

# SEQLinkage Documentation [Version 1.0 alpha]

Gao Wang, Di Zhang and Hang Dai

Last updated: April 15, 2014

# Contents

<b>1</b>	<b>SEQLinkage Reference Manual</b>	<b>2</b>
1.1	Introduction . . . . .	2
1.2	SEQLinkage Program Command Options . . . . .	2
1.2.1	Input files . . . . .	3
1.2.1.1	--vcf [required] . . . . .	3
1.2.1.2	--fam [required] . . . . .	4
1.2.1.3	--blueprint [default to RefSeq genes] . . . . .	4
1.2.2	Additional input options . . . . .	5
1.2.2.1	--freq [default to sample MAF calculated from founders in data] . . . . .	5
1.2.2.2	--maf-cutoff [default to 1.0] . . . . .	5
1.2.2.3	--chrom-prefix [default to empty] . . . . .	5
1.2.3	Collapsed haplotype pattern method coding options . . . . .	5
1.2.3.1	--bin [default to $R^2 > 0.8$ ] . . . . .	5
1.2.3.2	--single-markers [default to disabled] . . . . .	6
1.2.4	Linkage analysis options . . . . .	6
1.2.4.1	--prevalence [required] . . . . .	6
1.2.4.2	--moi [required] . . . . .	6
1.2.4.3	--wt-pen [required] / --mut-pen [required] . . . . .	6
1.2.4.4	--theta-max [default to 0.5] . . . . .	6
1.2.4.5	--theta-inc [default to 0.05] . . . . .	6
1.2.4.6	--run-linkage [default to disabled] . . . . .	6

1.2.4.7	--output-entries [default to 10]	7
1.2.5	Format conversion options	7
1.2.5.1	--format [default to LINKAGE]	7
1.2.5.2	--output [default to LINKAGE]	7
1.2.6	Runtime arguments	7
1.2.6.1	--jobs [default to 2]	7
1.2.6.2	--tempdir [default to system temporary folder]	7
1.2.6.3	--cache [default to disabled]	8
1.2.6.4	--quiet [default to disabled]	8
1.3	Linkage Analysis Results	8
1.3.1	Tables of LOD and HLOD scores	8
1.3.2	Heatmaps of LOD and HLOD scores	9
<b>2</b>	<b>Data Analysis Using SEQLinkage</b>	<b>10</b>
2.1	Introduction	10
2.2	Regional Markers from Sequence Data	11
2.2.1	Understanding terminal output and regional marker data	11
2.2.1.1	Terminal output	11
2.2.1.2	Regional marker data	12
2.2.2	Collapsing themes	12
2.2.2.1	LD based collapsing	12
2.2.2.2	Complete collapsing	13
2.2.2.3	No collapsing	13
2.2.2.4	Comparison between different themes	13
2.2.3	Population vs. founder allele frequencies	14
2.2.4	Recombination events	14
2.3	Two-point Linkage Analysis	14
2.4	SEQLinkage with Other Linkage Programs	14
2.5	Prepare Customized “Blueprint” of Regional Markers	14

# SEQLinkage Reference Manual

## 1.1 Introduction

This program implements a *combined haplotype pattern* (CHP) method to generate markers from sequence data for linkage analysis. The core concept is that instead of treating each variant as a separate marker, we create regional markers for variants in specified genetic regions (e.g. genes) based on haplotype patterns within families, and perform linkage analysis on markers thus generated. CHP method outperforms traditional single marker based approach for compound heterozygosity and allelic heterogeneity in genes. We recommend the use of CHP in conjunction with filtering based variant prioritization method in the analyses of sequence data of human pedigrees.

For details of the method and evaluation of performance using simulated data, please refer to our paper: ...

*Web resource:* please visit <http://bioinformatics.org/seqlink> for more information including download & installation instructions, software updates and supports from SEQLinkage user forum.

## 1.2 SEQLinkage Program Command Options

To display the command interface

```
seqlink -h
```

SEQLinkage interface

```
usage: seqlink [--bin FLOAT] [-b FILE] [--single-markers] --fam FILE --vcf
FILE [--freq INFO] [-c P] [--chrom-prefix STRING] [-o Name]
[-f FORMAT [FORMAT ...]] [-K FLOAT] [--moi STRING] [-W FLOAT]
[-M FLOAT] [--theta-max FLOAT] [--theta-inc FLOAT]
[--run-linkage] [--output-entries N] [-h] [-j N]
[--tempdir PATH] [--cache]
SEQLinkage, linkage analysis using sequence data
```

```

[1.0.alpha]
Collapsed haplotype pattern method arguments:
--bin FLOAT          Defines theme to collapse variants. Set to 0 for
                    "complete collapsing", 1 for "no collapsing", r2 value
                    between 0 and 1 for "LD based collapsing" and other
                    integer values for customized collapsing bin sizes.
                    Default to 0.8 (variants having r2 >= 0.8 will be
                    collapsed).
-b FILE, --blueprint FILE
                    Blueprint file that defines regional marker (format:
                    "chr startpos endpos name avg.distance male.distance
                    female.distance").
--single-markers      Use single variant markers. This switch will overwrite
                    "--bin" and "--blueprint" arguments.

Input / output options:
--fam FILE            Input pedigree and phenotype information in FAM
                    format.
--vcf FILE            Input VCF file, bgzipped.
--freq INFO           Info field name for allele frequency in VCF file.
-c P, --maf-cutoff P  MAF cutoff to define "common" variants to be excluded
                    from analyses.
--chrom-prefix STRING
                    Prefix to chromosome name in VCF file if applicable,
                    e.g. "chr".
-o Name, --output Name
                    Output name prefix.
-f FORMAT [FORMAT ...], --format FORMAT [FORMAT ...]
                    Output format. Default to LINKAGE.

LINKAGE options:
-K FLOAT, --prevalence FLOAT
                    Disease prevalence.
--moi STRING          Mode of inheritance, AD/AR: autosomal
                    dominant/recessive.
-W FLOAT, --wt-pen FLOAT
                    Penetrance for wild type.
-M FLOAT, --mut-pen FLOAT
                    Penetrance for mutation.
--theta-max FLOAT     Theta upper bound. Default to 0.5.
--theta-inc FLOAT     Theta increment. Default to 0.05.
--run-linkage         Perform Linkage analysis using FASTLINK program.
--output-entries N    Write the highest N LOD/HLOD scores to output tables.
                    Default to 10.

Runtime arguments:
-h, --help            Show help message and exit.
-j N, --jobs N        Number of CPUs to use.
--tempdir PATH        Temporary directory to use.
--cache              Load cache data for analysis instead of starting
                    afresh.
-q, --quiet           Disable the display of runtime MESSAGE.

Copyright (c) 2013 - 2014 Gao Wang <gaow@bcm.edu> and Di Zhang <di.zhang@bcm.edu>
Distributed under GNU General Public License
Home page: http://bioinformatics.org/seqlink

```

---

## 1.2.1 Input files

### ■ --vcf [required]

Input genotype data must be [bgzipped](http://samtools.sourceforge.net/tabix.shtml)<sup>1</sup> VCF file indexed by [tabix](http://samtools.sourceforge.net/tabix.shtml)<sup>2</sup>. To create such files from plain VCF file, e.g. data.vcf:

```

bgzip data.vcf
tabix -p vcf -f data.vcf.gz

```

You should end up with two files data.vcf.gz and data.vcf.gz.tbi. In SEQLinkage command

<sup>1</sup>bgzipped <http://samtools.sourceforge.net/tabix.shtml>

<sup>2</sup>tabix <http://samtools.sourceforge.net/tabix.shtml>

you can then use `--vcf data.vcf.gz` to load the genotype data.

#### ■ `--fam` [required]

This file contains information of pedigree structure, sample sex and disease status. It partially follows the [LINKAGE format](#)<sup>3</sup> convention: it has only 6 columns with each column being Family ID, Individual ID, Paternal ID, Maternal ID, Sex and Status.

#### ■ `--blueprint` [default to RefSeq genes]

A “blueprint” file can be supplied to define regional marker units. SEQLinkage has a default built-in blueprint which is suitable for WES studies when it is desired to group variants to create regional markers by genes. Customized blueprint file can be provided by users for specific studies. Even for WES studies one can provide alternative blueprint based on exome sequencing capture targets rather than genes. The file should contain 7 columns:

- Chromosome name, without leading chr character, e.g. “5” not “chr5”
- Start position of the genetic region
- End position of the genetic region
- Region name, e.g. gene names
- Average genetic map distance of the region on average
- Female genetic map distance of the region on average
- Male genetic map distance of the region on average

Genetic map distance will be useful for performing multi-point linkage analysis. Users can output regional markers from SEQLinkage to, for example, Merlin format and perform linkage analysis using Merlin. In the built-in blueprint file we use the map distance of the variant at the median position of a genetic region as a substitute for the map distance of the genetic region. Such information can be interpolated using [Rutgers Linkage-Physical Map](#)<sup>4</sup> database. If multi-point linkage analysis is not the aim of your study you can leave these columns with a placeholder symbol “.” (a dot) for missing data in the blueprint file you provide to SEQLinkage. Example lines of a blueprint file is shown below:

blueprint.txt						
...						
3	126111874	126113641	CCDC37-AS1	134.382	168.977	102.287
3	126113781	126155398	CCDC37	134.411	169.021	102.296
3	126156443	126194762	ZXDC	134.465	169.105	102.315
...						

<sup>3</sup>LINKAGE format <http://www.jurgott.org/linkage/LinkagePC.html>

<sup>4</sup>Rutgers Linkage-Physical Map <http://compgen.rutgers.edu/maps>

## 1.2.2 Additional input options

### ■ `--freq` [default to sample MAF calculated from founders in data]

Linkage analysis requires input of allele frequency for markers to control for type I error in the presence of missing genotypes. The INFO field name for population (minor) allele frequencies of variants in VCF file. For well defined populations we recommend using MAF for variants from publicly available data bases such as [Exome Variant Server](#) <sup>5</sup> or [1000 Genomes](#) <sup>6</sup>. For variants not presented in these data bases it is safe to assign a very small proportion, e.g. 0.00015 which is roughly the MAF for a singleton variant in 3000 samples ( $\frac{1}{3000 \times 2} = 0.000167$ ). You may use other bioinformatics tools such as [variant tools](#) <sup>7</sup> to obtain and update such information to your VCF file. If this option is left unset, MAF estimated from founders in the sample will be used for linkage analysis.

### ■ `--maf-cutoff` [default to 1.0]

When specified, variants having MAF (defined by `--freq` option) greater than this value will be excluded from analyses.

### ■ `--chrom-prefix` [default to empty]

This option specifies the prefix to chromosome names in VCF file. For example for VCF files having chromosome names such as “1”, “5” and “X” there is no need to specify this option. For files having names such as “chr1”, “chr5” and “chrX” you need to use `--chrom-prefix chr` in SEQLinkage command.

## 1.2.3 Collapsed haplotype pattern method coding options

The CHP method has been described in the SEQLinkage paper (see “Introduction” section of this chapter). This section introduces the usage of parameters involved in implementing the CHP method.

### ■ `--bin` [default to $R^2 > 0.8$ ]

This option defines the collapsing theme of variants in a genetic region, before computing haplotype patterns. Several collapsing themes are available via this option:

- “Linkage disequilibrium (LD) based collapsing”. The bin value takes a fraction number (between 0 and 1) as the  $R^2$  cutoff to define LD blocks. Variant sites having LD greater than  $R^2$  will be collapsed to binary codes.

---

<sup>5</sup>Exome Variant Server <http://evs.gs.washington.edu/EVS/>

<sup>6</sup>1000 Genomes <http://www.1000genomes.org/>

<sup>7</sup>variant tools <http://varianttools.sourceforge.net>

- “No collapsing”. Set `--bin 1` which literally means collapsing variants by units of 1 variant site, i.e., no collapsing is applied to variants before computing haplotype patterns.
- “Complete collapsing”. Set `--bin 0` to collapse variant in the entire region to a single binary code.
- “Arbitrary collapsing”. Set `--bin` to any arbitrary positive integer value  $N$  to collapse  $N$  variants to a single binary code.

#### ■ `--single-markers` [default to disabled]

When this switch is turned on, single variant markers will be generated from data instead of regional markers, and both `--bin` and `--blueprint` options will be ignored.

### 1.2.4 Linkage analysis options

SEQLinkage has a built-in two-point linkage analysis routine to analyze data generated via the CHP method. Below are options for configuring linkage model parameters and producing graphic / HTML format analysis reports.

#### ■ `--prevalence` [required]

Disease prevalence.

#### ■ `--moi` [required]

Mode of inheritance, choose from “AD” (autosomal dominant) and “AR” (autosomal recessive).

#### ■ `--wt-pen` [required] / `--mut-pen` [required]

Penetrance of wild type / mutation.

#### ■ `--theta-max` [default to 0.5]

Recombination rate value upper bound ( $\theta_{max}$ ) up to which the linkage analysis will evaluate.

#### ■ `--theta-inc` [default to 0.05]

Increment steps from 0 to  $\theta_{max}$ . At each step the  $\theta$  value will be used to calculate a LOD score.

#### ■ `--run-linkage` [default to disabled]

When this switch is on, two-point linkage analysis will be performed.



- `--output-entries` [default to 10]

Output to HTML file the best  $N$  markers in terms of LOD and HLOD scores respectively. When  $N = 0$ , no heatmap graph or HTML file will be generated.

### 1.2.5 Format conversion options

SEQLinkage supports output in some population linkage software format including LINKAGE, Merlin and MEGA2. Many more linkage software format can be converted from MEGA2 format using the MEGA2 software. With the format conversion feature, CHP coding of sequence data can be written to these file formats for use in various linkage analysis software.

- `--format` [default to LINKAGE]

Output format for CHP coded data.

- `--output` [default to LINKAGE]

Output file / folder name prefix.

### 1.2.6 Runtime arguments

- `--jobs` [default to 2]

Number of CPUs to use for SEQLinkage. SEQLinkage supports analyzing many markers in parallel and the more CPUs it is assigned the shorter the computational time will be.

- `--tempdir` [default to system temporary folder]

The linkage analysis routine in SEQLinkage performs analysis per marker per family, thus involving frequent file I/O operations which can be a computational bottleneck. By default such I/O operations take place in one of the system temporary folders, e.g. `/tmp`, `/var/tmp` in Linux system. To speed things up one can set the SEQLinkage temporary directory to some high speed hard drives, e.g. a solid state drive (SSD), or, if possible, a “RAM drive”. Below is an example to create a 5GB RAM drive in Linux:

```
sudo mkdir /tmp/ramdisk; sudo chmod 777 /tmp/ramdisk
sudo mount -t tmpfs -o size=5120M ramfs /tmp/ramdisk
```

With `--tempdir /tmp/ramdisk` option the newly created RAM drive will be used for the intensive file I/O in the analysis.

- **--cache [default to disabled]**

To speed up repeated runs of SEQLinkage on the same data set under similar parameter settings, data are archived to the cache folder under the work directory the first time SEQLinkage executes. With this switch on, SEQLinkage will use the archived data whenever appropriate to skip as many steps previously performed. For example in a repeated analysis under the same setting but only change `--output-entries` from 10 to 50, SEQLinkage will skip the CHP coding and linkage analysis step, only updating the result HTML table using archived analysis results.

Note that change of some input parameters will overwrite the effect of `--cache`. For example changing `--moi` will result in re-run of linkage analysis; changing `--vcf` or `--fam` input will result in re-run of CHP coding step.

- `--quiet` [default to disabled]

When this switch is on, the program will not display any log message during runtime. It will, however, display error message if an error occurs.

### 1.3 Linkage Analysis Results

SEQLinkage summarizes two-point linkage analysis results to heatmap plots and tables in HTML format, which can be viewed with a web browser program. Note that on the HTML document each section can be temporarily folded such that you can focus only on the section of interest (see the hand gesture on the heatmap screenshot below).

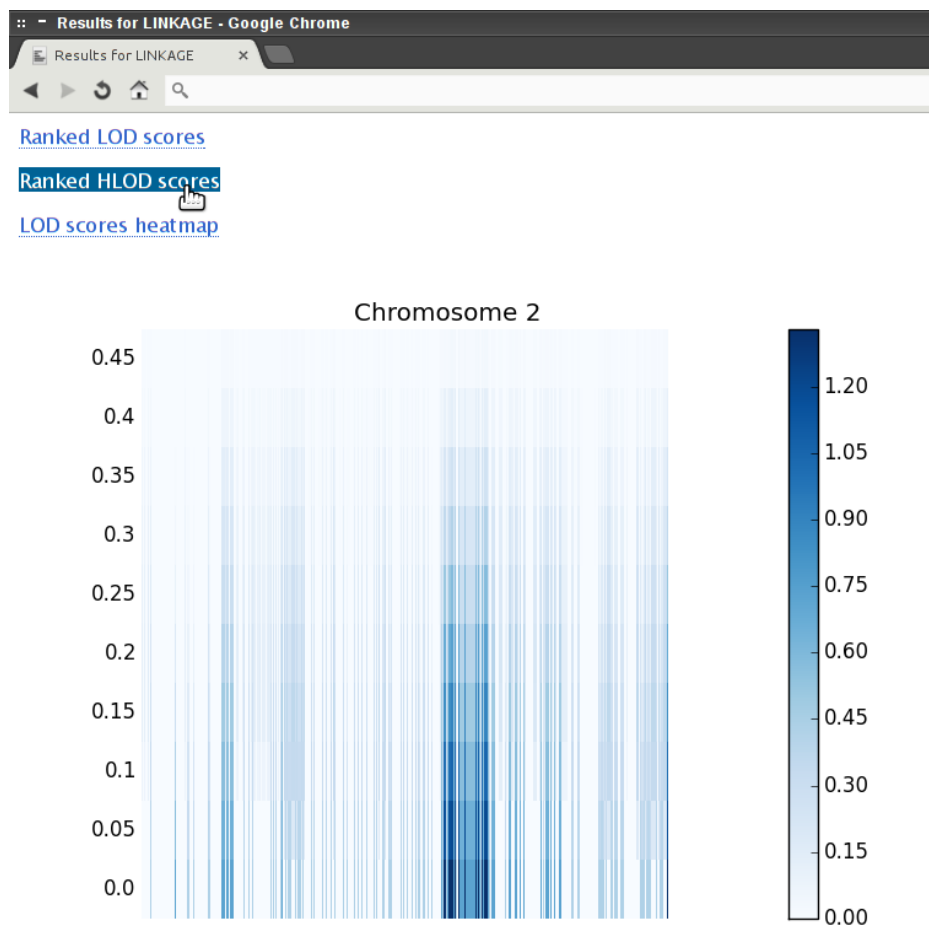
### 1.3.1 Tables of LOD and HLOD scores

Ranked LOD scores													Ranked HLOD scores														
0=0		0=0.05		0=0.1		0=0.15		0=0.2		0=0.25		0=0.3		0=0.0		0=0.05		0=0.1		0=0.15		0=0.2		0=0.25		0=0.3	
Lod	Marker name chrstart-end	Lod	Marker name chrstart-end	Lod	Marker name chrstart-end	Lod	Marker name chrstart-end	Lod	Marker name chrstart-end	Lod	Marker name chrstart-end	Lod	Marker name chrstart-	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-end	Hlod	Marker name chrstart-
1.329	UBEQ1 1154521050- 154931120	1.194	1154521050- 154931120	1.051	1154521050- 154931120	0.9	1154521050- 154931120	0.742	1154521050- 154931120	0.579	1154521050- 154931120	0.416	1154521050- 154931120	1.329	1154521050- 154931120	1.194	1154521050- 154931120	1.051	1154521050- 154931120	0.9	1154521050- 154931120	0.742	1154521050- 154931120	0.579	1154521050- 154931120	0.416	1154521050- 154931120
	UBEQ1		UBEQ1		UBEQ1		UBEQ1		UBEQ1		UBEQ1		UBEQ1		FANCF		SP3		SP3		SP3		SP3		SP3		SP3
1.329	171571856- 171574498	1.194	171571856- 171574498	1.051	171571856- 171574498	0.9	171571856- 171574498	0.742	171571856- 171574498	0.579	171571856- 171574498	0.416	171571856- 171574498	1.329	112264078- 174003430	1.194	171571856- 171574498	1.051	171571856- 171574498	0.9	171571856- 171574498	0.742	171571856- 171574498	0.579	171571856- 171574498	0.416	171571856- 171574498
	ITGA1		ITGA1		ITGA1		ITGA1		ITGA1		ITGA1		ITGA1		GRN1		SP3		SP3		SP3		SP3		SP3		SP3
1.329	171571856- 171574498	1.194	152084135- 52249485	1.051	152084135- 52249485	0.9	152084135- 52249485	0.742	152084135- 52249485	0.579	152084135- 52249485	0.416	152084135- 52249485	1.329	194003368- 140063214	1.194	171571856- 171574498	1.051	171571856- 171574498	0.9	171571856- 171574498	0.742	171571856- 171574498	0.579	171571856- 171574498	0.416	171571856- 171574498
	FBXL18		FBXL18		FBXL18		FBXL18		FBXL18		FBXL18		FBXL18		FOXQ1		MYO3B		MYO3B		MYO3B		MYO3B		MYO3B		MYO3B
1.329	152084135- 52249485	1.194	73515427- 5553399	1.051	73515427- 5553399	0.9	73515427- 5553399	0.742	73515427- 5553399	0.579	73515427- 5553399	0.416	73515427- 5553399	1.329	61312674- 1314993	1.194	2173034654- 171511674	1.051	2173034654- 171511674	0.9	2173034654- 171511674	0.742	2173034654- 171511674	0.579	2173034654- 171511674	0.416	2173034654- 171511674
	SWAP70		SDK1		SDK1		SDK1		SDK1		SDK1		SDK1		ITGA1		ITGA1		ITGA1		ITGA1		ITGA1		ITGA1		ITGA1
1.329	116865627- 4274507	1.194	73341079- 4308631	1.051	73341079- 4308631	0.9	73341079- 4308631	0.742	73341079- 4308631	0.579	73341079- 4308631	0.416	73341079- 4308631	1.329	1714771186- 52424461	1.194	152084135- 52249485	1.051	152084135- 52249485	0.9	152084135- 52249485	0.742	152084135- 52249485	0.579	152084135- 52249485	0.416	152084135- 52249485
	FBXL18		DPT		DPT		DPT		DPT		DPT		DPT		SP3		FBXL18		FBXL18		FBXL18		FBXL18		FBXL18		FBXL18
1.329	73515427- 5553399	1.194	1168664694- 168698442	1.051	1168664694- 168698442	0.9	1168664694- 168698442	0.742	1168664694- 168698442	0.579	1168664694- 168698442	0.416	1168664694- 168698442	1.329	171571856- 171574498	1.194	73515427- 5553399	1.051	73515427- 5553399	0.9	73515427- 5553399	0.742	73515427- 5553399	0.579	73515427- 5553399	0.416	73515427- 5553399
	SDK1		LTBP2		LTBP2		LTBP2		LTBP2		LTBP2		LTBP2		MYO3B		SDK1		SDK1		SDK1		SDK1		SDK1		SDK1
1.329	73341079- 4308631	1.194	174964885- 75079034	1.051	174964885- 75079034	0.9	174964885- 75079034	0.742	174964885- 75079034	0.579	174964885- 75079034	0.416	174964885- 75079034	1.329	2173034654- 171511674	1.194	73341079- 4308631	1.051	73341079- 4308631	0.9	73341079- 4308631	0.742	73341079- 4308631	0.579	73341079- 4308631	0.416	73341079- 4308631
	TPRN		RFXS		RFXS		RFXS		RFXS		RFXS		RFXS		ITGA1		ERIC2		ERIC2		ERIC2		ERIC2		ERIC2		ERIC2
1.329	1194006068- 140063	1.194	1151313115- 151319769	1.051	1151313115- 151319769	0.9	1151313115- 151319769	0.742	1151313115- 151319769	0.579	1151313115- 151319769	0.416	1151313115- 151319769	1.329	152084135- 52249485	1.194	1715627191- 171655481	1.051	1715627191- 171655481	0.9	1715627191- 171655481	0.742	1715627191- 171655481	0.579	1715627191- 171655481	0.416	1715627191- 171655481
	FAM134B		NRXN3		NRXN3		NRXN3		NRXN3		NRXN3		NRXN3		SWAP70		DPT		DPT		DPT		DPT		DPT		DPT
1.329	147637146- 16617167	1.194	178636715- 80334633	1.051	178636715- 80334633	0.9	178636715- 80334633	0.742	178636715- 80334633	0.579	178636715- 80334633	0.416	178636715- 80334633	1.329	119685627- 9774507	1.194	1168664694- 168698442	1.051	1168664694- 168698442	0.9	1168664694- 168698442	0.742	1168664694- 168698442	0.579	1168664694- 168698442	0.416	1168664694- 168698442
	DPT		PAPLN		PAPLN		PAPLN		PAPLN		PAPLN		PAPLN		FBXL18		RFXS		RFXS		RFXS		RFXS		RFXS		RFXS
1.329	1168664694- 168698442	1.194	173704204- 73741347	1.051	173704204- 73741347	0.9	173704204- 73741347	0.742	173704204- 73741347	0.579	173704204- 73741347	0.416	173704204- 73741347	1.329	73515427- 5553399	1.194	151313115- 151319769	1.051	151313115- 151319769	0.9	151313115- 151319769	0.742	151313115- 151319769	0.579	151313115- 151319769	0.416	151313115- 151319769

The top ranked  $N$  LOD and HLOD scores for each  $\theta$  value evaluated are summarized in HTML table as displayed on the screenshot above (left: LOD, right: HLOD). Results are annotated

with the names and the genomic coordinates of the regional markers. For HLOD scores the corresponding  $\alpha$  values are also displayed. The length of the table  $N$  is controlled by the `--output-entries` option.

### 1.3.2 Heatmaps of LOD and HLOD scores

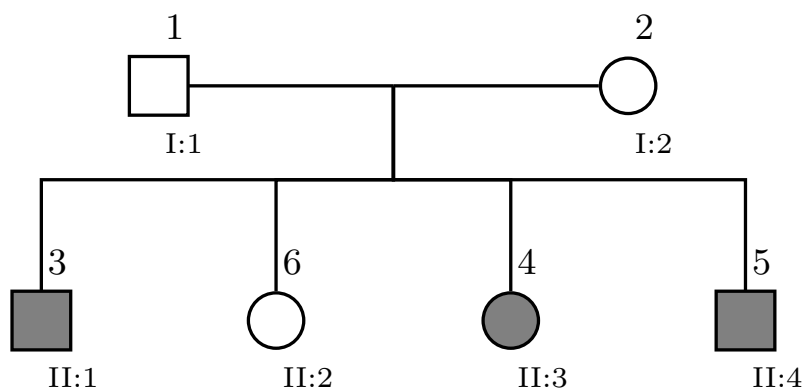


LOD and HLOD scores for all markers analyzed are displayed per chromosome as heatmaps, using a sequence of blue colors from light to dark for the score values; the darker the color the higher the score. Linkage regions of potential interest across the entire genome can be easily identified on the heatmaps. Notice that scales for per-chromosome heatmaps might differ from each other, as labeled on the right side of each heatmap.

# Data Analysis Using SEQLinkage

## 2.1 Introduction

Here we demonstrate the use of SEQLinkage to generate regional markers from sequence data and perform linkage analysis. For demonstration purpose we will use a [simulated example data set](#)<sup>1</sup> of two nuclear families of the same structure and phenotype pattern (see pedigree illustration below) containing sequence data of 18 variants. From the phenotypic pattern it is reasonable to assume the disease follows a recessive mode of inheritance. We will first generate regional markers using CHP method with various collapsing themes, then perform two-point linkage analysis using regional markers generated. We further demonstrate the usage of SEQLinkage in conjunction with other linkage programs by formatting the regional markers into MEGA2 and Merlin input, and perform additional analysis using those software. Finally we demonstrate how to create blueprint file for customized regional marker definition.



<sup>1</sup>simulated example data set <http://bioinformatics.org/seqlink/download/seqlinkage-example.zip>

## 2.2 Regional Markers from Sequence Data

### 2.2.1 Understanding terminal output and regional marker data

Here we perform a test run of SEQLinkage to generate regional marker data without performing linkage analysis. For now we (mostly) stick to default settings and focus on interpretation of terminal output generated by the program.

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz -f MERLIN
```

#### ■ Terminal output

Command above generates the following output:

---

terminal info

---

```
MESSAGE: Binary trait detected in [/ramcache/seqlinkage-example.fam]
MESSAGE: Checking local resources 5/5 ...
MESSAGE: 12 samples found in [/ramcache/seqlinkage-example.vcf.gz]
MESSAGE: 2 families with a total of 12 samples will be scanned for 25,305 pre-defined units
MESSAGE: 2 units (from 18 variants) processed; 3 Mendelian inconsistencies and 2 recombination events handled
MESSAGE: 25,302 units ignored due to absence in VCF file
MESSAGE: 1 units ignored due to absence of variation in samples
MESSAGE: Archiving regional marker data to directory [/ramcache/cache]
MESSAGE: 2 units will be converted to MERLIN format
MESSAGE: 2 units successfully converted to MERLIN format
MESSAGE: Archiving MERLIN format to directory [/ramcache/cache]
MESSAGE: Saving data to [/ramcache/LINKAGE]
```

---

**Line 2** of MESSAGE checks for some resource programs & files required for the execution of SEQLinkage. These files are stored in a hidden folder `~/.SEQLinkage` on your computer. SEQLinkage will automatically download these files the first time it is executed on your computer so please make sure your computer is **connected to Internet when running SEQLinkage for the first time!** Out of the 5 resource files only one of them is relevant to the generation of regional markers: the *blueprint* file that defines genetic regions to be considered as one *marker*. This *blueprint* is based on UCSC RefSeq database. We use genomic coordinates of RefSeq genes to determine start and end positions for regional markers. The genomic coordinates are based on UCSC hg19 (or NCBI build 37) reference genome. To convert to previous builds for your data we recommend running the [UCSC liftOver tool](http://genome.ucsc.edu/cgi-bin/hgLiftOver)<sup>2</sup> to get updated *blueprint* and use the `--blueprint` option to load the file.

**Line 3** checks samples from VCF file against FAM file. For our test data samples in VCF file matches those in FAM file. SEQLinkage allows for samples in VCF file but not in FAM file, or otherwise. For such cases only samples in both files will be analyzed and a warning message will be given if samples are found in FAM but not VCF file.

---

<sup>2</sup>UCSC liftOver tool <http://genome.ucsc.edu/cgi-bin/hgLiftOver>

**Line 4** summarizes data information, mostly from FAM file, VCF header and the blueprint file. In the example one family with six members are found in both VCF and FAM input; also there are 25,305 pre-defined genetic regions in the blueprint file.

**Line 5** is dynamic: it was a progress meter during runtime, and becomes a summary of runtime statistics after the CHP algorithm is complete for all regional marker units. “2 units (from 18 variants) were processed” is based on those variants in both VCF file and covered by the blueprint definition. **You should comparing the number of variants processed with the total number of variants in the VCF file to evaluate how much data was covered by the pre-defined regional marker positions in blueprint file**, and decide whether or not a customized blueprint should be provided. SEQLinkage performs Mendelian error check on the fly, ignoring genotype calls due to Mendelian inconsistency when there is not enough information to infer them correctly. It also deals with recombination events during haplotype construction and CHP coding process, and for those regions with recombination events the regional markers are divided into sub-units. This will be discussed in details later.

**Lines 6-7** provide additional information on variants and units ignored in the analysis. Note that values on lines 6 and 7 plus the number of units processed on line 5 equals the total number of pre-defined units in the blueprint (line 4).

The last line of MESSAGE displays the path of output data, which in our case is in Merlin format. We will examine next the content of the output.

## ■ Regional marker data

[FIXME: write this part after the Merlin format is ready ... marker name is gene name. allele numbering based on per family per unit]

### 2.2.2 Collapsing themes

## ■ LD based collapsing

The default collapsing theme is LD based with the  $R^2 > 0.8$  rule; variants within LD blocks thus defined will be collapsed to binary codes before haplotype patterns are computed. You may set `--bin` to other values  $R^2 \in (0, 1)$  for different LD block definition. The output result will be written to MERLIN format (via the `--format MERLIN` argument) for use later.

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --format MERLIN --output RMBPt8 --jobs 8
```

terminal info

```
MESSAGE: Binary trait detected in [/ramcache/seqlinkage-example.fam]
MESSAGE: Checking local resources 5/5 ...
MESSAGE: 12 samples found in [/ramcache/seqlinkage-example.vcf.gz]
MESSAGE: 2 families with a total of 12 samples will be scanned for 25,305 pre-defined units
MESSAGE: 2 units (from 18 variants) processed; 3 Mendelian inconsistencies and 2 recombination events handled
MESSAGE: 25,302 units ignored due to absence in VCF file
MESSAGE: 1 units ignored due to absence of variation in samples
```

```
MESSAGE: Archiving regional marker data to directory [/ramcache/cache]
MESSAGE: 2 units will be converted to MERLIN format
MESSAGE: 2 units successfully converted to MERLIN format
MESSAGE: Archiving MERLIN format to directory [/ramcache/cache]
MESSAGE: Saving data to [/ramcache/RMBPt8]
```

---

## ■ Complete collapsing

Setting `--bin 0` will collapse all variants in the region to generate one marker per region. Haplotype patterns are thus simply either “1” for all wild type or “2” for any mutation in the region.

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --format MERLIN --output RMB0 --jobs 8 --bin 0
```

---

terminal info

---

```
MESSAGE: Binary trait detected in [/ramcache/seqlinkage-example.fam]
MESSAGE: Checking local resources 5/5 ...
MESSAGE: 12 samples found in [/ramcache/seqlinkage-example.vcf.gz]
MESSAGE: 2 families with a total of 12 samples will be scanned for 25,305 pre-defined units
MESSAGE: 2 units (from 18 variants) processed; 3 Mendelian inconsistencies and 2 recombination events handled
MESSAGE: 25,302 units ignored due to absence in VCF file
MESSAGE: 1 units ignored due to absence of variation in samples
MESSAGE: Archiving regional marker data to directory [/ramcache/cache]
MESSAGE: 2 units will be converted to MERLIN format
MESSAGE: 2 units successfully converted to MERLIN format
MESSAGE: Archiving MERLIN format to directory [/ramcache/cache]
MESSAGE: Saving data to [/ramcache/RMB0]
```

---

## ■ No collapsing

Setting `--bin 1` will compute the haplotype pattern for the region as is, without collapsing.

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --format MERLIN --output RMB1 --jobs 8 --bin 1
```

---

terminal info

---

```
MESSAGE: Binary trait detected in [/ramcache/seqlinkage-example.fam]
MESSAGE: Checking local resources 5/5 ...
MESSAGE: 12 samples found in [/ramcache/seqlinkage-example.vcf.gz]
MESSAGE: 2 families with a total of 12 samples will be scanned for 25,305 pre-defined units
MESSAGE: 2 units (from 18 variants) processed; 3 Mendelian inconsistencies and 2 recombination events handled
MESSAGE: 25,302 units ignored due to absence in VCF file
MESSAGE: 1 units ignored due to absence of variation in samples
MESSAGE: Archiving regional marker data to directory [/ramcache/cache]
MESSAGE: 2 units will be converted to MERLIN format
MESSAGE: 2 units successfully converted to MERLIN format
MESSAGE: Archiving MERLIN format to directory [/ramcache/cache]
MESSAGE: Saving data to [/ramcache/RMB1]
```

---

## ■ Comparison between different themes

[FIXME: visualize the output after MERLIN format is done]

### 2.2.3 Population vs. founder allele frequencies

In the examples above we left `--freq` option unspecified, thus frequency from samples are used. The example VCF file contains an INFO field EVSEAAF, the allele frequency for European Americans in Exome Variant Server (EVS). We assume the simulated variant data is from European American samples, and we incorporated the information to the VCF file as long as the variant is found in EVS; for variants not found in EVS we use EVSEAAF=0.00015. Command below uses such frequency information from VCF file:

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --freq EVSEAAF --format MERLIN --jobs 8
```

Result comparison:

---

OUTPUT

---

FIXME: write this after Merlin format is done

---

### 2.2.4 Recombination events

For regional markers sub-divided into smaller units by recombination events, we use [1], [2], ..., [i] convention to mark different units (see below). They can be treated different regional markers. In the linkage analysis pipeline incorporated in SEQLinkage, we choose the one sub-unit that gives strongest evidence of linkage to represent the entire region under consideration.

---

OUTPUT

---

FIXME: write this after Merlin format is done

---

## 2.3 Two-point Linkage Analysis

[Examples, particularly choice of parameters ... ]

## 2.4 SEQLinkage with Other Linkage Programs

[give one example output to MEGA2 with `-bin 0`; and another example output to Merlin and run some analysis in Merlin]

## 2.5 Prepare Customized “Blueprint” of Regional Markers

Will Hang Dai write up this part? (upload the python script to bioinformatics.org and explain the workflow using that script)