Processing ddRAD for population history inference

April Wright

ISU and KU

01-06-2016

ddRAD data

- Reduced-representation genomic method

ddRAD data

- Reduced-representation genomic method
- Cheap

ddRAD data

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

ddRAD data

- 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

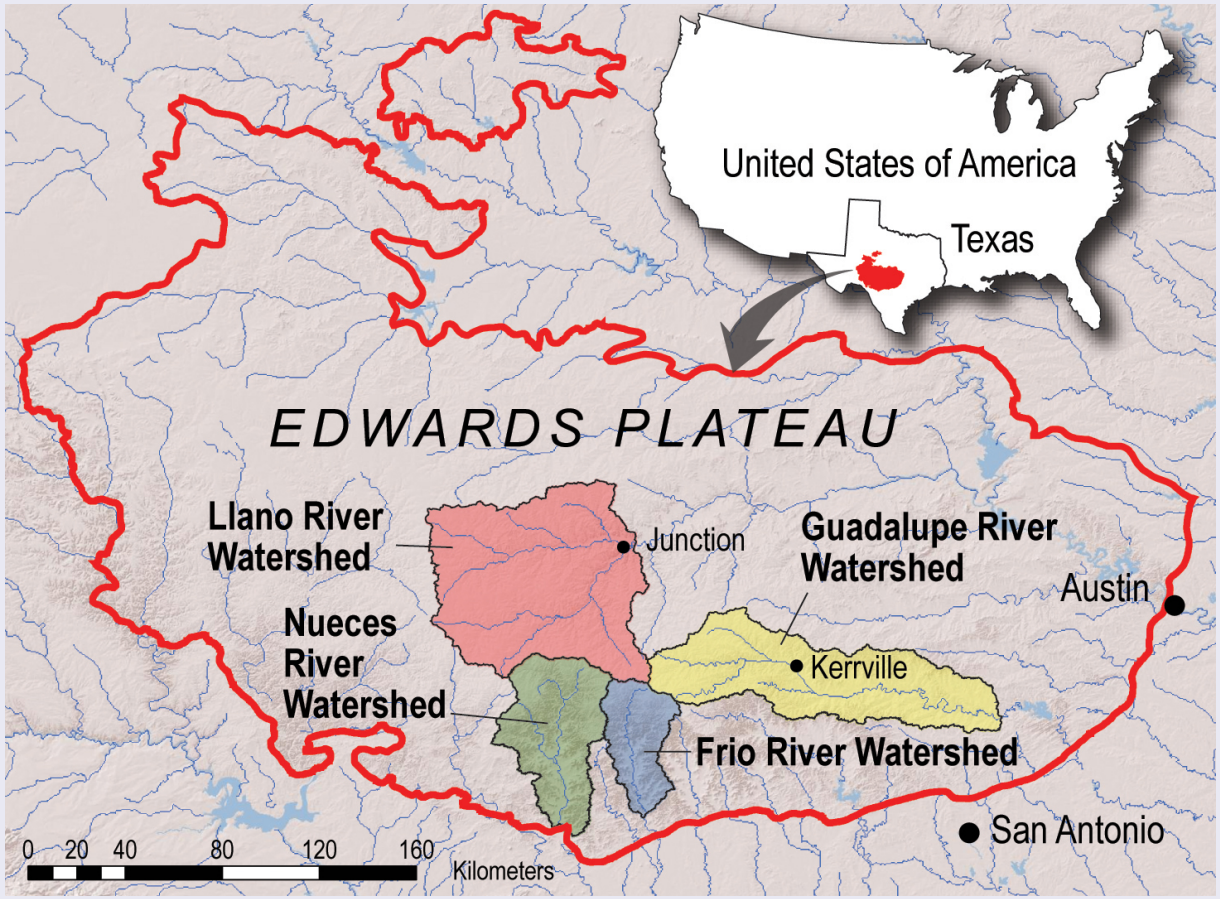


Figure 1: Image: AGU

Our Study

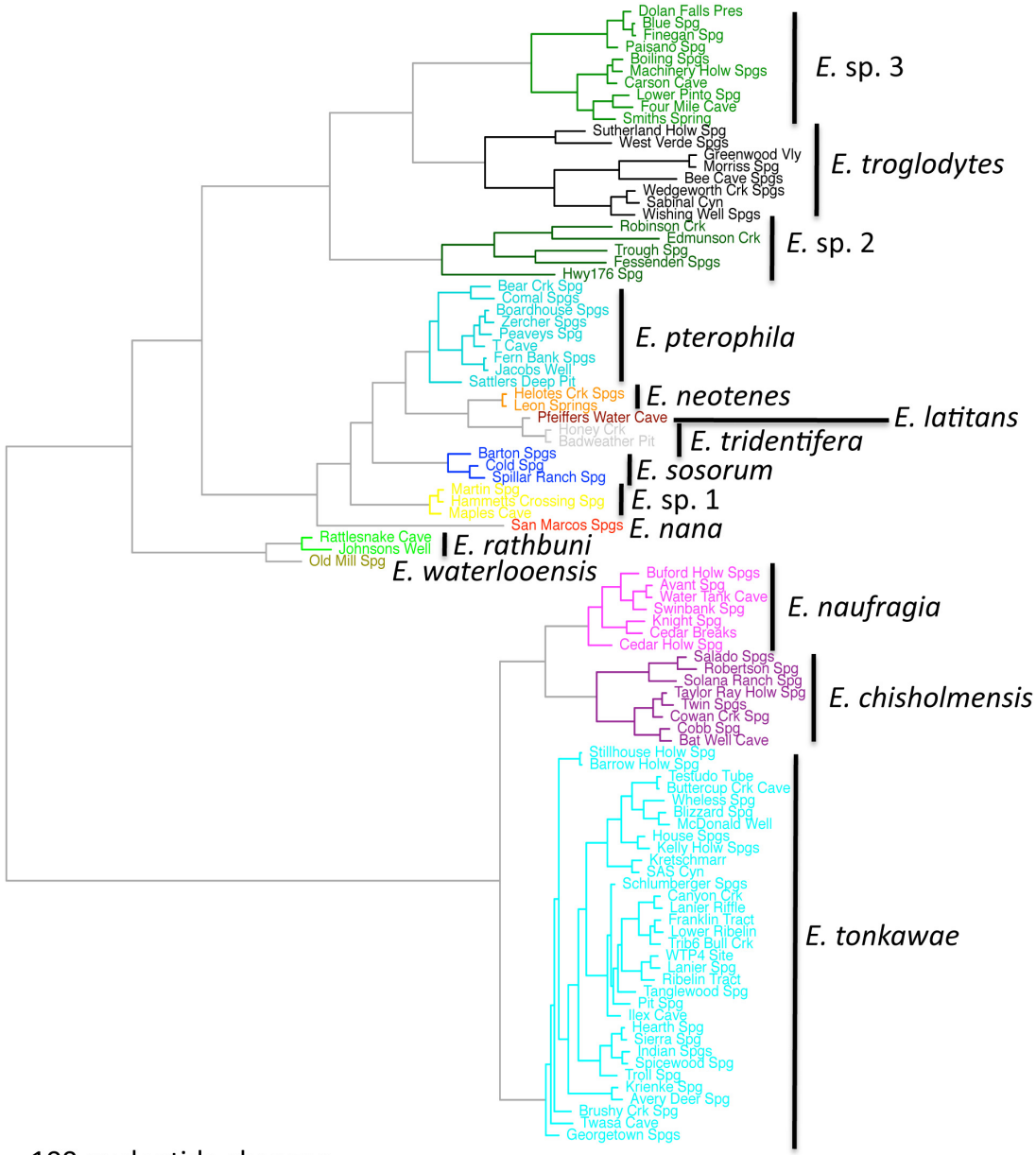
Our Study

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

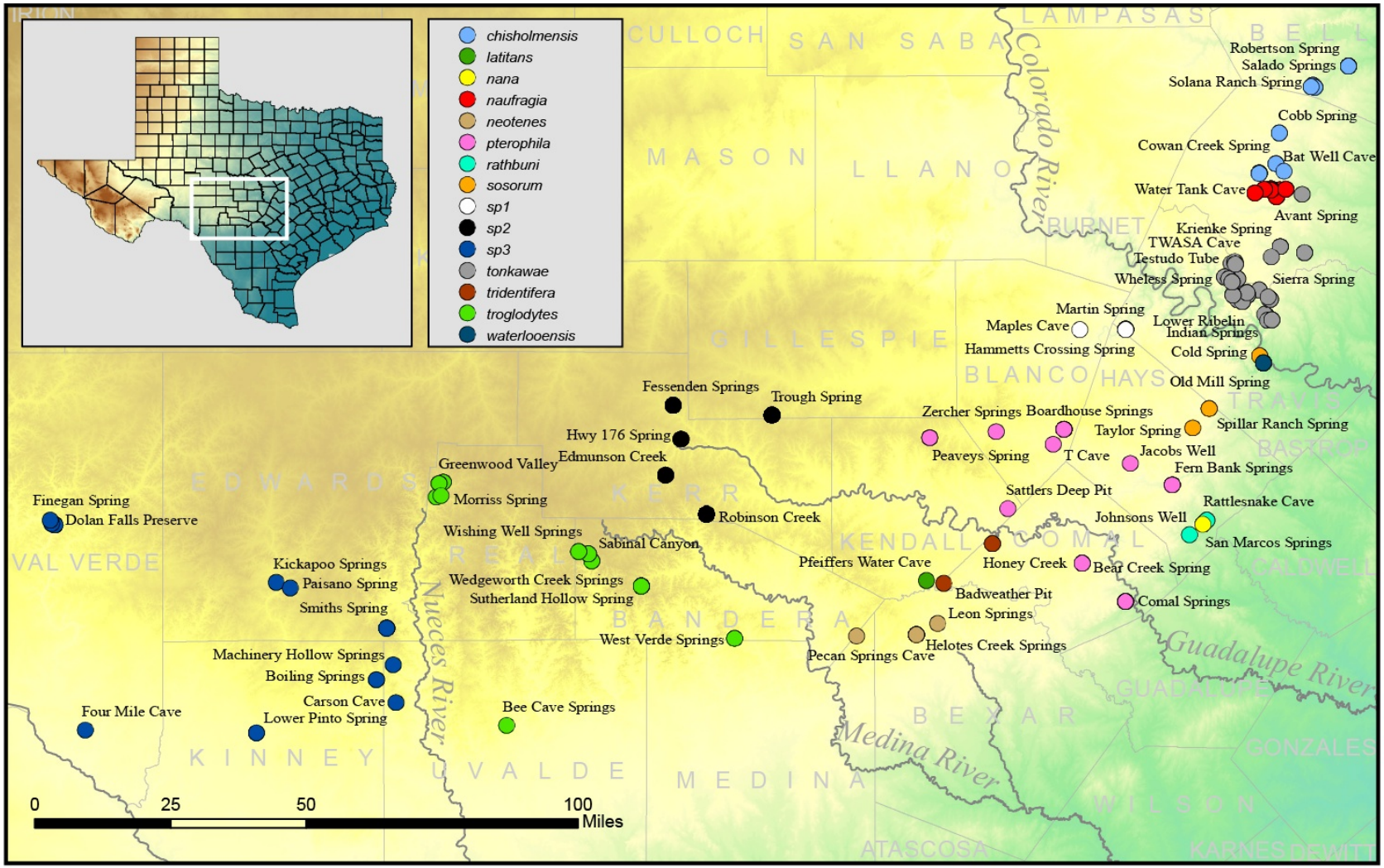
Our Study

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

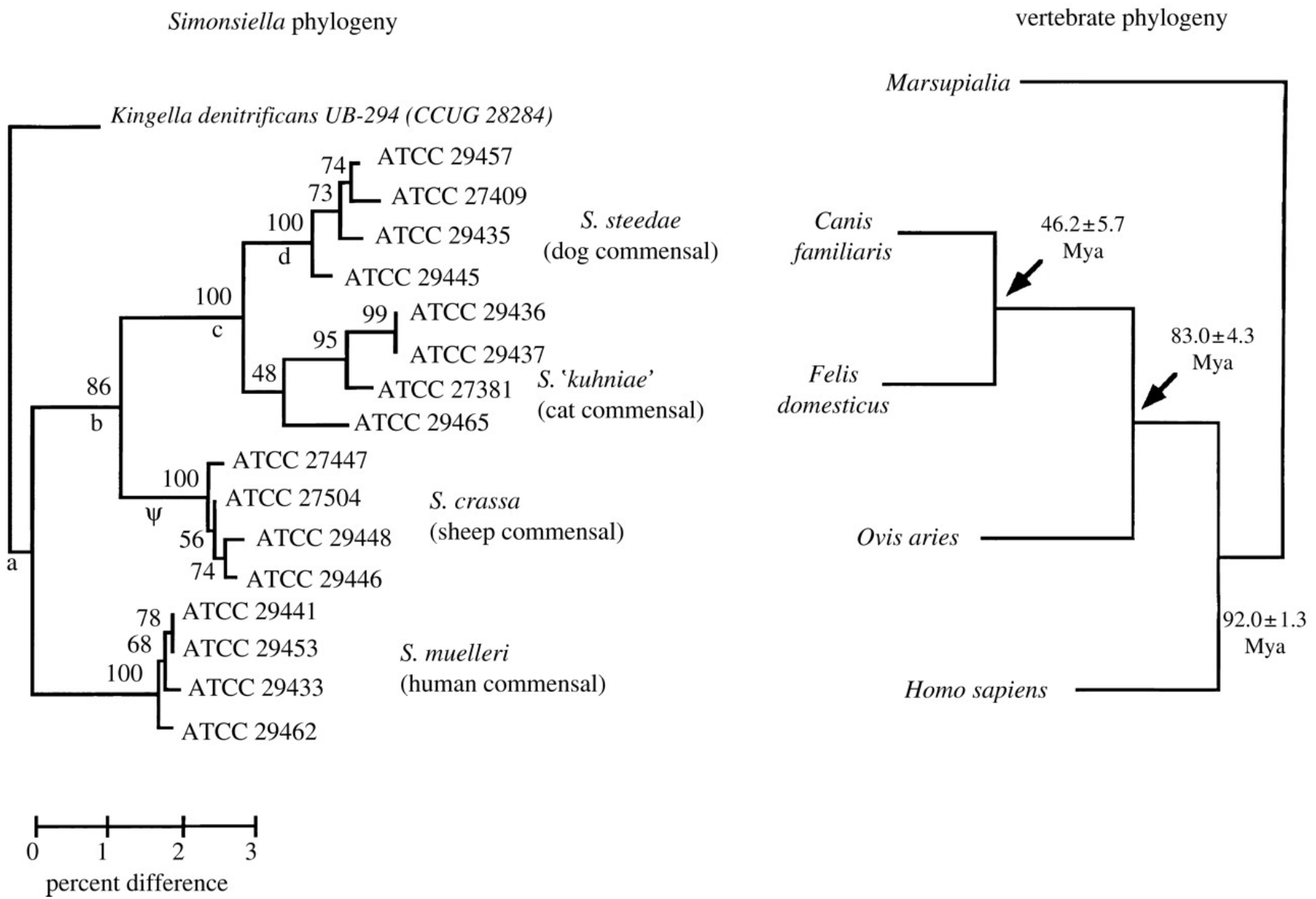
Our Study

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

Our Study


chisholmensis
latitans
nana
naufragia
neotenes
pterophila
rathbuni
sosorum
sp1
sp2
sp3
tonkawae
tridentifera
troglodytes
waterlooensis
Robertson Spring
Salado Springs
Solana Ranch Spring
Cobb Spring
Cowan Creek Spring
Bat Well Cave
Water Tank Cave
Avant Spring
Krienke Spring
TWASA Cave
Testudo Tube
Wheless Spring
Sierra Spring
Martin Spring
Maples Cave
Lower Ribelin
Indian Springs
Hammetts Crossing Spring
Cold Spring
Old Mill Spring
Fessenden Springs
Trough Spring
Zercher Springs
Boardhouse Springs
Taylor Spring
Spillar Ranch Spring
Hwy 176 Spring
Peaveys Spring
T Cave
Jacobs Well
Edmunson Creek
Greenwood Valley
Fern Bank Springs
Finegan Spring
Morriss Spring
Sattlers Deep Pit
Rattlesnake Cave
Dolan Falls Preserve
Robinson Creek
Johnsons Well
Wishing Well Springs
Sabinal Canyon
San Marcos Springs
Kickapoo Springs
Pfeiffers Water Cave
Honey Creek
Bear Creek Spring
Paisano Spring
Wedgeworth Creek Springs
Sutherland Hollow Spring
Badweather Pit
Comal Springs
Smiths Spring
Leon Springs
West Verde Springs
Helotes Creek Springs
Machinery Hollow Springs
Pecan Springs Cave
Boiling Springs
Carson Cave
Four Mile Cave
Lower Pinto Spring
Bee Cave Springs
0
25
50
100
Miles
Colorado River
Nueces River
Guadalupe River
Medina River
LAMPASAS
CULLOCH
SAN SABA
MASON
LLANO
BURNET
GILLESPIE
BLANCO
HAYS
TRAVIS
BASTROP
EDWARDS
KERR
REAL
KENDALL
COMAL
CALDWELL
VAL VERDE
BANDERA
GUADALUPE
BEXAR
GONZALES
KINNEY
UVALDE
MEDINA
WILSON
ATASCOSA
KARNES
DEWITT

Phylogenetics



Maximum likelihood

Phylogenetics

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

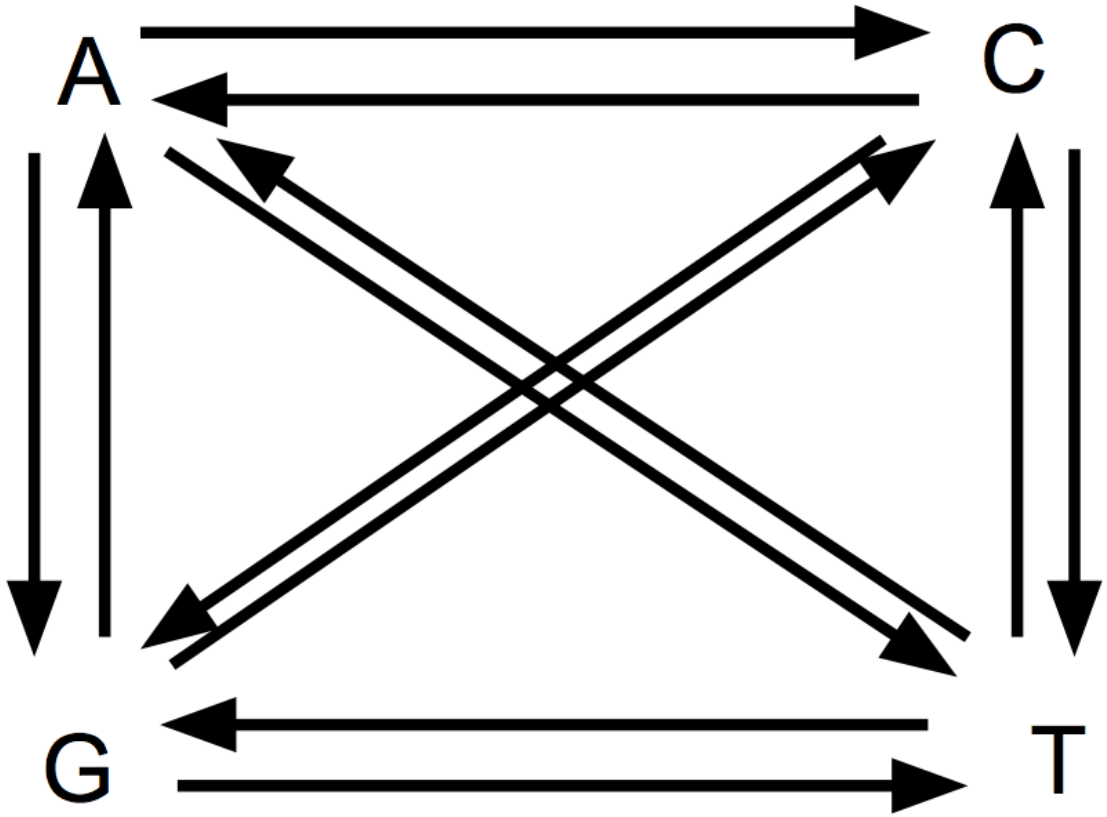
Phylogenetics

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

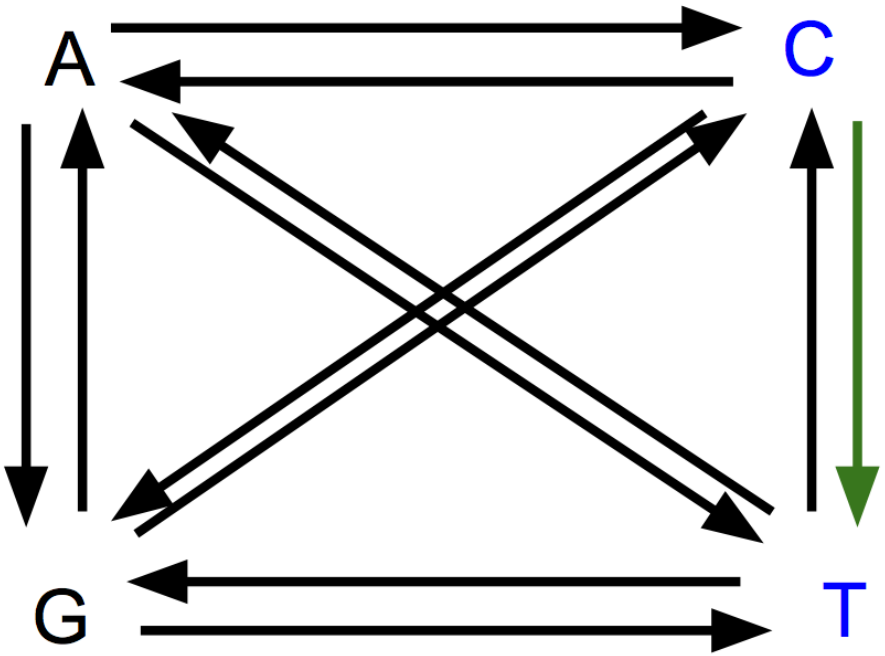
Phylogenetics

- **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

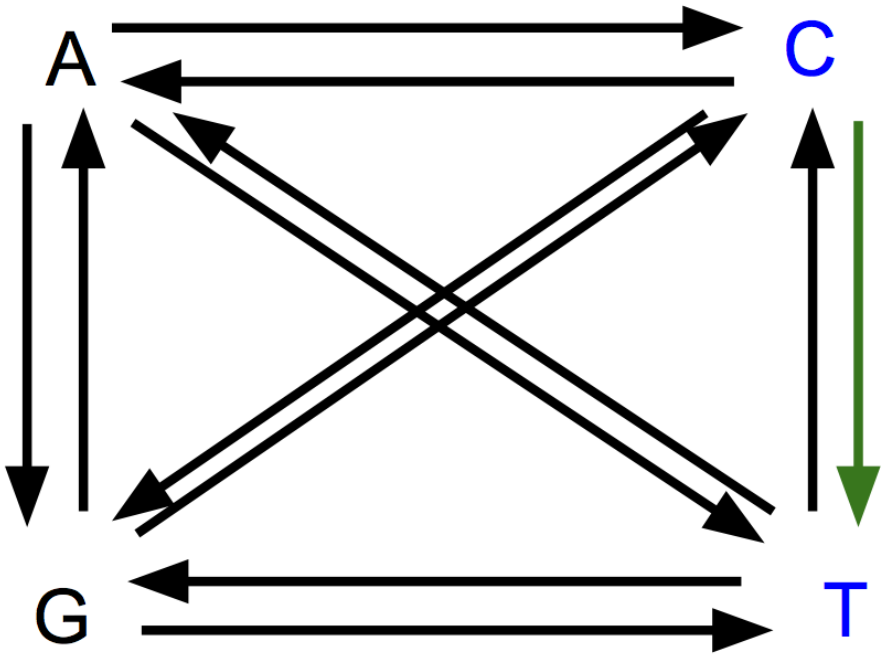


Phylogenetics



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

Phylogenetics

- **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

Phylogenetics

- **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

Phylogenetics

- **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

Phylogenetics

- **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

Phylogenetics

- **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

Phylogenetics

- **Problems**

Phylogenetics

- **Problems**
- Missing data

Phylogenetics

- **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

Phylogenetics

- **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

Phylogenetics

- **Problems**
- Model misspecification

Phylogenetics

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

Phylogenetics

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

Phylogenetics

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

For this step, we will use 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.

Initial Identification of SNPs

Key Parameters

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

Initial Identification of SNPs

Other Parameters

- -i: ID for this sample

Exercise

One of the issues we discussed was biased missing data

Catalog Building

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

Catalog Building

Key Parameters

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

Outputting Data for Phylogenetics

We use populations for this.

Outputting Data for Phylogenetics

A new file is needed, here: **the population map**

Outputting Data for Phylogenetics

Key Parameters

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

Exercise

Looking at this output is easy.

Exercise

Looking at this output is easy. But we can also look in a more complex way: countPhyloMissing.sh and plotPhyloMissing.py

RAxML approximate likelihoods

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

	323320	-287898.7874
	323340	-84289.0263
Garli	323360	-27384.6012
	323380	-10273.63021
	323390	-1697.6284