Genetic linkage analysis in the age of whole-genome sequencing

Jurg Ott[1,2], Jing Wang[1] and Suzanne M. Leal[3]

Abstract | For many years, linkage analysis was the primary tool used for the genetic mapping of Mendelian and complex traits with familial aggregation. Linkage analysis was largely supplanted by the wide adoption of genome-wide association studies (GWASs). However, with the recent increased use of whole-genome sequencing (WGS), linkage analysis is again emerging as an important and powerful analysis method for the identification of genes involved in disease aetiology, often in conjunction with WGS filtering approaches. Here, we review the principles of linkage analysis and provide practical guidelines for carrying out linkage studies using WGS data.

Genetic mapping
The ordering of loci on a chromosome and the determination of the distances between two adjacent loci. For short distances, the recombination fraction can serve as a measure of genetic distance, with the unit of measurement being the centimorgan (cM); 1 cM = 1% recombination fraction.

[1]*Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing 100101, China.*
[2]*Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.*
[3]*Center for Statistical Genetics, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.*
Correspondence to J.O. and S.M.L.
e-mails: ottjurg@psych.ac.cn; sleal@bcm.edu
doi:10.1038/nrg3908
Published online 31 March 2015

Linkage analysis was the predominant statistical genetic mapping approach used in the latter half of the twentieth century. More recently, the focus shifted to association studies of complex traits that analyse common variants, which have a modest effect. For such variants, association analyses are more powerful than linkage analyses, and genome-wide association studies (GWASs) using single-nucleotide polymorphism (SNP) marker loci became the preferred association mapping tool. However, an emerging view is that rare variants, which are not well interrogated by GWASs, could be responsible for a substantial proportion of complex human disease[1]. Importantly, the increased availability of exome and whole-genome sequence data has brought linkage analysis once again to the forefront owing to the development of powerful methods to detect rare variants involved in disease aetiology using family-based data; such an approach has many advantages over simply using filter methods to identify causal variants. Several reviews[2–5] and books[6–8] have been written on genetic linkage analysis, but none, to our knowledge, covers linkage analysis coupled with whole-genome sequencing (WGS).

Several recent studies have generated genome-wide association data for families. For example, the T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples) consortium has generated WGS data on 1,043 individuals from 20 Mexican families and reported analysis of risk variants for type 2 diabetes. However, for cost reasons, most studies currently only obtain WGS data for a small number of family members.

To date, most family-based WGS studies have therefore been analysed using filtering approaches, and only a few family members are prioritized for sequencing (FIG. 1). However, filtering approaches do not offer statistical evidence of a variant's involvement in disease susceptibility, whereas linkage analysis does provide this statistical support. With the decreasing cost of sequencing, it will become more commonplace to have WGS data available for every informative pedigree member.

This Review provides the reader with a practical guide for performing linkage analysis to identify variants that are responsible for Mendelian[9] trait aetiology. After briefly mentioning the relative merits of linkage and association analysis, we discuss linkage algorithms and their implementations in computer programs, with a special emphasis on the use of sequence data. We then outline a step-by-step approach to successful linkage analysis using WGS data.

Genome-wide linkage analysis

For all informative family members, genotypes can be generated using SNP arrays and analysed using genome-wide linkage analysis. This approach is beneficial in that it evaluates DNA sample quality; elucidates whether specified familial relationships are correct; allows the detection of mis-specification of affection status and locus heterogeneity; aids the selection of an individual (or individuals) to undergo WGS; and facilitates the mapping of the disease locus to a region (or regions) of the genome, thus reducing the number of variants that need to be followed up. Linkage analysis can also provide statistical evidence of the involvement of a variant or gene in disease aetiology and can be performed either directly using WGS data or after filtering using data on variants