MAPGD: Maximum-likelihood analysis of population genomic data

Matthew S Ackerman

Takahiro Maruki

Michael Lynch

Abstract

[100 words] Here we announce the initial release of MAPGD—a suite of software tools implementing maximum likelihood methods for estimating heterozygosity, linkage disequilibrium, inbreeding coefficients, genotypic correlation coefficients, and allele and genotypic frequencies. MAPGD estimates parameters more accurately than standard tools and implements novel methods of quality control that improve the accuracy of quantitative genomic analyses. MAPGD's high statistical performance is computationally costly, and non-maximum likelihood methods analyze data more quickly. However, efficient utilization of high performance computer environments offsets some of these computational cost, and allow MAPGD to outperform other statistically sophisticated programs such as ANGSD.

Source code for compiling MAPGD on Linux, OS X or Windows is available for download at https://github.com/LynchLab/MAPGD.

Both 32bit and 64bit windows binaries are available. Developer documentation can be obtained from https://lynchlab.github.io/MAPGD/.

Introduction

[750 words]

The problem.

Inferring parameters based on population variation, such allele frequency spectra, inbreeding coefficients, etc. from the reads generated by sequencing machines is complicated by the uncertainty of whether the variation in the reads represents true variation or errors introduced by the sequencing and alignment process. For example, if reads come from a single diploid individual, a parameter of interest might be heterozygosity— the fraction of sites where the two parental chromosomes differ. When heterogeneity exists in the reads, this may represent errors in the sequence, *or* it may represent a difference between the parental chromosomes.

Conversely, when there is *no* heterogeneity among reads, this may represent homogeneity of the parental chromosomes or a chance failure to sample both parental chromosomes.

There is one final potential source of errors. The reads which have been aligned to some particular region in some particular sample may not represent sequence from that sample in that region. This can happen because of mapping errors, structural differences between the sequence in population and the reference, contamination during sample preparation, or errors that occurring during de-multiplexing. Detecting and correcting for these errors is particularly difficult to address because the chain of events leading to the error can be complex and not have simple behaviors amenable to stochastic modeling.

Error rates and quality scores.

Programs that do offer genotypic likelihood statistics do so based on models that assume ...

The solution.

These problems can be addressed by incorporating the entire process of sampling and sequencing into a stochastic model.

Both sequencing errors and sampling are incorporated into likelihood models that include the parameters to be estimated, and thus maximum likelihood estimates of these parameters can be obtained. The third and final problem of reads arising from an un-modeled and nebulous process of mis-assigns/alignment is addressed by assessing the goodness of fit of the stochastic model to the observed data. This goodness of fit assessment is a powerful means of detecting and removing artifacts.

System requirements

MAPGD is written in C++ and can be compiled by any compiler implementing the C++11 standards. Some python scripts are provided for the simulation of test data. MAPGD uses the gnu scientific libraries (GSL) for likelihood maximization of the genotypic correlations (Ackerman et al., 2016a), and open source Message Passing Interface (MPI) and Open Multi-Processing (OMP) are used for parallel computation. The output generated by the program can optional be stored in an SQL database.

The build scripts for MAPGD use autoconf, allowing MAPGD to be compiled on many UNIX-like systems, including OS X. Precompiled binaries are available for windows users.

Design philosophy

Although there are a number of computationally challenging tasks in the analysis of sequence data, typically, the computational investment of calculating genotypic likelihoods and the statistics derived from them is minimal. A user may be able to obtain genotypic likelihoods for their $10,000 worth of sequence by running programs for a few hours on typical desktop computers. Because sequencing still represents such a substantial monetary investment, MAPGD emphasizes the efficient use of this sequencing data over minimization of the computational demands of analysis. For instance, we choose to represent genotypic likelihoods as 64-bit floating point numbers, preserving a high degree of precision, rather than a more typical 8-bit phred quality scores. In order to permit timely analysis using this computationally demanding approach, MAPGD makes full use of clustered computing environments.

Quality control

The computational efficiency and statistical accuracy of each of the methods implemented by MAPGD is ? These claims can be continually evaluated in compairison to other programs by a pipeline ...

Commands

allele: estimating allele and genotype frequencies.
The method MAPGD uses for estimating allele frequencies in diploid population (Maruki and Lynch, 2015)... yields essentially unbiased allele-frequency estimates with low root mean squared error (RMSE), even from low-coverage sequencing data. Allele frequency estimates on low coverage data using this method are substantially more accurate than those obtained from GATK (McKenna et al., 2010) or vcf-tools (Danecek et al., 2011), which do not implement a maximum likelihood method, and slightly more accurate that estimates from ANGSD (Korneliussen et al., 2014), which does (Figure 1). ... Uniquely, MAPGD achives this without utilizing the ...

filter: removing aberrant sites.
Several criteria can be used to remove potential artifacts. MAPGD is unique among genome analysis software in that it reports a goodness of fit statistic for each site in the genome and for each individual in a population. Goodness of fit statistics evaluate whether or not observed data are drawn from a specified model. Typically a χ^2 or a variant of the K-S test is used to assess goodness of fit. However, we make use of novel approach which uses standard scores of the probability of observations to assess goodness of fit. The details of this method are described on the MAPGD website at ....

Although very few sites have a poor fit in simulated data (1.8% of polymorphic sites at 25× coverage, and fewer at lower coverage) removing poorly fit sites can have a profound effect on real data (Figure 2). ... Details of calculation of the goodness of fit statistic are described in the supplemental material of chapter 2. Some care should be exercised with the interpretation of goodness of fit values, as they do not follow a normal distribution for simulated data.

linkage: estimation of linkage disequilibrium.

We have developed a maximum-likelihood (ML) method for estimating LD directly from sequence data (Maruki and Lynch, 2014). This method does not require phasing haplotypes, which may be difficult with low depth of coverage.

In our implementation, we assume that all sequence reads independently cover at most one of the two polymorphic sites of interest. We examined the effect of this assumption on LD estimation using computer simulations, and found the assumption reasonable. To adjust the known biases of ML estimates of LD (Weir and G., 1980), we multiply estimates of D (Lewontin and ichi Kojima, 1960) by $N_i/(N_i - 1)$ and subtract $1/N_i$ from estimates of r^2 (Hill and Robertson, 1968), where N_i is the effective number of sampled individuals (Maruki and Lynch, 2014). The application of this method to low-coverage high-throughput sequencing data shows that it enables smooth description of LD patterns even with low depths of coverage (Lynch et al., 2016).

relatedness: estimation of genotypic correlation coefficients.

Coefficients of identity or relatedness form the basis of quantitative genetics analysis. In Ackerman et al. (2016a) we extend and refine the concept of coefficients of identity and outline a maximum likelihood method of estimating these coefficients from short read data. These estimation procedures are implemented in MAPGD and aid in the detection of duplication and contamination among samples, as well as facilitating quantitative genetic analysis. Although estimates are accurate for reasonably low coverages, caution should be exercised in interpreting results from less that 5× coverage.

pool: population-genomic analyses of pooled sequencing data.

Finally, pooled sequencing is a cost-effective sequencing strategy that enables sequencing of many individuals with a single library. Because of its low cost and wide applicability, it is becoming increasingly popular in population-genomic research (Schlotterer et al., 2014). To use the rapidly accumulating data as accurately and efficiently as possible, we have developed an maximum likelihood allele-frequency estimator and statistical frameworks for carrying out subsequent population-genomic analyses (Lynch et al., 2014). These methods have been applied to experimental evolution data, and allow for the sensitive detection of polymorphic sites while simultaneously controlling for rates of false positives (Ackerman et al., 2016b).

genotype: estimating individual genotypes.

Parameter estimation can be improved by the use of genotypic likelihoods, rather than the calls of the most likely genotype (as shown in figure 1). Genotypic likelihoods can be reported using either uniform priors (which are appropriate for subsequent maximum likelihood calculations) or priors based off of the maximum likelihood estimates of genotypic frequencies.

VCFtools, GATK, ANGSD and MAPGD all use a similar models for calculating genotypic likelihoods. However, these programs differ in how error rates are assessed. VCFtools assumes that error rates reported by the Illumina machine are correct. One problem with this assumption is that the quality scores are 8-bit phred scaled values, so they are very coarse grained. Additionally, while the quality scores may represent accurate estimates of error rates under some conditions, actual error rates vary depending on properties of the DNA being sequenced, the quality of the library, and other conditions. GATK addressed this problem by re-calibrating the quality scores; this addresses some of the systematic ways that quality scores mis-estimate true error rates, however, these re-calibrated scores are still represented as 8-bit phred scaled values, and so are poor estimates of actual error rates. ANGSD implements both of these methods. MAPGD departs from other programs in that it estimates the error rate at each site separately, based of a maximum likelihood model, and represents these error rates as 64-bit floating point values.

Documentation

Input and output.

Currently, MAPGD begins all analysis from the mpileup files generated by SAMtools. These are converted into a 'pro', which summarize the number of reads of each nucleotide at each site, using the "mapgd proview" command, and then analysis begins with either the "mapgd allele" command or the "mapgd pool" command.

All data generated is output in a simple, tab delimited, plain text format with two header lines that explain the data presented in the file. The meaning of fields in the header lines is explained by the "mapgd help [field]" command, and documented in the README file available from https://github.com/LynchLab/MAPGD/.

read and **write**: In order to facilitate the long-term accessibility of the data generated with MAPGD, an SQL output format is available that documents the meaning of the data more extensively than the plain text format. This may prove useful if MAPGD becomes unavailable at some future time, since the widespread adaptation of SQL makes it unlikely that an SQL-formatted database will become inaccessible.

Developer documentation.

The developer documentation for MAPGD is generated directly from the source code using doxygen, in order to minimize the effort necessary to ensure that documentation is kept current with continuing development of the program.

Testing.

Statistical and basic functionality testing is performed automatically on all current methods implemented within MAPGD in order to ensure that the behavior of MAPGD remains consistent with future development. Additionally, Tavis CI is used to ensure that MAPGD can be automatically configured and installed properly on both OS X and Linux based systems. Windows binaries are kept current with the source code, but are not tested for functionality.

Availability.

The program is freely available from https://github.com/LynchLab/MAPGD under GPLv2.0. The program may be used, in part or in whole, in any software projects as long as the source code of MAPGD, in the form modified for use with the software project, is freely distributed.

Acknowledgments.

The research was made possible by ...

References

Ackerman, M. S., Johri, P., Spitze, K., Xu, S., Doak, T., , Young, K., and Lynch, M. (2016a). A general statistical model for coefficients of relatedness and its application to the analysis of population-genomic data. *Genetics*.

Ackerman, M. S., Miller, S. F., Nguyen, M., Behringer, M., Doak, T., and Lynch, M. (2016b). Mutation rate and transfer size do not influence the effective population size of experimentally evolving *Esherichia coli*. *in prep.*

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R., Lunter, G., Marth, G., Sherry, S. T., McVean, G., Durbin, R., and Group‘, . G. P. A. (2011). The variant call format and vcftools. *Bioinformatics*, 27:2156–2158.

Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, 38(6):226–231.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356.

Lewontin, R. C. and ichi Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:458–472.

Lynch, M., Ackerman, M. S., Spitze, K., Zhiqiang, Y., and Maruki, T. (2016). Population genomics of *Daphnia pulex*. *Genetics*.

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

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

Maruki, T. and Lynch, M. (2015). Genotype-frequency estimation from high-throughput sequencing data. *Genet*, 201:473–486.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20:1297–1303.

Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat Rev Genet*, 15.

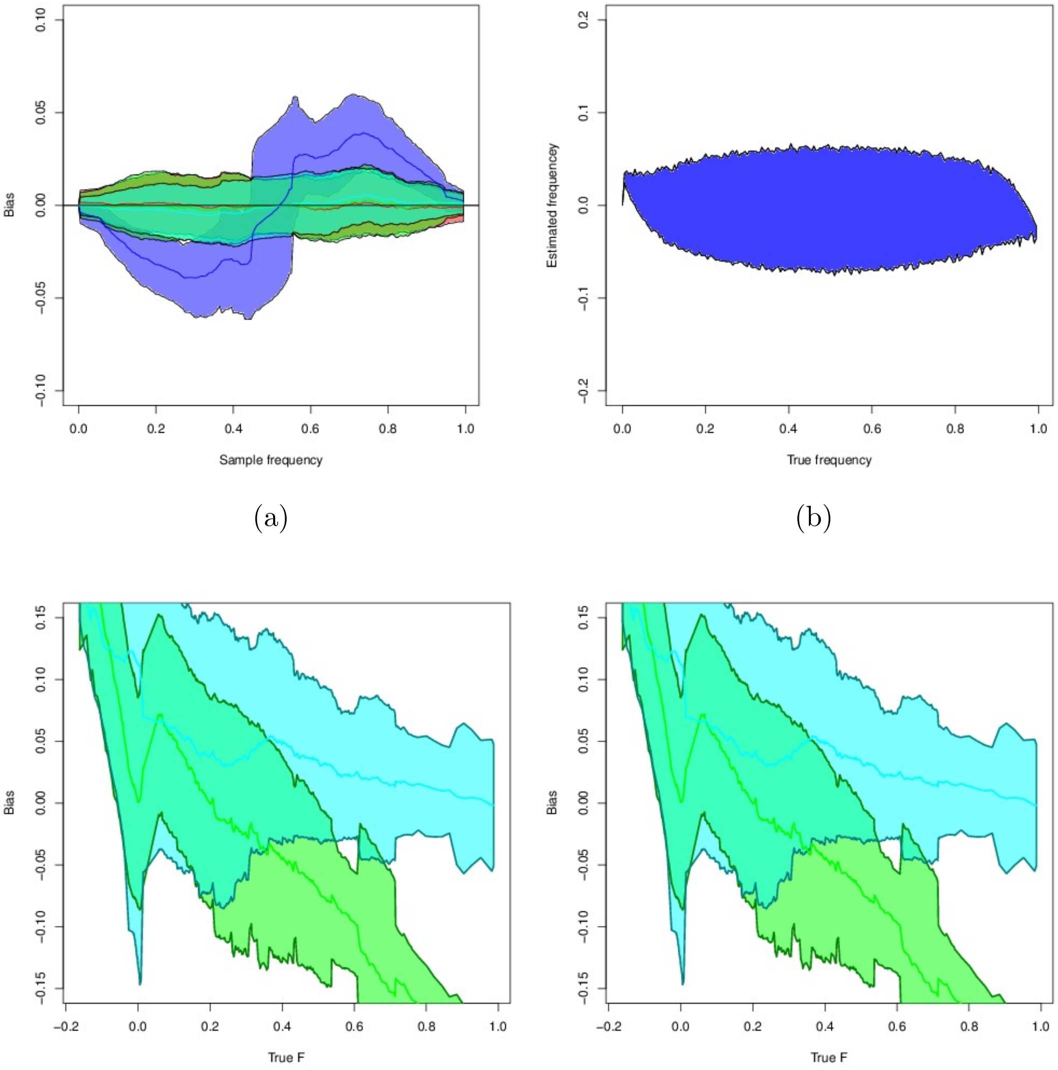
Weir, B. S. and G., H. W. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, 95:477–488.

Table 1: **Summary of commands implemented in MAPGD.**

Command	
allele	Estimates allele frequencies from labeled data
relatedness	Estimates genotypic correlation coefficients
pool	Estimates allele frequencies from pooled data
genotype	Produces the posterior probabilities of a genotype
linkage	Estimates LD between all snps
read	Reads data from an SQLite database
write	Writes data to an SQLite database
filter	Filters sites based on criteria
vcf	Writes output in a vcf format

Table 2: **Programs analyzed in this publication.** AF:Allele frequency (labeled), PL: Allele frequency (pooled), GC:Genotypic correlation, HWE:Departure form HWE(F), LD:Linkage disequilibrium

Program	Version	
MAPGD	0.5.0	AF, GC, LD, PL, HWE
GATK	3.6	AF, GC, HWE
BCFtools	0.1.19	AF, GC
ANGDSD	0.913-14	AF, GC, HWE
Plink	?	AF, GC, LD
Breseq		PL



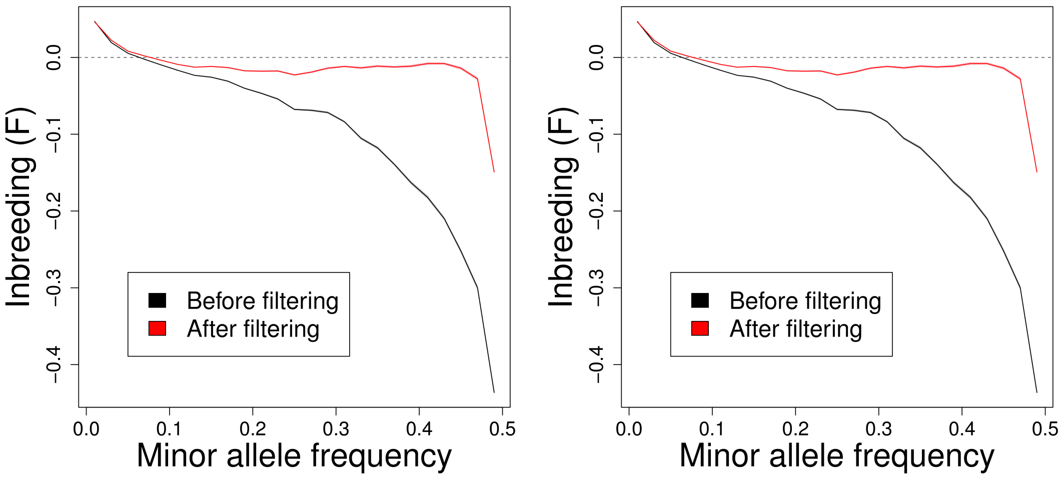
(a)

(b)

(c)

(d)

Figure 1: (a) Bias and RMSE of MAPGD, GATK, BCFtools and ANGSD in calculating allele frequencies from the sequences of 1,000 individuals with a mean coverage simulated of 2.5, and error rate was 0.01. (b) Bias and RMSE of MAPGD and BRESEQ on pooled sequences with a similar total depth of as a. (c) Bias and RMSE of MAPGD, GATK, BCFtools and ANGSD in calculating inbreeding (F).(d) Bais and RMSE of LD estimation in data simulated as in a) and b).



(a)

(b)

Figure 2: Effect of goodness of fit filtering on inbreeding calculation in a population of *Daphnia pulex*. (a) The effect of filtering on inbreeding coefficient estimation in simulated data. MAPGD is filtered on the basis of goodness of fit (GOF), GATK is filtered on ?, and ...Data from Lynch et al. (2016) Default values used for filtering (goodness of fit> -2 and individuals cut< 4). Goodness of fit and inbreeding are uncorrelated in simulated data($\rho = 0.004, p = 0.56, df = 16,000$).