

Processing ddRAD for population history inference

April Wright

ISU and KU

01-06-2016

- Reduced-representation genomic method

- Reduced-representation genomic method
- Cheap

- Reduced-representation genomic method
- Cheap
- Lots of data returned

- Reduced-representation genomic method
- Cheap
- Lots of data returned
- Stable software pipelines for using these data

A Quick Note

Slides that contain ddRAD specific info will be noted. Some steps can be used with multiple data sources.

Our Study

The Edwards Plateau

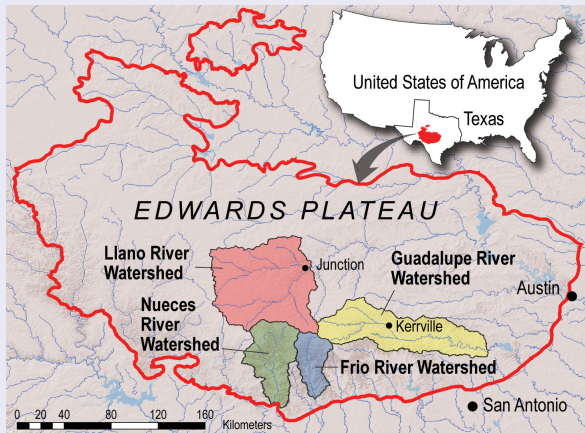


Image: AGU

Our Study



Our Study

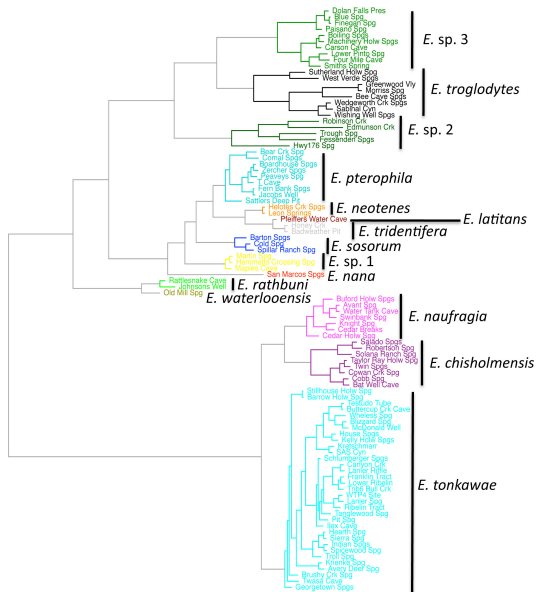
13 putative species of *Eurycea*

Our Study

13 putative species of *Eurycea*

All of which are fairly threatened by development

Our Study



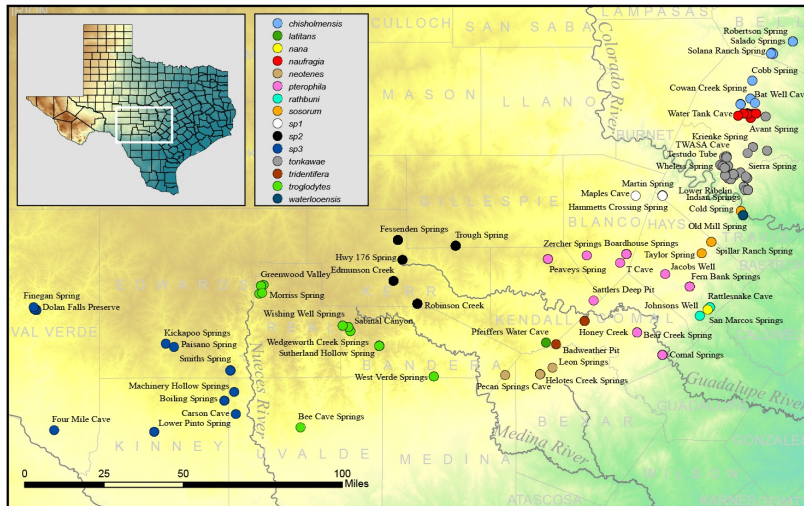
- 100 nucleotide changes

How many species of *Eurycea* are there, really?

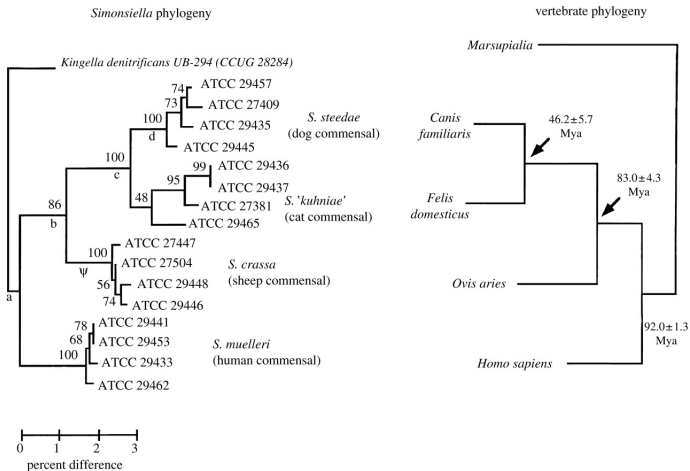
Our Study

How many species of *Eurycea* are there, really?
And is there introgression between them?

Our Study



Phylogenetics



JP Staley, 2006. Figure 2

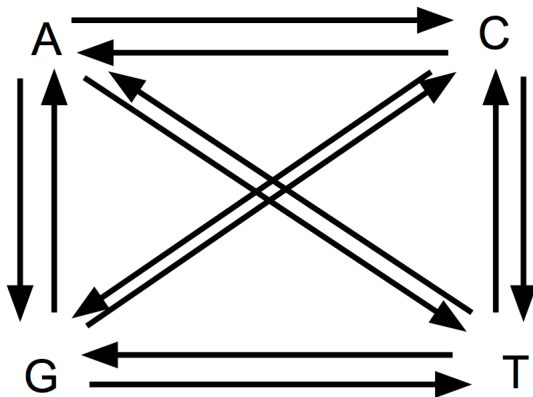
Maximum likelihood

Maximum likelihood is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data

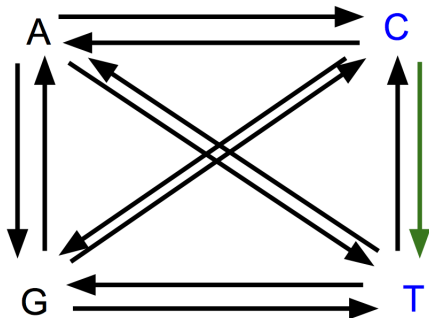
- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions

Phylogenetics



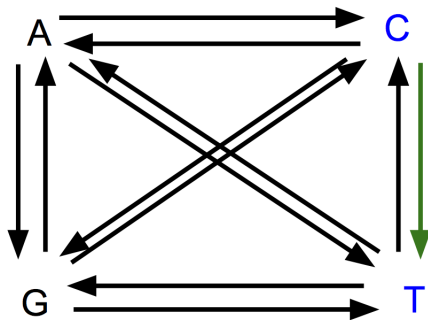
Phylogenetics



Probability of C to T change
Equilibrium frequency of C

$$0 * .25 = 0$$

Phylogenetics



Probability of C to T change

Equilibrium frequency of C

$$.75 * .25 = .1875$$

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions
- Statistically consistent

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions
- Statistically consistent: When we use the true model to analyze our data, we will eventually converge to the true answer as more data is added

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions
- Statistically consistent: When we use the true model to analyze our data, we will eventually converge to the true answer as more data is added
 - And because we are making mathematically-defined assumptions, we can use model-fitting to find the "true" model

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions
- Statistically consistent: When we use the true model to analyze our data, we will eventually converge to the true answer as more data is added
 - And because we are making mathematically-defined assumptions, we can use model-fitting to find the "true" model
- Superimposed changes

- **Maximum likelihood** is a framework for estimating phylogeny by modeling the process of evolution that generated our sequence data
- Model-based: We make mathematically explicit assumptions
- Statistically consistent: When we use the true model to analyze our data, we will eventually converge to the true answer as more data is added
 - And because we are making mathematically-defined assumptions, we can use model-fitting to find the "true" model
- Superimposed changes

- **Problems**

- **Problems**
- Missing data

- **Problems**
- **Biased Missing data**

Species 1	A	A	G	?	G	A	G	A	G
Species 2	G	?	C	A	C	?	C	?	C
Species 3	C	C	T	?	T	T	?	T	T
Species 4	T	G	A	T	A	?	T	C	?
Species 5	A	T	G	C	G	C	A	G	A

- **Problems**
- **Biased** Missing data

- Missing data concentrated in specific individuals

Species 1	?	?	G	G	G	A	G	A	G
Species 2	?	?	C	A	C	C	C	G	C
Species 3	?	?	T	G	T	T	C	T	T
Species 4	?	?	A	T	A	T	T	C	G
Species 5	A	T	G	C	G	C	A	G	A

Phylogenetics

- Missing data concentrated in specific individuals
- Missing data concentrated in certain loci in your data matrix

Species 1	?	?	?	?	?	?	?	A	G
Species 2	?	?	?	?	?	?	?	G	C
Species 3	G	C	T	G	T	T	C	T	T
Species 4	T	G	A	T	A	T	T	C	G
Species 5	A	T	G	C	G	C	A	G	A

- **Problems**
- Model misspecification

- **Problems**

- Model misspecification: when your data are not adequately described by your model

Today, we'll be visualizing our data at every step to try and minimize a bias in which individuals have missing data

We'll also look at ways to make sure we aren't overly-conservative in our choosing of SNPs (i.e., biasing our collection towards sites that exhibit little change)

The Demultiplex

One of the things that makes RADseq, and especially ddRADseq, so cheap is the pooling of samples

The Demultiplex

One of the things that makes RADseq, and especially ddRADseq, so cheap is the pooling of samples

The way we recover individual samples is via demultiplexing

The Demultiplex

This allows for the cost-saving properties of batching, without the cost-increasing properties of synthesizing oligonucleotides.

The Demultiplex

The STACKS step for this is called **Process RAD Tags**

The Demultiplex

Output

- FASTQ files

The Demultiplex

Let's look at the output

- FASTQ files
- Reads, grouped by individual

The Demultiplex

Let's look at the output

- FASTQ files
- Reads, grouped by individual
- We haven't done any SNP calling. This is just the step that gets our data ready to do that

Initial Identification of SNPs

The STACKS step for this is called **ustacks**

Initial Identification of SNPs

Each RAD tag has usually been sequenced multiply per-individual

Initial Identification of SNPs

Each RAD tag has usually been sequenced multiply per-individual
This allows us to sort tags into "stacks" of identical and unique reads

Initial Identification of SNPs

Each RAD tag has usually been sequenced multiply per-individual
This allows us to sort tags into "stacks" of identical and unique reads
From these sets of identical and unique reads, we do a first pass at identifying SNPs.

Key Parameters

- -m: Minimum stack depth
- -M: Maximum mismatches allowed between reads in a stack

Other Parameters

- -i: ID for this sample

One of the issues we discussed was biased missing data

Once we have our within-individual stacks, we build a catalog of loci across individual catalogs (**cstacks**)

Key Parameters

- -n: number of mismatches to allow between a putative tag, and a tag in the catalog

Outputting Data for Phylogenetics

The STACKS step for this is called **populations**

Key Parameters

- -r: Percentage of individuals in a population that must have a locus to output it
- -m: Minimum stack depth at a locus

Let's look at this output

Let's look at this output

But we can also look in a more complex way: `countPhyloMissing.sh` and `plotPhyloMissing.py`

RAxML approximate likelihoods

RAxML	
Dataset	L score
323320	-288336.664115
323340	-84377.460743
323360	-27407.770692
323380	-10281.525371
323390	-1699.210794

Lastly, let's build the tree

RAxML Approximate final L scores

Lastly, let's build the tree

Garli

Lastly, let's build the tree

Garli

Dataset	L score
323320	-287898.7874
323340	-84289.0263
323360	-27384.6012
323380	-10273.63021
323390	-1697.6284

Why

Tips	Number of unrooted (binary) trees	
4	3	
5	15	
6	105	
7	945	
8	10,395	
9	135,135	
10	2,027,025	
11	34,459,425	
12	654,729,075	
13	13,749,310,575	
14	316,234,143,225	
15	7,905,853,580,625	
16	213,458,046,676,875	
17	6,190,283,353,629,375	
18	191,898,783,962,510,625	
19	6,332,659,870,762,850,625	
20	22,164,309,5476,699,771,875	
21	8,200,794,532,637,891,559,375	
22	319,830,986,772,877,770,815,625	
23	13,113,070,457,687,988,603,440,625	> 21 moles of trees
24	563,862,029,680,583,509,947,946,875	

Why

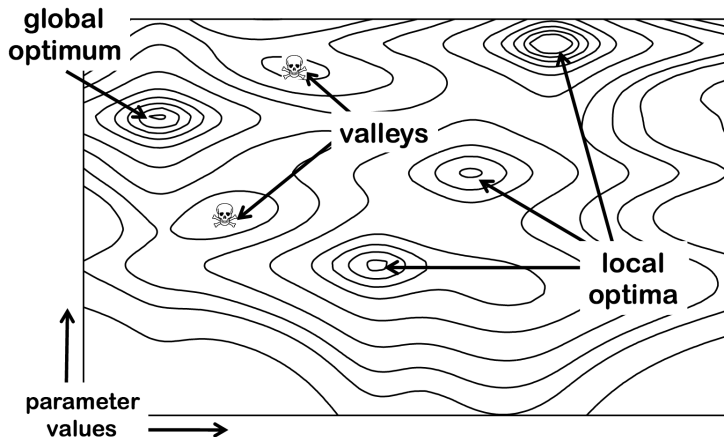


Image stolen, shamelessly, from Derrick Zwickl

Lastly, let's build the tree

Garli can take an input tree and optimize phylogenetic parameter estimates.

We actually had fantastic results doing this in a large scale phylogenetics paper ([Wright et al. 2015](#))

Lastly, let's build the tree

Second-round optimization

Dataset	L score
323320	-287897.3609
323340	-84281.7290
323360	-27384.6012
323380	-10273.6302
323390	-1697.6283