|  |
| --- |
| 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 SEQLinkage software package that uses the method described here to perform linkage analysis on sequence data. Additionally SEQLinkage can output marker data in formats compatible with LINKAGE, MERLIN and MEGA2 software, reviving many linkage analysis tools for use in NGS era.

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

**Contact:** sleal@bcm.edu

# introduction

The advent and advance of next generation sequencing technology (NGS) 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 array 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 results, such markers 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, creating markers that are more 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 across different loci among multiple families. We provide software package SEQLinkage implementing our method to facilitate linkage analysis using NGS data.

# Methods

The idea of CHP method is that instead of analyzing each variant separately, we create numeric representations of haplotypes within a given genetic region containing multiple variants, e.g., genes for WES data, resulting in regional markers which sufficiently reflect the transmission pattern of the entire region informative for linkage analysis, and are numerically compatible with the off-the-shelf linkage methods that are traditionally used to analyze microsatellite or SNP array markers. Such regional markers are often more heterozygous than single locus markers, and allelic heterogeneity within the region of interest across individuals from different families is naturally taken into consideration 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; inconsistencies detected are treated as missing data. An 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 “bins”, e.g., LD blocks, and collapse variants in a bin into an indicator variable of values 0 or 1 for having no minor allele or at least one minor allele within the bin, 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 value 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 recover 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 bins. Consider a haplotype divided into *K* bins with each bin having *J* variants. Cumulative MAF for bin *k* is  where is MAF for the *jth* locus in the bin. 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 for linkage analysis.

For variants having high population frequency it is not advisable to include them in regional marker generation, as their genotypes may predominate the marker pattern. We therefore exclude variants up to a specified MAF cutoff and treat those variants as standalone SNP markers.

To facilitate linkage analysis using sequence data in VCF format, we developed a software package SEQLinkage that incorporates the FASTLINK routine (whose execution time scales linearly with family size) (Cottingham *et al.*, 1993) and produces high quality graphical report on both LOD and HLOD scores for two-point linkage analysis on regional markers generated by the CHP method previously described. Additionally SEQLinkage supports output of regional genotype data into formats compatible with linkage software such as LINKAGE (Lathrop *et al.*, 1984), 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 Combined Linkage-Physical Map (Matise *et al.*, 2007) to approximate the genetic distance of the regional marker.

# Results

To demonstrate the usefulness of CHP method, we performed empirical power calculations of two-point linkage analysis on several known non-syndromic hearing loss genes (see Supplemental Methods for simulation details). Power analysis results are shown in Figures S3 ~ SXX. Although CHP is underpowered for dominant model without allelic heterogeneity, it is as good as single marker analysis method for recessive model without allelic heterogeneity, and is substantially more powerful for compound recessive model and for all models in the presence of allelic heterogeneity.

# discussion

In the haplotype reconstruction step for CHP method, loci that cannot be deterministically phased using available transmission information are assigned to the phase that results in smallest number of regional marker alleles, as shown in Fig. S2. Definition of regions for marker generation and the choice of bin size for variants collapsing can be arbitrary. For WES data one can use genes as regional markers. Within each region, commonly used bin size options for variants collapsing are LD based collapsing (use estimated LD blocks as bins), complete collapsing (bin size equals region length) and no collapsing (bin size equals one), although other arbitrary bin size options are also available in SEQLinkage software. 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 SEQLinkage 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 SEQLinkage in parallel to filtering methods on the same sequence data to take full advantage of the power of NGS in families.

acknowledgments

*Funding*: The authors would like to thank Dr. Regie Lyn Santos-Cortez for helpful discussions. 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.

Cottingham,R.W.,Jr *et al.* (1993) Faster sequential genetic linkage computations. *Am. J. Hum. Genet.*, **53**, 252–263.

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

Lathrop,G.M. *et al.* (1984) Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci.*, **81**, 3443–3446.

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)