bio-101

Likelihood Estimation of *Arabidopsis* Population Parameters with MAPGD

Yue Hao[1, $, *], Matthew S. Ackerman[2], and Michael Lynch[1]

[1]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ, USA
[2]Department of Biology, and National Center for Genome Analysis Support, Indiana University, Bloomington, IN, USA
[$]Current address: Translational Genomics Research Institute (TGen), Phoenix, AZ, USA
*For correspondence: yhao@tgen.org

Abstract

MAPGD (https://github.com/LynchLab/MAPGD) is a likelihood-based software for population genomics data analysis. Here, we demonstrate a simple tutorial for using MAPGD to analyze *Arabidopsis* population data. We estimated the allele and genotype frequencies, the average heterozygosity, relatedness, and other population parameters. This pipeline could be applied to further plant population genomic studies.

Keywords: MAPGD, Population genomics, *Arabidopsis*, SNP, Allele frequency, Heterozygosity

Background

MAPGD (Ackerman, 2016; Ackerman et al., 2022) is a series of related programs that estimate allele frequency, heterozygosity, Hardy-Weinberg disequilibrium, linkage disequilibrium, and identity-by-descent coefficients from population genomic data using a statistically rigorous maximum likelihood approach. The MAPGD software was primarily designed for analysis of planktonic crustacean *Daphnia* populations (Lynch et al., 2017; Maruki et al., 2019; Ye et al., 2019). It is most useful for the analysis of low coverage population genomic data or pooled data, where many individuals are used to prepare a single sample (Lynch et al., 2014). Since its release, MAPGD has been widely used in population genomic studies of various organisms, including the nematode *Caenorhabditis elegans* (Adams et al., 2022), stickleback fish (Jeffries et al., 2022), and human gut microbiota (Shoemaker, 2022). A benchmark study showed that MAPGD could outperform other single nucleotide polymorphism (SNP) callers and has great potential in different applications (Guirao-Rico and González, 2021). Here, we demonstrate the usage of MAPGD in the analysis of a publicly available *Arabidopsis* population dataset (The 1001 Genomes Consortium, 2016). We show that MAPGD could be used for plant population genomic studies.

Software and datasets

Software

1. SAMtools (Li et al., 2009; Li, 2011; Danecek et al., 2021; version 1.9; http://www.htslib.org/)
2. MAPGD (Lynch et al., 2014; Maruki and Lynch, 2014 and 2015; Ackerman et al., 2017 and 2022; version 0.4.40; https://github.com/LynchLab/MAPGD). The complete user manual is available at https://github.com/LynchLab/MAPGD/tree/master/docs/man. Users may also find the software performance testing results at https://github.com/LynchLab/genomics_simulation
3. R (R Core Team, 2022; 4.1.3, https://www.r-project.org/) for visualization

Input data

The sample bam files were downloaded from the *Arabidopsis* 1001 Genomes Project (The 1001 Genomes Consortium, 2016). All five bam files used were from the "JGIHeazlewood2011" subset. The subset project page is: https://1001genomes.org/projects/JGIHeazlewood2011/index.html.

Procedure

The analyses showcased here were performed on the Linux system.

A. Prerequisites

MAPGD depends on the standard GNU Scientific Library (GSL); if GSL is already installed on the system, you should load the GSL module prior to the MAPGD installation. Commonly, pre-installed available modules on high-performance computing clusters can be checked using "module avail." For this demonstration, we are using GSL version 2.6.

```
module load GSL/2.6-GCC-9.3.0
```

SAMtools will be used for bam file processing:

```
module load samtools/1.9
```

B. Installation

Using MAPGD version 0.4.40 for this case study:

```
wget https://github.com/LynchLab/MAPGD/archive/master.zip
unzip master.zip
cd MAPGD-master/
# To install software in your home directory
./configure --prefix=/home/directory
```

The configuration step will create “Makefiles” that are later used for software compilation. Once all the configuration is done, the package can then be compiled using the standard commands:

```
make
make install
```

Check to see if the software has been successfully installed in the home directory:

```
cd
mapgd –help
```

Check usage of a certain command; in this case, the allele command:

```
mapgd allele -h
```

Copy useful scripts to home directory bin:

```
cd /where/MAPGD/was/downloaded/MAPGD-master/extras
cp sub_sample.py /home/directory/bin
```

C. Test dataset download

The example bam files were downloaded from the *Arabidopsis* 1001 Genomes Project subset JGIHeazlewood2011 (The 1001 Genomes Consortium, 2016). Five *Arabidopsis thaliana* accessions were used, namely Ri-0, Jea, Sakata, Oy-0, and Alc-0. The reads were mapped to the TAIR10 reference genome (Berardini et al., 2015). The previously mapped results were used for subsequent analyses.
Data download command example:

```
wget https://1001genomes.org/data/JGI/JGIHeazlewood2011/releases/ \
current/TAIR10/strains/Alc-0/Alc-0.bam --no-check-certificate
```

D. Population analysis using MAPGD

1. Sort and filter bam files using SAMtools (http://www.htslib.org/doc/samtools.html). Unmapped reads, PCR duplicates, reads failed QC, and supplementary alignments that are not primary alignments were removed (SAM flag 3844). Create a .mpileup file using SAMtools.

```
samtools sort -o <file>.sort.bam <file>.bam
samtools view -q 20 -f 3 -F 3844 -b <file>.sort.bam \
> <file>.filtered.sort.bam
samtools index <file>.filtered.sort.bam
# Create header file
```

```
samtools view -H <file>.filtered.sort.bam > Arabidopsis.header
# Generate mpileup file
samtools mpileup -q 25 -Q 25 -B <file1>.filtered.sort.bam \
 <file2>.filtered.sort.bam <file3>.filtered.sort.bam \
<file4>.filtered.sort.bam          <file5>.filtered.sort.bam          >
Arabidopsis.mpileup
```

2. Make a .pro file of nucleotide-read quartets (counts of A, C, G, and T) from the .mpileup file.

```
mapgd proview -i Arabidopsis.mpileup -H Arabidopsis.header \
> Arabidopsis.pro
```

3. Run the allele command to estimate allele and genotype frequencies from the .pro file.

```
mapgd allele -i Arabidopsis.pro -o Arabidopsis -p Arabidopsis.clean
```

The output will be `Arabidopsis.map`, which contains a list of estimated genotypic frequency. Test statistics for polymorphism and Hardy-Weinberg disequilibrium, as well as sequencing error rate and population depth of coverage, are also stored in the .map file.

4. Extract the coverage information from `Arabidopsis.map` and filter the data based on the population coverage distribution using the filter command, with log-likelihood ratio of polymorphism greater than 20 (-p 20), minor allele frequency between 0.05 and 0.45 (-q 0.05 -Q 0.45), and the population coverage between 10 and 250 (-c 10 -C 250).

```
awk '{print $6}' Arabidopsis.map > Arabidopsis_coverage.txt
```

The minimum and maximum coverage cutoffs are determined by the coverage distribution (Figure 1).

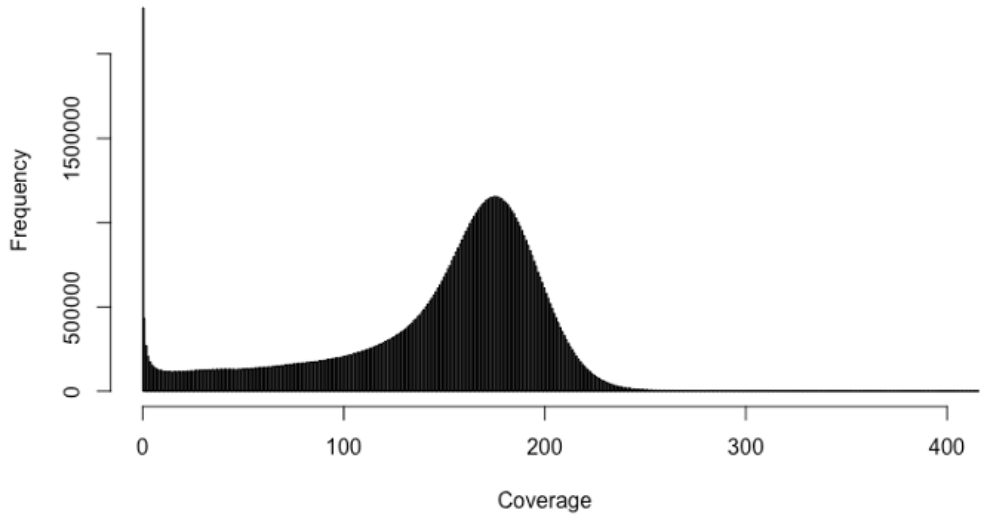

Figure 1. Histogram of the population coverage for all SNPs

```
mapgd filter -i Arabidopsis.map -p 20 -q 0.05 -Q 0.45 -c 10 -C 250 \
-o filtered_Arabidopsis
# -p: minimum value of the likelihood-ratio statistic for polymorphism
# -q: minimum minor-allele frequency estimate
# -Q: maximum minor-allele frequency estimate
# -c: minimum population coverage
```

```
# -C: maximum population coverage
```

5. Run the genotype command to generate a file of genotype likelihoods.

```
mapgd genotype -p Arabidopsis.clean.pro -m filtered_Arabidopsis.map \
 > Arabidopsis.genotype
```

6. For large datasets, the relatedness calculation will take longer. Users could subsample a group of SNPs using sub_sample.py, which will randomly sample N number of SNPs, with N being a user-defined number.

```
sub_sample.py F_Arabidopsis.genotype -N 200000 \
 > 200K_F_Arabidopsis.genotype
```

7. Run the relatedness command.

```
mapgd relatedness -i Arabidopsis.genotype -o Arabidopsis.rel
```

Data analysis

Result interpretation

The .map files contain a list of estimated genotypic frequency obtained with the allele command. These files store test statistics for polymorphism and Hardy-Weinberg disequilibrium, as well as useful statistics filtering variants, such as sequencing error rate and population depth of coverage (Figure 2). From the .map file, to extract and examine heterozygosity and allele frequency:

```
awk '{print $1, $2, $7, $16}' filtered_Arabidopsis.map > Arabidopsis_stats.txt
```

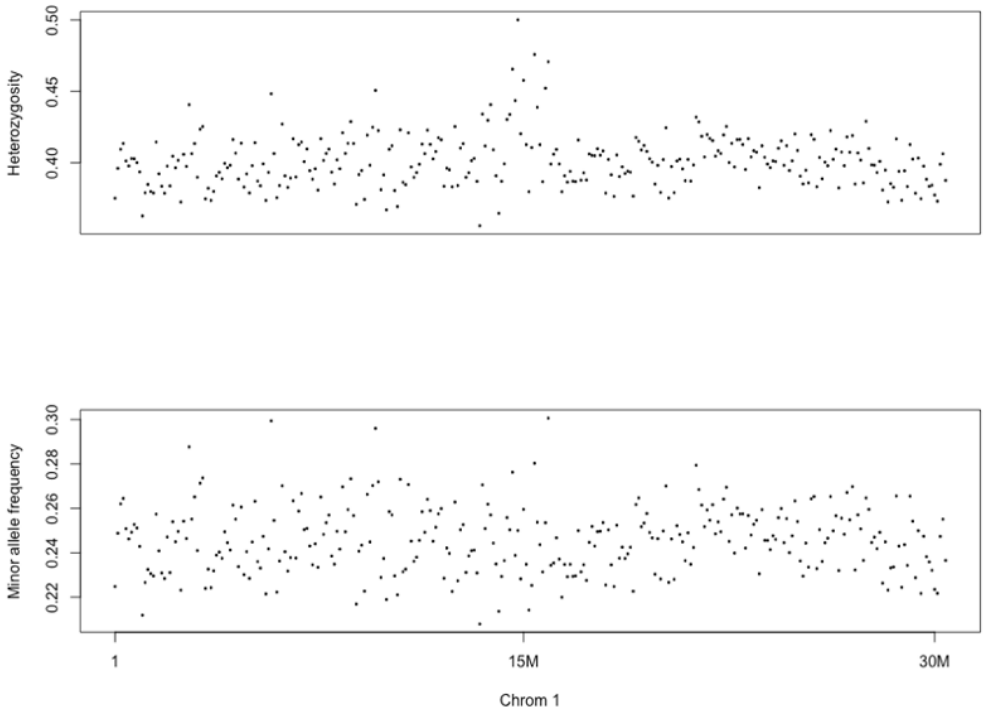

Figure 2. Average heterozygosity across every 100 kb window on chromosome 1

The output of the relatedness command contains the seven genotypic correlation coefficients for all pairs of individuals and some log likelihood ratio test statistics, as described in Ackerman et al. (2017). However, the current version of MAPGD might incur errors when running the relatedness command. Stable results are pending until the next software release on https://github.com/LynchLab/MAPGD. Users may also find example output figures and software performance testing figures at https://github.com/LynchLab/genomics_simulation/tree/master/figures.

Acknowledgments

M.L. is supported by National Science Foundation Division of Environmental Biology grants 1257806, 1927159, and Division of Biological Infrastructure grant 2119963. This protocol was partly adapted from the MAPGD user manual available at https://github.com/LynchLab/MAPGD.

Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Ackerman, M. S. (2016). Developing quantitative and population genomic techniques to test evolutionary theories. Indiana University.

Ackerman, M. S., Johri, P., Spitze, K., Xu, S., Doak, T. G., Young, K. and Lynch, M. (2017). Estimating Seven Coefficients of Pairwise Relatedness Using Population-Genomic Data. *Genetics* 206(1): 105–118.

Ackerman, M. S., Maruki, T. and Lynch, M. (2022). MAPGD: A program for the maximum likelihood analysis of population data (0.4.40).

Adams, P. E., Crist, A. B., Young, E. M., Willis, J. H., Phillips, P. C. and Fierst, J. L. (2022). Slow Recovery from Inbreeding Depression Generated by the Complex Genetic Architecture of Segregating Deleterious Mutations. *Mol Biol Evol* 39(1): e1093/molbev/msab330.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015). The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53(8): 474–485.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10(2): e1093/gigascience/giab008.

Guirao-Rico, S. and González, J. (2021). Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. *Mol. Ecol. Resour.* 21(4): 1216–1229.

Jeffries, D. L., Mee, J. and Peichel, C. L. (2022). Convergent recruitment of *Amh* as the sex determination gene in two lineages of stickleback fish. *bioRxiv*: e477894.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21): 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078–2079.

Lynch, M., Bost, D., Wilson, S., Maruki, T. and Harrison, S. (2014). Population-genetic inference from pooled-sequencing data. *Genome. Biol. Evol.* 6(5): 1210–1218.

Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T. and Jia, Z. (2017). Population Genomics of *Daphnia pulex*. *Genetics* 206(1): 315–332.

Maruki, T. and Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics* 197(4): 1303–1313.

Maruki, T. and Lynch, M. (2015). Genotype-Frequency Estimation from High-Throughput Sequencing Data. *Genetics* 201(2): 473–486.

Maruki, T., Ye, Z. and Lynch, M. (2019). Genomic analyses of population structure reveal metabolism as a primary driver of local adaptation in *Daphnia pulex*. *bioRxiv*: 807123.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Shoemaker, W. R. (2022). A macroecological perspective on genetic diversity in the human gut microbiome. *bioRxiv*: e487434

The 1001 Genomes Consortium. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis* thaliana. *Cell* 166(2): 481–491.

Ye, Z., Molinier, C., Zhao, C., Haag, C. R. and Lynch, M. (2019). *Genetic control of male production in Daphnia pulex*. *Proc. Natl. Acad. Sci. U. S. A.* 116(31): 15602–15609.

Supplementary information

1. Code availability: all code has been deposited into GitHub: https://github.com/Bio-protocol/MAPGD-for-plants
2. Data availability: intermediate and output files were uploaded to Figsuare: https://figshare.com/s/345244bc98fa0e8a6ab1