Collapsed Haplotype Pattern Method for Linkage Analysis of Next-Generation Sequencing Data

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**

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. 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. 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 demonstrated with simulation studies that the CHP method is substantially more powerful than performing linkage analysis using single variant markers. The CHP method was implemented in the SEQLinkage, a software package to perform linkage analysis using NGS data as well as to generate markers from NGS 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.

**Keywords**

Linkage analysis method, Next-generation sequence data, Mendelian disease mapping, Bioinformatics tool.**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 successful1, 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 becoming popular2. 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 arrays3.

Here we describe a 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-family allelic heterogeneity. We have developed the SEQLinkage software package implementing t Since SEQLinkage can calculate HLODs our approach remains powerful when there is locus heterogeneity, i.e. the underlying genetic etiology is not due to the same gene/region in all families.

**Materials and 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 pedigrees4. The resulting haplotypes are converted to markers for linkage analysis using CHP method illustrated in Figure 1. 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 variants5. 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 *j*th 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 FASTLINK6. 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 LINKAGE7 and Merlin8, with which two-point and multipoint linkage analysis can be performed. Additionally MEGA29 format is supported, which can be used to transform data to the required input for a number of linkage programs.

To evaluate performance of our method we performed empirical power calculations of two-point linkage analysis using data on four non-syndromic hearing impairment (NSHI) genes: two autosomal recessive genes *GJB2* and *SLC26A4*, and two autosomal dominant genes *MYO7A* and *MYH9*. We simulated 2-generational pedigrees, allowing for 3 up to 8 offspring in the last generation with the proportions determined by the distribution of number of children per family in the United States in 2012, rescaled such that these proportions add up to 100% (3 children: 69.34%, 4 children: 20.52%, 5 children: 6.84%, 6 children: 2.28%, 7 children 0.76%, 8 children 0.26%). Genotypes are simulated for the four genes based on the variant sites and the corresponding minor allele frequencies in European Americans recorded in Exome Variant Server. We annotate these variants using Deafness Variation Database (DVD) and NCBI ClinVar, labelling variants as “pathogenic” if they present in both databases as pathogenic. Disease status for individuals are determined by genotypes on those pathogenic sites under dominant mode of inheritance for *MYO7A* and *MYH9*, and recessive (compound heterozygotes and homozygotes) for *GJB2* and *SLC26A4*, assuming complete penetrance. Additionally for each mode of inheritance we allow for allelic heterogeneity among families, i.e., the causal variant site in a gene may not be the same for different families. We ascertain simulated families having two or more affected offspring for linkage analysis. To introduce locus heterogeneity we mix families having pathogenic mutations in one gene but not others, so that each simulated gene contributes to etiology of only a proportion of families in the entire dataset. We simulate 500 replicates under each different setting of sample size, modes of inheritance, presence of allelic heterogeneity and locus heterogeneity. For each replicate we compute LOD and HLOD (heterogeneity LOD) scores using regional markers generated by collapsed haplotype pattern markers (CHP) as well as the maximum LOD and HLOD scores analyzing individual SNV markers for comparison purposes. Power is estimated by  where the denominator is the total number of replicates and the numerator is the number of tests that successfully detected the simulated linkage signal, i.e. LOD score greater than 3.3 or HLOD score greater than 3.6.

**Results**

Empirical power calculations of two-point linkage analysis on several known non-syndromic hearing loss genes using CHP and individual SNV are summarized by contour plots as displayed in Figures 2. Power analysis on LOD and HLOD suggests that CHP is substantially more powerful for all models in the presence of intra- (Figure 2C) and inter-family allelic heterogeneity (Figures 2A – 2C). 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 at a significance level of α=0.05.

**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 error10. 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, freely available at <http://bioinformatics.org/seqlink>, 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 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 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.**Acknowledgements**

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, DC011651 and HG006493).

*Conflict of Interest: none declared.*

**References**

1 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42**: 30–35.

2 Santos-Cortez RLP, Lee K, Azeem Z, Antonellis PJ, Pollock LM, Khan S *et al.* Mutations in KARS, Encoding Lysyl-tRNA Synthetase, Cause Autosomal-Recessive Nonsyndromic Hearing Impairment DFNB89. *Am J Hum Genet* 2013; **93**: 132–140.

3 Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H *et al.* Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol* 2011; **12**: R85.

4 Abecasis GR, Wigginton JE. Handling Marker-Marker Linkage Disequilibrium: Pedigree Analysis with Clustered Markers. *Am J Hum Genet* 2005; **77**: 754–767.

5 Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.

6 Cottingham RW Jr, Idury RM, Schäffer AA. Faster sequential genetic linkage computations. *Am J Hum Genet* 1993; **53**: 252–263.

7 Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci* 1984; **81**: 3443–3446.

8 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.

9 Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* 2005; **21**: 2556–2557.

10 Freimer NB, Sandkuijl LA, Blower SM. Incorrect specification of marker allele frequencies: effects on linkage analysis. *Am J Hum Genet* 1993; **52**: 1102–1110.

**Web Resources**

America's Families and Living Arrangements, https://www.census.gov/prod/2013pubs/p20-570.pdf

Exome Variant Server (EVS), http://evs.gs.washington.edu/EVS

Deafness Variation Database (DVD), http://deafnessvariationdatabase.com

NCBI ClinVar, https://www.ncbi.nlm.nih.gov/clinvar

**Figure legends**

**Figure 1. The Collapsed Haplotype Pattern (CHP) method**. This figure illustrates the creation of regional markers for a group of 6 variants in pedigree with autosomal compound recessive disorder. Panel A displays the use of original haplotype patterns as regional markers, a special case of the CHP method using bins of size 1; panel B displays the *complete collapsing* theme with bin size equaling the length of the region; panel C displays a collapsing theme with bins of size 3 and panel D displays the *LD based collapsing* theme assuming the 2nd to 6th variant loci are in LD with each other.

**Figure 2. Power comparisons for LOD and HLOD statistics in two-point linkage analyses.** This figure shows the power for collapsed haplotype pattern markers (CHP) vs. single nucleotide variant (SNV) analysis under various modes of inheritance in the presence of inter-family allelic heterogeneity. X-axis is number of families, Y-axis is proportion of locus heterogeneity, i.e. the proportion of families with non-syndromic hearing impairment (NSHI) caused by detrimental mutations in the gene under investigation, i.e. either *MYO7A* or *MYH9* for dominant model, or *GJB2* or *SLC26A4* for recessive model. Contour curves on the graphs are power estimates, dark orange lines for CHP and light blue lines for SNV. Panel A displays the power for the LOD and HLOD statistics under dominant model; panel B displays the power for the LOD and HLOD statistics under recessive model; panel C displays the power for the LOD and HLOD statistics under recessive model in the presence of intra-family allelic heterogeneity, i.e. affected individuals are compound heterozygous. CHP method is more powerful for both LOD and HLOD at a significance level of α=0.05, but the absolute power of HLOD is not greater than LOD. This is because for most families the non-causal gene often has no variants at all and therefore are uninformative for HLOD calculation, due to the very low frequencies of the genes under investigation.