Performing Linkage Analysis using Sequence Data with SEQLinkage

Recent advances in next-generation sequencing (NGS) make it possible to directly sequence genomes and exomes of individuals with Mendelian diseases and screen sequence data for causal variants. 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, SEQLinkage program (http://www.bioinformatics.org/seqlink/) implements a collapsed haplotype pattern (CHP) method to generate marker alleles from sequence data for linkage analysis (Wang et al. 2015).

The CHP method collapses multiple variants within a genetic region, for example a gene, into a regional marker allele. To generate regional marker allele, haplotypes for the region must be obtained for all samples with sequence data. NGS data from family members are first checked for Mendelian errors. The Lander–Green algorithm for phasing is applied to reconstruct haplotypes in the pedigrees. For each pedigree, first cluster variants on regional haplotypes by 'bins', for example, 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 methods for rare variant association analysis. Then, each collapsed haplotype, composed by bins in a genetic region, will be assigned a single numeric value so that different patterns of collapsed haplotypes in each pedigree are uniquely represented. This process is illustrated in the figure below.

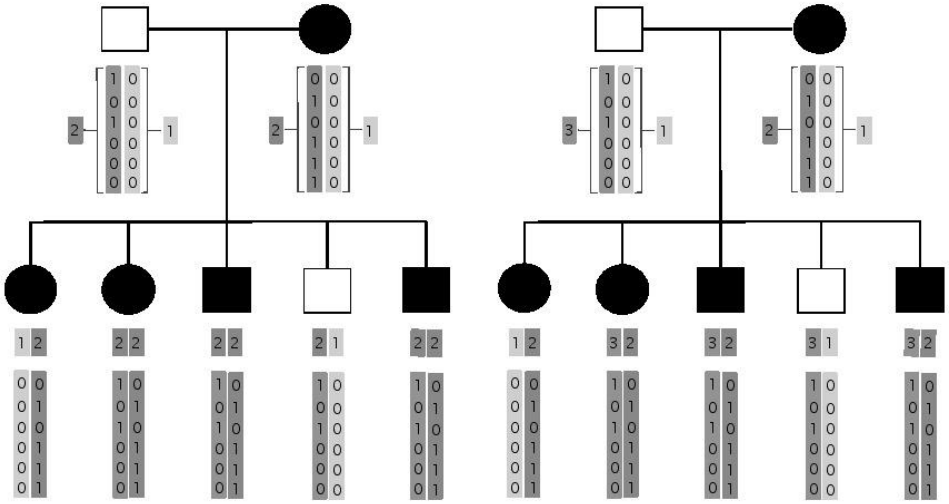

Figure 1. Coding of regional marker alleles using CHP method.

By the CHP method, the regional markers can incorporate allelic heterogeneity between and within families in a region and thus often have higher heterozygosity than SNVs, making them more informative and powerful to detect linkage.

This exercise will show the basic use of SEQLinkage. To display the command interface, run the command below. Through this exercise, the Linux commands will be in bold text and after the ">" prompt sign.

>**seqlink -h**

Please visit http://bioinformatics.org/seqlink for more information.

Here we demonstrate the use of SEQLinkage to generate regional markers from NGS data and perform linkage analysis. For demonstration purpose we will use simulated variants data for two genes *GJB2* and *SLC26A4* for nuclear pedigree. These genes are both involved in the etiology of autosomal recessive nonsyndromic hearing (ARNSHI). The variants in one gene were generated linked to the disease phenotype, ARNSHI, while the variants in the other gene were generated unlinked to the disease phenotype. (See pedigree illustration below). The data contains 10 variants of which 5 variants lie in the *GJB2* gene and 5 in the *SLC26A4* gene. These variants all have low frequencies of $\leq$ 0.01 except for two variants, which we will remove later before doing analysis. We will first learn how to generate regional markers using CHP method with various collapsing themes, then learn how to perform two-point linkage analysis using the generated "superloci".

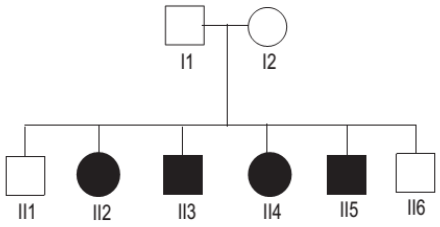

Figure 2. Simulated family DFNB44.

Generating regional marker data

We will first remove the two variants with high frequencies (>0.01) in the *hearing_with_high_freq.vcf* file, and generate a new *vcf* file called *hearing.vcf*. Run the following command:

>**awk '{if ($0~/^#/) {print} else {{maf=substr($8,12)}; if (maf<=0.01) {print}}}' hearing_with_high_freq.vcf >hearing.vcf**

Two variants were removed and now the *hearing.vcf* file only contains 8 variants. Note that the *vcf* file must be indexed by *tabix*. To index a *vcf* file, simply run the following commands. After that, you can also see a *hearing.vcf.gz.ibx* file.

>**bgzip -c hearing.vcf >hearing.vcf.gz**

>**tabix -p vcf hearing.vcf.gz**

Run the following command to generate regional marker data using sequencing data:

> **seqlink --fam hearing.fam --vcf hearing.vcf.gz --format MERLIN**

The "**--fam hearing.fam**" imports the pedigree information stored in *hearing.fam* file. *fam file* follows the commonly used format in linkage analysis, with 6 columns representing family ID, individual ID, father ID, mother ID, sex and phenotypes respectively.

The "**--vcf hearing.vcf.gz**" imports the sequencing data stored in *hearing.vcf.gz* file.

The "**--format MERLIN**" specifies the output format for regional marker data file is MERLIN. Actually in MERLIN, this format is called QTDT format. The QTDT stands for Quantitative Transmission Disequilibrium Tests. This format was used in the QTDT program and later used in MERLIN. We call it MERLIN format due to the popularity of MERLIN.

Task 1:

Notice the output on the screen. The program first detected one family with 8 samples; then it collapsed 8 variants into 2 units, among 25,305 predefined units. The 25,305 units are RefSeq genes. Each RefSeq gene region is considered as one regional marker, and its genetic position was interpolated using Rutgers Map (http://compgen.rutgers.edu/rutgers_maps.shtml). The gene boundaries (could also include other boundaries such as regulatory regions) and genetic positions for each gene/region are stored in a default blueprint file built in SEQLinkage. For exome data the blueprint file contains the gene boundaries and positions for all genes. If you want to analyze only a subset of genes, a blueprint file can be made with a subset of variants sites which will be analyzed. Now please take a look at the *dat* files and *map* files in the *LINKAGE* folder, identify the 2 markers and fill in the table below:

Marker name	Chromosome	Average genetic position

In the example above, default collapsing theme, LD (r^2>0.8) based collapsing, was used to generate regional marker data. Under the default theme, variants having LD with r^2 greater than 0.8 will be collapsed to binary codes, then the haplotype patterns will be computed.

The degree of association of variant sites within LD block can be set by using "**--bin r^2**" argument. For example, setting "**--bin 1**" means collapsing variant by unit of individual variant site, that is, no collapsing is applied to variants before computing haplotype pattern because no LD can have r^2 greater than 1. Run the following command:

> **seqlink --fam hearing.fam --vcf hearing.vcf.gz --format MERLIN --output nocollapsing --bin 1**

We can also set "**--bin 0**" to collapse all the variants in the marker region. In this way, the haplotype in the region will either be collapsed into 0 for all wild type or 1 for any variant, so the haplotype pattern can only be either 1 for all wild type or 2 for any variant. This collapsing theme is useful for generating data for use of linkage analysis software that only accept a limited number of alternative alleles. Run the following command:

> **seqlink --fam hearing.fam --vcf hearing.vcf.gz --format MERLIN --output completecollapsing --bin 0**

Task 2:

Check the files in *nocollapsing* folder and *completecollapsing* folder. Compare the files generated by the three different collapsing themes and fill in the table below. For the listed individuals, what are the alleles for the two markers?

	LD based collapsing		No collapsing		Complete Collapsing	
	SLC26AR	GJB2	SLC26AR	GJB2	SLC26AR	GJB2
I1						
I2						
II1						
II2						
II3						
II4						
II5						
II6						

We can see that LD-based collapsing theme and no collapsing theme generated identical alleles for both superloci. What is the possible explanation? Why did the third collapsing theme, complete collapsing, produced different results?

Performing two-point linkage analysis:

Run the following command to generate regional marker data using default CHP parameters and then perform linkage analysis :

> **seqlink --fam hearing.fam --vcf hearing.vcf.gz --freq EXACSASMAF -o linkageanalysis -K 0.002 --moi AR -W 0 -M 1 --run-linkage**

Variants in *vcf* file were already annotated by the MAF of South Asian in ExAC (Exome Aggregation Consortium), assuming the simulated samples are from South Asian population. The "**--freq EXACSASMAF**" uses such information.

"**-K 0.002**" specifies the nonsyndromic hearing impairment prevalence is 2 per 1000; "**--moi AR**" specifies the mode of inheritance is autosomal recessive; "**-W 0**" defines the penetrance for being both wild type individual and heterozygous carrier is 0; "**-M 1**" specifies the penetrance for carrying homozygous causal variant is 1. These parameters are used to define the parametric model used for linkage analysis.

"--run linkage" will let the program perform linkage analysis

Task 3:

Please go to the folder *linkageanalysis* where the results were output. Click the *linkageanalysis_Report.html* file. Fill in the LOD scores for the two markers under different recombination frequencies in the table below.

	θ=0	θ=0.05	θ=0.1	θ=0.2	θ=0.3	θ=0.4
SLC26A4						
GJB2						

We can conclude that the *GJB2* gene is linked with the disease phenotype. Filtering based variant prioritization method should be used to identify possible causal variant in this gene.

Please note that this exercise is rather artificial that for each gene there where multiple rare variants within the gene. Often there is only a single rare variant within a gene.

Answers:

Task 1:

Marker name	Chromosome	Average genetic position
SLC26A4	7	117.6978
GJB2	13	0.9031

Task 2:

	LD based collapsing				No collapsing				Complete Collapsing			
	SLC26AR		GJB2		SLC26AR		GJB2		SLC26AR		GJB2	
I1	1	3	2	3	1	3	2	3	1	2	2	2
I2	2	4	4	1	2	4	4	1	2	2	2	1
II1	1	4	3	1	1	4	3	1	1	2	2	1
II2	1	2	3	4	1	2	3	4	1	2	2	2
II3	1	4	3	4	1	4	3	4	1	2	2	2
II4	1	4	3	4	1	4	3	4	1	2	2	2
II5	1	4	3	4	1	4	3	4	1	2	2	2
II6	1	2	2	4	1	2	2	4	1	2	2	2

We can see that LD-based collapsing theme and no collapsing theme generated identical alleles for both superloci. What is the possible explanation? Why the third collapsing theme, complete collapsing, produced different results?

Collapsing based on LD and no collapsing gives identical results. This indicates that, in both superloci, there are no variants that are in strong LD with each other. The last theme gives different result because it can only generate 2 alleles: 1 for all wild type or 2 for any variant in the haplotype and there is a lost of information.

Task 3:

	$\theta=0$	$\theta=0.05$	$\theta=0.1$	$\theta=0.2$	$\theta=0.3$	$\theta=0.4$
SLC26A4	2.06	1.84	1.61	1.13	0.63	0.19
GJB2	$-4.34*10^{19}$	-0.43	-0.02	0.19	0.15	0.05

Reference:

Wang GT, Zhang D, Li B, Dai H, Leal SM. 2015. Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. Eur. J. Hum. Genet. EJHG.