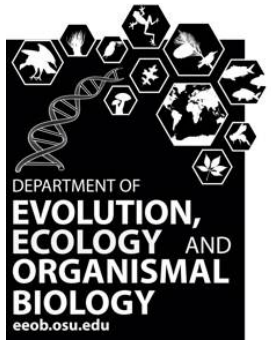


spedeSTEM tutorial

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.

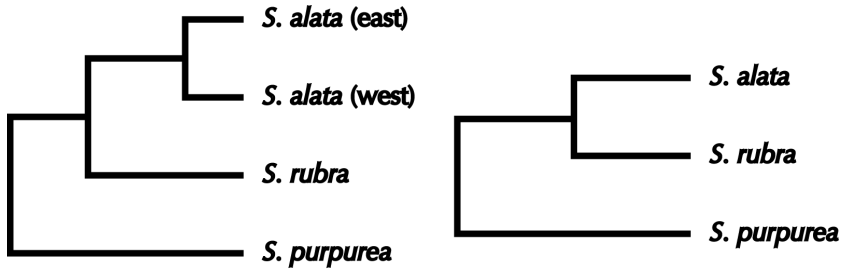
spedeSTEM: Introduction

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 $f(g_j \mid S, \tau)$ 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. Carstens[2] and L. Lacey Knowles[3]

[1]Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, [2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 and [3]Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

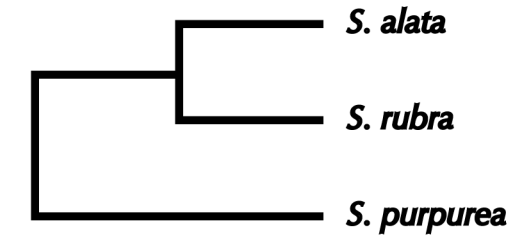
Received on November 28, 2008; revised and accepted February 04, 2009

Associate Editor: Martin Bishop

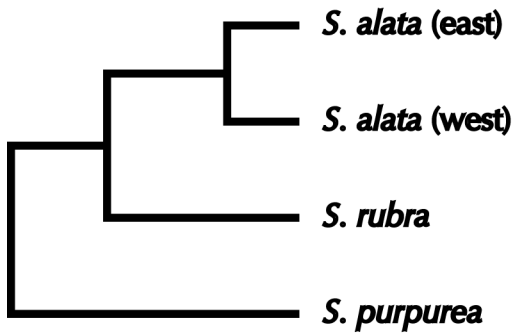
$$L(S,\tau) = \prod_{j=1}^{N} f(g_j|S,\tau)$$

spedeSTEM: background

-200.492



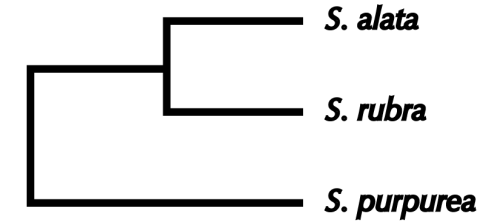
-198.737

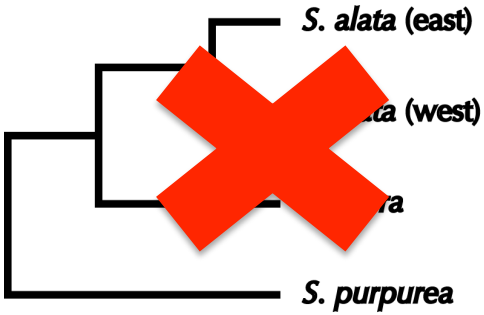


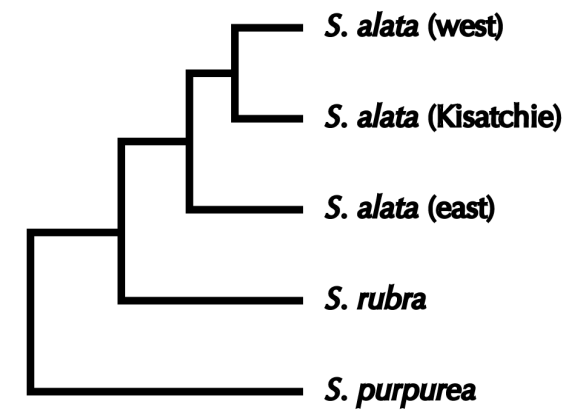
If we compare our two *a priori* models using a Likelihood ratio test, we find that we can not reject the model where all of *S. alata* constitutes a single evolutionary lineage ($\chi^2 = 3.512$, $p = 0.0609$)

spedeSTEM: background

-200.492





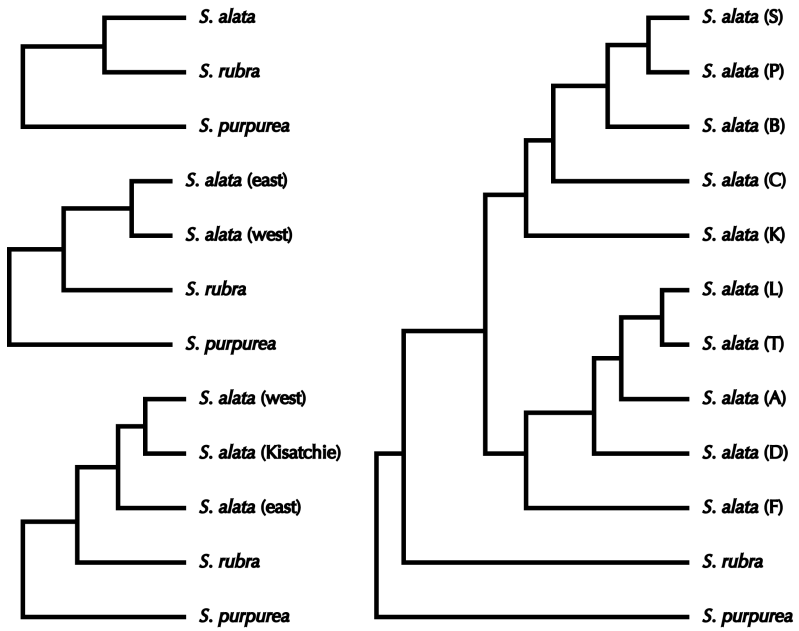


-188.186

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 (χ^2 = 24.054, p = < 0.0001)

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

spedeSTEM: background

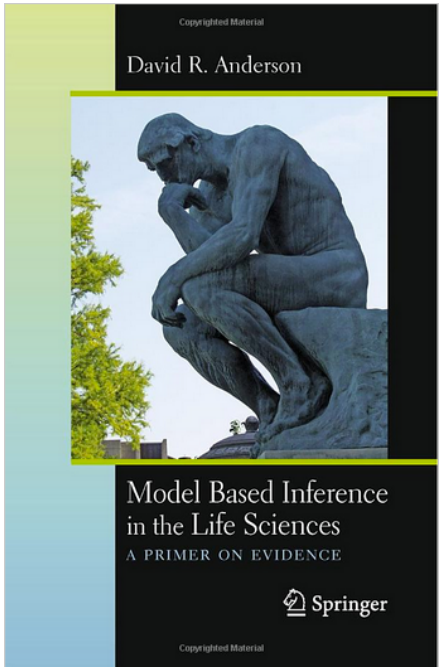


What is the appropriate null model for species delimitation? If we assume, geography and try to delimit 2 lineages, we do not reject the null. If we treat the results of the Structure analysis as the alternative hypothesis, then we do reject the null. Which of these is appropriate? It could be that neither 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

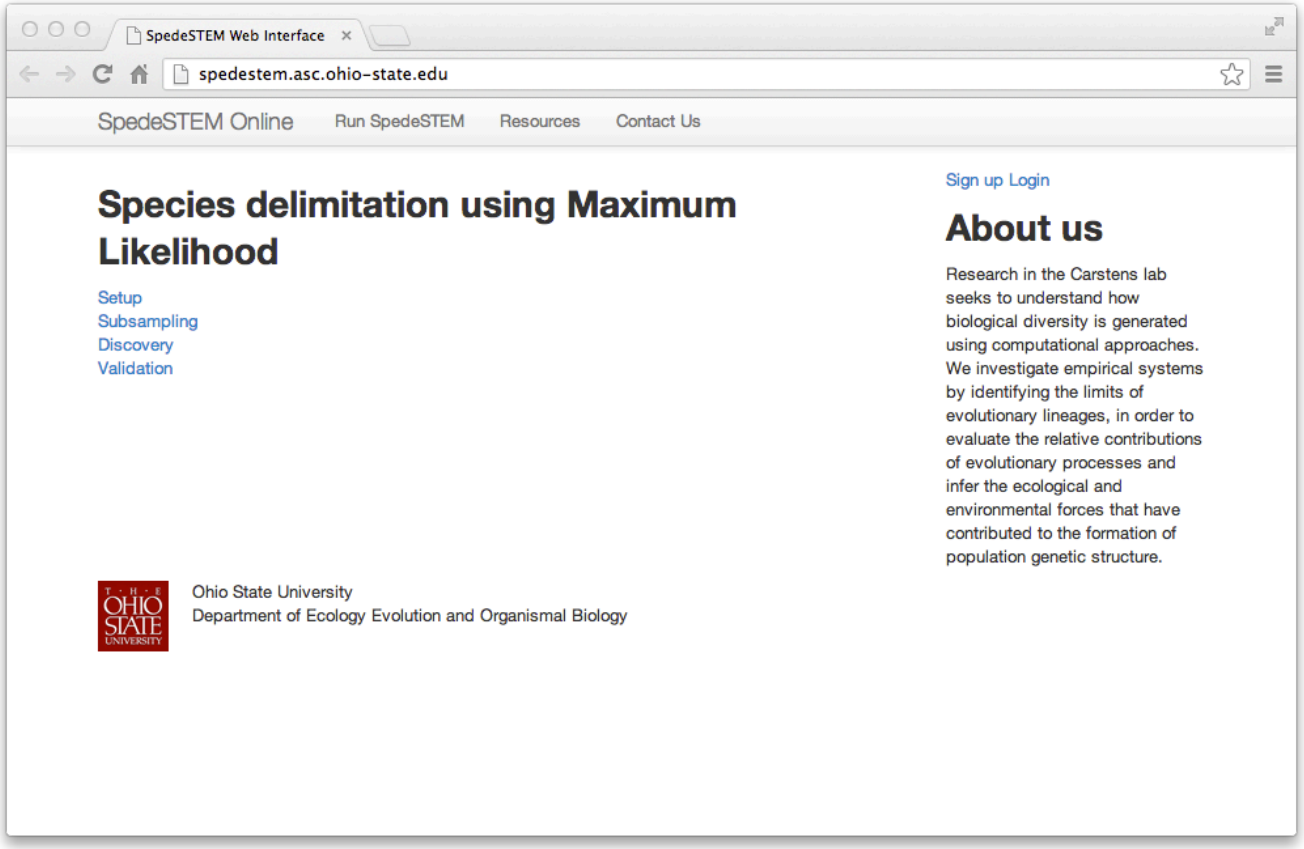
- spedeSTEM does not rely on hypothesis testing, and can consider any number of possible models (in theory). In practice, we have successfully considered >20000 for 10 putative lineages.
- Information theory is used to evaluate the probability of models (see Anderson 2008 for a readable introduction 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/

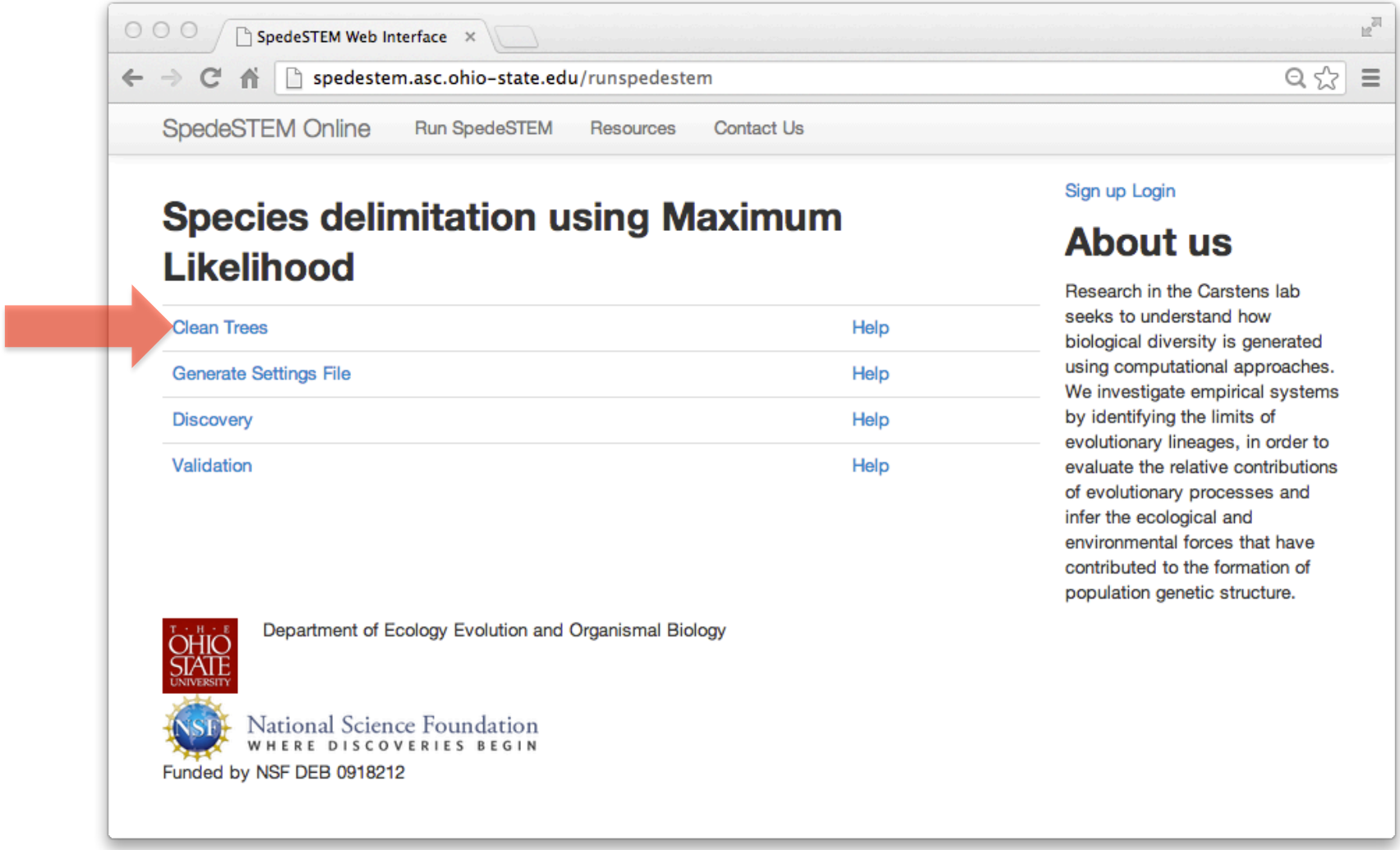
spedeSTEM: command line

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

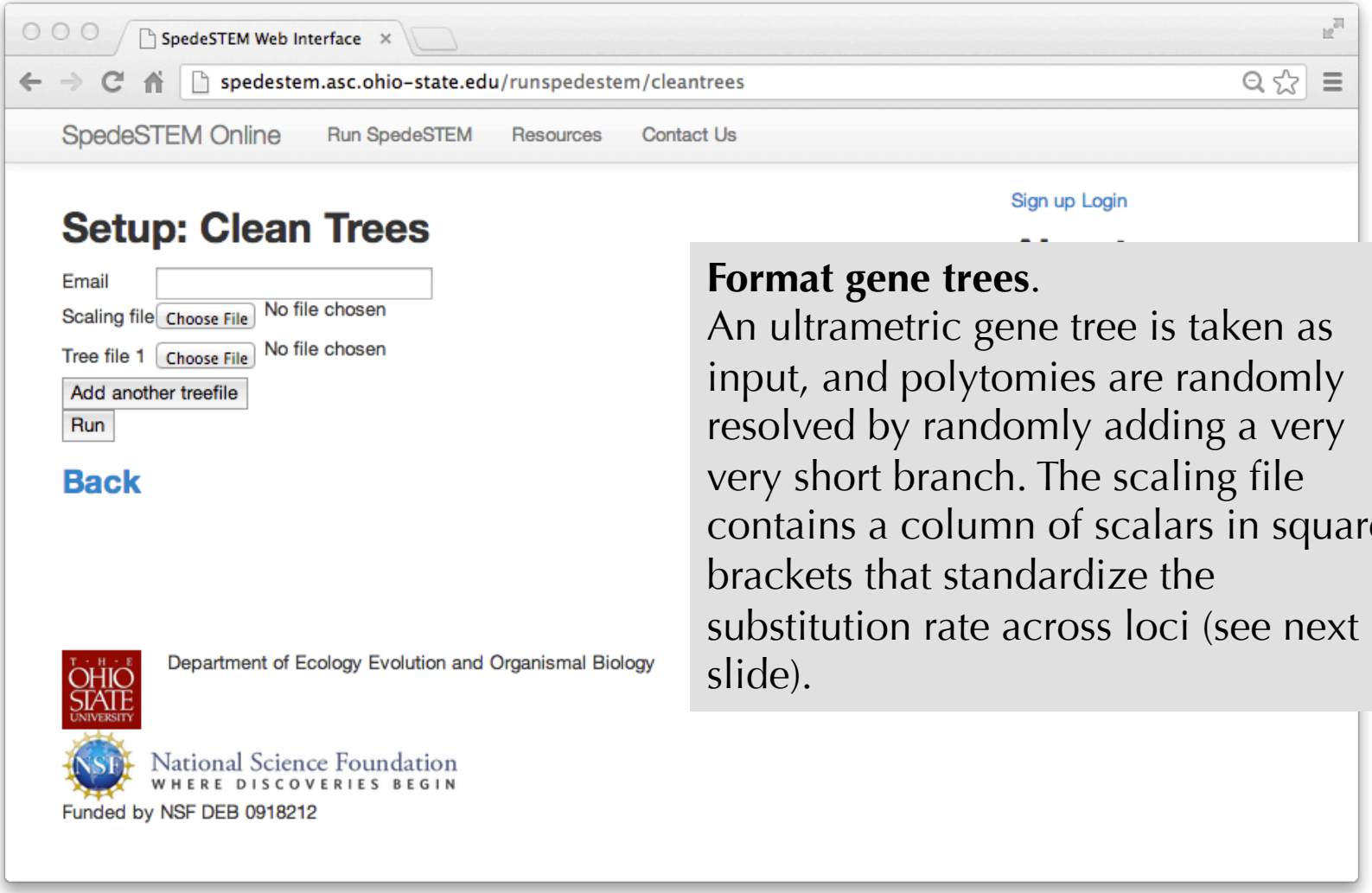
  http://carstenslab.org.ohio-state.edu/software.html

- It contains all example files used here.

Step 1: Clean your genetrees.

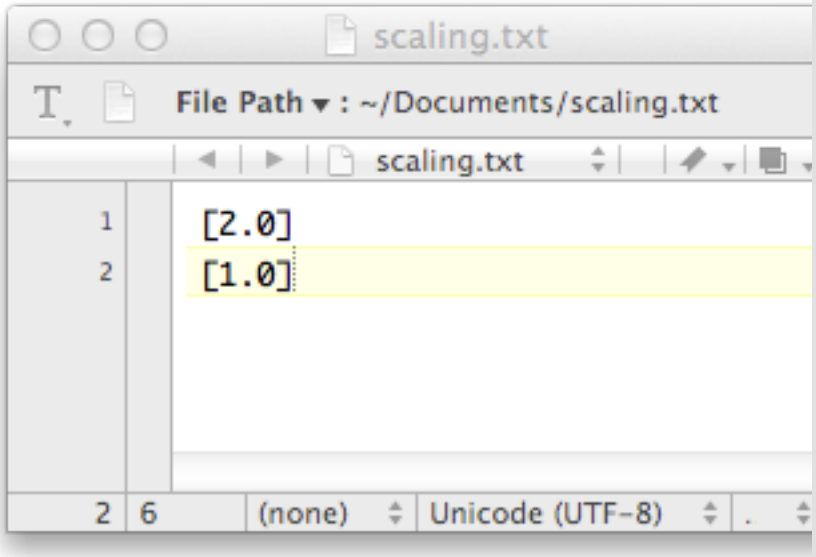


Step 1: Clean your genetrees.



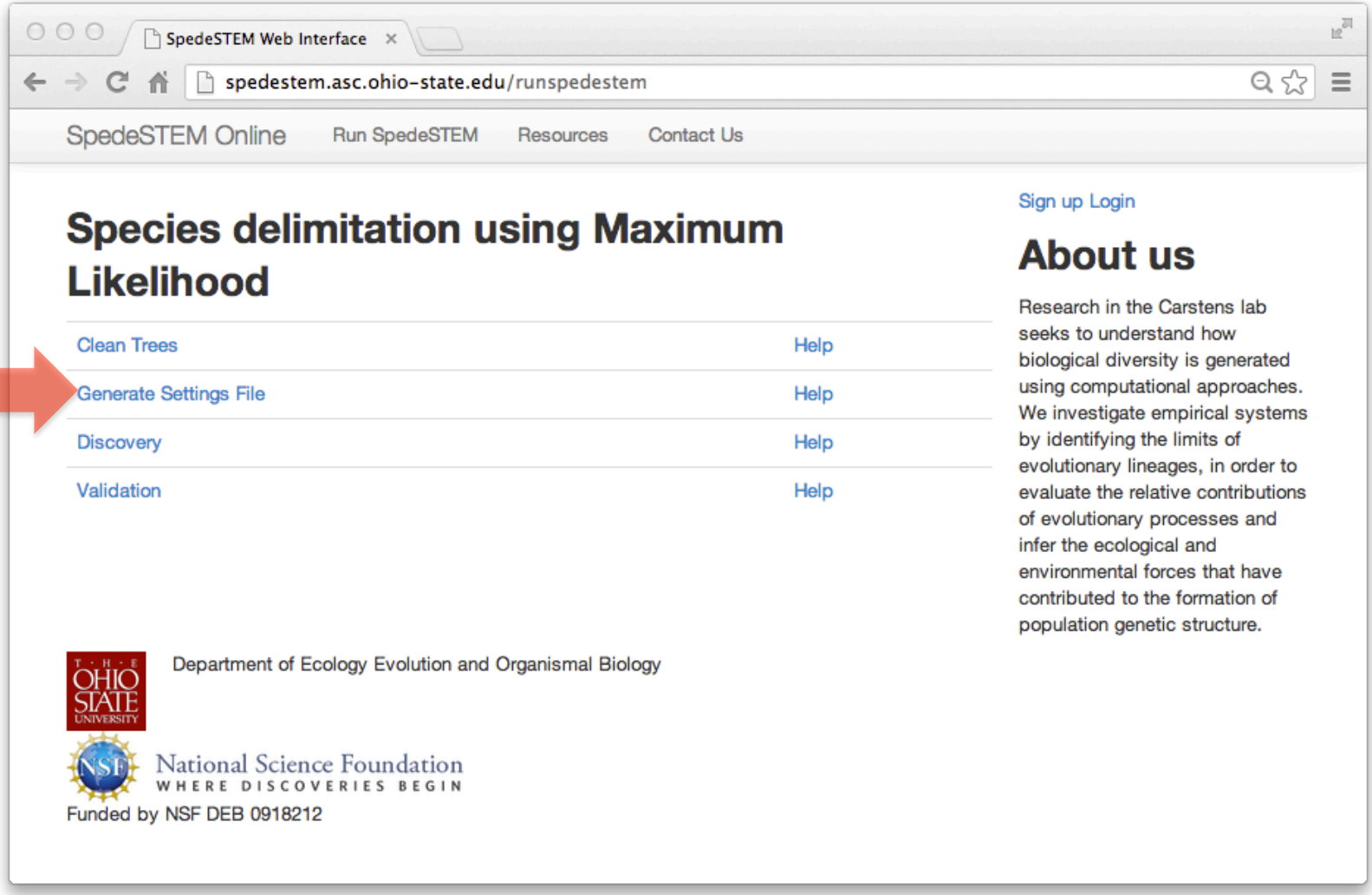
Format gene trees.
An ultrametric gene tree is taken as input, and polytomies are randomly resolved by randomly adding a very very short branch. The scaling file contains a column of scalars in square brackets that standardize the substitution rate across loci (see next slide).

Scaling Factors File

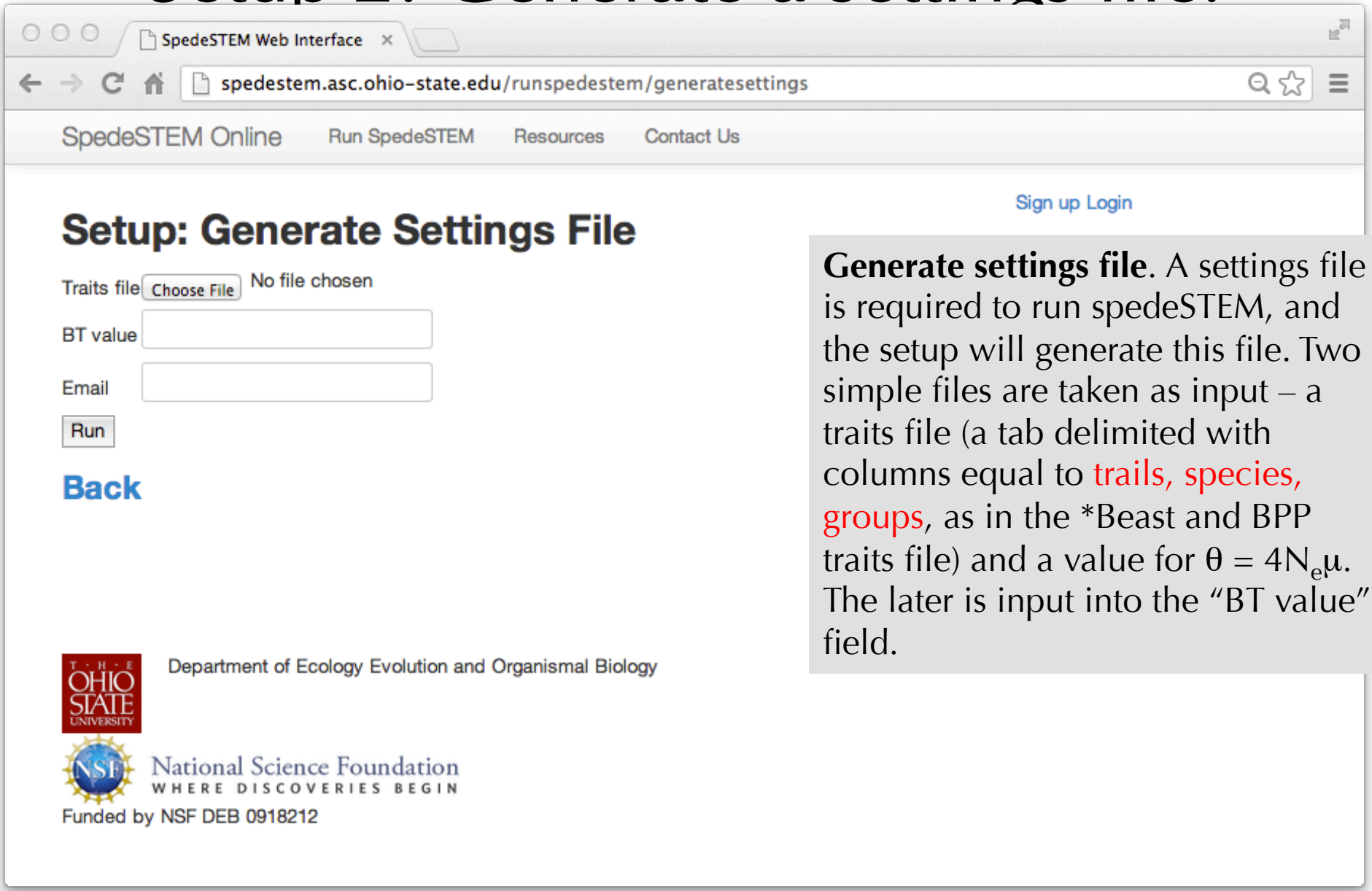


Scaling file. A scaling file is required to run spedeSTEM – this is simply a text file with scaling factors for each locus. Loci should be in the same order as entered. Scaling factors are used to normalize the length of the gene trees. Let the scaling factor of each locus be a number that normalizes 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*(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.



Setup 2: Generate a settings file.



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:

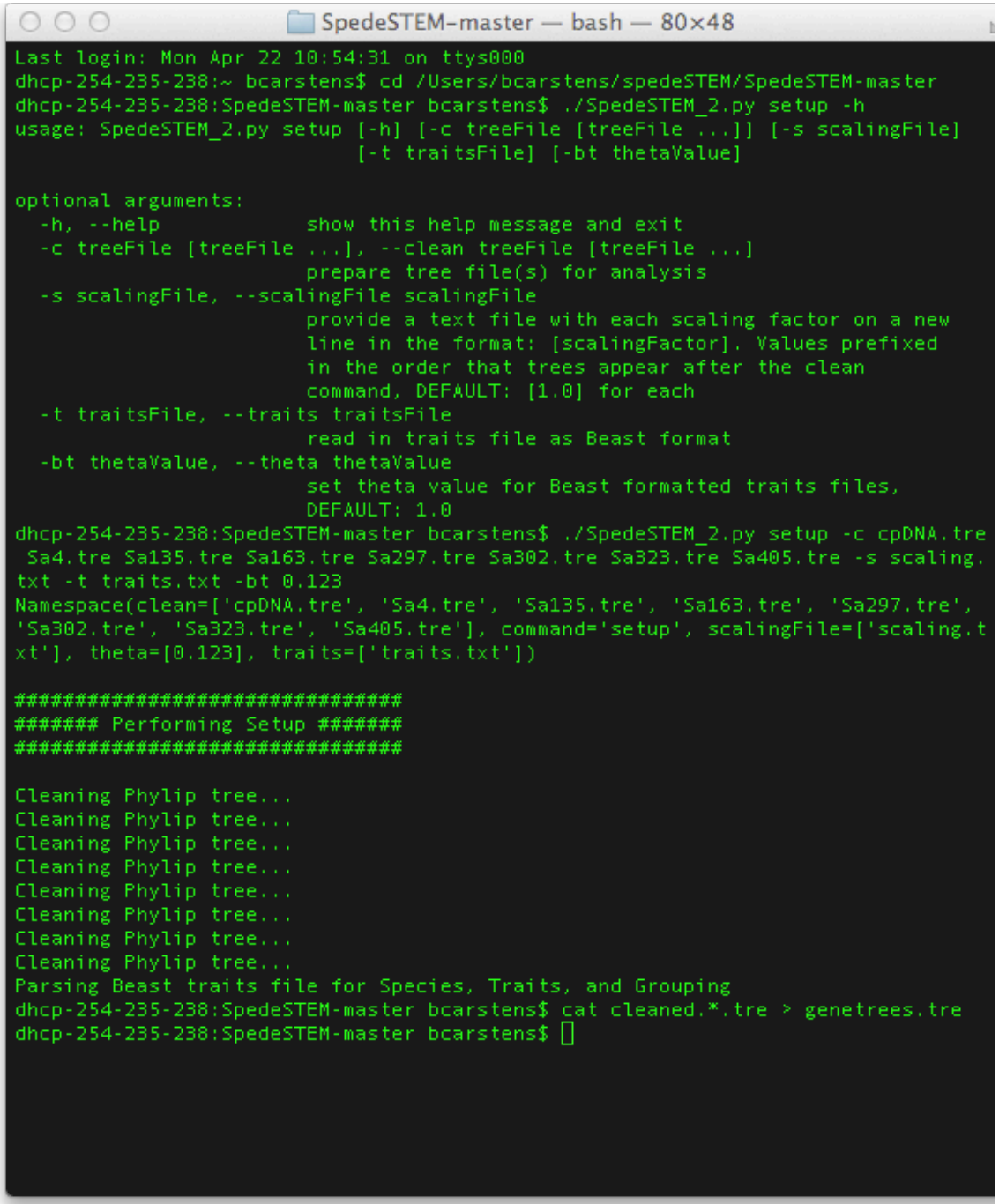
```
./SpedeSTEM_2.py setup –h
```

- generate settings file by entering:

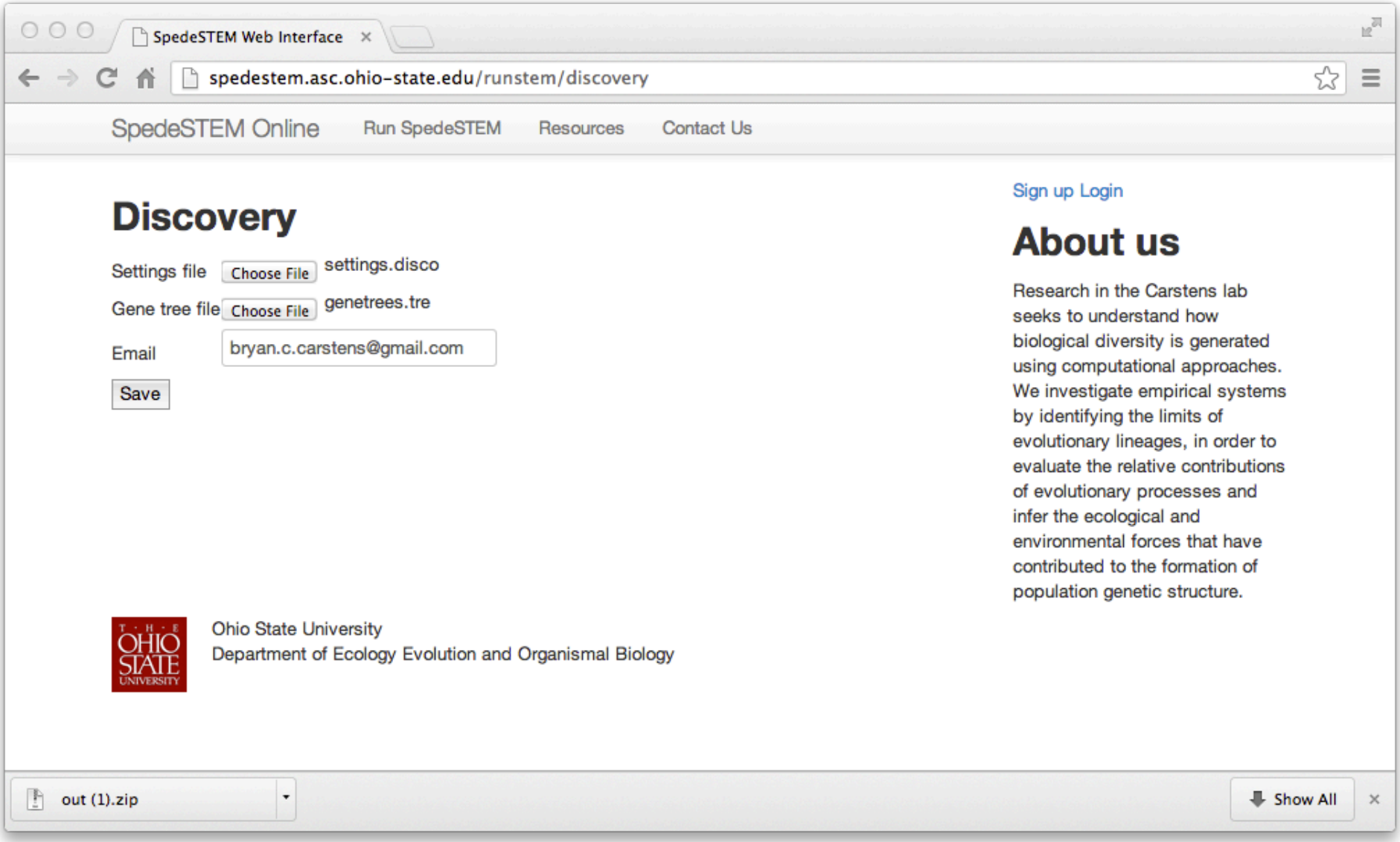
```
./SpedeSTEM_2.py setup -c
cpDNA.tre Sa4.tre Sa135.tre
Sa163.tre Sa297.tre Sa302.tre
Sa323.tre Sa405.tre -s
scaling.txt -t traits.txt -bt
0.123
```

- generate genetrees file by entering:

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



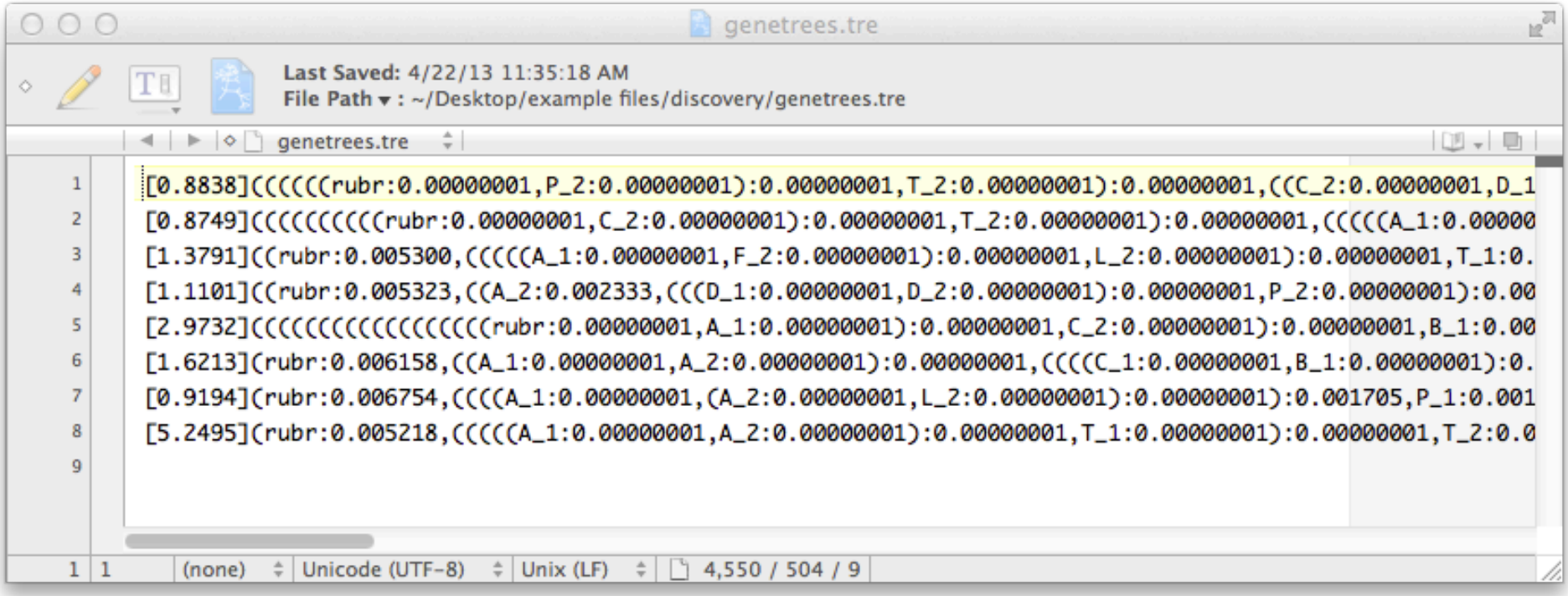
Discovery: available on web based server



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]
```

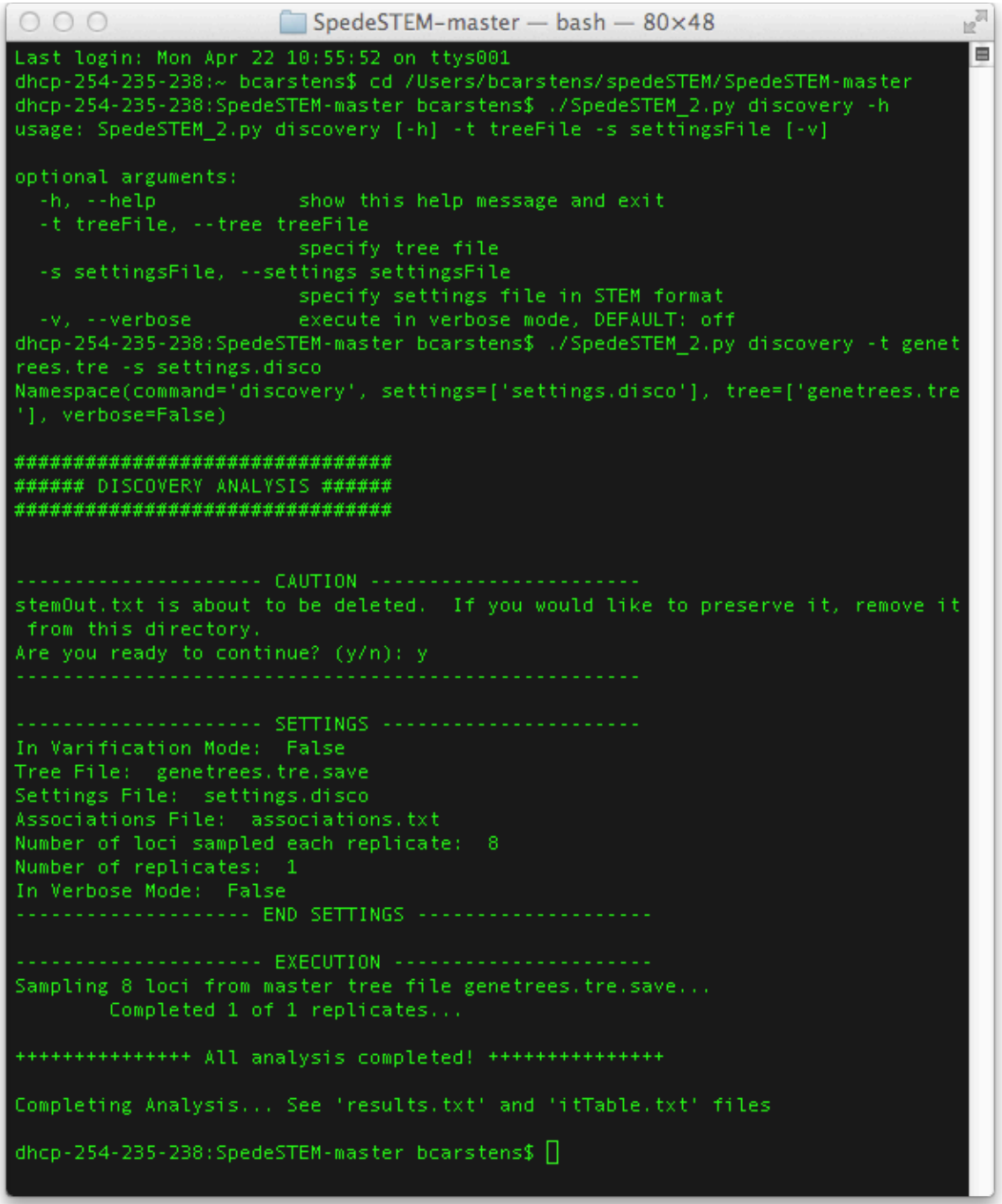
- access help by entering:

```
./SpedeSTEM_2.py discovery –h
```

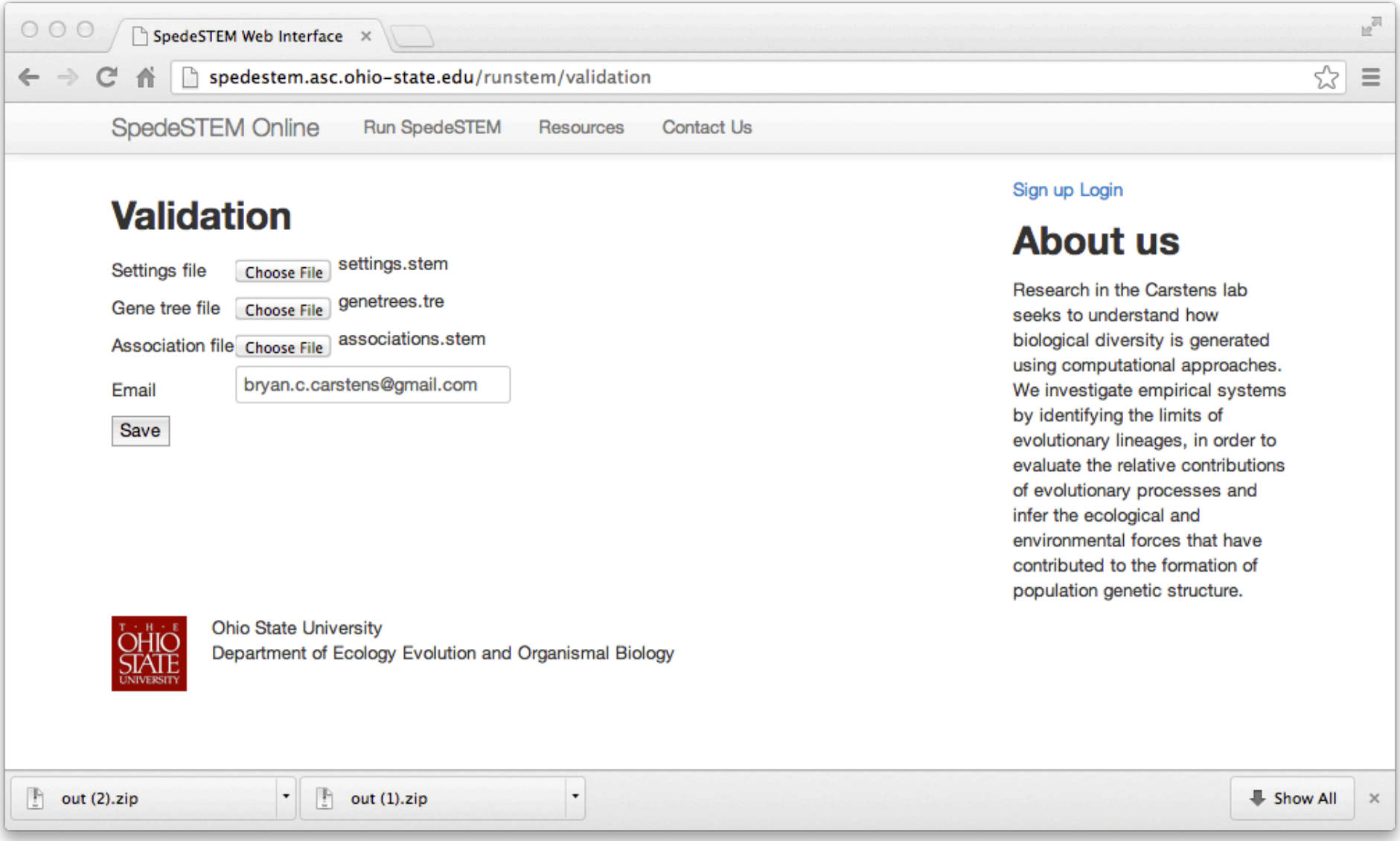
- generate settings file by entering:

```
./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***.



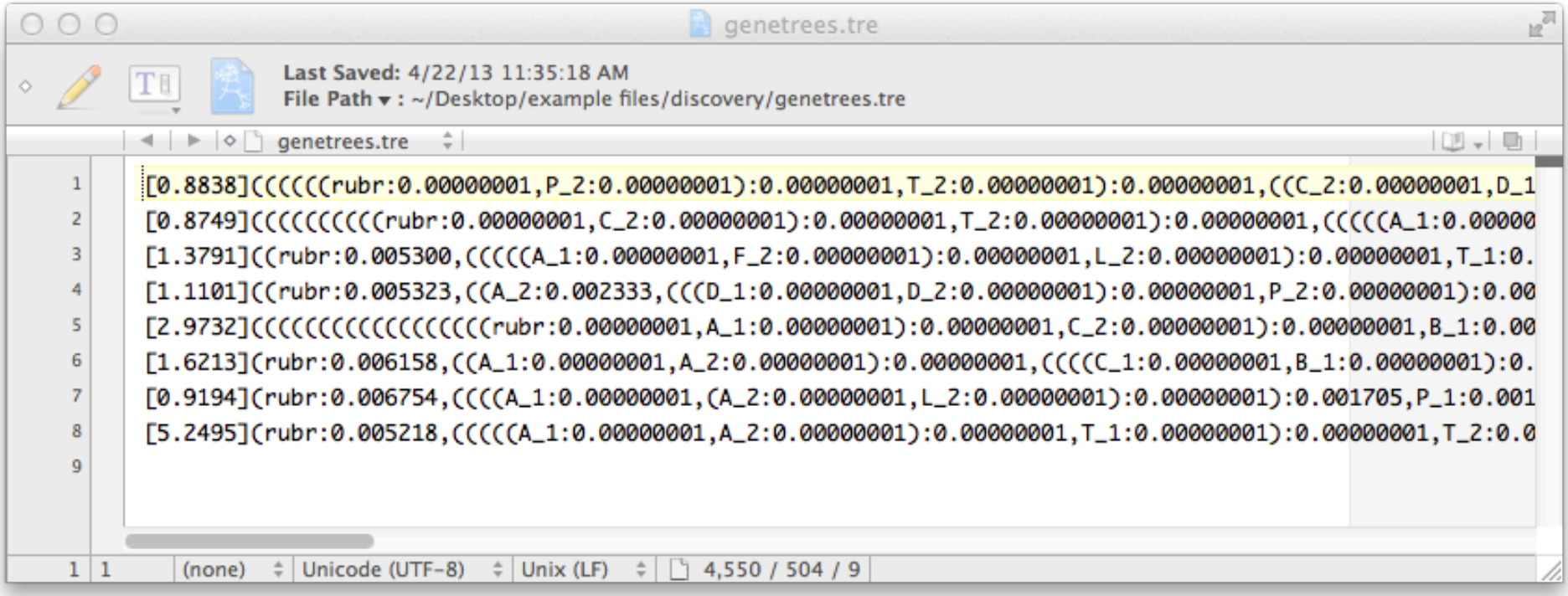
Validation: available on web based server



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.



validation

- navigate to spedeSTEM directory

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

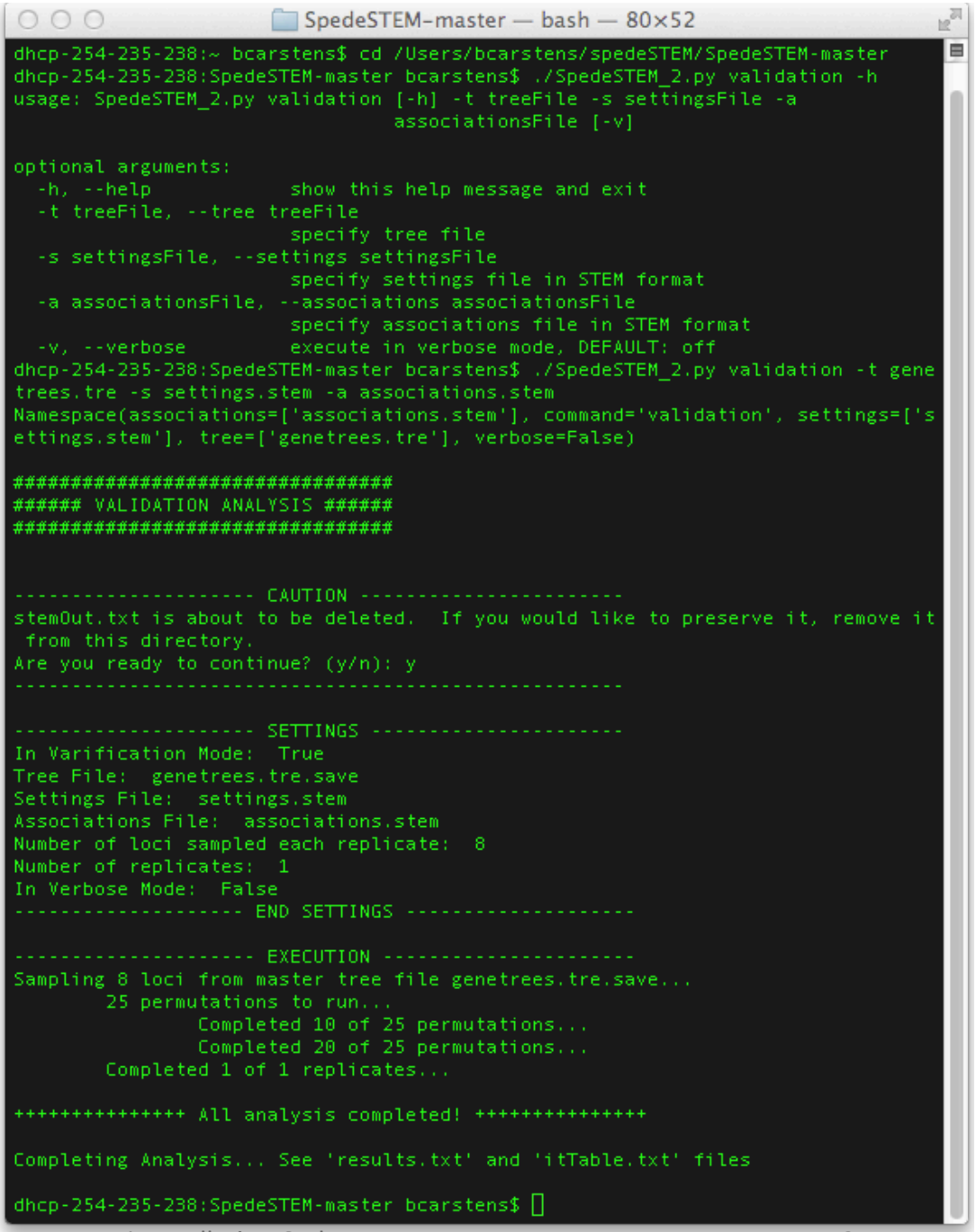
- access help by entering:

```
./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: subsampling

- Takes as input a settings and genetrees file.