|  |
| --- |
| 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, \*  1Center for Statistical Genetics, Department 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 was used to map Mendelian diseases and genes within linked regions were sequenced to identify the causal variants responsible for Mendelian disorders. Recent advances in next generation sequencing (NGS) make it possible to directly sequence genomes and exomes of individuals with Mendelian diseases and screen the sequence data for causal mutations. In order to reduce the number of variants which must be screened, and to increase the success of identifying the causal variant, results from linkage analysis which is performed with SNPs is used in conjunction with NGS. With the reduction of cost of NGS, entire families which segregate Mendelian traits can be sequenced and linkage analysis can be performed directly using NGS data.

**Summary:** Inspired by “burden” tests which are commonly used for complex trait rare variant association studies, we developed a collapsed haplotype pattern (CHP) method to generate markers from sequence data for linkage analysis. We demonstrate with simulation studies that the CHP method is substantially more powerful over linkage analysis using SNVs. The SEQLinkage software package that uses the CHP method was developed to perform linkage analysis using sequence data. Additionally, SEQLinkage can generate marker data in formats compatible with a number of programs including 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 a support forum, is available athttp://bioinformatics.org/seqlink

**Contact:** sleal@bcm.edu

# introduction

The advent and advance of NGS in recent years has led to the identification of a large number of Mendelian disease genes. The typical approach to identifying Mendelian disease causal variants using either whole-genome sequence (WGS) or exome sequence (ES) data is to filter variants in an affected individual or shared by affected family members, removing those which are found at higher frequencies, e.g. 0.1% in variant databases. Sometimes unaffected family member(s) are also used in the filtering process. While filtering is straightforward and has been successful (Ng *et al.*, 2010), such efforts rely on limited family information, e.g. mode of inheritance, sharing between a subset of family members and information from external resources on variant functional characterization and frequencies. On the other hand, linkage analysis, which incorporates information on mode of inheritance, penetrance, allele frequency and genetic map information, remains a powerful tool to identify 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; Santos-Cortez *et al.*, 2013). Although it has been shown that analyzing SNVs from ES data provide acceptable linkage results, due to their low heterozygosity they provide less power than the analysis of SNPs on genotyping arrays (Smith *et al.*, 2011).

We developed SEQLinkage package using collapsed haplotype pattern (CHP) method to create markers that are more heterozygous and informative for linkage analysis than individual SNVs. Unlike when SNPs are analyzed the CHP method does not require LD pruning and is particularly powerful in the presents of intra (e.g. compound heterozygotes) and inter familial allelic heterogeneity. Additionally because SEQLinkage can calculate HLODs it is remains powerful when there is locus heterogeneity, i.e. the underlying genetic etiology is not due to the same gene/region in all families.

# Methods

For the CHP method instead of analyzing each variant separately, multiple variants which form haplotypes within a genetic region, e.g. gene, are analyzed. This is done by constructing a marker which reflects the transmission pattern of the entire region and is numerically compatible with currently available linkage analysis methods and software. These markers incorporate allelic heterogeneity which is present between and within families in a region and often have higher heterozygosity than SNVs, making them more informative and powerful to detect linkage.

To generate regional 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 pedigrees (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 with 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 are 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 linkage 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 collapsed 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 above 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. Additionally SEQLinkage supports output of regional genotype data into formats compatible with linkage software such as LINKAGE (Lathrop *et al.*, 1984) and Merlin (Abecasis *et al.*, 2002), with which two-point and multipoint linkage analysis can be performed. Additionally MEGA2 (Mukhopadhyay *et al.*, 2005) format is supported, which can be used to transform the data to required input for a number of linkage programs including GeneHunter (Kruglyak *et al.*, 1996). 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 study design). Power analysis results suggests that CHP is slightly more powerful over single marker method under recessive and dominant models in the absence of allelic heterogeneity (Panel A of Figures 3S and 4S), and is substantially more powerful for all models in the presence of intra (Figure 5S) and inter-family allelic heterogeneity (Panel B of Figures 3S – 8S). Specifically for compound recessive model with inter-family allelic heterogeneity of 50%, it requires 12 families for CHP to achieve a power of 90% for gene *SLC26A4*, while single marker method requires 55 families to achieve the same power.

# 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 regional marker alleles most informative for linkage, as shown in Fig. S2. Definition of regions for marker generation and the choice of bin size for variants collapsing can be arbitrary. For ES data genes can be used as regional markers. Within each region, commonly used bin size options for variants collapsing are 1) LD based collapsing, which uses estimated LD blocks as bins, 2) complete collapsing, whose bin size equals gene/region length and 3) no collapsing, whose bin size equals one. Additional binning options are available and described in the supplemental materials. For regions with recombination events within a family, and the sub-unit giving the strongest evidence of linkage is 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 from 1000 genomes or Exome Variant Server. For some populations understudy allele frequency information may not be available and frequencies estimated from founders have 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. Unlike filtering, analysis with SEQLinkage provides statistical evidence that a gene regional is involved in disease etiology. 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.

Kruglyak,L. *et al.* (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.

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.

Santos-Cortez,R.L.P. *et al.* (2013) Mutations in KARS, Encoding Lysyl-tRNA Synthetase, Cause Autosomal-Recessive Nonsyndromic Hearing Impairment DFNB89. *Am. J. Hum. Genet.*, **93**, 132–140.

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.

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