LINKDATAGEN Documentation

Updated 3rd February 2016. (SVN revision 997)
For comments please contact Melanie Bahlo (bahlo@wehi.edu.au)

Developed by Melanie Bahlo and Catherine Bromhead, and with help from Tom Scerri, Katherine Smith, Rick Tankard and Luke Gandolfo.

As of February 2012, the three separate LINKDATAGEN scripts that were available for download, i.e. linkdatagen_affy.pl, linkdatagen_illumina.pl and linkdatagen_mps.pl have been amalgamated into a single script:

linkdatagen.pl

The **old scripts are now defunct** and should not be used any more.

The new LINKDATAGEN will continue being supported and further developed.

If you use LINKDATAGEN, please **acknowledge by citing**:

Bahlo M, Bromhead CJ. Generating linkage mapping files from Affymetrix SNP chip data. Bioinformatics 2009;25(15):1961-2.

If you use the LINKDATAGEN `-data m` option along with `vcf2linkdatagen.pl` then please also cite:

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biology 2011;12:R85.

PLEASE SEE NEXT PAGE(S) FOR IMPORTANT CHANGES

Recent Changes

Many of the command line options have had their names changed. Most changes involve turning their names into camel case (but option names are case-insensitive for LINKDATAGEN). This was done purely for consistency. However, other changes are to be aware of:

Old name	New name	Reason/Explanation
-annot_dir	-annotDir	Name changed for consistency.
-removeUninf	-removeWFHBS	Remove SNP markers that show "whole/within-family homozygosity-by-state". Default behaviour is not to remove such markers.
-snpChoice	-randomSNP	Select a "random SNP" from each bin. Default behaviour is to select the most heterozygous SNP.
-merr	-keepME	Request to "keep Mendelian error" SNPs. Default behaviour is to remove SNPs with Mendelian errors.

Inside the folder created for MERLIN is now:

1. The same "per chromosome" files as before. However, the "*.pre" files have been renamed "*.ped". Excessive tabs have been removed from these "*.ped" files.
2. A new "genome" folder with files (.ped|.dat|.map|.freq|.in) containing data for the whole genome, rather than per chromosome. Using these to run MERLIN speeds up and makes easier the linkage analysis.
3. README files explaining how to proceed with the linkage analysis in MERLIN, although inexperienced users should refer to the MERLIN website.

The -minDist { real number >= 0.0 } option can now accept a value for the minimum distance (in cM) between selected SNPs.

The -help option will display all options and their accepted values.

The -popHetTest option is now recommended to be run by selecting the "most heterozygous SNP" from each bin, rather than a "random SNP" per bin. SNPs are selected to have a MAF > 0.4 regardless of whether they are chosen randomly are as the most heterozygous.

The -popCol { integer >= 1 } option allows allele frequencies to be selected from an annotation file by column number rather than by specifying a population. Remember to count all columns; this includes any preceding columns such as the chromosome number, position, alleles etc…. The first column is 1.

The -bestPopTest option will now perform a goodness-of-fit test on each given sample for every population available in the annotation file, and will inform the user of the population allele frequencies that best match their sample(s).

1. Introduction

LINKDATAGEN is a PERL (5.8+) script that generates datasets for linkage analysis, relatedness checking, IBD and HBD inference and association analyses using genotypes generated from any of Affymetrix SNP chips, Illumina SNP chips or massively parallel sequencing (MPS; otherwise known as next generation sequencing or NGS) data.

LINKDATAGEN selects markers for linkage and association analyses and performs nuclear family Mendelian error detection. It also performs sex checks using both X and Y chromosome markers (if available). LINKDATAGEN supports all eleven HapMap populations (ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI and YRI) for the following platforms, unless otherwise specified:

- Affymetrix: 50K Xba, 50K Hind, 250K Sty, 250K Nsp, 5.0, 6.0 and 500K Sty+Nsp
- Illumina: 610Quad, Cyto12, Omni Express, 1M, plus others due to high overlap between Illumina platforms
- MPS data: Hapmap2 (for CEU, CHB, JPT and YRI populations) and HapMap3

However many other chips can be catered for (see -chip section).

LINKDATAGEN creates output files for the linkage mapping software ALLEGRO, MERLIN, MORGAN and PLINK, as well as for BEAGLE, FESTIM, PREST, fastPHASE and RELATE. Many of these programs are available through the Rockefeller website (http://linkage.rockefeller.edu/soft).

2. Download

Download of LINKDATAGEN is through the LINKDATAGEN website at: http://bioinf.wehi.edu.au/software/linkdatagen/. Please check periodically for updates.

LINKDATAGEN, the companion program VCF2LINKDATAGEN, test files, annotation files and documentation can now be downloaded together as `linkdatagen.tar.gz`.

To uncompress and extract the archive, type:

```
tar -xfvz linkdatagen.tar.gz
```

3. Usage

Typing linkdatagen.pl -help on the command line will provide a complete list of options available. However, please still read this manual first before using LINKDATAGEN.

The following section describes each of the options available when running LINKDATAGEN.

3.1. Mandatory options:

`-data { a , i , m }`

This indicates the genotype data source and is one of three options:

a = Affymetrix SNP chip data
i = Illumina SNP chip data
m = massively parallel sequencing (MPS; or NGS) data.

If Affymetrix or MPS data is used, then the -callFile option is required. If Illumina data is used, then the -callDir option is required instead.

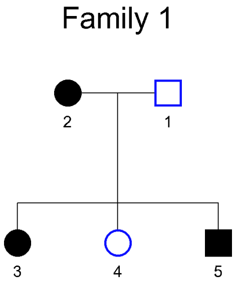
`-pedfile < filename >`

A standard (common) pedigree file (pedfile) with 6 columns:

1. FID: Family (pedigree) ID
2. IID: Individual's ID
3. PID: Paternal (father's) ID (0 if individual is a founder)
4. MID: Maternal (mother's) ID (0 if individual is a founder)
5. SEX: Sex of individual (0 = unknown, 1 = male, 2 = female)
6. AFF: Affection status (0 = unknown, 1 = unaffected, 2 = affected)

The columns of the pedigree file can be separated by tabs or spaces. A header row is optional. If the first row of the pedigree file does not contain any digits (numbers), then it will be assumed to be a header row — in other words, digits are not allowed in the header row. This file must be in the working directory that LINKDATGEN is run from.

As an example of a pedfile, consider the pedigree:



The pedfile for this family ("Family 1") is:

FID	IID	PID	MID	SEX	AFF
1	1	0	0	1	1
1	2	0	0	2	2
1	3	1	2	2	2
1	4	1	2	2	1
1	5	1	2	1	2

Further notes on how to code up pedigree files can be found at the MERLIN website under "Describing Relationships Between Individuals" at:

http://www.sph.umich.edu/csg/abecasis/Merlin/tour/input_files.html

LINKDATAGEN allows input of multiple families (pedigrees) in a single pedfile. In such a pedfile, each family (pedigree) will require a unique family ID (FID). Care needs to be taken with the `-removeWFHBS` option if this is the case.

```
-whichSamplesFile < filename >
-whichSamplesList < filename >
```

LINKDATAGEN requires one, and **only one**, of these two options. They are **mutually exclusive**. Both options work similarly in that they **connect the genotype data to the correct individuals in the pedfile**. They differ in their usage with regard to the data type. No matter which option is used, the file indicated by that option must be put in the working directory in which LINKDATAGEN is being run. If LINKDATAGEN complains of errors then it is worthwhile double-checking this file.

The "whichSamplesFile" file contains a single line of N values, where N equals the number of individuals in the pedfile. The values must be separated by tabs or spaces. The *k*th value in the "whichSamplesFile" indicates where to find genotypes for the individual listed in the *k*th row of the pedigree file. The value 0 indicates that the corresponding individual is not genotyped; otherwise:

- For Affymetrix SNP chip data and MPS data, specify the index of the column in the BRLMM file that lists that individual's genotypes. The first column of the BRLMM file lists the SNP names, so specify 1 for the first genotype column (second column overall), etc. The name (and path if necessary) of the BRLMM file must be specified when using `-callFile < filename >`
- For Illumina data the value X is given for each individual, where the file containing genotypes for that individual has a name "*_FinalReportX.txt" (**see next paragraph** for more complete details).

If using `-whichSamplesFile` for Illumina data, then the "Final Report" files must be named *_FinalReportX.txt, where X is the corresponding numeric entry in the “whichSamplesFile”. It is also required that the `-callDir` option specifies the location of the *_FinalReportX.txt files **and** simultaneously specifies the exact prefix (*) of the *_FinalReportX.txt files. The following example should help clarify these instructions.

Suppose we genotyped the children (3, 4 and 5) from “Family 1” (above), and that the genotype data are in a folder called “/GenoData/” with the filenames Fam1_FinalReport1.txt, Fam1_FinalReport3.txt and Fam1_FinalReport2.txt in that order.

We would have to specify “`-callDir` /GenoData/Fam1”, and then we would use the `-whichSamplesFile` option to specify a file containing the single line:

```
0 0 1 3 2
```

This means that the genotypes from the parents are not available and the genotypes for individuals 3, 4 and 5 are found in the files Fam1_FinalReport1.txt, Fam1_FinalReport3.txt and Fam1_FinalReport2.txt, respectively, in the folder “/GenoData/”.

For Illumina data, the alternative option `-whichSamplesList` specifies a file containing one line for each individual in the pedfile, with each line corresponding to an individual in the pedfile in the **same order**. Each line specifies the filename of the file containing genotypes for that individual, or 0 if no genotypes are available.

Specifying `-whichSamplesList` only makes sense for Illumina SNP chip data, where genotypes for different individuals are provided in different “Illumina Final Report” files. These generally have names of the form *_FinalReport1.txt, *_FinalReport2.txt etc..., in which case you could use `-whichSamplesFile` (**as described above**). Alternatively, `-whichSamplesList` can specify a file that lists these filenames explicitly and in the corresponding order of the pedfile. Following the example above for “Family 1”, this would be:

```
0
0
Fam1_FinalReport1.txt
Fam1_FinalReport3.txt
Fam1_FinalReport2.txt
```

```
-callDir < path to directory >
-callFile < filename >
```

LINKDATAGEN may take up to one of these two options. They are **mutually exclusive**. Both options work similarly in that they identify the location of the genotype data.

The `-callDir` option should **only** be used with Illumina data. It specifies the directory containing the genotype data files, with one output file per individual, named *_Final_ReportX.txt. This option can be ignored for Illumina data either if all the data sits in the working directory, or the full path is specified for the files listed in the file indicated by `-whichSamplesList`.

The `-callFile` option can be used with either Affymetix or MPS data types. It specifies a **single file** containing genotype data for all individuals in the pedfile.

Normally, Affymetrix genotyping data is generated by BRLMM (typically brlmm.calls.txt) or CRLMM (typically crlmm.calls.txt) and contains the genotype calls in a single file. BRLMM genotypes are coded as -1, 0, 1 and 2 while CRLMM genotypes are coded as 0, 1, 2 and 3. MPS data can also be coded in a similar way to BRLMM.

The coding of SNP alleles to these formats is derived by coding the A allele as the first alphabetical base, e.g. if a SNP is an A/T polymorphism AA becomes 11, TT 22 and AT 12. Using the HapMap annotation files we always refer to SNPs on the forward (+) strand. Allele frequencies always refer to the A (not B) allele and thus are a mix of major and minor allele frequencies. Hence, the BRLMM/CRLMM formats are shown in the table below.

Genotype	Birdseed/BRLMM	R/oligo/CRLMM
No call / missing data	-1	0
AA	0	1
AB	1	2
BB	2	3

The genotype call file contains one column of SNP identifiers and one column of genotype calls for each individual, for example:

```
rs12345678   0      0      1
rs56789012   2      0      1
rs13579135   1      1      1
```

```
-annotfile < filename >
-chip { 1 , 2 , 3 , 4 , 5 , 6 , 7 }
```

LINKDATAGEN requires one, and **only one**, of these two options. They are **mutually exclusive**. Both options work similarly in that they identify the appropriate **annotation file**. The annotation files include data on allele frequencies, and genetic and physical map positions.

The `-annotfile` option explicitly specifies the filename of the annotation file required.

The `-chip` option specifies the genotyping platform, relative to the data type, and hence selects the appropriate annotation file:

```
-data a -chip { 1 , 2 , 3 , 4 , 5 , 6 , 7 }
-data i -chip { 1 , 2 , 3 , 4 , 5 , 6 }
-data m -chip { 1 , 2 }
```

The choices above correspond to:

`-chip #`	`-data a`	`-data i`	`-data m`
1	50k Xba		HapMap2
2	50k Hind	610Quad	HapMap3
3	250k Sty		
4	250k Nsp	Cyto12	
5	5.0	Omni Express	
6	6.0	1M	
7	500k Sty+Nsp	HapMap2	
8		HapMap3	

Note: HapMap2 (only CEU, YRI, CHB, JPT, ~4 million SNPs)
HapMap3 (all 11 HapMap populations, ~1.5 million SNPs)

Illumina SNP chips have a lot of overlap. Hence, even if no annotation file currently exists, then one can use a related file and probably gain sufficient SNP coverage to allow at least linkage mapping. This includes the ability to specify the HapMap2 and HapMap3 annotation files with the -annotfile option that are universally compatible with MPS and Illumina SNP chip data, namely annotHapMap2U.txt and annotHapMap3U.txt. Overlaps of the various annotations (filtering sites without frequencies) and call files are given in the file on the LINKDATAGEN website linkdatagen_chip_overlaps.xlsx, in summary if you have one of the following Illumina chips then we suggest you perform your analysis with the corresponding annotation for your population:

	Annotation with greatest number of overlapping markers	
Illumina chip	**CEU, YRI, CHB, JPT populations only**	**All HapMap3 populations**
660Quad	HapMap2	610Quad
Omni1Quad	HapMap2	Omni Express
Omni2.5	Omni Express	Omni Express
Omni5	HapMap2	1m
HumanCoreExome	Omni Express	Omni Express

If using the `-chip` option, then the `-annotDir` option is also required.

```
-annotDir < path to directory >
```

Specify the directory containing the annotation files. This is mandatory if using the `-chip` option.

```
-prog { all | me | al | mo | pl | pr | cp | be | fe | re |
fp }
-popHetTest { summary , verbose , perChr , perChrVerbose
-freq
```

LINKDATAGEN requires **at least one of these three options** to be specified; else `-bestPopTes`t alone may be selected (see below). They all do very different things. In brief, -`prog` specifies the type of program specific output formats of files that LINKDATAGEN should create, `-popHetTest` performs a goodness-of-fit test of the selected population allele frequencies against the genotyped samples, and `-freq` outputs the allele frequencies for the founders of your genotyped samples.

The `-prog` option allows LINKDATAGEN to output files in different formats:

1. ALLEGRO { al } (Gudbjartsson, et al., 2000; Gudbjartsson, et al., 2005)
2. MERLIN { me } (Abecasis, et al., 2002; Abecasis and Wigginton, 2005)
3. MORGAN { mo } (Thompson, 1995; Thompson, 2000)
4. PREST { pr } (McPeek and Sun, 2000)
5. PLINK { pl } (Purcell, et al., 2007)
6. BEAGLE { be } (Browning 2006; Browning & Browning 2007)
7. FESTIM { fe } (Leutenegger 2006)
8. RELATE { re } (Albrechtsen et al. 2009, Albrechtsen et al. 2011)
9. fastPHASE { fp } (Scheet and Stephens, 2006)
10. Our COMPLETE genotype format { cp }
11. All of the above { all }

These can be selected individually or a comma separated list, e.g. to generate files for MERLIN, FESTIM and RELATE only, use "`-prog me,fe,re`".

ALLEGRO and MERLIN are exact multipoint linkage analysis programs, MORGAN allows a variety of Markov Chain Monte Carlo (MCMC) calculations but in particular is currently set up to perform MCMC multipoint linkage analysis.

PREST allows the identification of pedigree errors. PLINK is a program geared towards genome wide association analysis. BEAGLE can be used for genotype imputation, phasing and detection of IBD/HBD between individuals. FEstim can infer inbreeding coefficients and adjust linkage results based on inbreeding. RELATE also infers inbreeding and relatedness, but in the presence of linkage disequilibrium. Genotype imputation can be performed with fastPHASE.

LINKDATAGEN also outputs our own COMPLETE { cp } format for analysis of identity-by-state sharing (IBS). This format should be used with "`-binsize 0`" so that all data is displayed. The outputted data is in CRLMM/R-oligo format:

Genotype	CRLMM
No call / missing data	0
AA	1
AB	2
BB	3

Using "`-prog cp`" prints out individuals sorted by affection status with genotyping data displayed with SNPs in rows and individuals in columns. SNPs are ordered by genetic map distance. Three IBS sharing statistics modelled on the optimal S sharing statistics proposed by McPeek (McPeek, 1999) are implemented in the columns after the affecteds:

1. S_robdom — suitable for dominant pedigree but robust to a range of disease allele frequencies and penetrances
2. S_HBS — suitable for homozygosity mapping
3. S_pairs — suitable for recessive diseases not covered by homozygosity mapping

These have all been rescaled to lie between 0 and 1.

The `*.cp` files are best opened in EXCEL and annotated as desired by the researcher.

The `-popHetTest` performs a goodness-of-fit test comparing the **observed** genotype counts (AA, AB and BB) of SNPs from the genotyped sample(s) against the **expected** genotype counts given the selected population (e.g. CEU or JPT). This test is useful to establish whether or not the selected population allele frequencies are a good match for the genotyped sample(s). When using this option, it is **recommended** to select the most heterozygous SNPs from each bin (**do not use** `-randomSNP`) and not to filter out SNPs displaying homozygosity-by-state (**do not use** `-removeWFHBS`).

One of four **mutually exclusive** values can be supplied when running `-popHetTest`:

a) summary — (the default, not required to be specified) gives a small table of genome-wide and autosome-wide chi-square values for each sample; file generated called "popHetTest.txt".

b) verbose — gives the observed and expected counts used to derive the genome and autosome test statistics; file generated called "popHetTestDetails.txt".

c) perChr — gives the test statistics for each chromosome individually as well as for the genome and autosome; file generated called "popHetTestDetails.txt".

d) perChrVerbose — gives everything — all observed and expected counts for each chromosome, the genome and the autosome, used to derive the test statistics; file generated called "popHetTestDetails.txt".

The `-freq` option makes LINKDATAGEN output allele frequency estimates from the founders in the family. This option generates a file called "alleleFreqs.txt". This file contains data from all SNPs with no Mendelian errors and gives the allele frequency estimates based on founders alongside the specific population allele frequency estimates that have been selected.

If there are many founders (>30, i.e. >60 alleles) in the pedigree files then allele frequencies estimated from the data itself could be used as the population allele frequency in the generation of the linkage files. This could be useful for boutique populations. To do this, one would LINKDATAGEN twice, once to generate the frequency files with `-freq`, and then again with the "homemade" annotation data and the `-prog` option of choice. The "homemade" annotation data could be constructed by appending your column of allele frequencies to the end of an already existing annotation file, and then using the `-popCol` option to specify the exact column containing your custom allele frequencies.

`-bestPopTest`

This option is a **mutually exclusive** alternative to the options `-prog` , `-freq` , and `-popHetTest` . In effect, `-bestPopTest` cycles through `-popHetTest` and tests each of the available populations in the annotation file to see which population allele frequencies best fit the allele frequencies of the given sample(s).

3.2. Other options:

`-randomSNP`

If not defined, the default setting is to select the SNP with highest heterozygosity in the genetic map interval (size defined by `-binsize`), as calculated using the selected population allele frequencies. Specifying `-randomSNP` leads to random marker selection in any given genetic map.

High heterozygosity markers give a set of more highly informative markers for most analyses, such as linkage mapping. There would be little reason, or need, to use the `-randomSNP` option.

`-binsize { real number >= 0.0 }`

Default setting is 0.3 cM. The `-binsize` parameter exists to allow selection of markers in approximate linkage equilibrium. It works by selecting non-overlapping bins (or windows) of markers within a span of the `-binsize` length (cM). From these markers either a random (`-randomSNP`) or the highest heterozygosity marker (default) according to the stipulated HapMap population (default is the HapMap CEU dataset), is chosen from the bin to be its representative in the dataset. A minimum distance is imposed to make sure that markers are sufficiently distant (and hence not in linkage disequilibrium). Whilst MERLIN, RELATE and BEAGLE have internal methods of dealing with linkage disequilibrium, such as cluster analysis, ALLEGRO, MORGAN, FEstim, and PREST do not.

`-outputdir < prefix >`

The prefix to the name of the directories into which LINKDATAGEN will write all the files.

`-keepME`

Specify if markers found with Mendelian Errors are **NOT** to be removed from dataset. *The default is to perform Mendelian error checking and remove SNPs that are found to contain a Mendelian error.* Since we usually have so many markers to choose from this hardly makes a difference. It also enables creation of MORGAN and ALLEGRO input files without having to tediously remove Mendelian errors afterwards since neither of these programs allow for this.

LINKDATAGEN produces extensive output based on the Mendelian errors summarised per chromosome and by individual. Note that the by individual summarisation is only possible for individuals whose parents are at least partially genotyped. Thus high error rates in some individuals need to be interpreted with caution; that is, one of their parents could be the individual whose sample is problematic, not the actual individual in whom the high count of errors has been listed. For example if multiple siblings appear to have high Mendelian error rates then it is much more likely that the parent is the problem.

A log file is created called "mendelErrors.txt" which lists all SNPs with at least one Mendelian error detected, in genetic map position order. This is useful when there are gross cytogenetic abnormalities as these are reflected as runs of Mendelian errors.

`-minDist { real number >= 0.0 }`

Set the minimum distance between markers. This is the minimum distance that needs to be maintained between neighbouring SNPs chosen as representatives of their bin. This parameter ensures that markers don't get too close, possibly ending up in linkage disequilibrium. If this can't be satisfied no marker is chosen for that bin. This means that consecutive bins may end up with no marker data. We have found this to be of little consequence since the marker datasets are usually very large. However please consult the log file `missingMarkers.txt` to check if any large genomic regions have been left uncovered. *The default setting is 0.5*binsize, and since the default* ***`-binsize`*** *is 0.3 cM, this is 0.15 cM. The maximum this value can take when not explicitly defined is 0.2 cM (for any* ***`-binsize`*** *>= 0.4, the default value of* ***`-minDist`*** *is set to 0.2).*

`-removeWFHBS { i, u }`

Remove "within/whole-family homozygosity-by-state" (WFHBS) SNP markers. To be used with caution. If analysing multiple families, the default behaviour is remove the **union** (**u**) of WFHBS SNP markers; that is any SNP marker displaying WFHBS in **any** family. Alternatively, LINKDATAGEN can be forced to remove the **intersection** (**i**) of WFHBS SNP markers; that is any SNP displaying WFHBS in **all** families. For details see section 4.3.

`-noX`

Specify if you wish the X chromosome to be excluded from output.

`-crlmm`

CRLMM data option. Use when genotypes are coded 0 (missing), 1, 2 and 3. The default is BRLMM data when genotypes are coded -1 (missing), 0, 1 or 2. This option is only relevant for `-data a` (Affymetrix) and usually not needed.

`-pop { ASW , CEU , CHB , CHD , GIH , JPT , LWK , MEX , MKK , TSI , YRI }`

Population allele frequency choice (**see 4.4 below**). Default is CEU.

Value	**Population (description)**
ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western

	European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MEX	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscans in Italy
YRI	Yoruba in Ibadan, Nigeria (West Africa)

```
-popCol { integer >= 1 }
```

An alternative option to `-pop` that allows the exact column to be selected from an annotation file, without relying on choosing the actual population name.

```
-regions { #, chr#, #:#####-#####, chr#:#####-#####, … }
-regionsFile < filename >
```

Both options restrict analyses to particular regions. With `-regions` a comma separated list of regions (as above) are specified where the # following "chr" is a number from 1-22 or X, Y and Z, and the subsequent #####s are start and stop base-pair positions. Whole chromosomes can be specified by supplying "chr#" or just alone "#" alone. With `-regionsFile` a file is specified containing each region per line. The format can also be space-delimited:

```
chr#:####-####
chr# #### ####
chr#
```

```
-fileKeepSNPs < filename >
-fileRemoveSNPs < filename >
```

Both options specify a file containing a list of SNPs (e.g. rs#####) one-per-a-line:

```
rs1234512
rs2345641
rs4353534
rs3453434
```

Option `-fileRemoveSNPs` will remove these SNPs from the selection process, thereby preventing them from being selected.

Option `-fileKeepSNPs` will only keep those SNPs for the selection process, thereby allowing only those SNPs listed to be selected. That is not to say all these SNPs will be selected (unless `-binsize 0`), but simply this list constrains the choice of SNPs that can be selected from.

Hence, both `-fileRemoveSNPs` and `-fileKeepSNPs` could (theoretically) be used to produce the same list of SNPs to be selected from. One option may be preferential to the other.

A SNP in the "remove" list will **not** be kept if it is also in the "keep" list. That is, the "remove" list overrides the "keep" list when the same SNP appears twice.

The "remove" list might be used to exclude problematic SNPs that you can not otherwise easily remove from your dataset, or it may be used to pick a completely different list of SNPs (with no intersection) to run through linkage/FEstim a second time round to see that you get similar results with different SNPs.

The "keep" list might be useful if you want to try to replicate something with exactly the same list of SNPs, e.g. you might want to try different population allele frequencies, or different families with the same list of SNPs. In both cases you might first use `-binsize 0.3` to select your SNPs, and then subsequently use `-binsize 0` to make sure all the same SNPs are selected.

Any listed SNP that is not in the annotation file will be ignored — there is no warning to say it has been removed or kept.

3.3. Output files:

The output files depends on the `-prog/-popHetTest/-freq` option chosen, however there are several log files of interest that are written to the < outputdir >_tables folder regardless of choice. These include:

1. `missingMarkers.txt`

   Bins with no SNP representation.

2. `chrX_SNPs.txt`

   SNPs on the X chromosome.

3. `chrY_SNPs.txt`

   SNPs on the Y chromosome.

4. `chrMT_SNPs.txt`

   SNPs on the mitochondrial chromosome.

5. `mendelErrors.txt`

   SNPs with Mendelian errors ordered by chromosome and genetic map position. This file can identify gross cytological changes such as CNVs.

6. `selectedSNPs.txt`

   SNPs selected for analysis or output with `-prog` or `-popHetTest` .

4. Further Comments

4.1. X chromosome error detection

The X chromosome data is now examined in detail to detect any sex swaps. Furthermore all X chromosome data is pooled by sex per individual and a quick test statistic calculated to test for the likelihood of wrong sex. The test statistic is a Z score derived from the normal approximation of the observed number of heterozygotes for a particular sample compared to the mean of the heterozygote counts of all samples of the same sex. All males with heterozygous calls will have these calls set to missing. The file `chrX_SNPs.txt` lists all SNPs on the X chromosome, based on the information in the annotation file.

4.2. Y chromosome check

Y chromosome markers should have genotype calls for males and no calls for females. Linkdatagen prints a list of proportions of Y chromosomes with called alleles for each genotyped individual. This can be used to check that samples have not been swapped. Please note that Affymetrix SNP chips have either no or very few Y chromosome markers so this is only useful for `-data m` or `-data i`. The file `chrY_SNPs.txt` lists all SNPs on the Y chromosome, based on the information in the annotation file.

4.3. Within/whole-family homozygosity-by-state ("uninformative") markers

Stipulation of this option (`-removeWFHBS`) means that within-family homozygosity-by-state markers, i.e. markers that have missing or identical homozygous genotypes across all samples within a family, are removed prior to marker selection. Such markers are deemed "uninformative" for some analyses. To clarify, different families may be homozygous for different alleles for a marker to become uninformative. When dealing with multiple families, two values are available; "u" (for union) and "i" (for intersection). With value "u", a marker that is uninformative in ANY family is removed. With value "i", a marker that is uninformative in ALL families is removed. Value "u" is the default and will generally produce more informative markers across all families. Value "i" may be useful in a number of settings, in particular when quality of genotypes is poor such that using value "u" would wipe out the majority of genotypes in all families.

Option `-removeWFHBS` is a dangerous but really useful option. At its worst it can introduce significant bias, in particular it can remove homozygosity by descent (HBD) signals in small pedigrees. However in small to moderate sized pedigrees with a reasonable number of founders, usage of this option leads to a more optimal choice of markers for linkage mapping. Linkage mapping, for example implemented in MERLIN, removes "uninformative" markers prior to linkage calculation, as these do not contribute to the LOD score. In small to moderate sized pedigrees markers with high heterozygosities may still be "uninformative" and thus not useful for mapping. Problems also arise in pedigrees where there is some mismatch between the given population allele frequencies and the real population allele frequencies, which are not known. Here `-removeWFHBS` also leads to a more suitable marker selection.

The default setting is NO removal of WFHBS markers.

4.4. Annotation files

Different population allele frequencies can be specified through the `-pop` option. *The default is the Caucasian (CEU) population frequency data.* Not all SNPs have frequency data available for all populations. HapMap Phase III populations are:

ASW :	African ancestry in Southwest USA
CEU :	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB :	Han Chinese in Beijing, China
CHD :	Chinese in Metropolitan Denver, Colorado
GIH :	Gujarati Indians in Houston, Texas
JPT :	Japanese in Tokyo, Japan
LWK :	Luhya in Webuye, Kenya
MEX :	Mexican ancestry in Los Angeles, California
MKK :	Maasai in Kinyawa, Kenya
TSI :	Toscans in Italy
YRI :	Yoruba in Ibadan, Nigeria (West Africa)

For HapMap3 there are ~1.5 million SNPs available, but some populations have substantially fewer SNPs.

When using the `-data m` option you may wish to use the HapMap Phase 2 frequency data. This only allows the choice of the initial four HapMap populations, but it gives a much larger choice of SNP markers (~4 million SNPs to choose from).

It is also possible to include your own set of SNP allele frequencies. This is useful for linkage mapping in boutique populations where the researcher has been able to generate their own control allele frequency data.

Note: we have generated all annotation files for Illumina, Affymetrix and MPS data ourselves, using Hapmap2 and Hapmap3 data since these data sets provide good control allele frequency data. The genetic map and physical map positions are hg19 based. The genetic map positions are derived by linear interpolation or directly from the genetic map generated by the Hapmap consortium (http://www.hapmap.org).

5. General comments

- Multiple pedigrees can be analysed together but require the use of `-whichSamplesFile`

- Check that all your newline characters are UNIX style, not MAC style. This can cause LINKDATAGEN to fail to read in the file properly and produce errors. On UNIX/MAC systems this can be fixed with the command:
  ```
  cat filein | tr '\r' '\n' > fileout
  ```

- Input files may be compressed with gzip (.gz) or bzip2 (.bz2) with the appropriate extensions in the brackets. This feature requires either the PerlIO::gzip or PerlIO::via::Bzip2 Perl libraries to be installed, available at http://www.cpan.org/.

- Checking your pedigree. LINKDATAGEN does not perform any checking of your pedigree. We suggest you use HaploPainter (Thiele and Nürnberg 2004) to "draw" your pedigree and check it this way. Look out for large numbers of reported Mendelian errors in the LINKDATAGEN output. This may indicate the presence of a pedigree error.

- Marker choice by bin is simplistic and there are more sophisticated ways of choosing markers but they rely on HapMap linkage disequilibrium estimates and tagging markers. We would argue that our approach is more flexible when it comes to dealing with uninformative markers or markers with no information. In practice we have also had a lot of success in identifying linkage with this method so we don't think we lose much by not using a more sophisticated approach. However "holes" may appear in the linkage map. Consult the `missingMarkers.txt` file to check.

- Only SNPs with names beginning with "rs" are read in by the program, all other markers are mercilessly removed. No CNV probes will be included for example.

- For Illumina data all A/T and C/G SNPs are currently missing from the annotation files due to confusion about strandedness. These amount to a small proportion (<5%) of all SNPs on the chips.

- Only SNPs with A/T/C/G alleles are currently used. The Illumina 1M chip has SNPs with alleles I & D (insertion & deletion). These SNPs are currently ignored by LINKDATAGEN.

- Currently discards SNPs with no frequency annotation in the chosen population. A better option in future would be to keep these SNPs when output options that do not require allele frequencies i.e. PLINK (`-prog pl`), BEAGLE (`-prog be`) or `-prog cp` are used

- It is very useful to run parametric or non-parametric linkage analysis using one of ALLEGRO, MERLIN or MORGAN and then, once a peak has been identified, to eyeball the data further using the `-prog cp` option, in conjunction with `-binsize 0.0` (to get all the markers) to further narrow the linkage peak.

6. Marker frequency data across different populations

Chip	**CEU**	**ASW**	**CHB**	**CHD**	**GIH**	**JPT**	**LWK**	**MEX**	**MKK**	**TSI**	**YRI**
XBA50	1	0.595	0.549	0.540	0.581	0.535	0.581	0.564	0.592	0.582	0.568
HIND50	1	0.571	0.508	0.500	0.544	0.494	0.554	0.534	0.567	0.545	0.541
NSP250	1	0.923	0.804	0.775	0.833	0.783	0.897	0.834	0.916	0.840	0.898
STY250	1	0.918	0.788	0.763	0.833	0.764	0.896	0.842	0.906	0.838	0.883
5	1	0.950	0.825	0.799	0.863	0.803	0.924	0.870	0.940	0.867	0.919
6	1	0.923	0.784	0.756	0.820	0.761	0.899	0.823	0.909	0.829	0.895

Table 1: Proportion of marker data from the HapMap populations for Affymetrix chips

Chip	**CEU**	**ASW**	**CHB**	**CHD**	**GIH**	**JPT**	**LWK**	**MEX**	**MKK**	**TSI**	**YRI**
OmniEx	0.980	0.792	0.981	0.712	0.754	0.981	0.782	0.766	0.787	0.756	0.980
Annot610	0.982	0.932	0.983	0.858	0.920	0.982	0.908	0.920	0.925	0.919	0.982
Cyto12	0.917	0.715	0.917	0.672	0.701	0.916	0.706	0.702	0.710	0.700	0.909
1M Duo	0.952	0.853	0.950	0.744	0.799	0.949	0.840	0.818	0.846	0.802	0.956

Table 2: Proportion of marker data from the HapMap populations for Illumina chips

7. Test Datasets

7.1. Affymetrix

A small test sample data set is available for download from the website (testdata.tar.gz). The tarred zipped archive includes a small nuclear pedigree with 3 children (ped1.ped), a whichSamplesFile (wsf), a file with genotyping calls from the Affymetrix 6.0 chip (calls.txt) and a shell script to run LINKDATAGEN (runlinkdatagen.sh). NOTE: The user may need to change the location of annotation directories with `-annotDir` to somewhere other than the current working dir.

7.2. Next generation sequencing (NGS)

Three test data sets are available for download. These are the three datasets corresponding to three pedigrees described in Smith et al (2011). For each individual in each dataset there are three files:

(i) A VCF file containing all possible SNPs from HapMap Phase II SNPs
(ii) A FinalReport file and
(iii) A VCF file containing only the SNP genotypes from the MPS data at the location of the genotyping array SNPSs. (iii) is merely for concordance checks.

The three families are:

(i) Family A: Single affected individual, recessive family, homozygosity mapping
(ii) Family T: Single affected individual, recessive family, homozygosity mapping
(iii) Family M: Two affected siblings, dominant family.

7.3. Illumina

The three families above can be used as example datasets for Illumina data.

8. Bug Reports and Acknowledgements

Many present and past members of the Bahlo lab have contributed to the development of this software by suggesting improvements, writing code and documentations and most of all using it. Special thanks go to: Catherine Bromhead, Katherine Smith, Tom Scerri, Vesna Lukic and Rick Tankard.

For bug reports or other problems please email bahlo@wehi.edu.au.

9. References

[1] Bahlo M, Bromhead CJ. LINKDATAGEN - Generating linkage mapping files from Affymetrix SNP chip data. Bioinformatics 2009;25:1961-2.

[2] Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M (2011). Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biology 12:R85.

[3] Browning BL and Browning SR (2007) Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. Genet Epidemiol 31:365-375.

[4] Browning SR (2006) Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78:903-13.

[5] Leutenegger A, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darfoux F, Thompson E. Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics 2003 Sep; 73(3)516-523.Abecasis, G.R., Cherny,

[6] S.S., Cookson, W.O. and Cardon, L.R. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees, Nat Genet, 30, 97-101.

[7] Abecasis, G.R. and Wigginton, J.E. (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers, Am J Hum Genet, 77, 754-767.

[8] Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. and Kong, A. (2000) Allegro, a new computer program for multipoint linkage analysis, Nat Genet, 25, 12-13.

[9] Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G. and Ingolfsdottir, A. (2005) Allegro version 2, Nat Genet, 37, 1015-1016.

[10] McPeek, M.S. (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives, Genet Epidemiol, 16, 225-249.

[11] McPeek, M.S. and Sun, L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data, Am J Hum Genet, 66, 1076-1094.

[12] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. and Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, Am J Hum Genet, 81, 559-575.

[13] Thompson, E. (1995) Monte Carlo in Genetic Analysis. University of Washington.

[14] Thompson, E. (2000) Statistical Inference from Genetic Data on Pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics.

[15] Albrechtsen, A., Korneliussen, T.S., Moltke, I., Hansen., T.v.O., Nielsen, F.C. and Nielsen, R. (2009) Relatedness Mapping and Tracts of Relatedness for Genome-Wide Data in the Presence of Linkage Disequilibrium. Genetic Epidemiology 33:266–274

[16] Moltke, I., Albrechtsen, A., Hansen, T.v.O., et al. (2011) A method for detecting IBD regions simultaneously in multiple individuals -- with applications to disease genetics. Genome Res. April 2011

[17] Scheet, P and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet, 78(4):629–644.

[18] Thiele, H. and Nürnberg, P (2004) HaploPainter: a tool for drawing pedigrees with complex haplotypes Bioinformatics. 2005 Apr 15;21(8):1730-2. Epub 2004 Sep 17.