|  |
| --- |
| Genetics and Population Analysis  **A Novel Approach and Software to Use Next-Generation-Sequencing 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 the linked regions were sequenced to identify the causal variants. 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 causal variants, results from linkage analysis are used in conjunction with NGS. With the reduction in cost of NGS, DNA samples from entire families can be sequenced and linkage analysis can be performed directly using NGS data.

**Summary:** Inspired by “burden” tests which are 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 than performing linkage analysis using SNVs. The SEQLinkage software package, that uses the CHP method, was developed to perform linkage analysis using NGS data. Additionally, SEQLinkage can generate marker data in formats compatible with a number of programs including FASTLINK/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 (WES) 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 frequencies 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 again become popular (Santos-Cortez *et al.*, 2013). Although it has been shown that analyzing SNVs from WES data provides acceptable linkage results, due to the low heterozygosity of SNVs they provide less power than analysis of SNPs from genotyping arrays (Smith *et al.*, 2011).

We developed the collapsed haplotype pattern (CHP) method to create markers that are more heterozygous and informative for linkage analysis than individual SNVs, and a software package SEQLinkage implementing the CHP method. 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-family allelic heterogeneity. 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 with sequence data. NGS data from family members are first checked for Mendelian errors; Mendelian inconsistencies are treated as missing data. An improved version of the Lander-Green algorithm for genetic phasing and missing genotype imputation is then applied to reconstruct haplotypes in the 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 collapsing method for association analysis with rare variants (Li and Leal, 2008). We then assign each collapsed haplotype a single numeric value so 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 reconstruct genotypes for family members without sequence data, linkage analysis requires marker allele frequencies. 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 variants having high MAFs 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 these markers are analyzed individually.

To facilitate linkage analysis using sequence data in VCF format, we developed the SEQLinkage software that uses the Elston-Stewart algorithm as incorporated in FASTLINK (Cottingham *et al.*, 1993). It provides results in text format and high quality graphical reports for both LOD and HLOD scores. 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.

# results

To demonstrate the usefulness of the CHP method, we performed empirical power calculations of two-point linkage analysis on several known non-syndromic hearing loss genes (see Supplemental Methods). Power analysis on LOD and HLOD suggests that CHP 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 an autosomal recessive model with allelic heterogeneity, i.e. compound heterozygotes and also with locus heterogeneity of 50%, it requires 12 families for CHP to achieve a power of 90% for gene *SLC26A4*, while analyzing individual SNVs requires >50 families to achieve the same power.

# discussion

For WES 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 described in the supplemental materials. For regions with recombination events, the sub-unit that shows the strongest evidence of linkage among all sub-units created by recombination events is chosen to represent the entire region. For 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 sufficiently accurate 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 MAF 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 the commonly practiced filtering approaches used for sequence data, SEQLinkage provides statistical evidence of the involvement of genes/variants in the etiology of Mendelian diseases. Additionally because it incorporates inheritance information and penetrance models it is less likely than filtering to exclude causal genes/variants in the presents of phenocopies and/or reduced penetrance. 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

The authors would like to thank Regie Lyn Santos-Cortez, Daniel Weeks, Alejandro Schaffer, Jeffrey O’Connell and Jurg Ott for helpful discussions and support. This work is funded by National Institute of Health ([DC003594](http://projectreporter.nih.gov/project_info_details.cfm?aid=8577390&icde=20234922&ddparam=&ddvalue=&ddsub=&cr=1&csb=default&cs=ASC), [DC011651](http://projectreporter.nih.gov/project_info_details.cfm?aid=8642638&icde=20234922&ddparam=&ddvalue=&ddsub=&cr=2&csb=default&cs=ASC) and HG006493).

*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.

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)