|  |
| --- |
| Genetics and Population Analysis  **A Novel Approach and Software to Use Sequence Data for Linkage Analysis**  Gao T. Wang1, Di Zhang1, Biao Li1, Hang Dai1, Suzanne M. Leal1, \*  1Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Motivation:** 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 arrays 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.

**Summary:** 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 linkage analysis on sequence data. Additionally XXX can output marker data in formats compatible with XXX, XXX, and XXX software, reviving many linkage analysis tools for use in NGS era.

**Availability:** xxx, including source and multi-platform executables, documentation and examples, is available athttp://bioinformatics.org/xxx

**Contact:** sleal@bcm.edu

# introduction

The advent and advance of next generation sequencing technology in recent years has led to identifications of a large number of genes for rare Mendelian disorders. Typical approach to find Mendelian disease causing variants using whole-genome (WGS) or whole exome sequence (WES) data is to filter variants only shared by affected relatives in family. While the filtering method is straightforward and has been successful (Ng *et al.*, 2010), such efforts rely solely on obvious co-segregation of variants and pieces of information from external resources on variant functional characterization. Linkage analysis, on the other hand, takes advantage of inheritance information involving modeling of inheritance mode, reduced penetrance, allele frequency and genetic map information, and remains a powerful tool to detect Mendelian disease loci. As a result, combined SNP arrays based linkage analysis and sequence based filtering method is starting to become popular in human genetics community (Louis-Dit-Picard *et al.*, 2012; Yan *et al.*, 2013). Although it has been shown that single nucleotide variants (SNVs) in linkage equilibrium from WES data provides acceptable linkage analysis results, such SNVs are less heterozygous and less powerful compare to using SNP arrays (Smith *et al.*, 2011).

We developed a collapsed haplotype pattern (CHP) method to recode sequence data for linkage analysis that results in markers that are more likely to be heterozygous and informative for linkage analysis. Our method does not require LD pruning and is particularly powerful when disease causing mutations harbor in the same genetic region but different loci across multiple families. We provide software package XXX implementing our method to facilitate linkage analysis on NGS data.

# Methods

*Linkage Analysis with Collapsed Haplotype Patterns*

The idea of CHP method is that instead of analyzing each SNV separately, we create numeric representations of haplotypes within a given genetic region, e.g., genes for WES data. Such 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 microsatellite or SNP array 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 (Abecasis and Wigginton, 2005). The resulting haplotypes are converted to markers for linkage analysis using CHP method illustrated in Fig. S1. For each pedigree, we first cluster variants on regional haplotypes by “windows”, e.g., LD blocks, 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 (Li and Leal, 2008). 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. The sample haplotypes thus represented can be directly used for linkage analysis with many 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 (Abecasis *et al.*, 2002) and MEGA2 (Mukhopadhyay *et al.*, 2005), 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 (Matise *et al.*, 2007) 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 estimated 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 (Freimer *et al.*, 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 Exome Variant 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++, can efficiently extract genotypes from VCF files and use the CHP method described here to perform linkage analysis as well as data format conversion on sequence data. It provides a novel and effective approach that brings back well established linkage analysis techniques for use with the growing wealth of genomic data of human pedigrees. We recommend the use of XXX in parallel to filtering methods on the same sequence data to take full advantage of the power of NGS in families.

acknowledgments

*Funding*: This work is supported by National Institute of Health (XXX, XXX and XXX)

*Conflict of Interest: none declared.*

References

Abecasis,G.R. *et al.* (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.

Abecasis,G.R. and Wigginton,J.E. (2005) Handling Marker-Marker Linkage Disequilibrium: Pedigree Analysis with Clustered Markers. *Am. J. Hum. Genet.*, **77**, 754–767.

Freimer,N.B. *et al.* (1993) Incorrect specification of marker allele frequencies: effects on linkage analysis. *Am. J. Hum. Genet.*, **52**, 1102–1110.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Louis-Dit-Picard,H. *et al.* (2012) KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nat. Genet.*, **44**, 456–460.

Matise,T.C. *et al.* (2007) A second-generation combined linkage physical map of the human genome. *Genome Res.*, **17**, 1783–1786.

Mukhopadhyay,N. *et al.* (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics*, **21**, 2556–2557.

Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.*, **42**, 30–35.

Smith,K.R. *et al.* (2011) Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.*, **12**, R85.

Yan,J. *et al.* (2013) Combined linkage analysis and exome sequencing identifies novel genes for familial goiter. *J. Hum. Genet.*, **58**, 366–377.

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)