LDMAP manual

Reuben J. Pengelly and Andrew Collins
University of Southampton

Contact: R.J.Pengelly@soton.ac.uk or A.R.Collins@soton.ac.uk

Contents

1 Overview

LDMAP is intended to be used for the generation of linkage disequilibrium (LD) maps from genotype data. For a description of the scientific basis of LDMAP, see Kuo *et al.*, 2007 [1]. In brief, LDMAP generates a cumulative map of LD distances between markers, based upon the Malécot model of separation by distance:

$$\rho = (1 - L)\, M e^{-\epsilon d} + L \tag{1}$$

where ρ is the empirically observed correlation between two markers in a population, L is the component of ρ not due to LD, but due to confounding factors such as recent founder effects, M is the anticipated linkage between the two markers at 0 distance, ϵ is the rate of decline in the association between the markers and d is the physical distance between the markers [2].

The product generated utilising the Malécot model are maps in cumulative linkage disequilibrium units (LDU), which are broadly analogous to a population form of centimorgans; these LDU $= \epsilon d$. It should be noted that LDMAP is reference agnostic, not directly referring to a reference assembly to run; this provides flexibility to apply LDMAP to any species and using non-standard genome assemblies.

2 Implementation

The software is predominantly implemented in C, with accessory shell and Perl scripts. It is intended for use in Linux environment, though alternative platforms may work, the software has not been designed or tested on these. LDMAP should be compiled on your system prior to use. A makefile is provided to facilitate this, allowing compilation using just the `make` command; remove all existing `*.o` files prior to compilation to ensure a fresh build.

Hardware requirements are strongly dependant upon the scale of data which is to be processed. As an indicator, processing of a 12,000 marker region will require approximately 4 GiB of memory and 5 hours of CPU-time. If you have insufficient resources, files can be broken down into smaller regions and run separately. Where pan-genome LD maps are desired, particularly using high-density genotyping data, the use of a parallelised computing resource is strongly recommended.

3 Input data preparation

As always, the ability to generate quality results is strongly dependant upon the quality of the data on which analyses are run. Input data is genotypes for multiple unrelated samples from a single homogenous population. Some key considerations are discussed below.

3.1 Sample selection

3.1.1 Sample size

The required sample size is strongly dependant upon the population (and species) to be analysed. For example, populations with a recent genetic bottleneck, and thus reduced haplotypic diversity will require far fewer samples for quality LD map generation than older populations. 50 individuals will generally be sufficient to generate informative maps, though more samples is always preferable. LDMAP is only currently suitable for use in diploid species.

3.1.2 Sample homogeneity

Outliers from a population are likely to skew the resulting maps. As such, we recommend that multidimensional scaling (e.g. as implemented in PLINK [3]), or similar, be performed in order to identify and exclude outliers. Additionally, closely related samples must be excluded.

3.1.3 Sample genders

Unlike in family based linkage maps, gender is not of significance in the generation and interpretation of LD maps of autosomes due to the population based nature of the maps. However, heterogametic individuals (i.e. XY males in human) should be excluded from analyses of the sex chromosomes.

3.1.4 Haplotype phasing

Phasing of haplotypes is not required for use of data in LDMAP.

3.2 Reference genome assembly

The quality of your genome assembly can have significant impact upon the quality of your final maps. Incorrect ordering of contigs and other erroneous regions may lead

to artefacts. Known low quality regions should be masked, or at least interpreted with due care. It is of note that artefacts arising from incorrect assembly orders have been shown to useful in the determination of the correct assembly order out of multiple possibilities [4].

3.3 Marker selection

3.3.1 Marker QC

Quality control (QC) should be performed on marker (e.g. SNP) genotypes prior to LD map generation. Low-quality genotype calls should be excluded at a minimum. Hardy-Weinberg based marker QC should also be performed.

3.3.2 Allele frequencies

Rare variation is generally non-informative in terms of LD patterns, as such, only common variants should be used as input. Where possible, an alternate-allele (AF) frequency cut-off which results in at least two minor-alleles in the cohort should be used. Note that AF cut-offs should provide both a floor and a ceiling value, as a marker with an AF of 0.005 or 0.995 are equally uninformative for our purposes.

3.4 Genotype file format

The input genotype file format for LDMAP is the numeric `.tped` format as used by PLINK (described at `http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#tr`) [3] for a single chromosome, or segment of.. An example file of three individuals for five diploid loci is shown below:

```
1 snp1 0 5000653 1 1 2 2 1 2
1 snp2 0 5000837 2 2 0 0 1 2
1 snp3 0 5000975 1 1 1 1 1 2
1 snp4 0 5001149 1 1 1 2 1 1
1 snp5 0 5001576 2 2 2 2 1 2
```

In the `.tped` format, the space delimited columns should contain:

1. Chromosome (non essential)
2. Marker name (non essential)
3. Genetic position (e.g. cM, non essential)
4. Physical position (in bp)
5. onwards - genotypes at this loci across population, two digits for each diploid individual
    - 0 - missing genotypes

- 1 - reference genotypes
- 2 - alternate genotypes

To generate the required `.tped` files using PLINK, the command:
`./plink --file [source] --recode12 --transpose --out [name]`
should be used. **N.B.** `.tped` file names should be ≤ 15 bytes in length; excessively long file names will result in errors in the first stage of file processing using LDMAP.

4 Generation of LD maps

4.1 Running LDMAP

Interaction with the core LDMAP binary, `ldmapper1` is most user friendly through the shell wrapper script `ldmap`. Running `ldmap` brings up the following menu:

```
 L D M A P
Construction of linkage disequilibrium maps from diplotype data...
 VERSION 2.0 November 2014


>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
Enter number to select function
    1             Construct LD map in segments from intermediate
       file
    2             quit
```

The wrapper script `ldmap` works by requesting parameters from the user, and writing a small executable job files (`.cp`), which then runs `ldmapper1` with these desired parameters when executed. If automation of processes is desired, the required commands may be entered into a `.cp` file directly. An example `.cp` file may look like:
`./ldmapper1 input.tped int job example.map example.log 0.05 0.001`
where `int` is a temporary file used in LD map generation, `job` is a file containing run parameters (eponymous file included with source code), `example.map` will contain the final LD map, and `example.log` will contain detailed run information, `0.05` is the desired AF cutoff, and `0.001` is the desired HWE cutoff.

4.2 Output files

Generation of the final LD map is a computationally intensive process, iteratively fitting values for the unknown variables L, M and ϵ to the Malécot model. Utilisation of parallel computing for this stage will greatly increase throughput for LD map generation. The final `.map` file format is shown below:

```
#Reading the intermediate data file: example.int
#Writing the map file: example.map lnlk=  53850.17171
# N(number of pairs)= 1194501  m(number of SNPs)= 11938 df=
   1194499.0 V(error variance)= 0.04508
#
#           Locus     kb map       LDU map
#
     1  22:34135432  34135.43200     0.000000
     2  22:34135472  34135.47200     0.147823
     3  22:34135508  34135.50800     0.181625
     4  22:34135756  34135.75600     0.181625
     5  22:34135929  34135.92900     0.181625
```

For convenience, the Perl script `ldmap_to_csv.pl` is provided to convert the `.map` file to CSV format for downstream analyses if desired.

References

[1] T.-Y. Kuo, W. Lau, and A. R. Collins. "LDMAP: the construction of high-resolution linkage disequilibrium maps of the human genome". In: *Linkage Disequilibrium and Association Mapping*. Ed. by A. R. Collins. Vol. 376. Methods in Molecular Biology. Humana Press, 2007, pp. 47–57. DOI: 10.1007/978-1-59745-389-9_4.

[2] W. Tapper, A. Collins, J. Gibson, N. Maniatis, S. Ennis, and N. E. Morton. "A map of the human genome in linkage disequilibrium units". In: *Proc Natl Acad Sci U S A* 102.33 (2005), pp. 11835–9. DOI: 10.1073/pnas.0505262102.

[3] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *Am J Hum Genet* 81.3 (2007), pp. 559–75. DOI: 10.1086/519795.

[4] S. Ennis, A. Collins, W. Tapper, A. Murray, J. MacPherson, and N. Morton. "Allelic association discriminates draft orders". In: *Ann Hum Genet* 65.5 (September 2001), pp. 503–504.