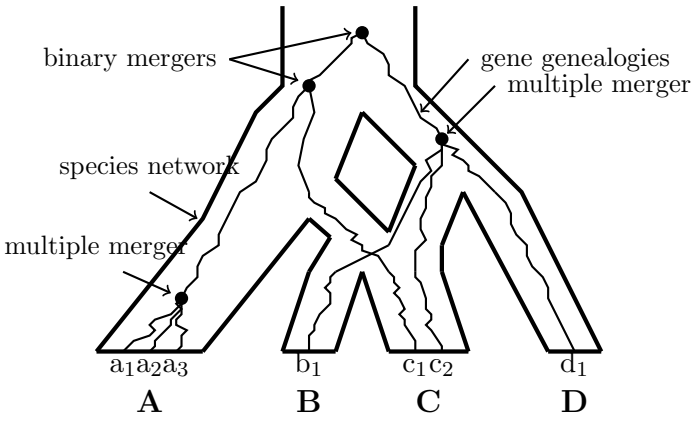
# Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees

Sha (Joe) Zhu

`Hybrid-Lambda` is a software package that can simulate gene trees within a rooted species network or a rooted species tree under the coalescent process. The main feature of this program is that users can choose to use the standard Kingman coalescent process, which produces bifurcating genealogies, or two other Λ-coalescent processes, which produce multifurcating genealogies. The other feature is that `Hybrid-Lambda` uses extended Newick formatted strings to make it easier to represent hybridization events between species.




## 1 Download and installation

`Hybrid-Lambda` can be downloaded from https://github.com/shajoezhu/hybrid-Lambda. Extract the source code by executing the following command:

```
tar -xf hybrid-Lambda-VERSION.tar.gz.
```

It is fairly standard to compile `hybrid-Lambda` on UNIX-like systems. In the directory `hybrid-Lambda-VERSION`, execute the following command:

```
$./bootstrap
$make
```

**Note:** The command `make doxygen-run` will generate `HTML` documentation of the source code.

### 1.1 Citation

Please cite our program using the following entry:

Zhu, Sha and Degnan, James H and Goldstien, Sharyn J and Eldon, Bjarki (2015). Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *bioRxiv*. doi: http://dx.doi.org/10.1101/023465.

## 2 Notation

### 2.1 Coalescent parameters

Under the coalescent process, the waiting time for lineages to coalesce is an exponential random variable. The Kingman coalescent process allows only two lineages to coalesce at a time. Thus the mean of the waiting time for $b$ lineages to coalesce into $b-1$ lineages is $\binom{b}{2}$ per unit of time. However, for the $\Lambda$-coalescent, if the coalescent parameter $\psi$ is between 0 and 1 [3], the rate $\lambda_{bk}$ at which $k$ out of $b$ active ancestral lineages merge is:

$$\lambda_{bk} = \binom{b}{k} \psi^{k-2}(1-\psi)^{b-k}, \quad \psi \in [0,1]. \tag{1}$$

If the coalescent parameter $\alpha$ is between 1 and 2, the rate is:

$$\lambda_{bk} = \binom{b}{k} \frac{B(k-\alpha, b-k+\alpha)}{B(2-\alpha,\alpha)}, \quad \alpha \in (1,2), \tag{2}$$

where $B(\cdot,\cdot)$ is the beta function [8].

### 2.2 Input/output formats

The input file for `Hybrid-Lambda` is a character string that describes the relationships among species. Standard Newick format [6] is used for inputting species trees and outputting gene trees, e.g.:

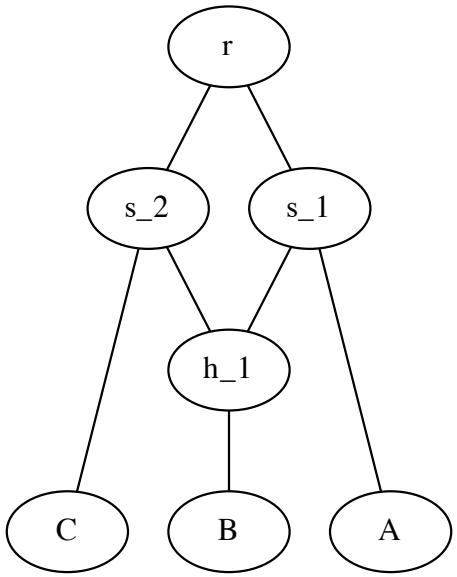
$$((A:t_A, B:t_B):t_{AB}, C:t_C), \tag{3}$$

where $t_i$ denotes the branch length from $i$ to its parent node. `Hybrid-Lambda` uses the values of $t_i$ to assign parameters for different inputs. In Expression (3), the interior nodes are not labelled. However, this is not essential. `Hybrid-Lambda` can also recognise input Newick string whose nodes are all labelled.

Extended Newick formatted strings [1, 4] label all internal nodes, and are used for inputting species networks. In the network string, the descendants of a hybrid node are recorded before the hybrid node the first time the hybrid node appears, otherwise is written as a tip node. For example,

$$(((B:t_B)h\#\gamma:t_h^{s1},A:t_A)s1:t_{s1},(h\#\gamma:t_h^{s2},C:t_C)s2:t_{s2})r, \tag{4}$$

where # identifies the hybrid node.




At a hybrid node, lineages travel to either parent nodes with given probabilities. The parameter $\gamma$ denotes the probability of that the lineage goes to the first parent node. Since the hybrid node has two parent nodes, the branch length needs to be specific, i.e. $t_i^j$ denotes the branch length from $i$ to $j$.

Normally, standard Newick formatted and extended Newick formatted strings do not include branch lengths at the root node. However, in this program, this is required, as the input strings assign population size or the coalescent parameter at the root. Thus, Expressions (3) and (4) are associated with Expressions (5) and (6) respectively:

$$((A:t_A,B:t_B):t_{AB},C:t_C):t_{root}, \tag{5}$$

$$(((B:t_B)h\#\gamma:t_h^{s1},A:t_A)s1:t_{s1},(h\#\gamma:t_h^{s2},C:t_C)s2:t_{s2})r:t_r. \tag{6}$$

# 3 Commands for simulation:

## 3.1 Simulating gene trees

```
hybrid-Lambda -spcu INPUT [-num N] [-seed SEED] [-o PREFIX].
```

`INPUT` is a(n) (extended) Newick formatted string (see Section 2.2), which can be entered through the command line or from a text file. If the input is followed by the flag `-spcu`, its branch lengths must be in coalescent units. The value `N` following the flag `-num` is the number of gene trees simulated. Users can specify a random seed for simulation by declaring it after `-seed`. By default, the branch lengths of the output trees are in coalescent units. They are saved in the file `PREFIX_coal_unit`. The flag `-o` enables the users to specify the file name prefix. For example:

```
hybrid-Lambda -spcu '((1:1,2:1):1,3:2);' -num 3 -seed 2 -o example1
```

will print the following message:

```
Default Kingman coalescent on all branches.
Default population size of 10000 on all branches.
Random seed: 2
Produced gene tree files:
example1_coal_unit
```

The following gene trees are saved in the file `example1_coal_unit`:

```
(3_1:2.20467,(2_1:1.57269,1_1:1.57269):0.631976);
((2_1:1.02627,1_1:1.02627):3.65525,3_1:4.68152);
((2_1:1.79776,1_1:1.79776):3.15359,3_1:4.95134);
```

## 3.2 Gene tree output options and user-defined mutation rate

```
hybrid-Lambda -spcu INPUT [-o PREFIX] [-mu MU] [option]
```

By default, the mutation rate $\mu = 0.00005$ is used. The flag `-mu` makes it possible for users to define a constant mutation rate. Moreover, the options in `[ ]` enable more manipulations of the output gene trees. These options include:

| | |
|---|---|
| `-sim_mut_unit` | Convert the simulated gene tree branch lengths to mutation units. |
| `-sim_num_gener` | Convert the simulated gene tree branch lengths to number of generations. |
| `-sim_num_mut` | Simulate the number of mutations on each branch of the simulated gene trees. |
| `-sim_Si_num` | Generate a table, which includes the number of segregating sites and the total branch length of the gene tree, as well as the TMRCA. |

For example, suppose the input network string in the file `4_tax_sp_nt1_para` is

```
((((B:1,C:1)s1:1)h1#.5:1,A:3)s2:1,(h1#.5:1,D:3)s3:1)r.
```

```
hybrid-Lambda -spcu trees/4_tax_sp_nt1_para -o example2 -num 2 \
              -mu 0.00003 -sim_mut_unit -sim_num_mut
```

will generate the following files:

```
$ cat example2_coal_unit
((B_1:1.9099,C_1:1.9099):2.82957,(A_1:4.05317,D_1:4.05317):0.686292);
((D_1:3.77974,(C_1:1.2291,B_1:1.2291):2.55064):0.369812,A_1:4.14956);
$ cat example2_mut_unit
((B_1:0.57297,C_1:0.57297):0.848871,(A_1:1.21595,D_1:1.21595):0.205888);
((D_1:1.13392,(C_1:0.36873,B_1:0.36873):0.765192):0.110944,A_1:1.24487);
$ cat example2_num_mut
((B_1:1,C_1:1):2,(A_1:1,D_1:1):0);
((D_1:0,(C_1:1,B_1:1):0):0,A_1:3); .
```

## 3.3 User-defined population sizes

```
hybrid-Lambda -spcu INPUT-1 -pop INPUT-2
```

By the default setting, the population sizes for each species are assumed to be equal and unchanged at any time, which is 10,000. This can be reassigned to other constant values followed by `-pop`. As a result, the branch lengths of the gene trees in number of generations will change. This can be observed though the option `-sim_num_gener`. For example, to simulate gene trees within a species network/tree with a population size of 25,000, we use the following:

```
hybrid-Lambda -spcu INPUT -num N -pop 25000 -sim_num_gener.
```

This command will also produce gene trees for which the branch lengths are in number of generations, saved in the file `OUT_num_gener`.

**Note:** The population size refers to the number of gene copies, not the number of individuals.

Instead of inputting a species network with branch lengths in coalescent units, input strings can have branch lengths representing the number of generations. Moreover, in the following example, we demonstrate that if the population sizes are assumed to vary, the input strings in Expression (5) can specify the population sizes on all branches and the root.

```
hybrid-Lambda -spng '(A:50000,B:50000)r;' -pop '(A:50000,B:50000)r:40000;'.
```

## 3.4 Simulating multiple samples per species

```
hybrid-Lambda -spcu INPUT -S n1 n2 ...
```

`Hybrid-Lambda` sorts the taxa names in a particular order. At each character of a taxon name, it sorts:

- numerics in ascending order,
- letters in alphabetical order,
- numerics then letters,
- upper-case letters then lower-case letters.

To sample multiple individuals, the order of the sample sizes needs to follow the order of the taxa names.

For example:

```
hybrid-Lambda -spcu '((((A:1.1,B:1.1):2.1,a:2.2):1.1,13D:.2):.3,4:.3);' \
                                -S 2 4 3 6 5 .
```

The order of the taxon names is `13D`, `4`, `A`, `B` and `a`. Thus, the program will sample 2 individuals in taxon `13D`, four samples from taxon `4`, three samples from taxon `A`, six samples from taxon `B` and five samples from taxon `a`.

## 3.5 Simulating gene trees with multiple merger coalescents

```
hybrid-Lambda -spcu INPUT-1 -mm INPUT-2
```

The Kingman coalescent is assumed by default. To use the Λ-coalescent, the coalescent parameter needs to be specified after `-mm`. For details, see Equation (1) and (2). Moreover, similar to assigning particular population sizes on branches (see the examples in Section 3.3), coalescent parameters can be specified as well. In this case, to assume the Kingman coalescent within some population, the multiple merger parameter needs to be set to 2. For example:

```
hybrid-Lambda -spcu '(A:1,B:1)r;' -mm '(A:1.9,B:.2)r:2;' -S 3 4.
```

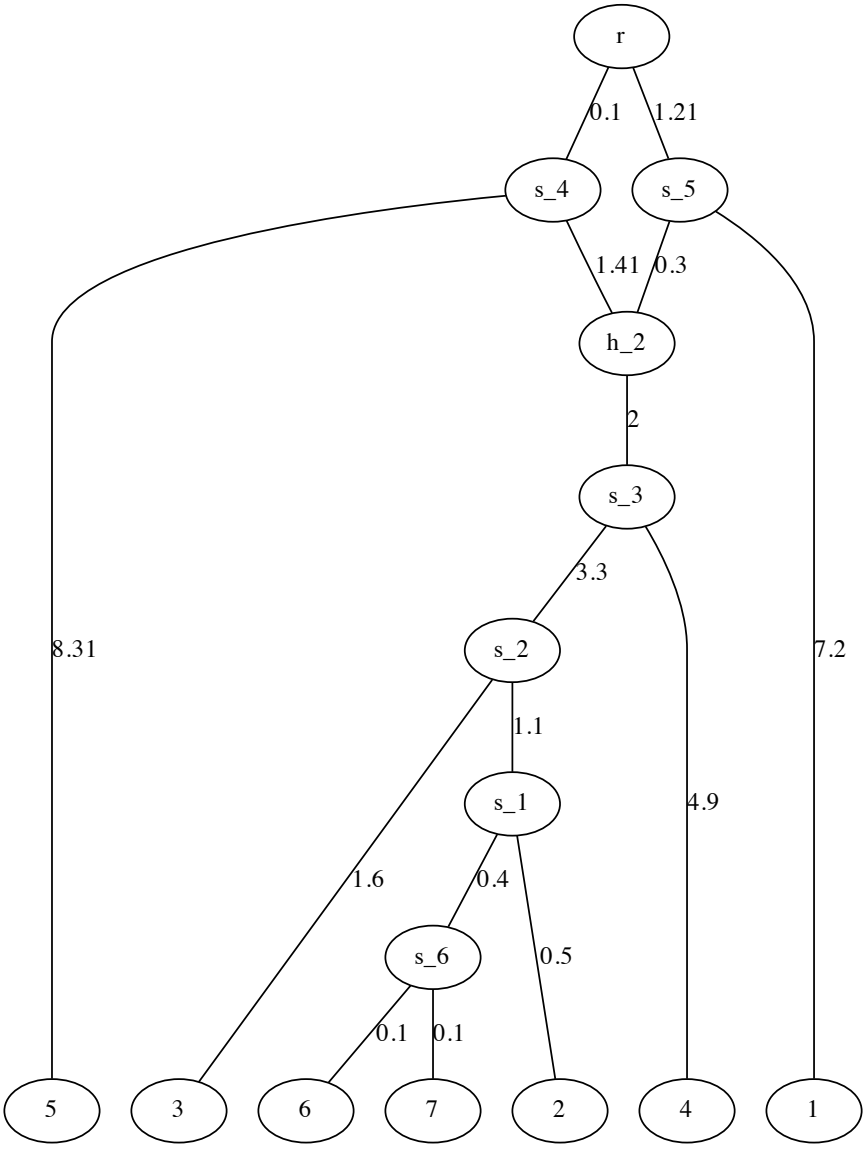
## 3.6 Commands for other features:

### 3.6.1 Plot

```
hybrid-Lambda -spcu INPUT -dot -o OUTPUT [-branch]
```

`Hybrid-Lambda` uses the program `dot` to generate figures. The option `[-branch]` will label the branch lengths in the figure, e.g.:

```
hybrid-Lambda -spcu trees/7_tax_sp_nt1_para -dot -o branch -branch
```




Alternatively, by replacing `-dot` with `-plot`, `Hybrid-Lambda` can generate LaTeX code for plotting a network or tree.

### 3.6.2 Analysing the frequencies of gene trees

`Hybrid-Lambda` can generate a topology frequency table for the simulated gene trees by:

```
hybrid-Lambda -spcu INPUT -num N -f.
```

The command "`hybrid-Lambda -gt INPUT -f -o OUTPUT`" reads trees in the file `INPUT`, and generates a topology frequency table in the file `OUTPUT_frequencies`.

### 3.6.3 Simulating and analysing the monophyly topology of the gene trees

```
hybrid-Lambda -spcu INPUT-1 -S n1 n2 -mono [-mm INPUT-2]
```

Recent studies on the shape of genealogies between two species investigated the probabilities of monophyletic taxa [2, 7]. The option `-mm` generates a frequency table of the gene trees whose taxa are monophyletic in one population, both populations, or are paraphyletic and polyphyletic. For example:

```
hybrid-Lambda -spcu '(A:5,B:5)r;' -mono -num 100 -mm .1 -S 4 4
```

will print the following message:

```
   A mono     B mono Recip mono     A para     B para  Polyphyly
     0.02       0.01           0       0.02       0.01       0.97
Random Seed  1342238826  used
Produced gene tree files:
GENE_TREE_coal_unit
100 trees simulated.
```

# 4 Summary of command line options

| | |
|---|---|
| `-h` or `-help` | Help. List the following content. |
| `-spcu STR` | Input the species network/tree string through command line or a file. Branch lengths of the INPUT are in coalescent unit. |
| `-spng STR` | Input the species network/tree string through command line or a file. Branch lengths of the INPUT are in number of generation. |
| `-pop STR/FLT` | Population sizes are defined by a single numerical constant, or a string which specifies the population size on each branch. The string can be input through command line or a file. By default, population size 10,000 is used. |
| `-mm STR/FLT` | Multiple merger parameters are defined by a single numerical constant, or a string which speifies the parameter on each branch. The string can be input through command line or a file. By default, Kingman coalescent is used. |
| `-S INT INT ...` | Specify the number of samples for each taxon. |

| | |
|---|---|
| `-num INT` | The number of gene trees to be simulated. |
| `-seed INT` | User defined random seed. |
| `-mu FLT` | User defined constant mutation rate per locus. By **default** mutation rate 0.00005 is used. |
| `-o STR [option]` | Specify the file name prefix for simulated gene trees. Prefix is set as "OUT" by default. When options are not specified, only output trees with branch lengths are in coalescent unit. |
| `-sim_mut_unit` | Convert the simulated gene tree branch lengths to mutation units. |
| `-sim_num_gener` | Convert the simulated gene tree branch lengths to number of generations. |
| `-sim_num_mut` | Simulate the number of mutations on each branch of the simulated gene trees. |
| `-sim_Si_num` | Generate a table, which includes the number of segregating sites and the total branch length of the gene tree, as well as the TMRCA. |
| `-f` | Generate a topology frequency table of a set of input trees or simulated gene trees. |
| `-gt STR` | Specify the FILE NAME of trees to analyse tree topology frequencies. |
| `-seg` | Generate segregating site data |
| `-mono` | Generate a frequency table of monophyletic, paraphyletic and polyphyletic trees. |
| `-plot/-dot [option]` | Use LaTeX(`-plot`) or `Dot` (`-dot`) to draw the input (defined by `-spcu`) network (tree). |
| `-branch` | Branch lengths will be labelled in the figure. |

# Acknowledgements

**Funding:** New Zealand Marsden Fund (Sha Zhu and James Degnan), Engineering and Physical Sciences Research Council (Bjarki Eldon). This work was partly conducted while JD was a Sabbatical Fellow at the National Institute for Mathematical and Biological Synthesis, an institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville.

**Included files:** C++ Mersenne twister pseudo-random number generator [5].

# References


[1] Gabriel. Cardona, Francesc. Rossell, and Gabriel. Valiente. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, 9(532-540), 2008.

[2] Bjarki Eldon and James H. Degnan. Multiple merger gene genealogies in two species: monophyly, paraphyly, and polyphyly for two examples of Lambda coalescents. *Theoretical Population Biology*, 82:117–130, 2012.

[3] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633, 2006.

[4] D.H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, Cambridge, UK, 2010.

[5] M. Matsumoto and T. Nishimura. Mersenne Twister: A 623-dimensionally equidistributed uniform Pseudo-Random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, 1998.

[6] Gary Olsen. Gary Olsen's interpretation of the "Newick's 8:45" tree format standard, 1990. http://evolution.genetics.washington.edu/phylip/newick_doc.html. Feb 21, 2013.

[7] Noah A. Rosenberg. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. *Evolution*, 57(7):1465–1477, 2003.

[8] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications*, 106:107–139, 2003.