Post-Doctoral Research Fellowship from Leukaemia Research. We are grateful to two reviewers for their helpful comments.

REFERENCES


