SEQLinkage Documentation [Version 1.0 alpha]

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

Last updated: May 8, 2014

Contents

Chapter 1

SEQLinkage Reference Manual

1.1 Introduction

This program implements a *collapsed 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:

- Gao T. Wang, Di Zhang, Biao Li, Hang Dai and Suzanne M. Leal, *SEQLinkage: A Novel Linkage Analysis Method for Next-Generation Sequencing Data*. [under review]

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 [1] VCF file indexed by tabix [2]. To create such files from plain VCF file, e.g. `data.vcf`:

[1] bgzipped http://samtools.sourceforge.net/tabix.shtml

[2] tabix http://samtools.sourceforge.net/tabix.shtml

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

`--fam` [required]

This file contain information of pedigree structure, sample sex and disease status. It partially follows the LINKAGE format [3] 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 [4] database. If multi-point linkage analysis is not the aim of your study you can leave these columns with a place holder symbol "." (a dot) for missing data in the blueprint file you provide to SEQLinkage. Example lines of a blue print file is shown below:

[3] LINKAGE format http://www.jurgott.org/linkage/LinkagePC.html

[4] Rutgers Linkage-Physical Map http://compgen.rutgers.edu/maps

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
...
```

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 [5] or 1000 Genomes [6]. 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 [7] 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.

[5] Exome Variant Server http://evs.gs.washington.edu/EVS/

[6] 1000 Genomes http://www.1000genomes.org/

[7] variant tools http://varianttools.sourceforge.net

- `--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.
- "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 (θ_{max}) up to which the linkage analysis will evaluate.

- `--theta-inc` **[default to 0.05]**

Increment steps from 0 to θ_{max}. At each step the θ 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 foldes, 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 used 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.

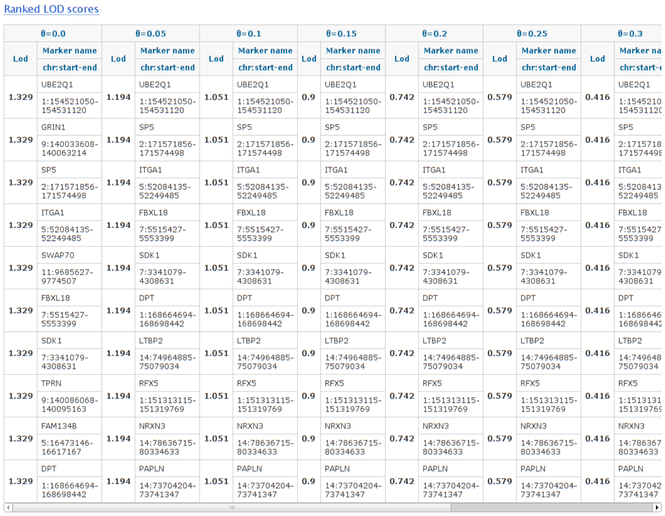
- --quiet **[default to disabled]**

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

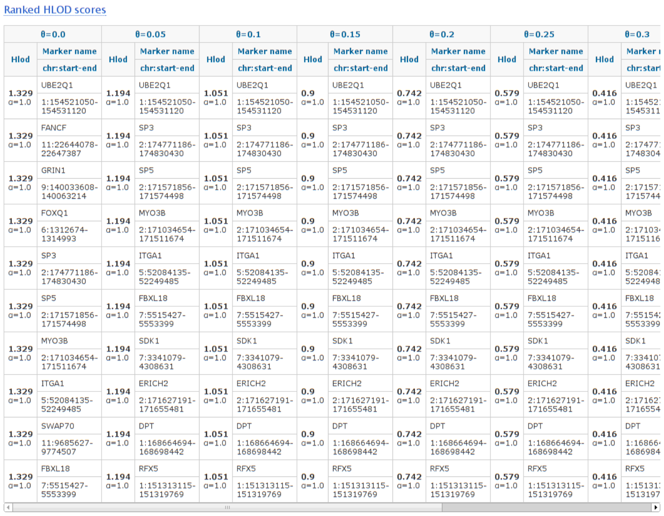
SEQLinkage summarizes two-point linkage analysis results to heatmap plots and tables in HTML format, which can be viewed with a web browser program. Note that on the HTML document each section can be temporarily folded such that you can focus only on the section of interest (see the hand gesture on the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

1.3.1 Tables of LOD and HLOD scores



Ranked LOD scores

θ=0.0		θ=0.05		θ=0.1		θ=0.15		θ=0.2		θ=0.25		θ=0.3	
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	Lod	Marker name chr:start-end	Lod	Marker na chr:start-e
1.329	UBE2Q1 1:154521050-154531120	1.194	UBE2Q1 1:154521050-154531120	1.051	UBE2Q1 1:154521050-154531120	0.9	UBE2Q1 1:154521050-154531120	0.742	UBE2Q1 1:154521050-154531120	0.579	UBE2Q1 1:154521050-154531120	0.416	UBE2Q1 1:1545210 15453112
1.329	GRIN1 9:140033608-140063214	1.194	SP5 2:171571856-171574498	1.051	SP5 2:171571856-171574498	0.9	SP5 2:171571856-171574498	0.742	SP5 2:171571856-171574498	0.579	SP5 2:171571856-171574498	0.416	SP5 2:1715718 17157449
1.329	SP5 2:171571856-171574498	1.194	ITGA1 5:52084135-52249485	1.051	ITGA1 5:52084135-52249485	0.9	ITGA1 5:52084135-52249485	0.742	ITGA1 5:52084135-52249485	0.579	ITGA1 5:52084135-52249485	0.416	ITGA1 5:5208413 52249485
1.329	ITGA1 5:52084135-52249485	1.194	FBXL18 7:5515427-5553399	1.051	FBXL18 7:5515427-5553399	0.9	FBXL18 7:5515427-5553399	0.742	FBXL18 7:5515427-5553399	0.579	FBXL18 7:5515427-5553399	0.416	FBXL18 7:5515427 5553399
1.329	SWAP70 11:9685627-9774507	1.194	SDK1 7:3341079-4308631	1.051	SDK1 7:3341079-4308631	0.9	SDK1 7:3341079-4308631	0.742	SDK1 7:3341079-4308631	0.579	SDK1 7:3341079-4308631	0.416	SDK1 7:3341079 4308631
1.329	FBXL18 7:5515427-5553399	1.194	DPT 1:168664694-168698442	1.051	DPT 1:168664694-168698442	0.9	DPT 1:168664694-168698442	0.742	DPT 1:168664694-168698442	0.579	DPT 1:168664694-168698442	0.416	DPT 1:1686646 16869844
1.329	SDK1 7:3341079-4308631	1.194	LTBP2 14:74964885-75079034	1.051	LTBP2 14:74964885-75079034	0.9	LTBP2 14:74964885-75079034	0.742	LTBP2 14:74964885-75079034	0.579	LTBP2 14:74964885-75079034	0.416	LTBP2 14:749648 75079034
1.329	TPRN 9:140086068-140095163	1.194	RFX5 1:151313115-151319769	1.051	RFX5 1:151313115-151319769	0.9	RFX5 1:151313115-151319769	0.742	RFX5 1:151313115-151319769	0.579	RFX5 1:151313115-151319769	0.416	RFX5 1:1513131 15131976
1.329	FAM134B 5:16473146-16617167	1.194	NRXN3 14:78636715-80334633	1.051	NRXN3 14:78636715-80334633	0.9	NRXN3 14:78636715-80334633	0.742	NRXN3 14:78636715-80334633	0.579	NRXN3 14:78636715-80334633	0.416	NRXN3 14:786367 80334633
1.329	DPT 1:168664694-168698442	1.194	PAPLN 14:73704204-73741347	1.051	PAPLN 14:73704204-73741347	0.9	PAPLN 14:73704204-73741347	0.742	PAPLN 14:73704204-73741347	0.579	PAPLN 14:73704204-73741347	0.416	PAPLN 14:737042 73741347

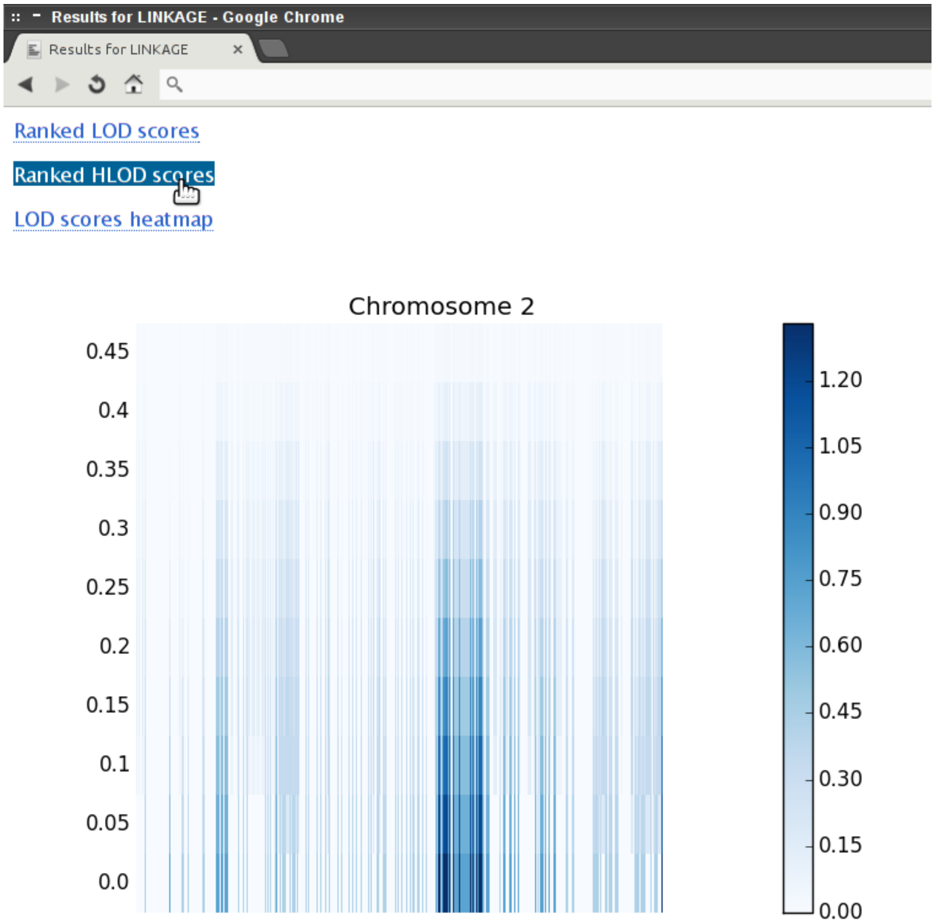


Ranked HLOD scores

θ=0.0		θ=0.05		θ=0.1		θ=0.15		θ=0.2		θ=0.25		θ=0.3	
Hlod	Marker name chr:start-end	Hlod	Marker name chr:start-end	Hlod	Marker name chr:start-end	Hlod	Marker name chr:start-end	Hlod	Marker name chr:start-end	Hlod	Marker name chr:start-end	Hlod	Marker r chr:start
1.329 α=1.0	UBE2Q1 1:154521050-154531120	1.194 α=1.0	UBE2Q1 1:154521050-154531120	1.051 α=1.0	UBE2Q1 1:154521050-154531120	0.9 α=1.0	UBE2Q1 1:154521050-154531120	0.742 α=1.0	UBE2Q1 1:154521050-154531120	0.579 α=1.0	UBE2Q1 1:154521050-154531120	0.416 α=1.0	UBE2Q1 1:15452 1545311
1.329 α=1.0	FANCF 11:22644078-22647387	1.194 α=1.0	SP3 2:174771186-174830430	1.051 α=1.0	SP3 2:174771186-174830430	0.9 α=1.0	SP3 2:174771186-174830430	0.742 α=1.0	SP3 2:174771186-174830430	0.579 α=1.0	SP3 2:174771186-174830430	0.416 α=1.0	SP3 2:17477 1748304
1.329 α=1.0	GRIN1 9:140033608-140063214	1.194 α=1.0	SP5 2:171571856-171574498	1.051 α=1.0	SP5 2:171571856-171574498	0.9 α=1.0	SP5 2:171571856-171574498	0.742 α=1.0	SP5 2:171571856-171574498	0.579 α=1.0	SP5 2:171571856-171574498	0.416 α=1.0	SP5 2:17157 1715744
1.329 α=1.0	FOXQ1 6:1312674-1314993	1.194 α=1.0	MYO3B 2:171034654-171511674	1.051 α=1.0	MYO3B 2:171034654-171511674	0.9 α=1.0	MYO3B 2:171034654-171511674	0.742 α=1.0	MYO3B 2:171034654-171511674	0.579 α=1.0	MYO3B 2:171034654-171511674	0.416 α=1.0	MYO3B 2:17103 1715116
1.329 α=1.0	SP3 2:174771186-174830430	1.194 α=1.0	ITGA1 5:52084135-52249485	1.051 α=1.0	ITGA1 5:52084135-52249485	0.9 α=1.0	ITGA1 5:52084135-52249485	0.742 α=1.0	ITGA1 5:52084135-52249485	0.579 α=1.0	ITGA1 5:52084135-52249485	0.416 α=1.0	ITGA1 5:52084 5224948
1.329 α=1.0	SP5 2:171571856-171574498	1.194 α=1.0	FBXL18 7:5515427-5553399	1.051 α=1.0	FBXL18 7:5515427-5553399	0.9 α=1.0	FBXL18 7:5515427-5553399	0.742 α=1.0	FBXL18 7:5515427-5553399	0.579 α=1.0	FBXL18 7:5515427-5553399	0.416 α=1.0	FBXL18 7:55154 5553399
1.329 α=1.0	MYO3B 2:171034654-171511674	1.194 α=1.0	SDK1 7:3341079-4308631	1.051 α=1.0	SDK1 7:3341079-4308631	0.9 α=1.0	SDK1 7:3341079-4308631	0.742 α=1.0	SDK1 7:3341079-4308631	0.579 α=1.0	SDK1 7:3341079-4308631	0.416 α=1.0	SDK1 7:33410 4308631
1.329 α=1.0	ITGA1 5:52084135-52249485	1.194 α=1.0	ERICH2 2:171627191-171655481	1.051 α=1.0	ERICH2 2:171627191-171655481	0.9 α=1.0	ERICH2 2:171627191-171655481	0.742 α=1.0	ERICH2 2:171627191-171655481	0.579 α=1.0	ERICH2 2:171627191-171655481	0.416 α=1.0	ERICH2 2:17162 1716554
1.329 α=1.0	SWAP70 11:9685627-9774507	1.194 α=1.0	DPT 1:168664694-168698442	1.051 α=1.0	DPT 1:168664694-168698442	0.9 α=1.0	DPT 1:168664694-168698442	0.742 α=1.0	DPT 1:168664694-168698442	0.579 α=1.0	DPT 1:168664694-168698442	0.416 α=1.0	DPT 1:16866 1686984
1.329 α=1.0	FBXL18 7:5515427-5553399	1.194 α=1.0	RFX5 1:151313115-151319769	1.051 α=1.0	RFX5 1:151313115-151319769	0.9 α=1.0	RFX5 1:151313115-151319769	0.742 α=1.0	RFX5 1:151313115-151319769	0.579 α=1.0	RFX5 1:151313115-151319769	0.416 α=1.0	RFX5 1:15131 1513197

The top ranked N LOD and HLOD scores for each θ 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 α 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



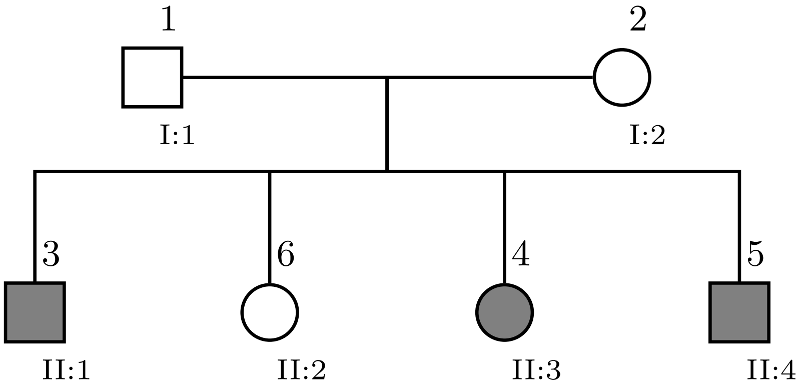
LOD and HLOD scores for all markers analyzed are displayed per chromosome as heatmaps, using a sequence of blue colors from light to dark for the score values; the darker the color the higher the score. Linkage regions of potential interest across the entire genome can be easily identified on the heatmaps.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

Chapter 2

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 [1] of two nuclear families of the same structure and phenotypic 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.



[1] 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 running linkage analysis. For now we (mostly) stick to default settings and focus on interpretation of terminal output and marker data 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 [2] 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.

[2] 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

In the example above, regional marker data generated via the CHP method is written to a folder called `LINKAGE` (this is default value for parameter `--output`). Data are saved per file bundle per chromosome and for Merlin format the data bundle consists of 3 files: the PED file, the MAP file and the DAT file.

contents in LINKAGE/MERLIN folder

```
LINKAGE.chr16.dat  LINKAGE.chr16.map  LINKAGE.chr16.ped  LINKAGE.chr1.dat  LINKAGE.chr1.map  LINKAGE.chr1.ped
```

The PED file `LINKAGE.chr1.ped` is displayed below:

LINKAGE.chr1.ped

```
1       II:2      I:1       I:2       2       2       1       4
1       I:2     0       0       2       1       2       1
1       I:1     0       0       1       1       3       4
1       II:3      I:1       I:2       1       2       1       4
1       II:4      I:1       I:2       2       1       2       3
1       II:1      I:1       I:2       1       2       1       4
2       II:B      I:A       I:B       2       2       4       1
2       I:B     0       0       2       1       2       4
2       I:A     0       0       1       1       1       3
2       II:C      I:A       I:B       1       2       4       1
2       II:D      I:A       I:B       2       1       2       1
2       II:A      I:A       I:B       1       2       2       3
```

The first 6 columns of the PED file contains information in the input FAM file. Since the test data have only one gene on chromosome 1, there is only one regional marker generated for this chromosome and the genotypes are found at the last two columns of this file.

The MAP file `LINKAGE.chr1.map` is displayed below. *PAPPA2* is the marker name and the genetic distances are extrapolated from Rutgers linkage-physical map.

LINKAGE.chr1.map

```
CHROMOSOME        MARKER              POSITION          FEMALE_POSITION         MALE_POSITION
1        PAPPA2        186.278964324        136.402021932        238.991541401
```

The DAT file `LINKAGE.chr1.dat` is displayed below.

LINKAGE.chr1.dat

```
A        Trait
M        PAPPA2
```

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 definitions. 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

To see the outcome of different collapsing themes, for example we look at marker *PAPPA2* on chromosome 1. The shell command below extracts the genotypes from the three themes and concatenated them for visual convenience – the first 2 columns of integer values are LD 0.8 collapsing, the middle 2 columns are no collapsing and the last 2 columns are complete collapsing.

```
paste RMBPt8/MERLIN/RMBPt8.chr1.ped RMB1/MERLIN/RMB1.chr1.ped RMB0/MERLIN/RMB0.chr1.ped | cut -f 2,3,4,7,8,15,16,23\
,24
```

```
concatenated genotypes from different collapsing themes
II:2        I:1        I:2        1        4        1        4        1        2
I:2        0        0        2        1        2        1        2        1
I:1        0        0        3        4        3        4        2        2
II:3        I:1        I:2        1        4        1        4        1        2
II:4        I:1        I:2        2        3        2        3        2        2
II:1        I:1        I:2        1        4        1        4        1        2
II:B        I:A        I:B        4        1        4        1        2        1
I:B        0        0        2        4        2        4        2        2
```

```
I:A       0       0       1       3       1       3       1       2
II:C      I:A     I:B     4       1       4       1       2       1
II:D      I:A     I:B     2       1       2       1       2       1
II:A      I:A     I:B     2       3       2       3       2       2
```

From the output we see that the first two collapsing themes are identical, indicating there are no variants in this gene that are in strong LD with each other. The last theme is different and has only two alleles: this collapsing theme is useful for generating data for use of linkage analysis software that only accept a limited number of alternative alleles.

2.2.3 Recombination events

For regional markers sub-divided into smaller units by recombination events, we use `[1]`, `[2]`, `...`, `[i]` convention to label different units (see below). They can be treated different regional markers. In the linkage analysis pipeline incorporated in SEQLinkage, we choose the one subunit that gives strongest evidence of linkage to represent the entire region under consideration when computing combined LOD scores and HLOD scores for all samples.

LINKAGE.chr16.map

```
CHROMOSOME        MARKER              POSITION        FEMALE_POSITION        MALE_POSITION
16        MC1R[1]        133.689089888        111.195245154        159.050776809
16        MC1R[2]        133.689089888        111.195245154        159.050776809
```

2.3 Two-point Linkage Analysis

SEQLinkage incorporates FASTLINK [3] routine to perform two-point linkage analysis. It is difficult to handle the original FASTLINK routine to analyze data for our purpose, due to the nature of both the FASTLINK software and the way SEQLinkage generates markers. The FASTLINK routine involves running 4 standalone programs sequentially, some of which are "interactive mode" only (i.e. users will have to follow a series of "questions" from the program, entering input parameters only when prompted). The coding and allele frequencies of regional markers are unique to each family; as a result only one marker from one family can be processed with each FASTLINK analysis, which means numerous input files are required. To ease such efforts, SEQLinkage automatically generates markers from VCF file input, executes the FASTLINK analysis workflow hiding under the hook issues discussed above, and generates user friendly table and graphical summaries to present analysis results. This document only focus on the SEQLinkage command interface and output results. The FASTLINK website offers a number of documentations explaining details involved in linkage analysis, to which users should resort for questions regarding choice of parameters, analysis algorithms and interpretation of LOD scores from linkage analysis.

In addition to reporting LOD scores computed by FASTLINK, for analysis involving multiple families SEQLinkage computes HLOD scores based on methods described in Terwilliger and Ott

[3] FASTLINK http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html

(1994) [4], but instead of using grid search method we applied numerical optimizations to find the α value that maximize the HLOD score.

2.3.1 Regional marker linkage analysis

We use default CHP parameters to generate regional markers and perform linkage analysis with command below:

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --freq EVSEAAF -o LinkageAnalysis \
-K 0.001 --moi AR -W 0 -M 1 --theta-max 0.5 --theta-inc 0.05 -j 8 --run-linkage
```

Parameters `-K/--moi/-W/-M` are important user input that defines the parametric model for linkage analysis. Here we assume a prevalence of $\frac{1}{1000}$, autosomal recessive disease with complete penetrance. In practice users should carefully set these parameters for specific studies. The terminal output is as follows:

regional marker linkage

```
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 LINKAGE format
MESSAGE: 2 units successfully converted to LINKAGE format
MESSAGE: Archiving LINKAGE format to directory [/ramcache/cache]
MESSAGE: Linkage analysis succesfully performed for 2 units
MESSAGE: Reports in html format generated
```

Results are written to file `LinkageAnalysis/LinkageAnalysis_Report.html`.

Note

> In examples in previous sections we left `--freq` option unspecified, thus MAF from founders in samples are used to compute the regional marker frequencies. In the example above, to demonstrate the use of external MAF reference we annotated the example VCF file with allele frequency for European Americans in Exome Variant Server (EVS), assuming the simulated variant data is from European American samples. We incorporated the information to the VCF file INFO field `EVSEAAF` as long as the variant is found in EVS; for variants not found in EVS we set `EVSEAAF=0.00015`. Please refer to the SEQLinkage paper for details in the motivation and calculation of regional marker allele frequencies using external annotation sources.

[4] Terwilliger and Ott (1994) Joseph Douglas Terwilliger and Jurg Ott, Handbook of Human Genetic Linkage, Johns Hopkins Univ Pr, 1994 (ISBN: 9780801848032)

2.3.2 Single variant marker linkage analysis

It is also possible to apply linkage analysis directly on the original sequence data, taking advantage of the user friendly linkage analysis feature in SEQLinkage. Simply add `--single-markers` switch to the command above:

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --freq EVSEAAF --single-markers \
-o LinkageAnalysisSNV -K 0.001 --moi AR -W 0 -M 1 --theta-max 0.5 --theta-inc 0.05 -j 8 --run-linkage
```

single variant linkage

```
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 18 pre-defined units
MESSAGE: 17 units (from 18 variants) processed; 3 Mendelian inconsistencies and 0 recombination events handled
MESSAGE: 1 units ignored due to absence of variation in samples
MESSAGE: Archiving regional marker data to directory [/ramcache/cache]
MESSAGE: 17 units will be converted to LINKAGE format
MESSAGE: 17 units successfully converted to LINKAGE format
MESSAGE: Archiving LINKAGE format to directory [/ramcache/cache]
MESSAGE: Linkage analysis succesfully performed for 17 units
MESSAGE: Reports in html format generated
```

Note that the total number of units analyzed is now 17, which is the number of polymorphic sites in sample. Results are written to `LinkageAnalysisSNV/LinkageAnalysisSNV_Report.html`

2.3.3 Accessing archived analysis results

The HTML output only displays top N signals of LOD and HLOD but the result for all markers can be found at the output directory. For example for regional marker analysis above the text files of LOD and HLOD are stored to `LinkageAnalysis/heatmap/*.lods` and `LinkageAnalysis/heatmap/*.hlods` respectively. Similar is the case for single variant marker analysis. The default $N = 10$ signals are displayed on the HTML file for single marker analysis but you may want to set $N = 20$ so that all 17 units can be displayed. To do this without re-running the entire analysis (which may be very time consuming for real world WES or WGS data), you can add the `--cache` switch to the command, for example:

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --freq EVSEAAF --single-markers \
-o LinkageAnalysisSNV -K 0.001 --moi AR -W 0 -M 1 --theta-max 0.5 --theta-inc 0.05 -j 8 --run-linkage \
--cache --output-entries 20
```

load results from cache

```
MESSAGE: Binary trait detected in [/ramcache/seqlinkage-example.fam]
MESSAGE: Checking local resources 5/5 ...
MESSAGE: Loading regional marker data from archive ...
MESSAGE: Loading LINKAGE data from archive ...
MESSAGE: Loading linkage analysis result from archive ...
MESSAGE: Reports in html format generated
```

The terminal output suggests that archived results are used, but the HTML table now contains complete entries for all markers in data analyzed.

2.4 SEQLinkage with Other Linkage Programs

We have previously shown the use of SEQLinkage to write regional markers generated to LINKAGE and MERLIN formats. Here we demonstrate one more supported program format, the `Mega2` format.

2.4.1 Output regional marker data to `Mega2` format

`Mega2` [5] is a versatile data-handling program for facilitating genetic linkage analysis. It can be used to convert data formats between various linkage software. SEQLinkage supports output of regional markers to `Mega2` format which can then be converted to input for a number of other linkage analysis programs via `Mega2`. For example,

```
seqlink --fam seqlinkage-example.fam --vcf seqlinkage-example.vcf.gz --freq EVSEAAF -o example --bin 0 --format MEG\
A2
```

Mega2 format

```
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 MEGA2 format
MESSAGE: 2 units successfully converted to MEGA2 format
MESSAGE: Archiving MEGA2 format to directory [/ramcache/cache]
MESSAGE: Saving data to [/ramcache/example]
```

Note

> The collapsing theme `--bin 0` is used for `Mega2` format, because unlike with the `LINKAGE` program it can only handle bi-allelic variants.

2.5 Prepare Customized “Blueprint” of Regional Markers

A utility program [6] can be used to find the genetic positions (in cM) for genomic regions provided by users. To use the program, you must have `Python` and `tabix` programs installed to your system, and the human linkage-physical map [7] downloaded to your hard drive. The first 4 columns of input file should be chromosome, start position (hg19 coordinate), end position (hg19 coordinate), region name delimited by white space (tab or multiple spaces). The columns of output file are chromosome, start position, end position, region name, genetic position of the gene on average, in female and in male, delimited by tab. The command is:

[5] Mega2 http://watson.hgen.pitt.edu/docs/mega2_html/mega2.html

[6] utility program downloadable via command `wget http://www.bioinformatics.org/seqlink/uploads/genetic_pos_searcher`

[7] the human linkage-physical map http://compgen.rutgers.edu/maps

```
./genetic_pos_searcher /path/to/input/file
```

The RutgersMap provides the physical and genetic positions of large amount of SNPs, allowing determination of cM-scale linkage-based map positions for any marker, given its physical position.

The utility program works as follows. For a given region, it will take the middle position of the region as the position of the region. Two closest SNP markers to this position will be searched in RutgersMap. Then the genetic position of the region will be interpolated using the relative distance from the position to the two markers. Under the circumstances that the position of the region is outside of the boundary of the smallest or largest positions provided by RutgersMap, the genetic position of the region will be interpolated in scale with respect to the smallest or largest positions.

For chromosome X, if the position falls in [175751,2800677] or [155045646,155150393], i.e. pseudo-autosomal regions (hg19 coordinate), genetic position in both male and female will be interpolated; otherwise only genetic position in female are available. No genetic position on average can be obtained for chromosome X.