Collapsed haplotype pattern method for linkage analysis of next-generation sequencing data

Gao T. Wang, Di Zhang, Biao Li, Hang Dai and Suzanne M. Leal†

†Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

Center for statistical genetics,

Motivation

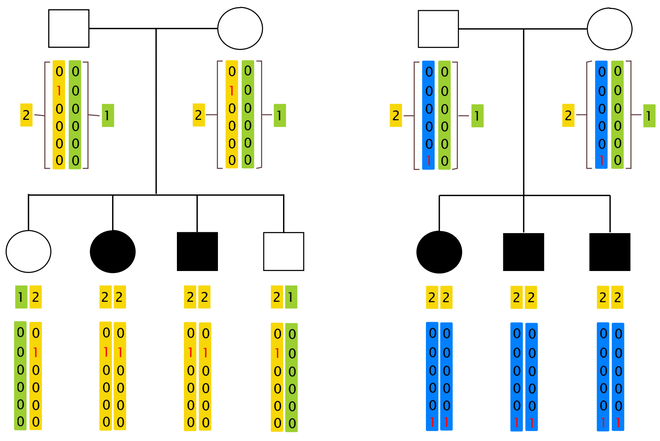
- Variant filtering using next-generation sequencing (NGS) data of families has successfully identified many causal mutations for Mendelian diseases, yet such approach
  - ▷ May result in many variants to follow-up
  - ▷ Is sensitive to mis-classification
    - ▷ sample swaps, phenocopies, reduced penetrance
  - ▷ Do not provide statistical significance for variants
- Linkage analysis has been a powerful approach to map Mendelian disease loci using genetic marker data, but is under-powered when applied to sequence data, mainly due to lack of heterogeneity in single-nucleotide variants
- We developed a ***Collapsed Haplotype Pattern*** (CHP) method and a **SEQLinkage** software to utilize NGS data of multiple families for powerful linkage analysis

The CHP Method

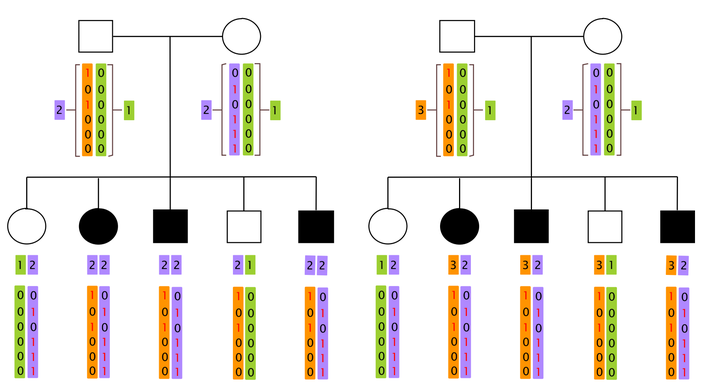
Inspired by rare variant "burden test" for association, the CHP method tests for linkage with a genetic region rather than with individual variants

- Rare variants from NGS data are analyzed in aggregate
- **Regional markers** are generated for given genomic units
  - ▷ *e.g.* genes for exome sequence data
- Regional markers are **more heterozygous** than single variants, thus more informative in tracking the transmission of disease mutations within families
- **Tolerates missing data**; no LD pruning required

Complete Collapsing



Complete vs. No Collapsing



Implementation

1. Family NGS data preprocessing

- Mendelian error check and genotype imputation
- Haplotype reconstruction via genetic phasing

2. Regional marker construction

- Complete collapsing and LD based collapsing themes
- Sub-regional markers by recombination events

3. Annotation and statistics for linkage analysis

- Genetic distance interpolation via *Rutgers Map*
- Calculation of marker allele frequencies for linkage analysis with samples having no genotype data

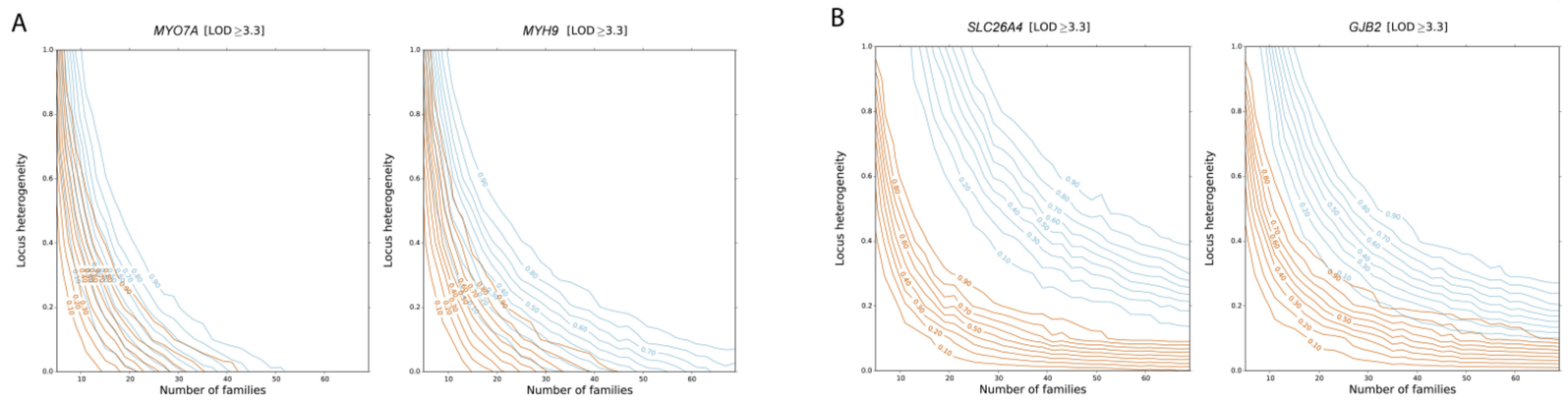
4. Parametric test for linkage using regional markers

- LOD/HLOD scores computed over multiple pedigrees

Power analyses and sample size estimations

Simulation methods

4 nonsyndromic hearing impairment gene sequences simulated, using sequences from European American samples in NHLBI Exome Sequencing Project; causal variants determined by NCBI-Clinvar database; **two-generational pedigrees** generated with 3 to 8 offspring based on USA population demographic data; **full-penetrance** for causal variants, allowing for **allelic heterogeneity** and varying degrees of **locus heterogeneity**; two-point linkage analysis performed comparing **CHP vs. single variant linkage** methods; empirical power evaluated via 500 replicates



Power comparisons

CHP (orange curves) outperforms single variant linkage method (blue curves) under both dominant (A) and compound recessive (B) models. With 50% locus heterogeneity it requires **12 families** for CHP to achieve a power of 90% for *SLC26A4* at a genome-wide α level of 0.05, while single variant linkage method requires over **50 families**

Sample size estimations

Number of families required to achieve desired power are evaluated; **50% locus heterogeneity** is assumed for all scenarios; power calculations are based on HLOD score instead; impact of **missing causal variant** in families is evaluated

Required Power	Gene	MOI	CHP[a]	SNV[b]	CHP-M75%[c]
0.8	*SLC26A4*	recessive	11	40	39
0.9	*SLC26A4*	recessive	13	45	46
0.8	*SLC26A4*	compound recessive	11	50	39
0.9	*SLC26A4*	compound recessive	13	55	46
0.8	*GJB2*	recessive	12	23	44
0.9	*GJB2*	recessive	14	28	52
0.8	*GJB2*	compound recessive	12	25	44
0.9	*GJB2*	compound recessive	14	34	52
0.8	*MYO7A*	dominant	12	16	31
0.9	*MYO7A*	dominant	14	20	36
0.8	*MYH9*	dominant	11	13	32
0.9	*MYH9*	dominant	14	18	41

a. minimum number of families required to achieve desired power for CHP method
b. minimum number of families required to achieve desired power for single variant linkage method
c. minimum number of family requirement for CHP when the causal variant in 75% families are missing

The SEQLinkage software

- Written in `C++` with a `Python` parallel computing interface to rapidly scan through markers genome-wide
- Implements the CHP method for linkage analysis with sequence data of pedigrees in VCF format
- Supports output of CHP coded markers to formats compatible with other linkage programs
  - ▷ *FASTLINK, MEGA2, Merlin, PLINK*
- Performs two-point linkage analysis involving multiple families, maximizing linkage signals across families to allow for locus heterogeneity (HLOD score calculation)
- Provides results in text, graphical and table format output, organized under a user-friendly webpage interface
- Can be used in conjunction with variant filtering method to analyze sequence data of human pedigrees

SEQLinkage is freely available at

`http://bioinformatics.org/seqlink`

Acknowledgments

We would like to thank Regie Lyn Santos-Cortez, Daniel Weeks, Alejandro Schaffer, Jeffrey O'Connell and Jürg Ott for helpful discussions and support. This work was funded by National Institute of Health grants DC003594, DC011651 and HG006493.

SEQLinkage analysis output

Whole exome sequences of 2 nuclear families each with 4 offspring (3 affected) are analyzed using **SEQLinkage**

θ=0.0		θ=0.02		θ=0.04		θ=0.06		θ=0.08	
Lod	Marker name chr:start-end	Lod	Marker name chr:start-end	Lod	Marker name chr:start-end	Lod	Marker name chr:start-end	Lod	Marker name chr:start-end
2.658	GRIP2 3:14530618-14583588	2.536	GRIP2 3:14530618-14583588	2.412	GRIP2 3:14530618-14583588	2.286	GRIP2 3:14530618-14583588	2.158	GRIP2 3:14530618-14583588
1.329	UBE2Q1 1:154521050-154531120	1.276	UBE2Q1 1:154521050-154531120	1.222	UBE2Q1 1:154521050-154531120	1.166	UBE2Q1 1:154521050-154531120	1.11	UBE2Q1 1:154521050-154531120
1.329	FANCF 11:22644078-22647387	1.276	SP3 2:174771186-174830430	1.222	SP3 2:174771186-174830430	1.166	SP3 2:174771186-174830430	1.11	SP3 2:174771186-174830430
1.329	GRIN1 9:140033608-140063214	1.276	SP5 2:171571856-171574498	1.222	SP5 2:171571856-171574498	1.166	SP5 2:171571856-171574498	1.11	SP5 2:171571856-171574498
1.329	FOXQ1 6:1312674-1314993	1.276	MYO3B 2:171034654-171511674	1.222	MYO3B 2:171034654-171511674	1.166	MYO3B 2:171034654-171511674	1.11	MYO3B 2:171034654-171511674
1.329	SP3 2:174771186-174830430	1.276	ITGA1 5:52084135-52249485	1.222	ITGA1 5:52084135-52249485	1.166	ITGA1 5:52084135-52249485	1.11	ITGA1 5:52084135-52249485