`EGGS`(v1.0): a software for the efficient numerical computation of genetic summary statistics under generalized demographic models

Feng Gao and Alon Keinan

July 7, 2015

Contents

1 Background

This manual describes the options and use of the `EGGS` software (Efficient computation of Generalized Growth demographic models' summary Statistics, v1.0). A demographic model considered here is a function $N(T)$ of diploid population size N against time T in generations, which describes the population size changes from present to the past (thus time increases backward in time). More specifically, we consider the model is made up of several epochs in sequence, where each epoch in the model describes the population size change during some time period from starting time T_i to ending time T_f. As a result, the first epoch always has $T_\text{i} = 0$ and the last epoch always has $T_\text{f} = \infty$, and T_f of the previous epoch is the same as T_i of the immediate next epoch. We break up the demographic model to a minimum number of epochs such that each epoch can be described by the following differential equation [1,2]:

$$\frac{dN}{dT} = -rN^b, \tag{1}$$

where N is the diploid population size, T is the time in units of generation and b is the parameter that controls the speed of the growth: if $b > 1$, the growth is faster than exponential; if $b = 1$, the growth is of exponential; if $b < 1$, the growth is slower than exponential. If $r = 0$, then the epoch is of constant population size. This kind of piece-wise defined models is similar to that considered in [3]. Notice that a demographic model that *only* consists of constant population size epochs and exponential epochs is a subset of the models considered here. As a result, they can be easily specified in this software (see Section 2 for relevant options).

This software is aimed at numerically generating five kinds of population genetic summary statistics given a specified generalized growth model. The five summary statistics considered are: (1) the total number of segregating sites (S), (2) the time to the most recent common ancestor (T_MRCA), (3) the site frequency spectrum (SFS, both relative version and absolute version), (4) the average pairwise difference between chromosomes per site (π) [4] and (5) the burden of private mutation (α) [5]. All generations of these summary statistics are based on Kingman's coalescent [6]. We give a short description of each summary statistic below:

1. Total Number of Segregating Sites (S)
   The count of sites in the sampled sequences that are polymorphic (sequences have different genotypes), excluding the sites where all sequences have the same genotype.

2. Time to the Most Recent Common Ancestor (T_MRCA)
   The time (measured in generations) when all of the sequences sampled at present coalesce to the same ancestor.

3. Site Frequency Spectrum (SFS)
   Suppose we have n sequences sampled at present. The *absolute unfolded* SFS $\boldsymbol{\xi}^\text{A}$ has $(n-1)$ entries, and entry ξ_i^A records the number of segregating sites where i sequences bear the *derived* allele at this site and $(n-i)$ sequences have the *ancestral* allele. The *absolute folded* SFS $\boldsymbol{\eta}^\text{A}$ is used when we don't have information about the derived allele and ancestral allele. It has $\left[\frac{n}{2}\right]$ entries, with entry η_i^A being the number of segregating sites where i sequences bear one type of allele at this site and $(n-i)$ sequences have the other allele. The following relationship holds: $\eta_i^\text{A} = \frac{\xi_i^\text{A}+\xi_{n-i}^\text{A}}{1+\delta(i,n-i)}$. The *relative unfolded* SFS $\boldsymbol{\xi}^\text{R}$ is $\boldsymbol{\xi}^\text{A}$ after normalization: $\boldsymbol{\xi}^\text{R} = \frac{\boldsymbol{\xi}^\text{A}}{\sum_{i=1}^{n-1}\xi_i^\text{A}}$. And similarly the *relative folded* SFS $\boldsymbol{\eta}^\text{R} = \frac{\boldsymbol{\eta}^\text{A}}{\sum_{i=1}^{\left[\frac{n}{2}\right]}\eta_i^\text{A}}$.
   Sometimes we also consider *binning* the SFS. Suppose an SFS contains l entries. By binning

from the m^{th} entry, it means that we only keep the first $(m-1)$ terms of the SFS and calculate the accumulated sum of the entries from m to l. As a result, the binned SFS contains m entries. This technique is often used in demographic inference works to reduce the noise and unreliability contained in the later parts of the SFS.

4. Pairwise Difference Between Chromosomes Per Site (π)
   This quantity is calculated by comparing every two different sequences, counting the number of differences between them, calculating the average of the total differences and normalizing by the total number of sites (L). Suppose there are n sequences, this quantity has the following relationship with the relative SFS:

$$\pi = \frac{S}{\binom{n}{2}L}\sum_{i=1}^{n-1} i(n-i)\xi_i^{\mathrm{R}} = \frac{S}{\binom{n}{2}L}\sum_{i=1}^{\left[\frac{n}{2}\right]} i(n-i)\eta_i^{\mathrm{R}}$$

5. Burden of Private Mutation (α)
   Suppose we have n diploid individuals sequenced (thus there are $2n$ sequences). This quantity stands for the proportion of heterozygous positions in a newly sequenced $(n+1)^{\text{th}}$ individual that are novel (all of the previous n individuals have the same genotype, but this newly sequenced individual have a different genotype).

Similar to the work in [7,8], the summary statistics are computed by evaluating explicit expressions and the output are *expectations*, which is unlike simulation approaches such as `ms` [9] or `FTEC` [1]. For more details, please refer to:

Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Submitted.*

2 Downloading EGGS

The software can be freely downloaded from the Keinan lab website:

`http://keinanlab.cb.bscb.cornell.edu/content/tools-data`

Please download the corresponding version for MAC OS and LINUX.

3 Options

In this section, we describe all the options that can be used in the `EGGS` software. Note that a common option is `-silent`, which specifies that only the results will be output, without a header that lists the genetic summary statistics in the order of the output (see the next section for the detail of the header).

3.1 Demographic Model Options

We first list out the options for specifying the demographic model under consideration. The demographic model is essentially a combination of the following options. The input population size is always in *diploids* and the input time is always the *duration* (not starting or ending time) of the epoch in *generations*. Notice that because time increases backward in time, the starting time of an epoch is *closer* to present and the ending time is more *ancient*. For example, if an epoch is between 400 generations ago and 800 generations ago with population size $N(400) = 1,000$ and $N(800) = 2,000$, then the starting time of the epoch is 400 and the starting population size is 1,000. Similarly the ending time of the epoch is 800 and the ending population size is 2,000. Also notice that the demographic history is constructed by connecting all the epochs specified by these `-g`, `-e` and `-c` options in turn, which means that these pieces must be input in order (from present to the past).

`-n` h The sample size (number of chromosomes or haploid individuals) is h. This option is required.

`-g` N_i N_f b τ Specifying an epoch with a generalized growth model, where N_i is the starting population size of this epoch *in diploids*, N_f is the ending population size of this epoch, b is the growth speed parameter in Equation (1) and τ is the *duration* of the epoch. Notice that unlike simulation softwares such as `ms` or `FTEC`, the parameter r in Equation (1), which is typically hard to calculate, doesn't need to be input. This is because a single epoch can be uniquely determined by N_i, N_f, b and τ. The parameter r is a "middle product" and is fully handled by the software.

`-e` N_i N_f τ Specifying an epoch with an exponential growth model, where N_i is the starting population size of this epoch, N_f is the ending population size of this epoch and τ is the *duration* of the epoch. This corresponds to $b = 1$ in Equation (1), which means that this option is equivalent to [`-g` N_i N_f `1` τ].

`-c` N τ Specifying an epoch with a constant size model, where N is the population size of this epoch and τ is the *duration* of the epoch. This corresponds to $r = 0$ in Equation 1, which means that this option is equivalent to [`-g` N N b τ] (N_i and N_f are the same, and it doesn't matter which value you choose for b, because it is not used) or [`-e` N N τ].

`-c` N `inf` Specifying an ending epoch with population size N. `inf` stands for the limit as time extends infinitely into the past. This option is required as the last part in specifying a demographic model (closes the model).

3.2 Summary Statistic Options

We then list out the options for the summary statistics. Please refer to Section 1 for explanatory information of these summary statistics. At least one of these options must be specified (one of the summary statistics must be output).

`-mu0` μ_0 Mutation rate per site per generation μ_0. This option can *only* be accompanied with `-pdiff` option (see below for the description of this options). If `-mu0` option is not specified with `-pdiff`, the default value of $\mu_0 = 1.2 \times 10^{-8}$ is used.

`-mu` μ Locus-based mutation rate per generation $\mu = \mu_0 L$, where μ_0 is the mutation rate per base pair per generation and L is the length of the locus. This option can *only* be accompanied with either `-nsites` option or `-asfs` option (see below for the description of the two options). If `-mu` option is not specified with `-nsites` or `-asfs`, the default value of $\mu = 1$ is used.

`-nsites` The expected number of total segregating sites (S).

`-tmrca` The expected time to the most recent common ancestor (T_{MRCA}), in units of generation.

`-pdiff` The expected average pairwise difference between chromosomes per site (π).

`-rsfs` The expected relative site frequency spectrum (relative SFS), with all terms summing up to 1.

`-asfs` The expected absolute site frequency spectrum (absolute SFS).

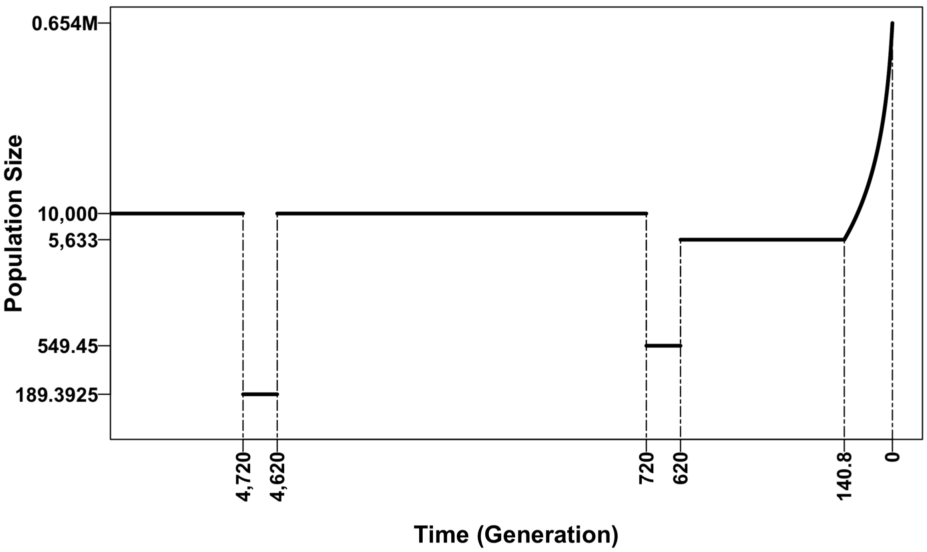
`-burden` The expected burden of private mutation (α).

`-fold` To output the folded site frequency spectrum, instead of the unfolded. This option must be accompanied by either `-rsfs` or `-asfs`.

`-bin` t To bin the terms of the site frequency spectrum starting from the t^{th} item, which means that values of the first $(t-1)$ terms will be output, together with the sum of the rest.

4 Example

In this section we give a detailed example of the use of the software. Consider the demographic history shown in the following figure.



This two-bottleneck demographic history is identical to that inferred for the European population in [10], except that we changed the growth speed of the recent growth epoch to $b = 1.5$ (faster than exponential growth), instead of the original $b = 1$ (exponential growth). Suppose we consider a sample size of 1,000 chromosomes (500 diploid individuals). Further suppose that the loci under consideration have a total length of L = 100,000 base pairs and has a mutation rate of $\mu_0 = 1.2 \times 10^{-8}$ per generation per base pair. Then the command line for this demographic model will be:

```
./EGGS -n 1000 -rsfs -asfs -tmrca -nsites -burden -pdiff -mu0 1.2e-8 -mu
1.2e-3 -fold -g 0.654e6 5633 1.5 140.8 -c 5633 479.2 -c 549.45 100 -c 10000
3900 -c 189.3925 100 -c 10000 inf -bin 11
```

The explanation of this command is as follows:

1. `-n 1000` specifies that the sample size is 1,000 chromosomes or 500 diploid individuals.
2. `-rsfs -asfs -tmrca -nsites -burden -pdiff` specifies that the required summary statistics are: relative site frequency spectrum, absolute site frequency spectrum, time to the most recent common ancestor, the total number of segregating sites, the burden of private mutation and the pairwise difference per site.
3. `-mu0 1.2e-8` specifies that the mutation rate $\mu_0 = 1.2 \times 10^{-8}$.
4. `-mu 1.2e-3` specifies that the locus-based mutation rate $\mu = 1.2 \times 10^{-3}$. This is calculated by $\mu = \mu_0 L = 1.2 \times 10^{-8} \times$ 100,000 $= 1.2 \times 10^{-3}$.
5. `-fold` specifies that we require the site frequency spectrum to be folded.
6. `-g 0.654e6 5633 1.5 140.8` specifies a generalized growth epoch, with a starting population size of 0.654×10^6 diploids, an ending population size of 5,633 diploids, a growth speed parameter $b = 1.5$ and the duration of the growth epoch is 140.8 generations.
7. `-c 5633 479.2` specifies a constant-size epoch, with population size 5,633 and a duration of 479.2 generations. The duration is calculated as $620 - 140.8 = 479.2$.
8. `-c 549.45 100` specifies the recent 100-generation bottleneck with population size 549.45. The duration is calculated as $720 - 620 = 100$.
9. `-c 10000 3900` specifies the recovery from the bottleneck with population size 10,000 and it lasts 3,900 generations. The duration is calculated as 4,620 − 720 = 3,900.
10. `-c 189.3925 100` specifies the ancient 100-generation bottleneck with population size 189.3925. The duration is calculated as 4,720 − 4,620 = 100.
11. `-c 10000 inf` specifies an ending epoch with population size 10,000 (it lasts to the indefinite past).
12. `-bin 11` specifies that only the values of the first 10 terms of the site frequency spectrum will be output, together with the sum of the rest.

The output result from this command is as follows:

```
The results are shown in the following order:  TMRCA, S, ALPHA at n = 499,
PI, RELATIVE SFS, ABSOLUTE SFS
3.2750e+04
3.0065e+02
5.6170e-03
3.5654e-04
3.3306e-01 7.2554e-02 3.6757e-02 2.4247e-02 1.8075e-02 1.4425e-02 1.2014e-02
1.0300e-02 9.0153e-03 8.0160e-03 4.6154e-01
1.0013e+02 2.1814e+01 1.1051e+01 7.2899e+00 5.4342e+00 4.3369e+00 3.6121e+00
3.0966e+00 2.7105e+00 2.4100e+00 1.3876e+02
```

The explanation of this output per-line is as follows:

1. The header line, which specifies the order of the results, with other explanatory information.
2. The expected time to the most recent common ancestor is 3.2750×10^4 generations.
3. The total number of segregating sites is 3.0065×10^2.

4. The burden of private mutation at sample size $n = 499$ diploids is 5.6170×10^{-3}.

5. The average pairwise difference between chromosomes per site is 3.5654×10^{-4}.

6. The first 10 terms of the folded relative site frequency spectrum, with the sum of the rest as 0.46154.

7. The first 10 terms of the folded absolute site frequency spectrum, with the sum of the rest as 1.3876×10^{2}.

Notice that the output values are all *expectations*, unlike a single-run simulation output for simulation softwares. And because the values are expectations, they might have fractional numbers for the total number of segregating sites and the absolute SFS.

5 Bug Reporting

If you find any bugs while using the software, please email to `keinanlab.eggs@gmail.com` with detailed information.

6 References

1. Reppell M, Boehnke M, Zöllner S. FTEC: a coalescent simulator for modeling faster than exponential growth. Bioinformatics 2012;28:1282-3.

2. Reppell M, Boehnke M, Zöllner S. The impact of accelerating faster than exponential population growth on genetic variation. Genetics 2014;196:819-28.

3. Bhaskar, A and Song, YS. Descartes' Rule of Signs and the Identifiability of Population Demographic Models from Genomic Variation Data. Ann Stat 2014;42:2469-2493.

4. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics 1983; 105:437-60.

5. Gao F, Keinan A. High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics 2014;15 Suppl 4:S3.

6. Kingman, JFC. The coalescent. Stochastic Process Appl 1982;13:235-248.

7. Bhaskar A, Clark AG, Song YS. Distortion of genealogical properties when the sample is very large. Proc Natl Acad Sci U S A 2014;111:2385-90.

8. Bhaskar A, Wang YX, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res 2015;25:268-79.

9. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 2002;18:337-8.

10. Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs RA, Sing CF, Clark AG, Keinan A. Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci U S A 2014;14;111:757-62.