**gfselector: a $\mathcal{G}/\mathcal{F}$-class selector for the site-and-branch-heterogeneous GFmix model**
**Version 1.0**
**April 1, 2026**

**Charley McCarthy**
*Department of Biochemistry and Molecular Biology, Dalhousie University*

## Introduction

The main program `gfselector` attempts to find the optimal $\mathcal{G}$ and $\mathcal{F}$ classes for fitting the GFmix profile mixture model to a dataset/tree. `gfselector` can be run at the command line as follows:

```
$ gfselector -aln alnfile -fmt format -crit criterion [-iqtree iqtreefile] [-
tree treefile] [-root rootfile] [-mix mixture] [-conv] [-out outfile] [-cpu numthreads]
[-iter num] [-group groupfile]
```

A description of all options and output is given below.

`-aln alnfile`: Multiple sequence alignment file.

`-fmt format`: Format of multiple sequence alignment file. Can be one of FASTA ("fasta"), PHYLIP-interleaved ("phylip-interleaved") or PHYLIP-sequential ("phylip" or "phylip-sequential"). Compatible with relaxed PHYLIP formats (i.e. taxon names greater than 10 characters in length). Default: "phylip".

`-crit criterion`: Criterion to optimize. Determines the type of $\mathcal{G}/\mathcal{F}$ optimization that `gfselector` attempts. Can be based on a single binomial test of two proportions ("bi"), optimization based on $\chi^2$ test statistic ("chi") or optimization based on log-likelihood under the original GFmix model ("gfmix"). Default: "bi".

`-iqtree iqtreefile`: Required for GFmix-based optimization. An IQ-Tree report file estimated under a base branch-homogeneous model. Used to assign mixture weights and $\alpha$.

`-tree treefile`: Required for GFmix-based optimization. A Newick tree estimated under a base branch-homogeneous model. Used to assign branch lengths.

`-root rootfile`: Required for GFmix-based optimization. A file with the names of taxa on one side of the root split.

`-mix mixture`: Required for GFmix-based optimization. A file containing the frequencies for each frequency class in a mixture model as rows.

`-conv`: Usually required for GFmix-based optimization. Will convert taxon names in alignment and tree files to 0th-index integers compatible with the strict PHYLIP alignment format for use with the original GFmix model. May not be required if taxon names are already compatible with strict PHYLIP format.

`-out outfile`: Optional. A file to which is output the best $\mathcal{G}/\mathcal{F}$ class from each iteration of an optimization procedure alongside information for that iteration of the procedure (or, the results

of the binomial test of two proportions if criterion = "bi"). The final line of the output is the optimized $\mathcal{G}$ and $\mathcal{F}$ classes. Default: print output of optimization procedure to screen.

`-cpu numthreads`: Optional. The number of threads to use, meaning the number of $\mathcal{G}/\mathcal{F}$ classes to be assessed simultaneously per iteration of the optimization procedure. Default: 1.

`-iter num`: Optional. The number of iterations the optimization procedure is allowed to run for. If not specified, optimization runs until criterion cannot be improved. Default: Inf.

`-group groupfile`: Optional. A file with the names of taxa in one of the two groups required for the binomial test of two proportions. All other taxa are assumed to comprise the other group. If not provided, `gfselector` defines these two groups by clustering taxa based on $\chi^2$ test residuals.

The output to screen is the status of each iteration of the optimization procedure and, if an output file is not specified, a table detailing information for each iteration of the optimization procedure.

## Usage examples

We illustrate usage with the Amborella dataset from Leebens-Mack et al. (2005). All required input files are available with this software, and also with the related `gfmixf` software.

### *Binomial test of two proportions*

The simplest method of assigning $\mathcal{G}/\mathcal{F}$ classes is to perform a binomial test of two proportions, following Baker et al. (2024). This test requires taxa to be classified into two groups. These groups can be based on *a priori* information, provided by the `-group` option detailed above. For example, Baker et al. (2024) divided taxa in halophilic and non-halophilic groups. If a groupfile is not provided, `gfselector` will cluster taxa by constructing a dendogram of taxa based on $\chi^2$ test residuals and divide taxa into two groups based on the deepest split in the dendrogram. For each amino acid, the composition bias between the two taxon groups is computed as a Z-score in the following equation:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$
$$\hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}, \hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad (1)$$

where $X_1$, $X_2$ are the total numbers of that amino acid and $n_1$, $n_2$ are the total number of all 20 amino acids across the two taxon groups respectively. This test assumes that the proportions of an amino acid across taxa is approximately normal, with the null hypothesis that $p_1 = p_2$. $|Z| > 1.96$ indicates rejection of the null hypothesis at significance level $p < 0.05$. Amino acids are divided into the classes $\mathcal{G}$ ($Z > 1.96$), $\mathcal{F}$ ($Z < -1.96$) or $\mathcal{O}$ ($|Z| \leq 1.96$) on the basis of this $Z$-score.

Assigning $\mathcal{G}/\mathcal{F}$ classes based on the binomial test can be performed as follows:

```
gfselector -aln amborella-interleave.phy -fmt phylip-interleaved -crit bi
```

which outputs the following to screen:

```
Group 1:  Acorus, Calycanthu, Typha, Yucca, Oryza, Spinacia, Nuphar, Ginkgo, Medicago, Pinus, Psilotum, Saccharum, Arabidopsi, Atropa, Amborella, Oenothera, Nymphaea, Nicotiana, Zea, Lotus, Ranunc, Triticum
Group 2:  Physcomitr, Marchantia
G O F Criterion Optimizer Iterations Iteration time Optimization time Swap times
RSGHVDP CETMAWY LFQINK 0 Binomial 0 0 0 0
```

The first two lines are the taxon group assignments required for the binomial test. The last line is the output for a given iteration of the "optimization" procedure, which for this method consists of a single test step. The columns mean the following:

- G/O/F: Assigned $\mathcal{G}$, $\mathcal{O}$ and $\mathcal{F}$ classes at this iteration.
- Criterion: Value of criterion corresponding to current $\mathcal{G}/\mathcal{O}/\mathcal{F}$, only applicable for $\chi^2$- and GFmix-based optimization methods.
- Optimizer: The value being optimized, or the method of $\mathcal{G}/\mathcal{F}$ class assignment.
- Iterations: Number of iterations optimization procedure has been running for.
- Iteration time: Time (in seconds) taken for full round of $\mathcal{G}/\mathcal{F}$ assessments for a given iteration of the optimization procedure.
- Optimization time: Time (in seconds) taken for optimization procedure up to a given iteration.
- Swap times: Time (in seconds) taken for each individual $\mathcal{G}/\mathcal{F}$ assessment in a given iteration of the optimization procedure.

The binomial test assigns $\mathcal{G}$ and $\mathcal{F}$ as RSGHVDP and LFQINK, respectively.

*Optimization based on $\chi^2$ test statistic*

A more complex method of assigning $\mathcal{G}/\mathcal{F}$ classes is to optimize $\mathcal{G}$ and $\mathcal{F}$ with respect to the $\chi^2$ test statistic, following Williamson et al. (2025). The $\chi^2$ test is a standard measure of compositional homogeneity in phylogenetics - the larger $\chi^2$ is, the most heterogeneous the data (Foster, 2004). This method attempts to identify $\mathcal{G}$ and $\mathcal{F}$ which maximizes the $\chi^2$ test statistic for a given alignment – in other words, the most compositionally-heterogeneous $\mathcal{G}$ and $\mathcal{F}$. It does so using an optimization procedure detailed in McCarthy et al. (2025).

Assigning $\mathcal{G}/\mathcal{F}$ classes by optimizing the $\chi^2$ test statistic can be performed as follows:

```
gfselector -aln amborella-interleave.phy -fmt phylip-interleaved -crit chi -out AmborellaChi.txt
```

which outputs the following to screen:

```
Group 1:  Acorus, Calycanthu, Typha, Yucca, Oryza, Spinacia, Nuphar, Ginkgo, Medicago, Pinus, Psilotum, Saccharum, Arabidopsi, Atropa, Amborella, Oenothera, Nymphaea, Nicotiana, Zea, Lotus, Ranunc, Triticum
```

```
Group 2:  Physcomitr, Marchantia
Iteration 1:  Assessing 28 swaps from starting point RSGHVDP/CETMAWY/LFQINK.
Iteration 1:  Found new starting point RSGHVDP/CETMAWYL/FQINK. Iteration time
(s):  0.031, Current optimization time (s):  0.031.
Iteration 2:  Assessing 29 swaps from starting point RSGHVDP/CETMAWYL/FQINK.
Iteration 2:  Found new starting point RSGHVDP/CETMAWYLI/FQNK. Iteration time
(s):  0.017, Current optimization time (s):  0.049.
...
Iteration 6:  Assessing 31 swaps from starting point RSGHVDPC/ETMAWYLIQF/NK.
Iteration 6:  G/O/F at optimization limit:  RSGHVDPC/ETMAWYLIQF/NK. Iteration
time:  0.009, Final optimization time:  0.152.
```

Examining the last line of the output file shows the following:

```
$ tail -n 1 AmborellaChi.txt
RSGHVDPC ETMAWYLIQF NK 366.710972392341 Chi2 6 0.00932216644287109 0.0869698524475098
0.000104904174804688...
```

The optimization procedure assigns $\mathcal{G}$ and $\mathcal{F}$ as RSGHVDPC and NK, respectively. These classes yield a $\chi^2$ result $\approx$ 367. It took six iterations of the optimization procedure to maximize $\chi^2$ for this dataset.

*Optimization based on GFmix model log-likelihood*

The most complex method of assigning $\mathcal{G}/\mathcal{F}$ classes is to assess the log-likelihood estimated for $\mathcal{G}/\mathcal{F}$ of an alignment and tree using the GFmix model. This is performed using the original implementation of the GFmix model detailed in Muñoz-Gómez et al. (2022) and modified in Baker et al. (2024) to accommodate any $\mathcal{G}$ and $\mathcal{F}$ class. This method attempts to identify $\mathcal{G}$ and $\mathcal{F}$ which maximizes the log-likelihood under GFmix for a given alignment/tree, following McCarthy et al. (2025).

First, we infer a phylogenetic tree for this dataset under a branch-homogeneous model of evolution such as LG+C20+Γ using IQ-Tree (Wong et al. 2025) as follows:

```
$ iqtree3 -s amborella-interleave.phy -te amborella-no-edge-lengths.treefile -m
LG+C20+G -pre amborella -T 20
```

This will output two files required for `gfselector`: `amborella.iqtree` containing information on mixture weights and $\alpha$, and `amborella.treefile` containing branch lengths estimated under the branch-homogeneous model. The optimization procedure is then run as follows:

```
/gfselector -aln amborella-interleave.phy -fmt phylip-interleaved -crit gfmix -
iqtree amborella.iqtree -tree amborella.treefile -root amborella.rootfile -mix
C20.aafreq.dat -cpu 20 -conv -out AmborellaGF.txt
```

which outputs the following to screen:

```
Group 1:  Acorus, Calycanthu, Typha, Yucca, Oryza, Spinacia, Nuphar, Ginkgo, Med-
icago, Pinus, Psilotum, Saccharum, Arabidopsi, Atropa, Amborella, Oenothera, Nymphaea,
Nicotiana, Zea, Lotus, Ranunc, Triticum
Group 2:  Physcomitr, Marchantia
```

```
Converted IDs in input alignment, tree and rootfile to 0-index integers.  File
paths:  amborella-interleave.phy_INT,amborella.treefile_INTandamborella.rootfile_INT.
Iteration 1:  Assessing 28 swaps from starting point RSGHVDP/CETMAWY/LFQINK.
Iteration 1:  Found new starting point RSGHVDPA/CETMWY/LFQINK. Iteration time
(s):  22.461, Current optimization time (s):  22.462.
Iteration 2:  Assessing 27 swaps from starting point RSGHVDPA/CETMWY/LFQINK.
Iteration 2:  Found new starting point RSGHVDPA/CETMWYQ/LFINK. Iteration time
(s):  22.538, Current optimization time (s):  45.001.
Iteration 3:  Assessing 28 swaps from starting point RSGHVDPA/CETMWYQ/LFINK.
Iteration 3:  Found new starting point RSHVDPA/GCETMWYQ/LFINK. Iteration time
(s):  22.678, Current optimization time (s):  67.68.
...
Iteration 14:  Assessing 29 swaps from starting point HVDAQM/SRCETYNK/PWGLFI.
Iteration 14:  Found new starting point HVDAQME/SRCTYNK/PWGLFI. Iteration time
(s):  22.39, Current optimization time (s):  323.352.
Iteration 15:  Assessing 28 swaps from starting point HVDAQME/SRCTYNK/PWGLFI.
Iteration 15:  G/O/F at optimization limit:  HVDAQME/SRCTYNK/PWGLFI. Iteration
time:  23.424, Final optimization time:  346.778
```

Examining the last line of the output file shows the following:

```
$ tail -n 1 AmborellaGF.txt
HVDAQME SRCTYNK PWGLFI -179736.846433632 GFmix 15 23.4244914054871 346.777663946152
14.4738280773163...
```

The optimization procedure assigns $\mathcal{G}$ and $\mathcal{F}$ as HVDAQME and PWGLFI, respectively. These classes yield a log-likelihood under LG+C20+Γ+GFmix ≈ −179,737. It took fifteen iterations of the optimization procedure to maximize the log-likelihood for this dataset.

*Use with `gfmixf`*

The `gfmixf` software implements the full maximum-likelihood implementation of the GFmix model described in McCarthy et al. (2025). The result of any optimization method can be used as input for `gfmixf` using the `-g` and `-k` options. For example, the following command will fit $\mathcal{G}$ and $\mathcal{F}$ as HVDAQME and PWGLFI with full optimization of all parameters:

```
$ gfmixf -f C20.aafreq.dat -t amborella.treefile -r amborella.rootfile -s amborella-
interleave.phy -o 6 -p 1 -b 10 -c 20 -g HVDAQME -k PWGLFI
```

**Installation and other details**

`gfselector` is an R script which requires the `parallel` package which is included in standard R installations. The script calls another R script, `gfmix_for_opt`, which implements the GFmix model of Muñoz-Gómez et al. (2022) and Baker et al. (2024) for the optimization procedure. This script requires several C programs - `rert`, `treecns` and `alpha_est_mix_rt` which are provided pre-compiled with this script. If the pre-compiled programs do not run correctly, the user is advised to download the source code for the older GFmix implementation from `https://www.mathstat.dal.ca/~tsusko/software.html` and follow the compilation and installation instructions from the document provided at the same address.

The C-series models (C10, C20, ..., C60) are also provided with these scripts. See the documentation for the GFmix implementations for further details.

`gfselector` expects `gfmix_for_opt` to be available in your PATH environmental variable. In turn, `gfmix_for_opt` expects `rert, treecns and alpha_est_mix_rt` to also be available in PATH. If the locations for these files are not in your PATH, you can modify the expression

```
bindir <- ""
```

in `gfselector` to point to the location where `gfmix_for_opt` is located, and the same expression in `gfmix_for_opt` to the location where `rert, treecns and alpha_est_mix_rt` are located. Both expressions are at the top of both R scripts.

## References


Baker, B. A., A. Gutiérrez-Preciado, Rodríguez del Río, C. G. P. McCarthy, P. López-García, J. Huerta-Cepas, E. Susko, A. J. Roger, L. Eme, and D. Moreira (2024). "Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments". *Nature Microbiology* **9**, 964–975. DOI: 10.1038/s41564-024-01647-4.

McCarthy, C. G. P., E. Susko, and A. J. Roger (2025). "Modeling site-and-branch-heterogeneity with GFmix". *bioRxiv*. DOI: 10.1101/2025.08.07.669136.

Muñoz-Gómez, S. A., E. Susko, K. Williamson, L. Eme, C. H. Slamovits, D. Moreira, P. López-García, and A. J. Roger (2022). "Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria". *Nature Ecology and Evolution* **6**, 253–262. DOI: 10.1038/s41559-021-01638-2.

Williamson, K., L. Eme, H. Baños, C. G. P. McCarthy, E. Susko, R. Kamikawa, R. J. S. Orr, S. A. Muñoz-Gómez, B. Q. Minh, A. G. B. Simpson, and A. J. Roger (2025). "A robustly rooted tree of eukaryotes reveals their excavate ancestry". *Nature* **640**, pp. 974–981. DOI: 10.1038/s41586-025-08709-5.

Wong, T., N. Ly-Trong, H. Ren, H. Baños, A. Roger, E. Susko, C. Bielow, N. De Maio, N. Goldman, M. Hahn, G. Huttley, R. Lanfear, and B. Q. Minh (2025). "IQ-TREE 3: phylogenomic inference software using complex evolutionary models". *EcoEvoRxiv*. DOI: 10.32942/x2p62n.