A Novel Approach and Software to Use Sequence Data for Linkage Analysis

Gao T. Wang1, Di Zhang1, Biao Li1, Hang Dai1, Suzanne M. Leal1, \*

1Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

\*Correspondence:

Suzanne M. Leal

Department of Molecular and Human Genetics

Baylor College of Medicine

One Baylor Plaza, 700D

Houston, TX 77030

713-798-4011

[sleal@bcm.edu](mailto:sleal@bcm.edu)**Abstract**

Traditionally, linkage analysis has been the main approach to elucidate causes of Mendelian disorders in families with multiple affected individuals. Recent advances in next generation sequencing (NGS) technology has made it standard approach to prioritize and screen for causal mutations in whole-genome or whole-exome sequenced individuals suffering from Mendelian diseases. Linkage analysis using SNP markers in conjunction with sequence analysis is gaining popularity in the human genetics community, yet there are few applications of directly performing linkage analysis using sequence data. Inspired by the “aggregation analysis” commonly practiced in rare variants complex disease association studies, we have developed a collapsed haplotype pattern method to generate markers from sequence data for linkage analysis. We demonstrate with a range of simulation studies of two-point linkage analysis that our method is substantially more powerful over linkage analysis using single nucleotide variants. We developed the XXX software package that uses the method described here to perform two-point linkage analysis on sequence data. Additionally XXX can output marker data in formats compatible with XXX, XXX, and XXX software, reviving these linkage analysis tools for use in NGS era. To illustrate the possibilities we examined a data set … and detected linkage of XXX (disease) to XXX (gene). **Introduction**

Long version (4 paragraphs) 1. Review of linkage methods, 2. Review of variants prioritization and screening, 3. The motivation to use sequence data for linkage analysis and 4. Introduce our method.

Short version (2 paragraphs) 1. The motivation to use sequence data for linkage analysis and 2. Introduce our method.

**Methods**

*Linkage Analysis with Collapsed Haplotype Patterns*

The idea of collapsed haplotype pattern method is that instead of analyzing each SNV separately, we create numeric representations of haplotypes within a given genetic region, e.g., genes for whole-exome sequences. These numeric representations contain information on the transmission of variants in the entire region informative for linkage analysis, and are mathematically compatible with the off-shelf linkage methods that are traditionally used to analyze RFLP or SNP genotype markers. The advantage of analyzing multiple SNVs collectively as regional markers is that allelic heterogeneity within the region of interest across multiple samples from different families is naturally handled by the coding method.

To generate regional genotype markers, haplotypes for the region must be obtained for all samples. Samples of the same family are first checked for Mendelian errors and inconsistencies detected are treated as missing data. The Merlin’s improved version of the Lander-Green algorithm for genetic phasing and missing genotype imputation was then applied to reconstruct haplotypes in pedigree. The resulting haplotypes are converted to markers for linkage analysis using the collapsed haplotype pattern method (CHP) illustrated in Fig. S1. For each pedigree, we first cluster variants on regional haplotypes by “windows” and collapse variants in a window into an indicator variable of values 0 or 1 for having no minor allele or at least one minor allele within the window, which is similar to the collapsing method for association analysis with rare variants {citation CMC}. We then assign each collapsed haplotype a single numeric code such that different patterns of collapsed haplotypes in all samples are uniquely represented. The choice of coding for patterns is arbitrary, although we use continuous positive integers and assign a smaller value for collapsed haplotypes having more 0’s than 1’s (Fig. S1). The sample haplotypes thus represented can be directly used for linkage analysis with existing software packages.

In order to reconstruct genotypes of unavailable individuals, linkage analysis requires that allele frequencies of markers be provided. Frequencies of regional markers generated by CHP method can be derived from the cumulative minor allele frequency (MAF) within collapsing windows. Consider a haplotype divided into *K* windows with each window having *J* variants. Cumulative MAF for window *k* is  where is MAF for the *jth* locus in the *kth* window. The collapsed haplotype *h* contains *K* elements, and the allele frequency for *h* is given by. For collapsed haplotype patterns within a pedigree, the allele frequencies will be normalized such that they sum to 1. The normalized collapsed haplotype pattern frequencies are then used as the allele frequencies for the corresponding regional genotype markers in linkage analysis.

To facilitate linkage analysis using sequence data in VCF format, we developed a software package XXX that incorporates the MLINK routine and produces high quality graphical report on both LOD and HLOD scores for two-point linkage analysis on regional genotype markers generated by the CHP method previously described. Additionally XXX supports output of regional genotype data into formats compatible with linkage software such as XXX, Merlin and MEGA2, with which multipoint and non-parametric linkage analysis can be performed. For multipoint linkage analysis, the median of map distance for variants in the region of interested is extracted from the Rutgers Map database {citation} as the surrogate for genetic distance of the regional marker.

*Simulation studies*

**Discussion**

In haplotype reconstruction for the CHP method, loci that cannot be deterministically phased using available transmission information are assigned an arbitrary phase since all possible phases will yield to the same regional marker coding, as shown in Fig. S2. Definition of regions for marker generation and the choice of window size for variants collapsing can be arbitrary. For whole-exome sequence data we recommend using genes as regional markers. Within each region, commonly used window size options for variants collapsing are complete collapsing (window size equals region length), no collapsing (window size equals one) and LD based collapsing (use LD blocks as windows), although other arbitrary window size options are also available from XXX. For regions with recombination events, the region is divided into sub-units by loci where a recombination event is observed in any sample in family, and in linkage analysis the sub-unit giving the strongest evidence of linkage will be chosen to represent the entire region. In model based linkage analysis, correct specification of marker allele frequency is crucial to controlling for type I error and reducing type II error {citation Freimer 1993}. The number of founders with available genotypes in data for linkage analysis might often be too small to obtain a decent allele frequency estimate, thus we recommend the input VCF file be annotated with external source of MAF information such as 1000 genomes or EVS server. It may be difficult to estimate allele frequencies in the particular population under study if such a population is not well defined, in which case frequencies estimated from the founders has to be used.

The XXX package, written in Python and C++, is a user friendly software that uses the CHP method described here to perform linkage analysis as well as data format conversion on sequence data. It is a useful tool to bring back the long established linkage analysis

**Web Resources**

**Acknowledgements**

We would like to thank … This work is supported by National Institute of Health grants …

**References**

**Tables**

**Figure legends**