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

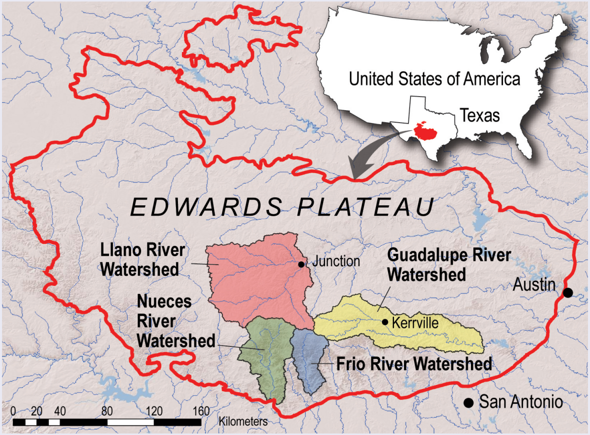

Figure 1: Image: AGU

Our Study

Our Study

13 putative species of *Eurycea*

Our Study

13 putative species of *Eurycea*
All of which are fairly threatened by development

Phylogenetics

- Maximum likelihood
- Statistically consistent

Phylogenetics

- Maximum likelihood
- Statistically consistent
- Superimposed changes

Phylogenetics

- Maximum likelihood
- Statistically consistent
- Superimposed changes
- Model-based

Phylogenetics

- **Problems**

Phylogenetics

- **Problems**
- Missing data

Phylogenetics

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

Phylogenetics

- Missing data concentrated in specific individuals

Phylogenetics

- Missing data concentrated in specific individuals
- Missing data concentrated in certain sites in the alignment

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

We'll be using STACKS for this step

The Demultiplex

We'll be using STACKS for this step The STACKS step for this is called Process RAD Tags

The Demultiplex

- **Key Parameters**

The Demultiplex

Key Parameters

- -b: A path to your barcodes file

The Demultiplex

Key Parameters

- -b: A path to your barcodes file
- -o: A path to where you want to put your output

The Demultiplex

Key Parameters

- -b: A path to your barcodes file
- -o: A path to where you want to put your output
- -q: Discard low-quality reads
- -D: capture the discarded reads in a file

The Demultiplex

Parameters You Will Get From the Sequencing Center

- –inline/index: How are the combinatorial barcodes stored in the data?
- Restriction enzymes
- -f: Name of the file. Either this will be the file you downloaded, or something you renamed

The Demultiplex

Putting it all together: processrad.sh

The Demultiplex

Let's look at the output

The Demultiplex

Let's look at the 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 depth of coverage
- -M: Maximum mismatches allowed between reads in a stack

Initial Identification of SNPs

Other Parameters

- -i: ID for this sample
- -f: filename

Exercise

Try it

- Script ustacks.sh
- Choose a different value for -m

Exercise

So now we have output

Exercise

I've included a script, calculateMissing.sh, and another, plotMissing.py

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

Catalog Building

Key Parameters

- -m: Maintain tags that match more than one RAD tag
- -n: number of mismatches to allow between a putative tag, and a tag in the catalog

Catalog Building

Exercise

- cstacks.sh
- Choose a different value for -n

Catalog Building

Exercise

- Run the two error-checking scripts

Check Individuals Against Catalog

We use sstacks for this

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 that must have a locus to output it
- -m: Minimum stack depth at a locus

Exercise

Run the populations script.

Exercise

Looking at this output is easy.

Lastly, let's build the tree

Run the tree building script like so:
treebuild.sh file.phylip Email your tree to me, titled with your group number

Examples

Some examples of commonly used commands and features are included, to help you get started.

Tables and Figures

- Use `tabular` for basic tables — see Table 1, for example.
- You can upload a figure (JPEG, PNG or PDF) using the files menu.
- To include it in your document, use the `includegraphics` command (see the comment below in the source code).

Item	Quantity
Widgets	42
Gadgets	13

Table 1: An example table.

Readable Mathematics

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.