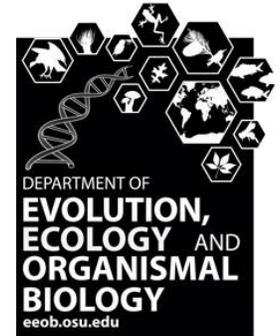


spedeSTEM tutorial

Bryan Carstens
Matthew Demarest
Maxim Kim
Tara Pelletier
Jordan Satler

Acknowledgements



Development of spedeSTEM was funded via a grant from the National Science Foundation (DEB-0918212).

Thanks to the OSU College of Arts and Sciences Technology Services, particularly Sanford Shew and Timothy Smith.

Thanks to Dan Ence, Sarah Hird, John McVay and Noah Reid for discussions about spedeSTEM.

Thanks to all users who have analyzed their data using our program.



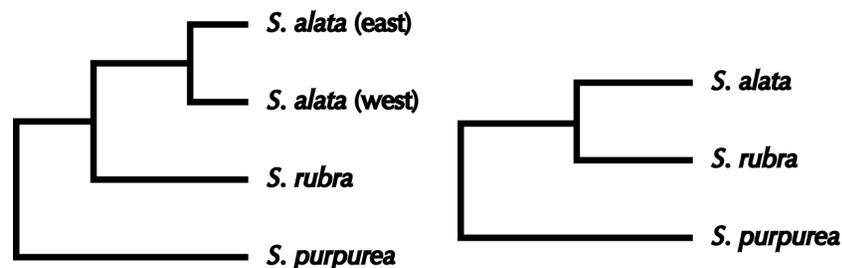
spedeSTEM: Introduction

spedeSTEM is a program that delimits species using maximum likelihood and information theory. Specifically, the probabilities of multiple permutations of putative evolutionary lineages are calculated using STEM 2.0 (Kubatko et al. 2009) and ranked by model probability (see Anderson 2004). **spedeSTEM** takes as input ultrametric gene trees from multiple loci and a user-supplied estimate of theta, and returns a table of models ranked by model probability. The web-based software here conducts both discovery and validation analyses, and also generates the set up files and allows the users to subsample alleles from large nexus files. **spedeSTEM** does not estimate gene trees; for this we suggest PAUP (Swofford 2002) or Garli (Zwickl 2006).

spedeSTEM also includes modules for subsampling, simulation testing. Currently these are limited to the python version, but we hope to implement them into the web-based application at some point in the future.

spedeSTEM: background

Species delimitation using species trees operates by comparing the probability of models where putative lineages are separate to the probability of models where they are the same. For example, consider one motivating example, the carnivorous plant *Sarracenia alata* that has a disjunct distribution in eastern and western regions of its range (Koopman & Carstens 2010; Zellmer *et al.* 2012; Carstens & Satler 2013). If we computed the probabilities of each of the models below, we might use a likelihood ratio test as suggested by Knowles and Carstens (2007).



spedeSTEM: background

Species delimitation using species trees: We could compute the probability of each of these models using STEM (Kubatko et al. 2009). Stem computes the maximum likelihood species tree using a coalescent model that accounts for the loss of ancestral polymorphism due to genetic drift. The probability is calculated using the gene tree density $f(g_j | S, \tau)$ given by Rannala & Yang (2003).

BIOINFORMATICS APPLICATIONS NOTE 2009, pages 1–3
doi:10.1093/bioinformatics/btp079

Phylogenetics

STEM: species tree estimation using maximum likelihood for gene trees under coalescence

Laura S. Kubatko^{1,*}, Bryan C. Carstens² and L. Lacey Knowles³

¹Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, ²Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 and

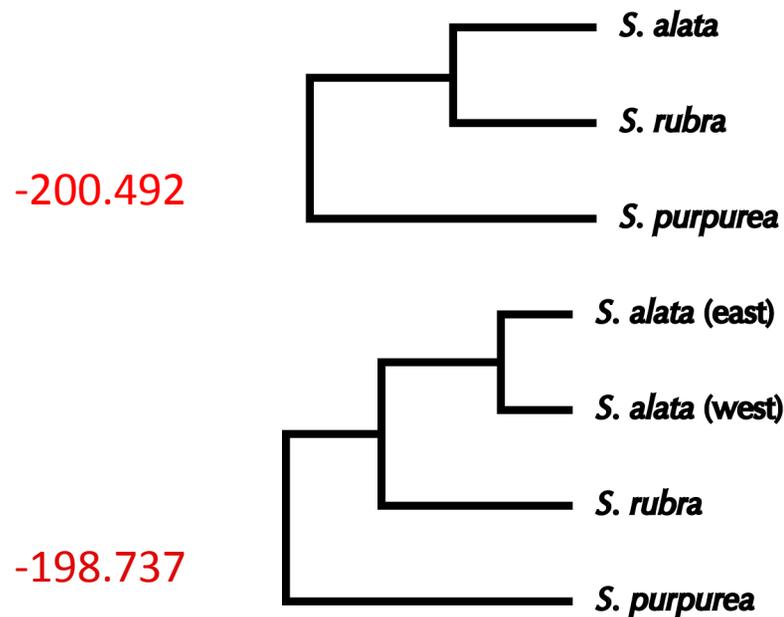
³Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Received on November 28, 2008; revised and accepted February 04, 2009

Associate Editor: Martin Bishop

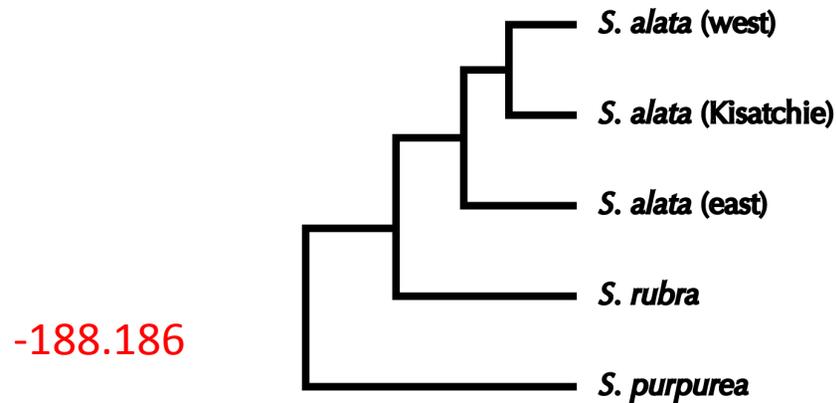
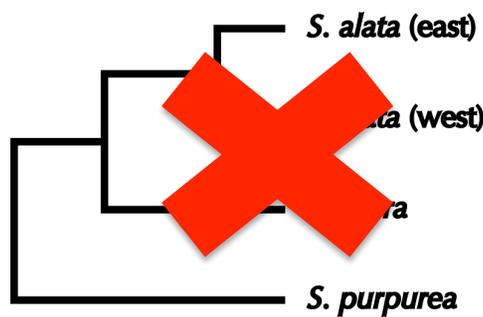
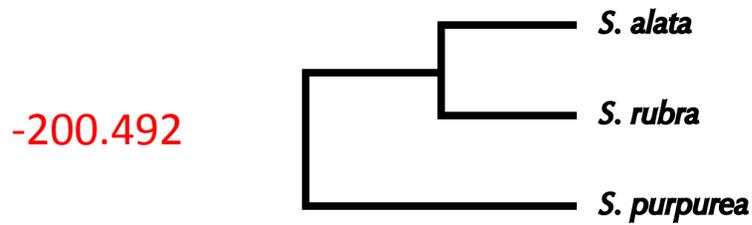
$$L(S, \tau) = \prod_{j=1}^N f(g_j | S, \tau)$$

spedeSTEM: background



If we compare our two *a priori* models using a Likelihood ratio test, we find that we can not reject the model where all of *S. alata* constitutes a single evolutionary lineage ($\chi^2 = 3.512, p = 0.0609$)

spedeSTEM: background

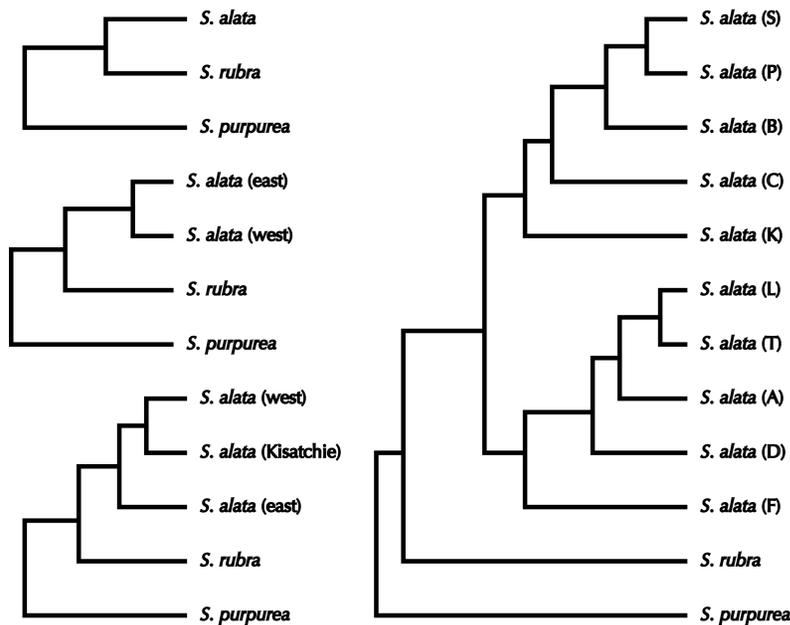


However, what if we consider a third model, as suggested by the optimal results from a Structure analysis? Here, we find that we can reject the null ($\chi^2 = 24.054, p = < 0.0001$)



Optimal partitioning of the data suggested by STRUCTURE results . . .

spedeSTEM: background

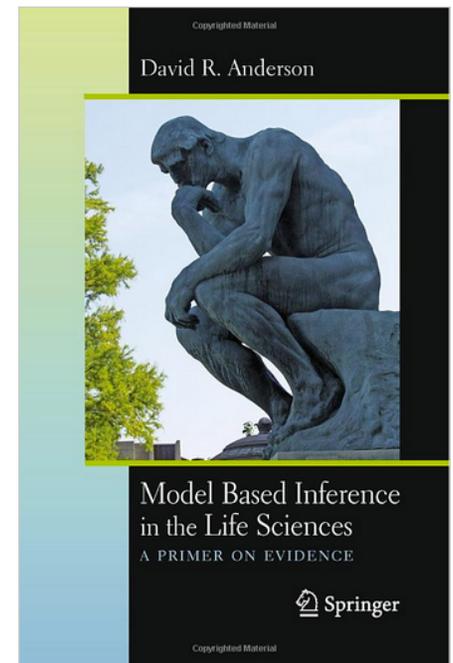


What is the appropriate null model for species delimitation? If we assume, geography and try to delimit 2 lineages, we do not reject the null. If we treat the results of the Structure analysis as the alternative hypothesis, then we do reject the null. Which of these is appropriate? It could be that neither is; there are dozens of possible alternative models, up to an including a model that treats all sampling localities as putative lineages.

This question is inherently related to your philosophy of statistics. We argue that null hypothesis testing is not particularly useful for a historical discipline such as phylogeography that lacks experimental replication or controls.

spedeSTEM: background

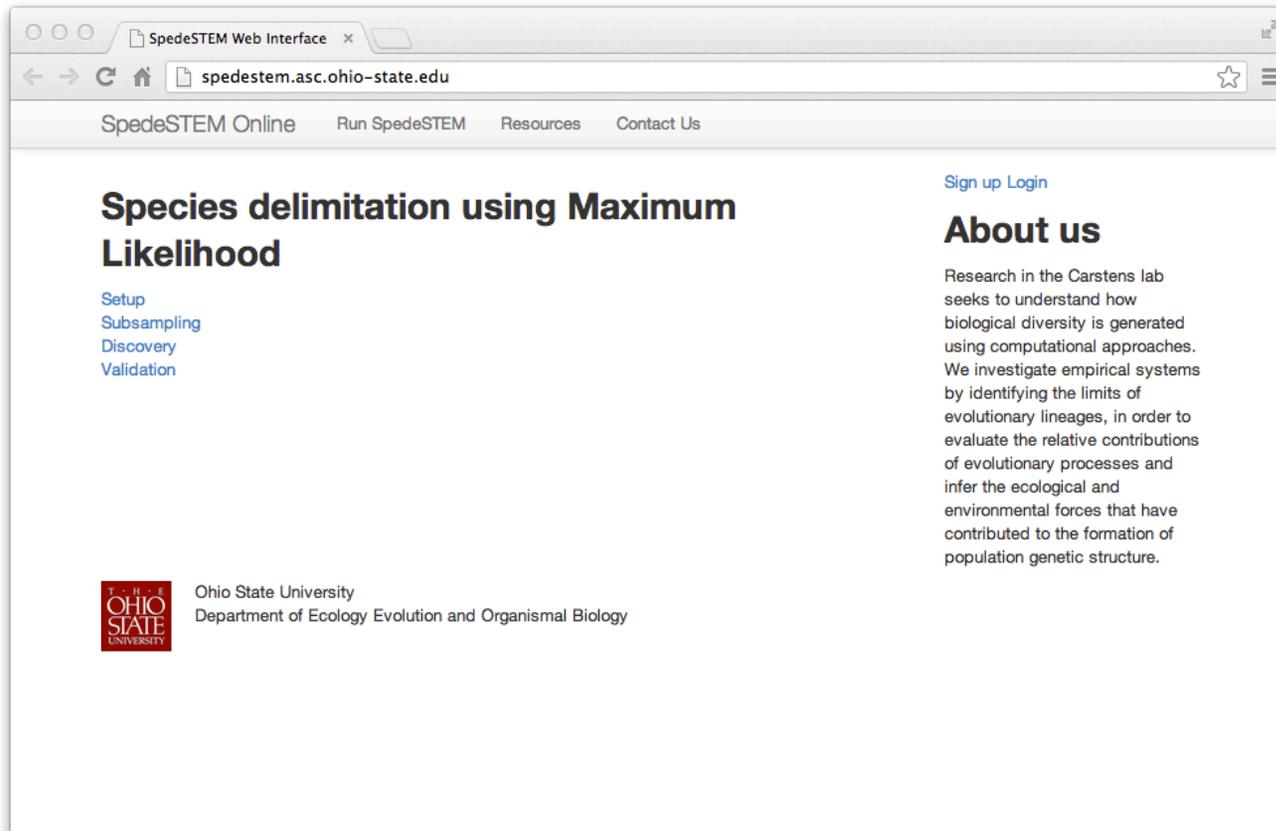
- spedeSTEM does not rely on hypothesis testing, and can consider any number of possible models (in theory). In practice, we have successfully considered >20000 for 10 putative lineages.
- Information theory is used to evaluate the probability of models (see Anderson 2008 for a readable introduction to information theory).
- All models are ranked using information theory, and the model probabilities are computed.
- STEM computes the probability of the model given the data in seconds, so thousands of models can be compared.



spedeSTEM: background

- SpedeSTEM 2.0 takes as input ultrametric gene trees from multiple loci.
- SpedeSTEM 2.0 includes validation and discovery approaches to species delimitation.
- SpedeSTEM 2.0 contains commands to assist the user in data formatting and allows for the subsampling of data sets.
- This tutorial covers both the **command line** and **web-based** versions of spedeSTEM 2.0. Both are written in Python, but the former has more functionality.

spedeSTEM: web-based



<http://spedestem.asc.ohio-state.edu/>

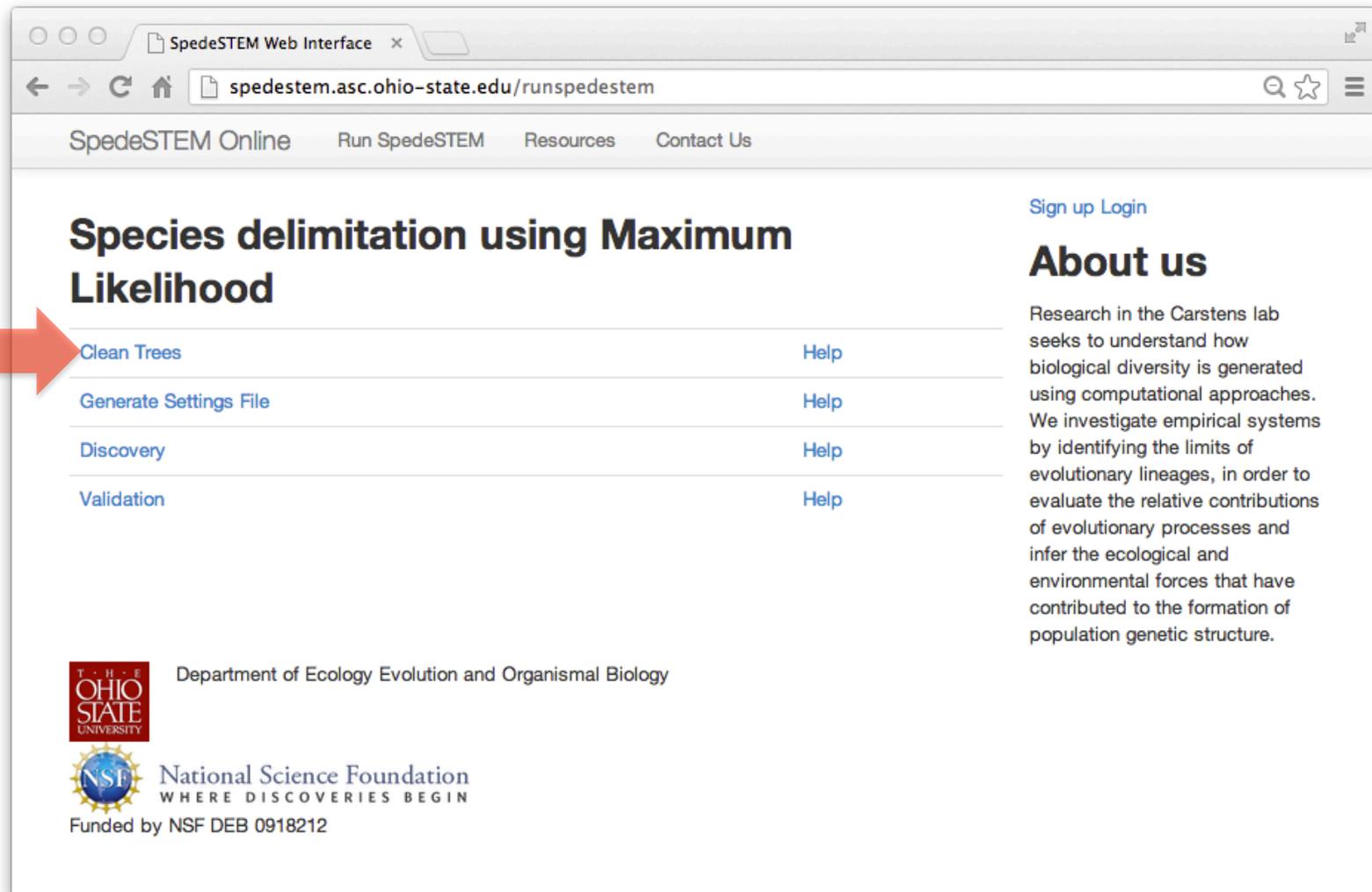
spedeSTEM: command line

- Download the python version of spedeSTEM 2.0 from the following URL.

<http://carstenslab.org.ohio-state.edu/software.html>

- It contains all example files used here.

Step 1: Clean your genetrees.



The screenshot shows a web browser window with the URL `spedestem.asc.ohio-state.edu/runspedestem`. The page title is "SpedeSTEM Online" and the navigation menu includes "Run SpedeSTEM", "Resources", and "Contact Us". The main heading is "Species delimitation using Maximum Likelihood". Below this heading is a table with four rows, each containing a link and a "Help" link. A red arrow points to the "Clean Trees" link.

Link	Help
Clean Trees	Help
Generate Settings File	Help
Discovery	Help
Validation	Help

Sign up Login

About us

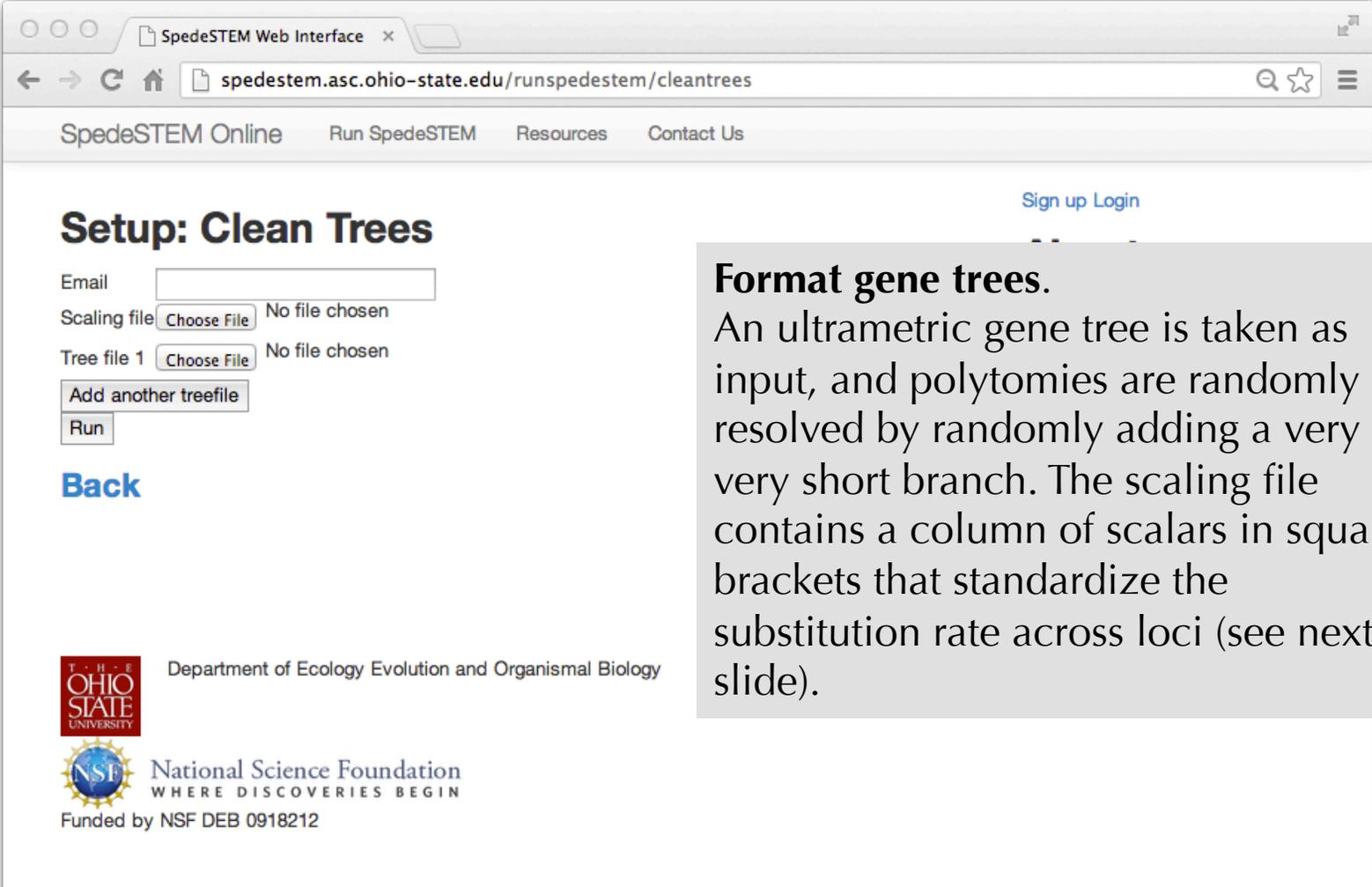
Research in the Carstens lab seeks to understand how biological diversity is generated using computational approaches. We investigate empirical systems by identifying the limits of evolutionary lineages, in order to evaluate the relative contributions of evolutionary processes and infer the ecological and environmental forces that have contributed to the formation of population genetic structure.

 Department of Ecology Evolution and Organismal Biology

 National Science Foundation
WHERE DISCOVERIES BEGIN

Funded by NSF DEB 0918212

Step 1: Clean your genetrees.

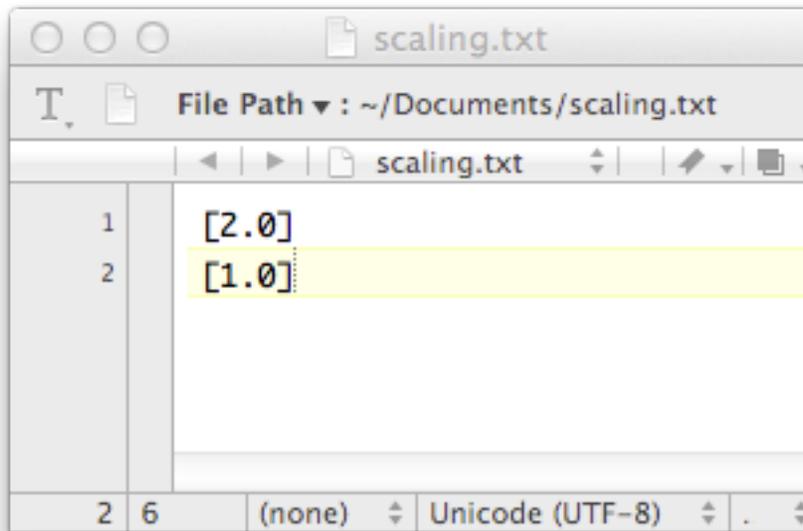


The screenshot shows a web browser window with the URL `spedestem.asc.ohio-state.edu/runspedestem/cleantrees`. The page title is "SpedeSTEM Online" and the navigation menu includes "Run SpedeSTEM", "Resources", and "Contact Us". There are links for "Sign up" and "Login". The main heading is "Setup: Clean Trees". The form includes an "Email" input field, a "Scaling file" field with a "Choose File" button and "No file chosen" text, a "Tree file 1" field with a "Choose File" button and "No file chosen" text, an "Add another treefile" button, and a "Run" button. A "Back" link is also present. At the bottom, there are logos for The Ohio State University and the National Science Foundation (NSF), along with the text "Department of Ecology Evolution and Organismal Biology" and "Funded by NSF DEB 0918212".

Format gene trees.

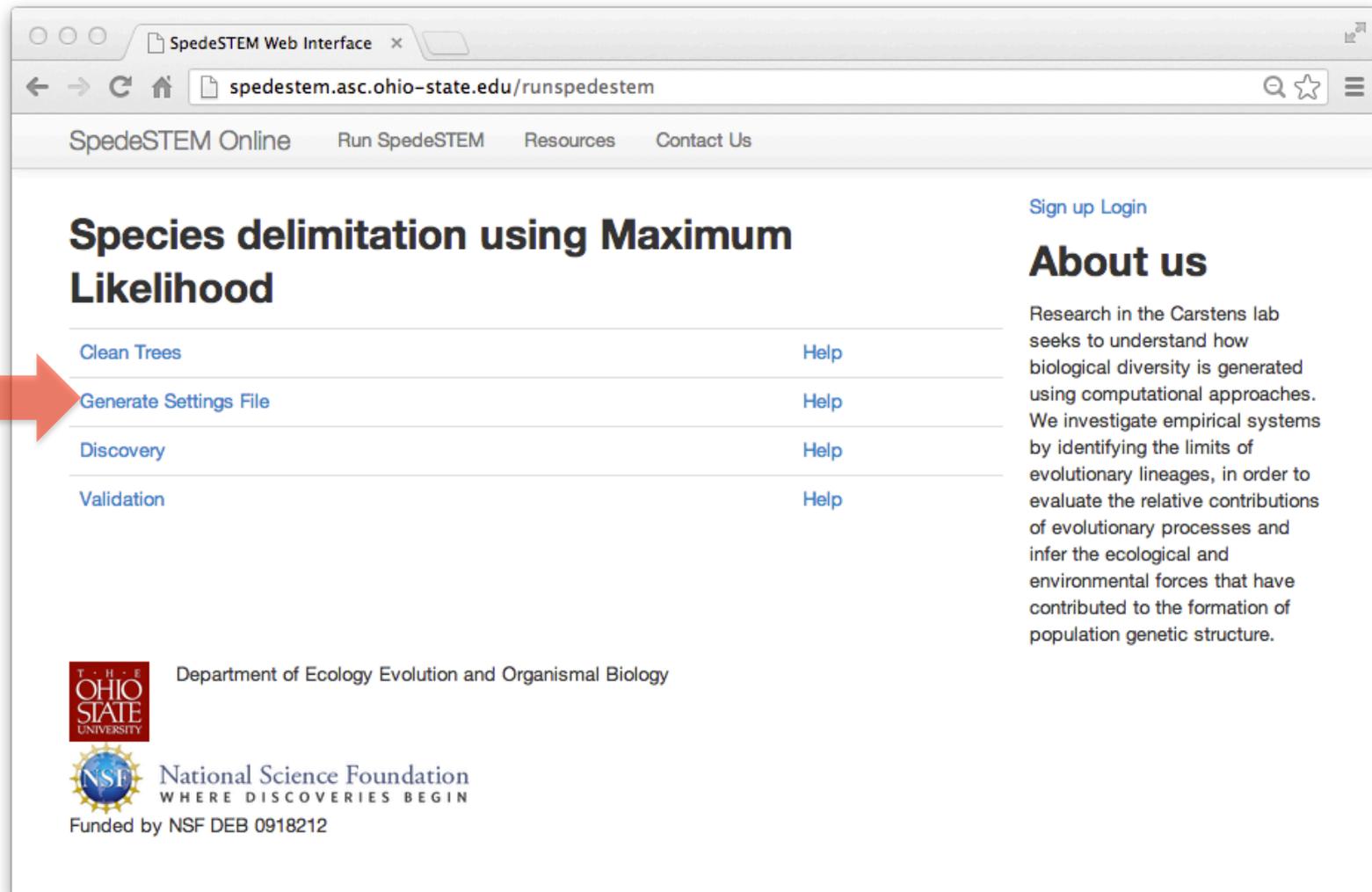
An ultrametric gene tree is taken as input, and polytomies are randomly resolved by randomly adding a very very short branch. The scaling file contains a column of scalars in square brackets that standardize the substitution rate across loci (see next slide).

Scaling Factors File



Scaling file. A scaling file is required to run spedeSTEM – this is simply a text file with scaling factors for each locus. Loci should be in the same order as entered. Scaling factors are used to normalize the length of the gene trees. Let the scaling factor of each locus be a number that normalizes the (# segregating sites across loci / length of the locus). So, if you have two loci, locus A with 10 snps / 1000 bp and locus B with 10 snps / 500 bp, your scaling factor of locus A would be [2.0] (because $2 \cdot (10/1000) = (10/500)$). It does not matter which of the loci are chosen as the locus with a scaling factor of [1.0] (we used locus B in this example). Place these in a simple text file as shown to the left.

Step 2: Generate settings file.



The screenshot shows a web browser window with the URL `spedestem.asc.ohio-state.edu/runspedestem`. The page title is "SpedeSTEM Online" and the navigation menu includes "Run SpedeSTEM", "Resources", and "Contact Us". The main heading is "Species delimitation using Maximum Likelihood". Below this heading is a table with four rows: "Clean Trees", "Generate Settings File", "Discovery", and "Validation". Each row has a "Help" link to its right. A red arrow points to the "Generate Settings File" link. To the right of the table is an "About us" section with a "Sign up Login" link. At the bottom, there are logos for The Ohio State University and the National Science Foundation, along with the text "Department of Ecology Evolution and Organismal Biology" and "Funded by NSF DEB 0918212".

Link	Help
Clean Trees	Help
Generate Settings File	Help
Discovery	Help
Validation	Help

About us
Sign up Login
Research in the Carstens lab seeks to understand how biological diversity is generated using computational approaches. We investigate empirical systems by identifying the limits of evolutionary lineages, in order to evaluate the relative contributions of evolutionary processes and infer the ecological and environmental forces that have contributed to the formation of population genetic structure.

THE OHIO STATE UNIVERSITY
Department of Ecology Evolution and Organismal Biology

NSF National Science Foundation
WHERE DISCOVERIES BEGIN
Funded by NSF DEB 0918212

Setup 2: Generate a settings file.

SpedeSTEM Web Interface

spedestem.asc.ohio-state.edu/runspedestem/generatesettings

SpedeSTEM Online Run SpedeSTEM Resources Contact Us

[Sign up](#) [Login](#)

Setup: Generate Settings File

Traits file No file chosen

BT value

Email

[Back](#)

 Department of Ecology Evolution and Organismal Biology

 National Science Foundation
WHERE DISCOVERIES BEGIN

Funded by NSF DEB 0918212

Generate settings file. A settings file is required to run spedeSTEM, and the setup will generate this file. Two simple files are taken as input – a traits file (a tab delimited with columns equal to **trails**, **species**, **groups**, as in the *Beast and BPP traits file) and a value for $\theta = 4N_e\mu$. The later is input into the “BT value” field.

Cmd line setup

- navigate to spedeSTEM directory

```
cd [path to directory]
```

- access help by entering:

```
./SpedeSTEM_2.py setup -h
```

- generate settings file by entering:

```
./SpedeSTEM_2.py setup -c  
cpDNA.tre Sa4.tre Sa135.tre  
Sa163.tre Sa297.tre Sa302.tre  
Sa323.tre Sa405.tre -s  
scaling.txt -t traits.txt -bt  
0.123
```

- generate genetrees file by entering:

```
cat cleaned.*.tre > genetrees.tre
```

```
SpedeSTEM-master — bash — 80x48
Last login: Mon Apr 22 10:54:31 on ttys000
dhcp-254-235-238:~ bcarstens$ cd /Users/bcarstens/spedeSTEM/SpedeSTEM-master
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py setup -h
usage: SpedeSTEM_2.py setup [-h] [-c treeFile [treeFile ...]] [-s scalingFile]
                             [-t traitsFile] [-bt thetaValue]

optional arguments:
  -h, --help                show this help message and exit
  -c treeFile [treeFile ...], --clean treeFile [treeFile ...]
                             prepare tree file(s) for analysis
  -s scalingFile, --scalingFile scalingFile
                             provide a text file with each scaling factor on a new
                             line in the format: [scalingFactor]. Values prefixed
                             in the order that trees appear after the clean
                             command, DEFAULT: [1.0] for each
  -t traitsFile, --traits traitsFile
                             read in traits file as Beast format
  -bt thetaValue, --theta thetaValue
                             set theta value for Beast formatted traits files,
                             DEFAULT: 1.0
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py setup -c cpDNA.tre
Sa4.tre Sa135.tre Sa163.tre Sa297.tre Sa302.tre Sa323.tre Sa405.tre -s scaling.
txt -t traits.txt -bt 0.123
Namespace(clean=['cpDNA.tre', 'Sa4.tre', 'Sa135.tre', 'Sa163.tre', 'Sa297.tre',
'Sa302.tre', 'Sa323.tre', 'Sa405.tre'], command='setup', scalingFile=['scaling.t
xt'], theta=[0.123], traits=['traits.txt'])

#####
##### Performing Setup #####
#####

Cleaning Phylip tree...
Parsing Beast traits file for Species, Traits, and Grouping
dhcp-254-235-238:SpedeSTEM-master bcarstens$ cat cleaned.*.tre > genetrees.tre
dhcp-254-235-238:SpedeSTEM-master bcarstens$
```

Discovery: available on web based server

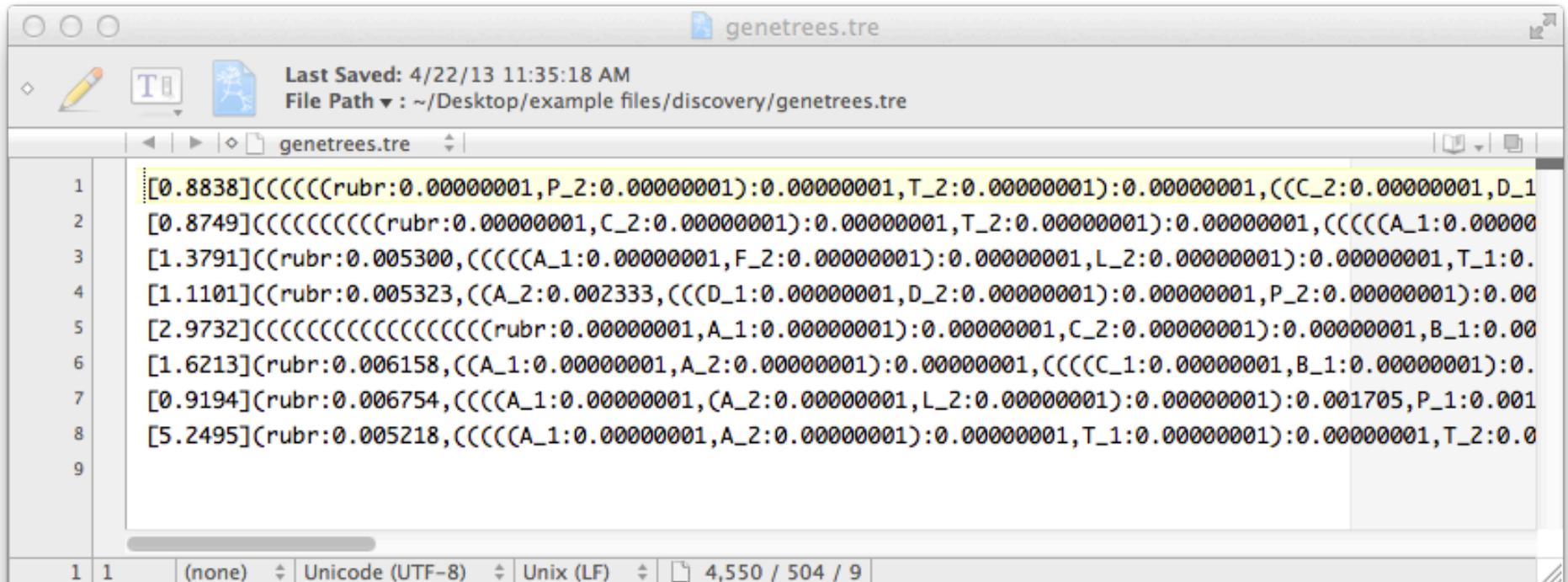
The screenshot shows a web browser window with the following elements:

- Browser Tab:** SpedeSTEM Web Interface
- Address Bar:** `spedestem.asc.ohio-state.edu/runstem/discovery`
- Navigation:** Back, Forward, Refresh, Home icons.
- Page Header:** SpedeSTEM Online | Run SpedeSTEM | Resources | Contact Us
- Main Content:**
 - Discovery Section:**
 - Settings file: settings.disco
 - Gene tree file: genetrees.tre
 - Email:
 -
 - About us Section:**
 - About us**
 - Research in the Carstens lab seeks to understand how biological diversity is generated using computational approaches. We investigate empirical systems by identifying the limits of evolutionary lineages, in order to evaluate the relative contributions of evolutionary processes and infer the ecological and environmental forces that have contributed to the formation of population genetic structure.
- Footer:**
 -  Ohio State University
Department of Ecology Evolution and Organismal Biology
 - Download area:

Discovery Required files

Settings file. Discovery analyses require a settings file formatted in the setup step.

Genetrees file. A gene trees files is also required. This file includes a cleaned ultrametric tree for each locus preceded by a scaling factor. This file is the same as the one used in the validation step.



Discovery

- navigate to spedeSTEM directory

`cd [path to directory]`

- access help by entering:

`./SpedeSTEM_2.py discovery -h`

- generate settings file by entering:

`./SpedeSTEM_2.py discovery -t
genetrees.tre -s settings.disco`

- Output: species trees and likelihoods in **results.txt** and information theoretic table in **itTable.txt**.

```
SpedeSTEM-master — bash — 80x48
Last login: Mon Apr 22 10:55:52 on ttys001
dhcp-254-235-238:~ bcarstens$ cd /Users/bcarstens/spedeSTEM/SpedeSTEM-master
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py discovery -h
usage: SpedeSTEM_2.py discovery [-h] -t treeFile -s settingsFile [-v]

optional arguments:
  -h, --help            show this help message and exit
  -t treeFile, --tree treeFile
                        specify tree file
  -s settingsFile, --settings settingsFile
                        specify settings file in STEM format
  -v, --verbose         execute in verbose mode, DEFAULT: off
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py discovery -t genetrees.tre -s settings.disco
Namespace(command='discovery', settings=['settings.disco'], tree=['genetrees.tre'], verbose=False)

#####
##### DISCOVERY ANALYSIS #####
#####

----- CAUTION -----
stemOut.txt is about to be deleted.  If you would like to preserve it, remove it
from this directory.
Are you ready to continue? (y/n): y
-----

----- SETTINGS -----
In Varification Mode: False
Tree File: genetrees.tre.save
Settings File: settings.disco
Associations File: associations.txt
Number of loci sampled each replicate: 8
Number of replicates: 1
In Verbose Mode: False
----- END SETTINGS -----

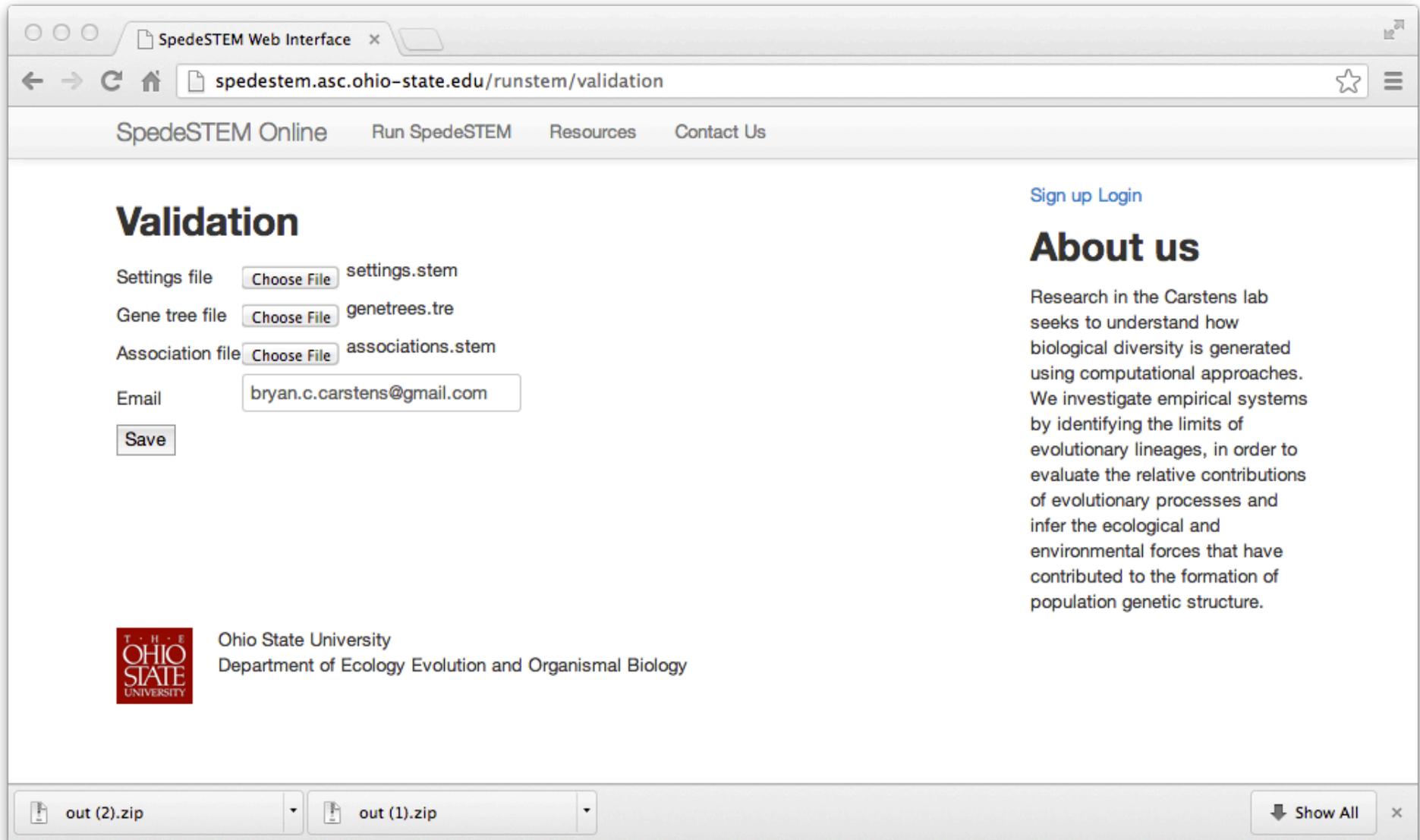
----- EXECUTION -----
Sampling 8 loci from master tree file genetrees.tre.save...
Completed 1 of 1 replicates...

+++++ All analysis completed! +++++

Completing Analysis... See 'results.txt' and 'itTable.txt' files

dhcp-254-235-238:SpedeSTEM-master bcarstens$
```

Validation: available on web based server



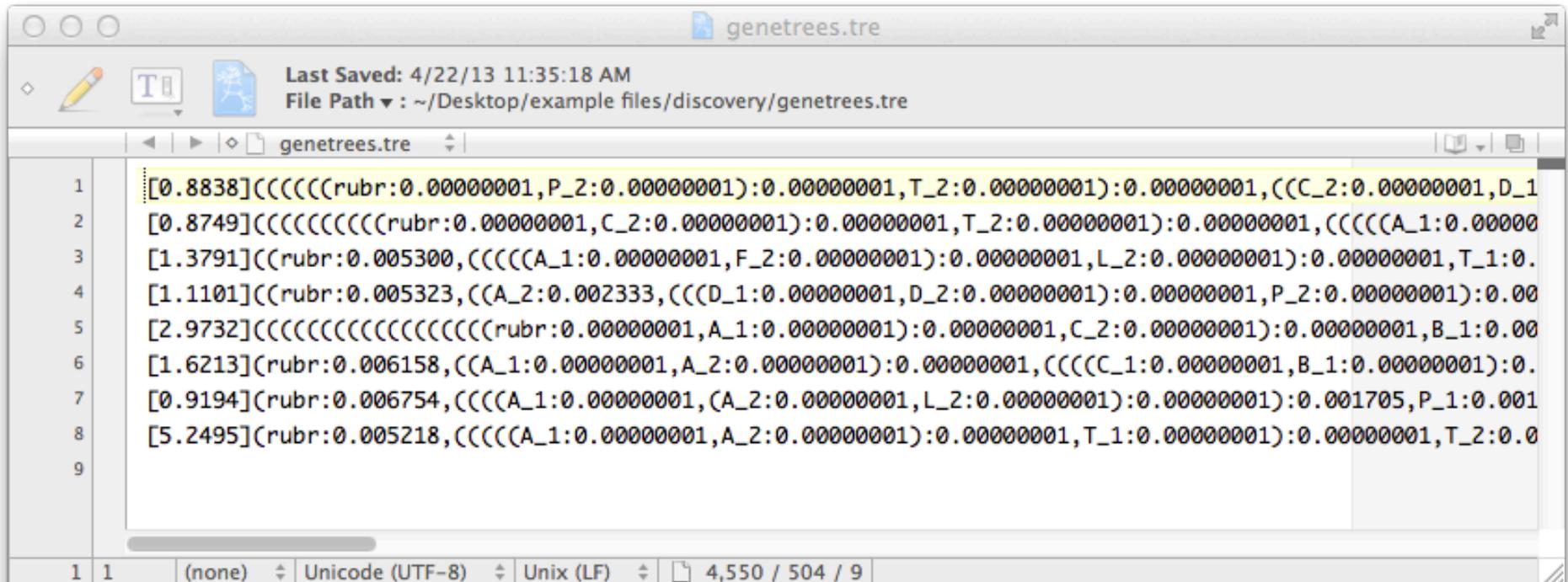
The screenshot shows a web browser window with the following elements:

- Browser Tab:** SpedeSTEM Web Interface
- Address Bar:** `spedestem.asc.ohio-state.edu/runstem/validation`
- Navigation:** Back, Forward, Refresh, Home icons.
- Page Header:** SpedeSTEM Online | Run SpedeSTEM | Resources | Contact Us
- Main Content:**
 - Validation Section:**
 - Settings file: settings.stem
 - Gene tree file: genetrees.tre
 - Association file: associations.stem
 - Email:
 -
 - About us Section:**
 - [Sign up](#) [Login](#)
 - ## About us
 - Research in the Carstens lab seeks to understand how biological diversity is generated using computational approaches. We investigate empirical systems by identifying the limits of evolutionary lineages, in order to evaluate the relative contributions of evolutionary processes and infer the ecological and environmental forces that have contributed to the formation of population genetic structure.
- Footer:**
 -  Ohio State University
Department of Ecology Evolution and Organismal Biology
 - Download area:

Validation Required files

Settings file. Discovery analyses require a settings file formatted in the setup step.

Genetrees file. A gene trees files is also required. This file includes a cleaned ultrametric tree for each locus preceded by a scaling factor – this file is the same as the one used in the Discovery step.



```
1 [0.8838]((((((rubr:0.00000001,P_2:0.00000001):0.00000001,T_2:0.00000001):0.00000001,((C_2:0.00000001,D_1
2 [0.8749]((((((((rubr:0.00000001,C_2:0.00000001):0.00000001,T_2:0.00000001):0.00000001,(((A_1:0.00000
3 [1.3791]((rubr:0.005300,(((A_1:0.00000001,F_2:0.00000001):0.00000001,L_2:0.00000001):0.00000001,T_1:0.
4 [1.1101]((rubr:0.005323,((A_2:0.002333,((D_1:0.00000001,D_2:0.00000001):0.00000001,P_2:0.00000001):0.00
5 [2.9732]((((((((((((((((rubr:0.00000001,A_1:0.00000001):0.00000001,C_2:0.00000001):0.00000001,B_1:0.00
6 [1.6213](rubr:0.006158,((A_1:0.00000001,A_2:0.00000001):0.00000001,(((C_1:0.00000001,B_1:0.00000001):0.
7 [0.9194](rubr:0.006754,(((A_1:0.00000001,(A_2:0.00000001,L_2:0.00000001):0.00000001):0.001705,P_1:0.001
8 [5.2495](rubr:0.005218,(((A_1:0.00000001,A_2:0.00000001):0.00000001,T_1:0.00000001):0.00000001,T_2:0.0
9
```

validation

- navigate to spedeSTEM directory

`cd [path to directory]`

- access help by entering:

`./SpedeSTEM_2.py validation -h`

- generate settings file by entering:

`./SpedeSTEM_2.py validation -t
genetrees.tre -s settings.stem -a
associations.stem`

- Output: species trees and likelihoods in **results.txt** and information theoretic table in **itTable.txt**.

```
SpedeSTEM-master — bash — 80x52
dhcp-254-235-238:~ bcarstens$ cd /Users/bcarstens/spedeSTEM/SpedeSTEM-master
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py validation -h
usage: SpedeSTEM_2.py validation [-h] -t treeFile -s settingsFile -a
associationsFile [-v]

optional arguments:
  -h, --help            show this help message and exit
  -t treeFile, --tree treeFile
                        specify tree file
  -s settingsFile, --settings settingsFile
                        specify settings file in STEM format
  -a associationsFile, --associations associationsFile
                        specify associations file in STEM format
  -v, --verbose         execute in verbose mode, DEFAULT: off
dhcp-254-235-238:SpedeSTEM-master bcarstens$ ./SpedeSTEM_2.py validation -t gene
trees.tre -s settings.stem -a associations.stem
Namespace(associations=['associations.stem'], command='validation', settings=['s
ettings.stem'], tree=['genetrees.tre'], verbose=False)

#####
##### VALIDATION ANALYSIS #####
#####

----- CAUTION -----
stemOut.txt is about to be deleted.  If you would like to preserve it, remove it
from this directory.
Are you ready to continue? (y/n): y
-----

----- SETTINGS -----
In Verification Mode: True
Tree File:  genetrees.tre.save
Settings File:  settings.stem
Associations File:  associations.stem
Number of loci sampled each replicate:  8
Number of replicates:  1
In Verbose Mode:  False
----- END SETTINGS -----

----- EXECUTION -----
Sampling 8 loci from master tree file genetrees.tre.save...
  25 permutations to run...
    Completed 10 of 25 permutations...
    Completed 20 of 25 permutations...
  Completed 1 of 1 replicates...

+++++ All analysis completed! +++++

Completing Analysis... See 'results.txt' and 'itTable.txt' files

dhcp-254-235-238:SpedeSTEM-master bcarstens$
```

spedeSTEM: subsampling

- Takes as input a settings and genetrees file.