RADseq

April Wright

Outline

- What are RADseq data?
- How do we analyze these data for population history and phylogenetics?
- How is RADseq data usually collected here on campus?

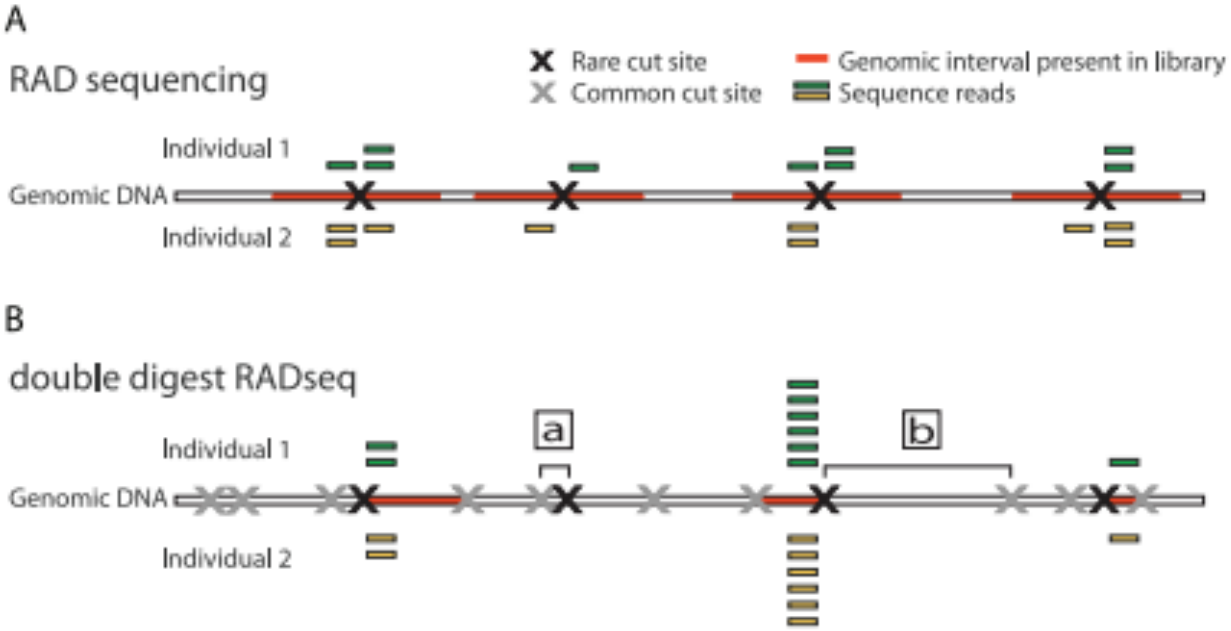
RADseq data

- Genome reduction technology
- Aim: Obtain thousands of variable sites that could be used for QTL, genotyping population history

RADseq data

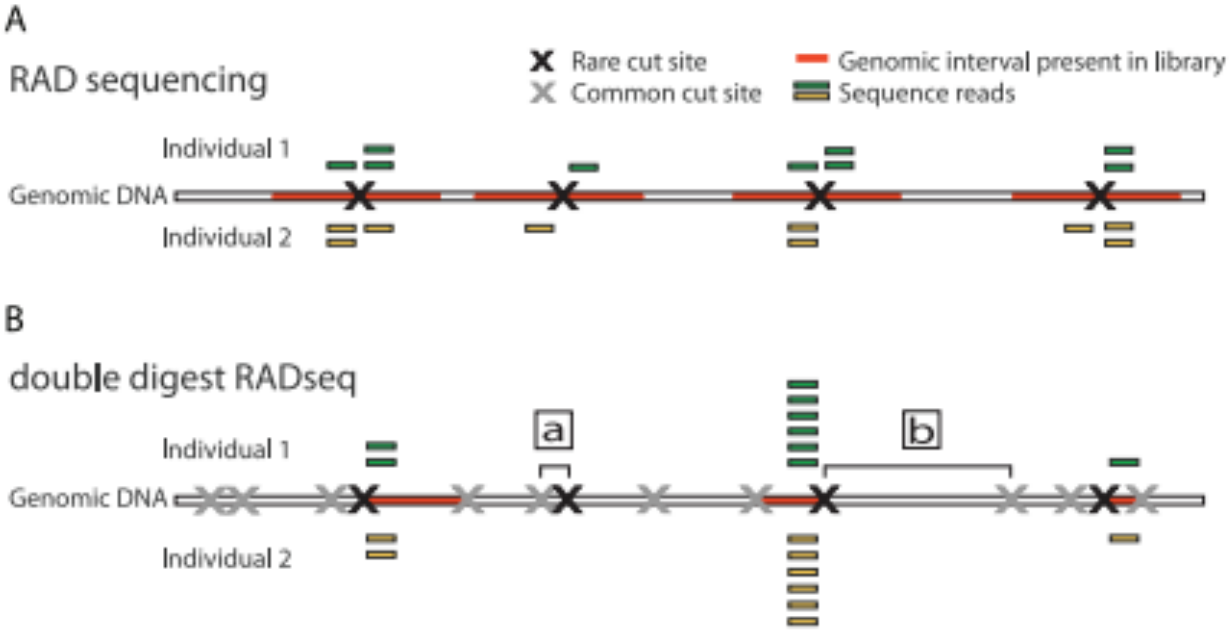
- Genome reduction technology
- Aim: Obtain thousands of variable sites that could be used for QTL, genotyping population history
  - Especially for non-model organisms, as it requires no reference genome (though you can use one for mapping)

RADseq data



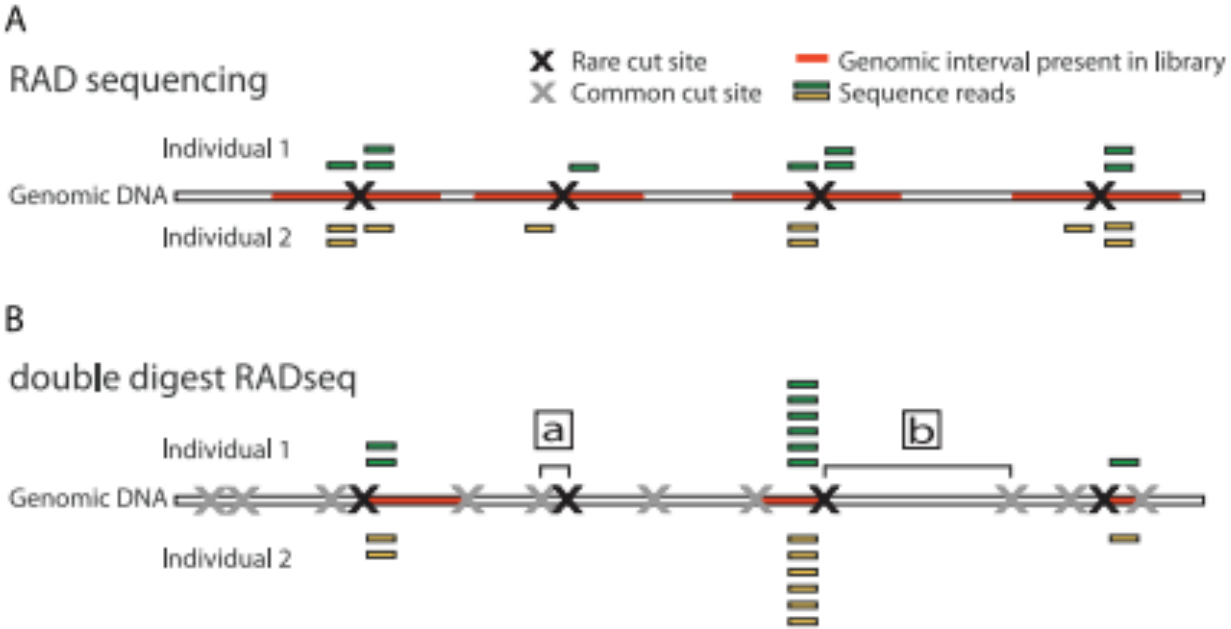
Peterson et al. 2012 PLoS ONE

RADseq data



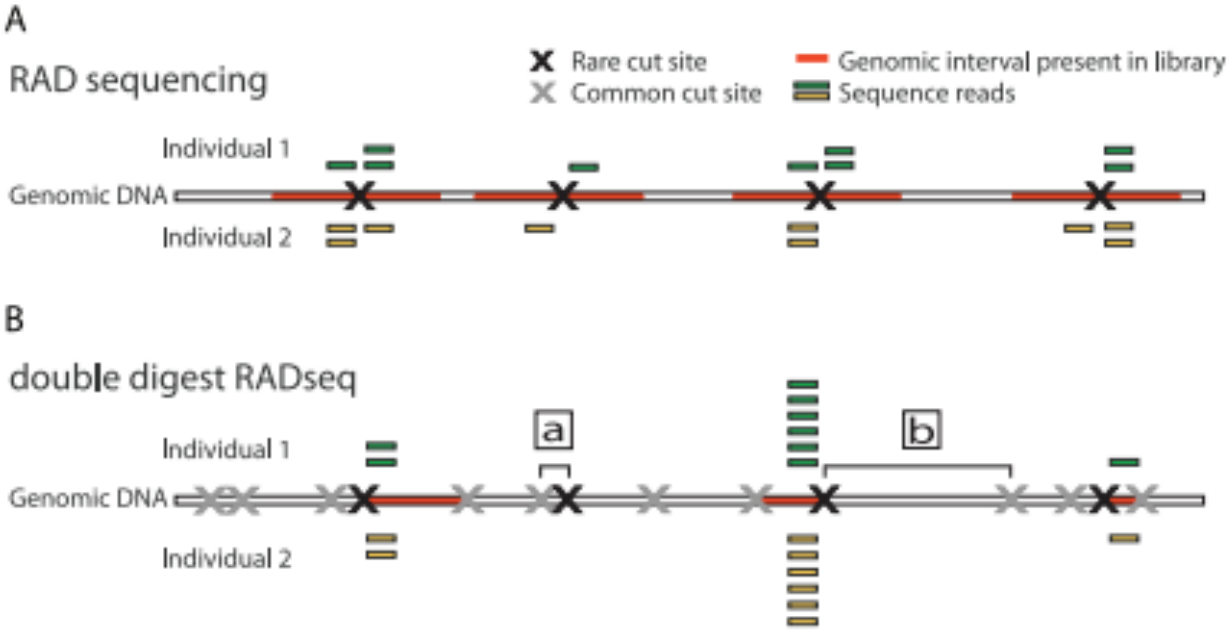
Restriction Site: 4-8 nucleotide sequences in genome; recognized by restriction enzymes

RADseq data



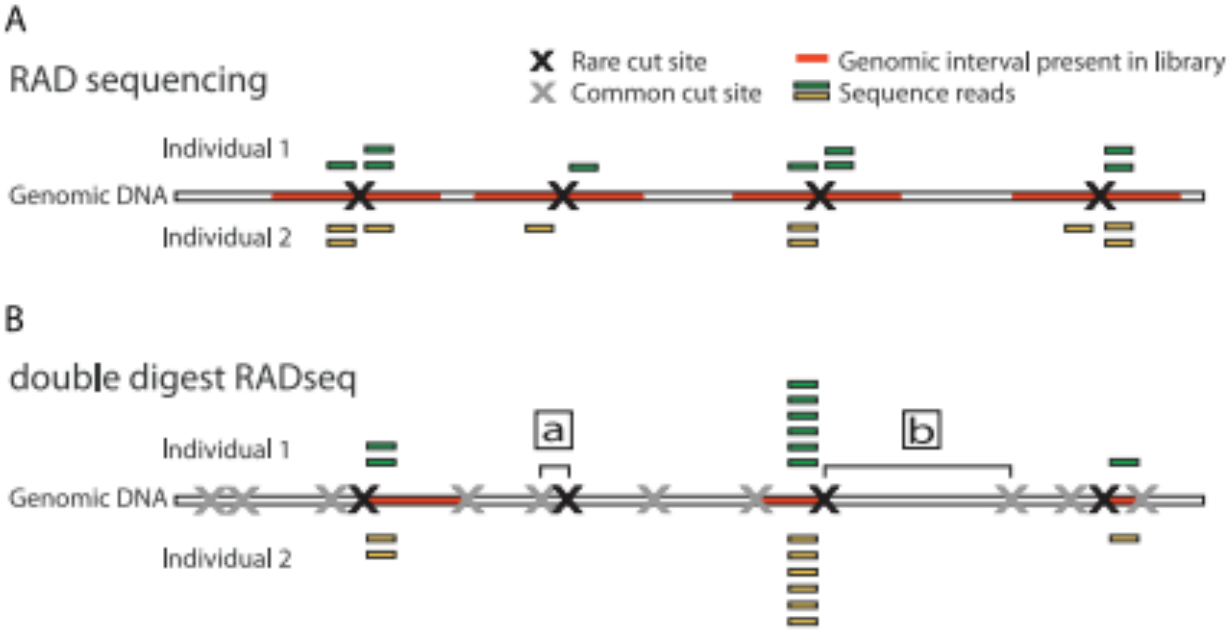
Restriction Enzyme: Enzyme to cut at a restriction site

RADseq data



Read: a set of bp obtained via RADseq. Of a target size

RADseq data



Barcode: Added sequence of nucleotides so samples can be identified

What you get

- Fastq files filled with reads.
- Refresher from last week: Fastq encodes data and quality information

What you get

- Fastq files filled with reads.
- Refresher from last week: Fastq encodes data and quality information
- Identified by the lane in the machine
- Coded with barcodes

Processing RADseq

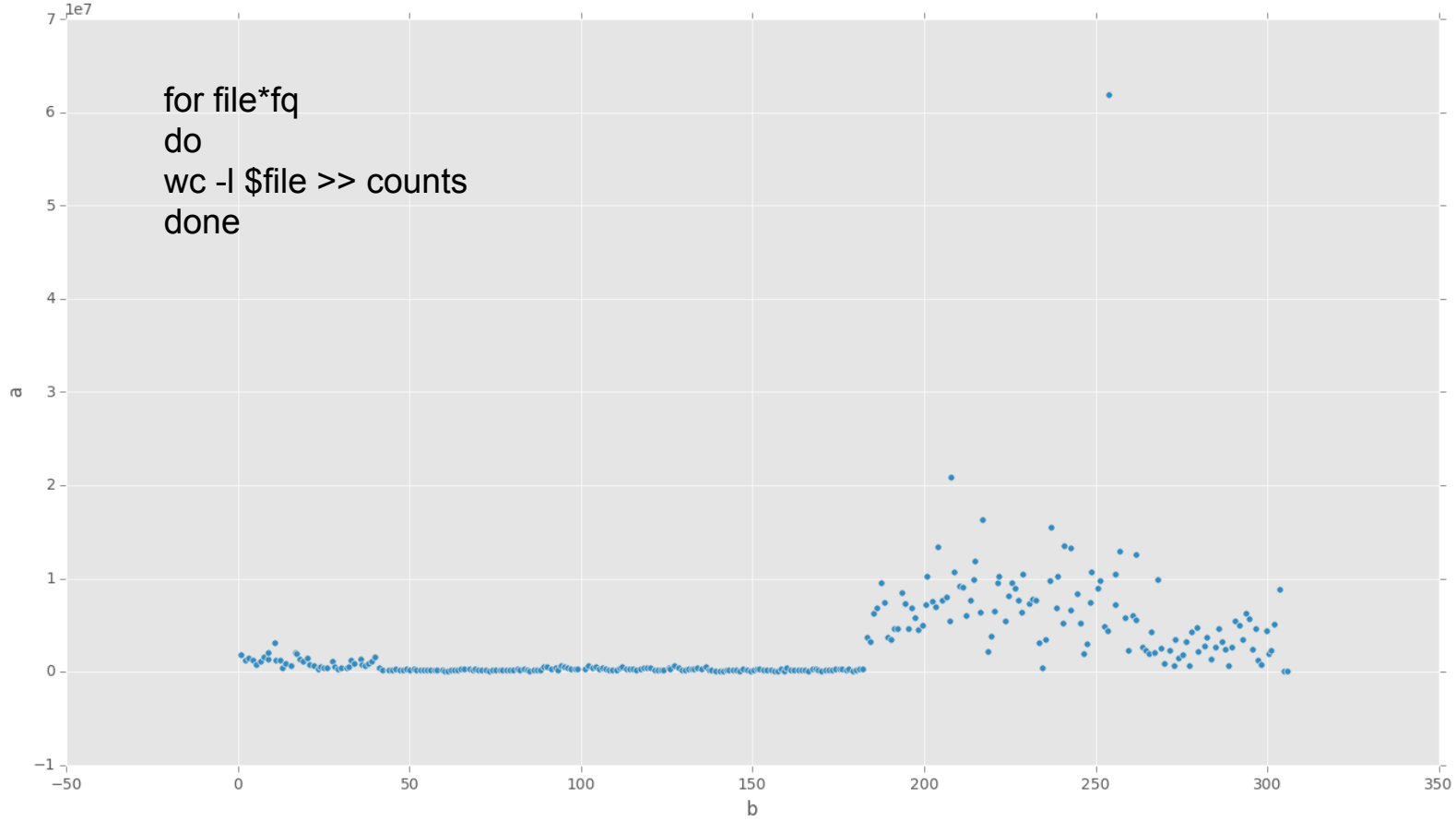
- Process Fastq files into individual samples
  - ‘Demultiplex’

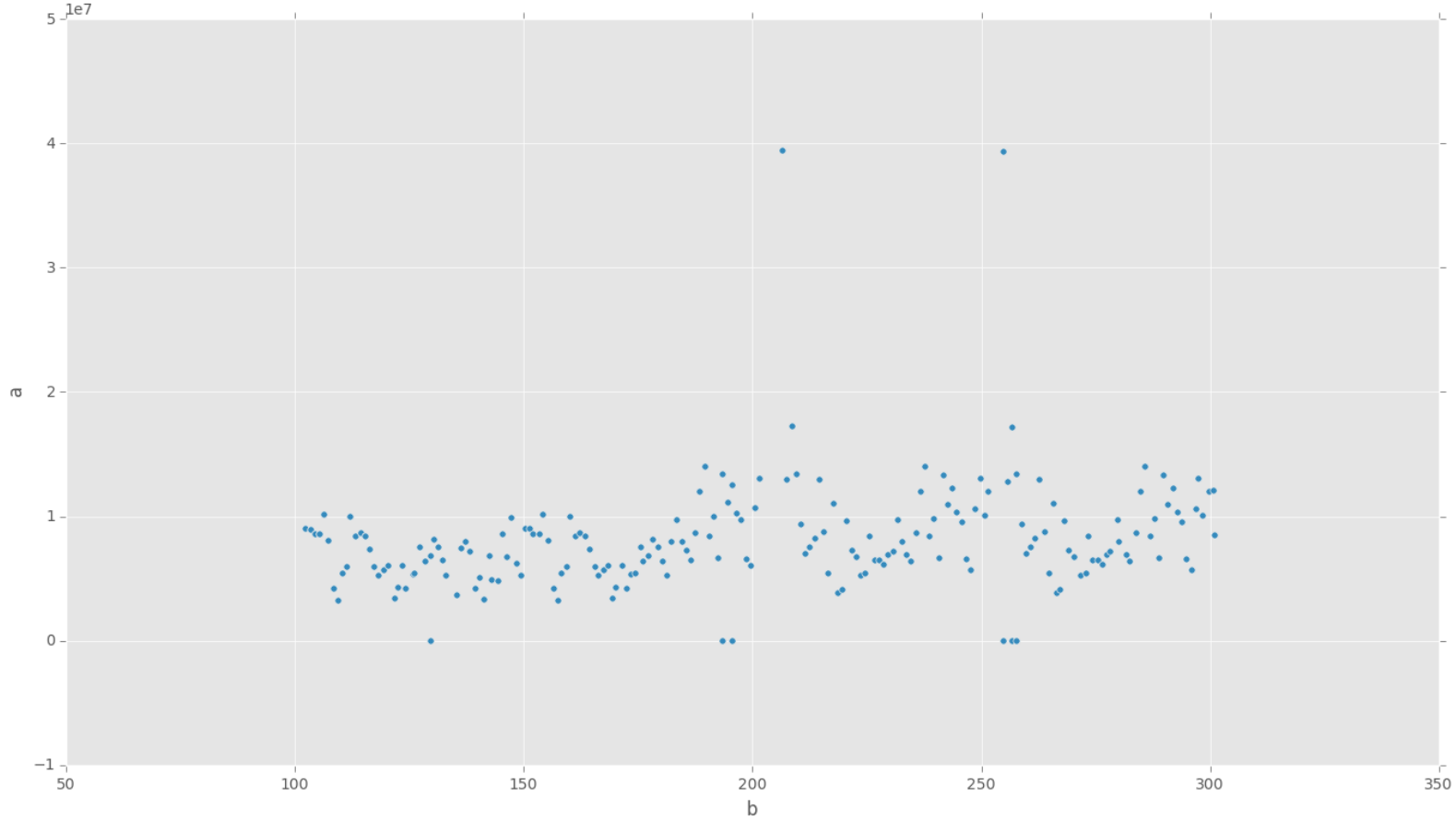
Processing RADseq

- Process Fastq files into individual samples
  - ‘Demultiplex’
  - Result: sequence data are now tagged to individuals

Processing RADseq

- Process Fastq files into individual samples
  - ‘Demultiplex’
  - Result: sequence data are now tagged to individuals
  - Process check: Do individuals have the same amount of data?


for file*fq
do
wc -l $file >> counts
done
1e7
7
6
5
4
3
2
1
0
−1
a
−50
0
50
100
150
200
250
300
350
b


1e7
5
4
3
2
1
0
−1
a
50
100
150
200
250
300
350
b

Two Major Pipelines

Stacks and pyRAD

Locus-building

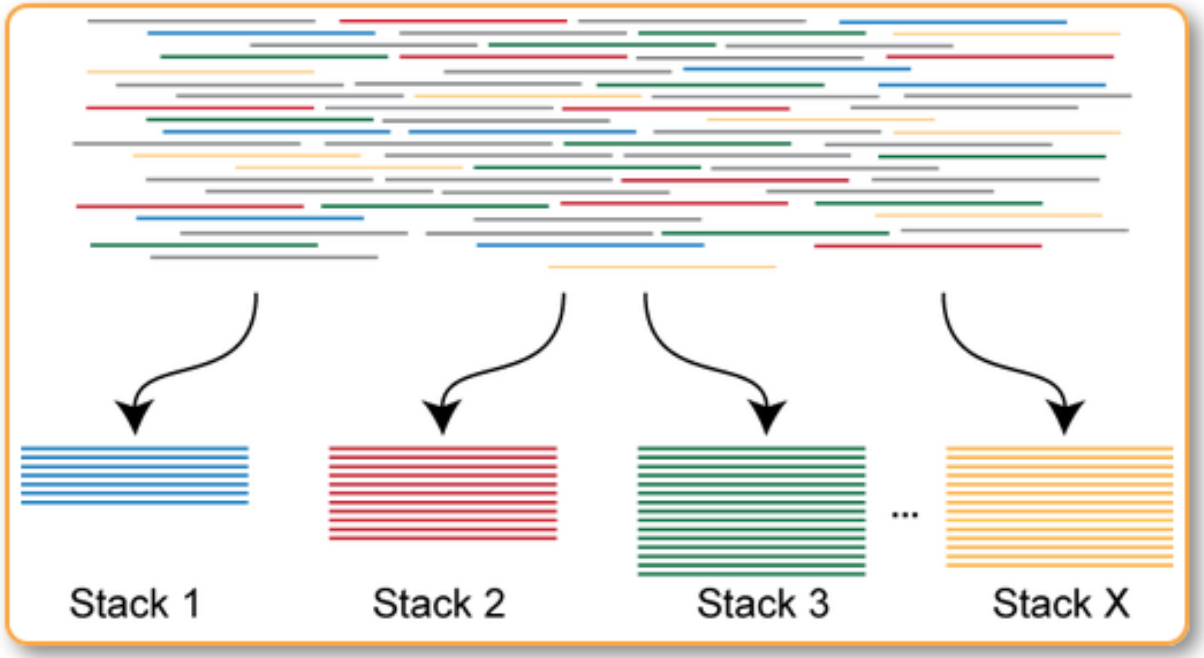
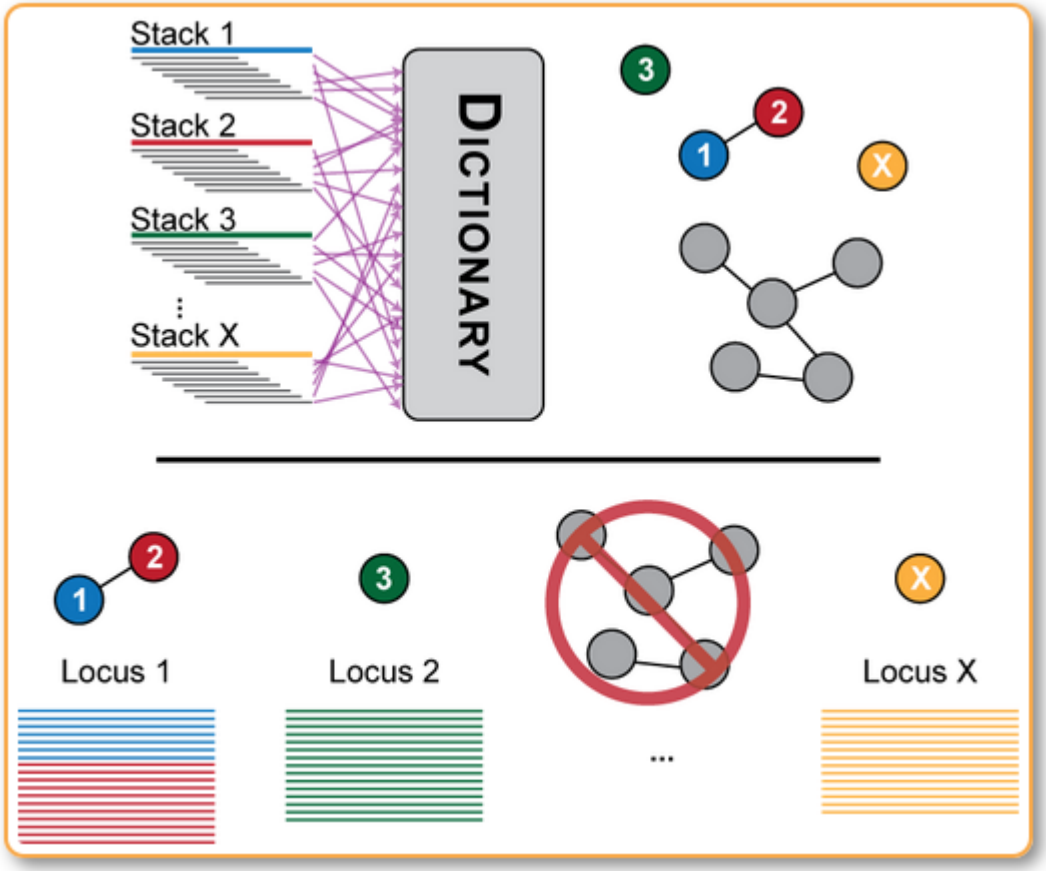
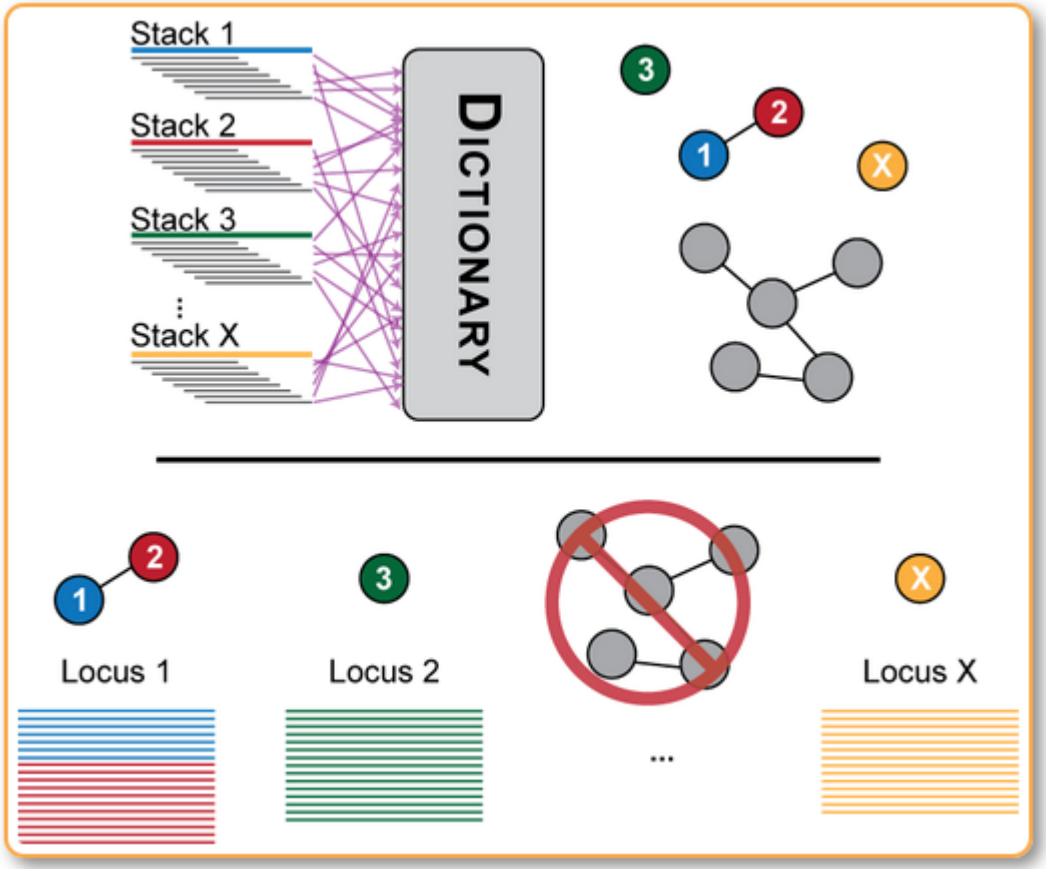


Figure 1. The initial stage of the **ustacks** *de novo* assembly algorithm forms exactly matching stacks from raw short-reads.

Locus-building

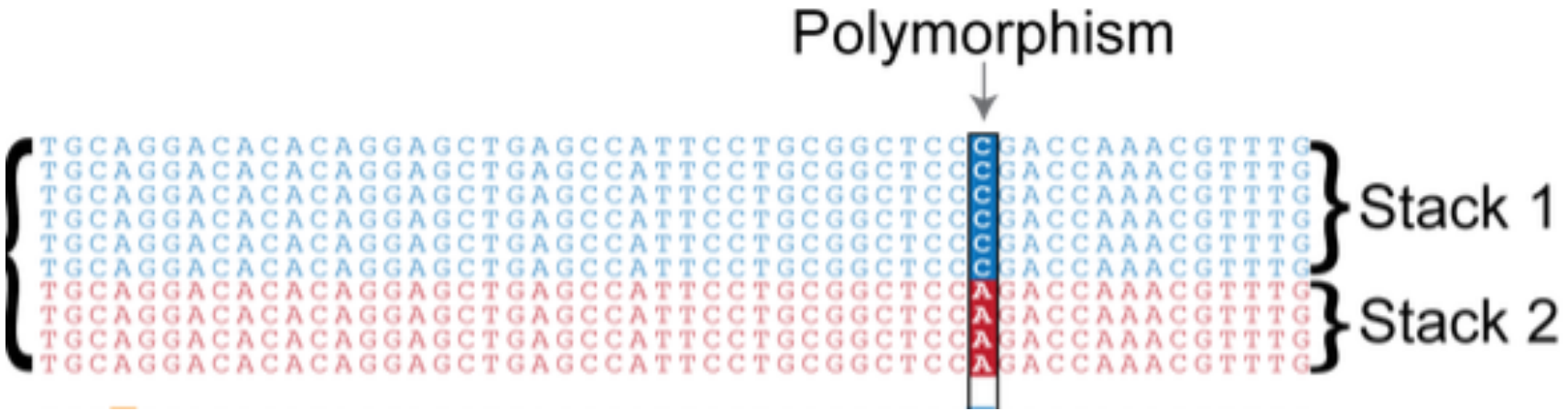


Locus-building



# of loci ↓

# of reads at a locus ↑


Polymorphism
Stack 1
Stack 2

Locus-building

- So far, *within* sample
- pyRAD has an additional alignment step and estimate error rate and heterozygosity. It uses these measures to build consensus sequences for the individual

Locus-building: across individuals

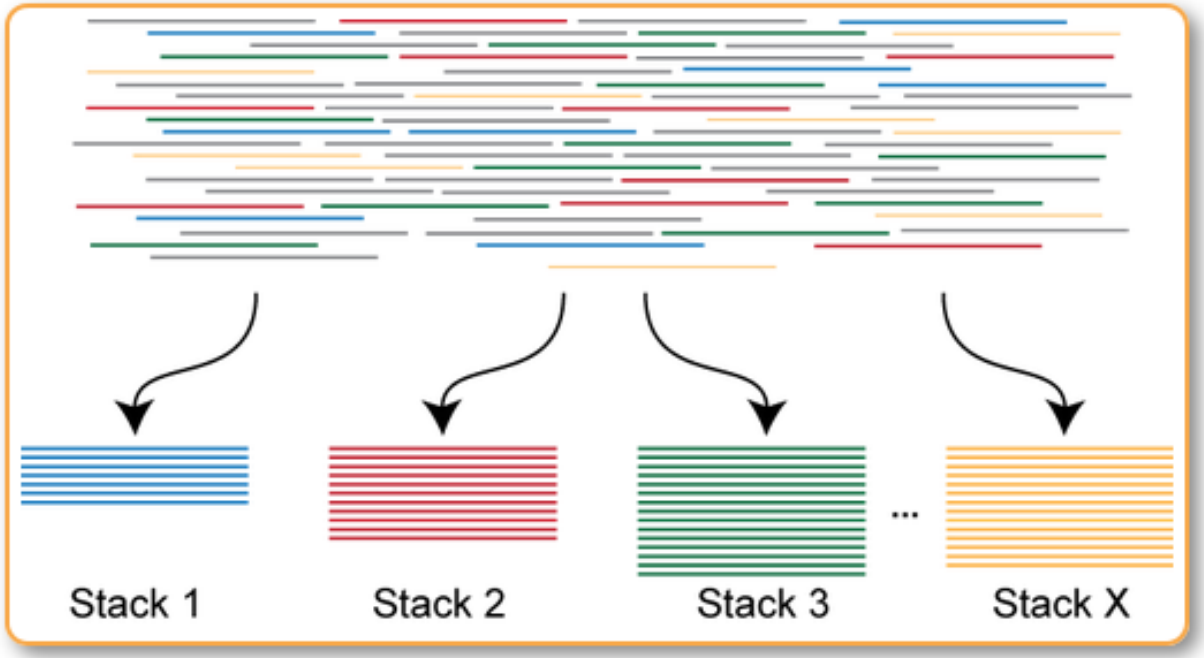
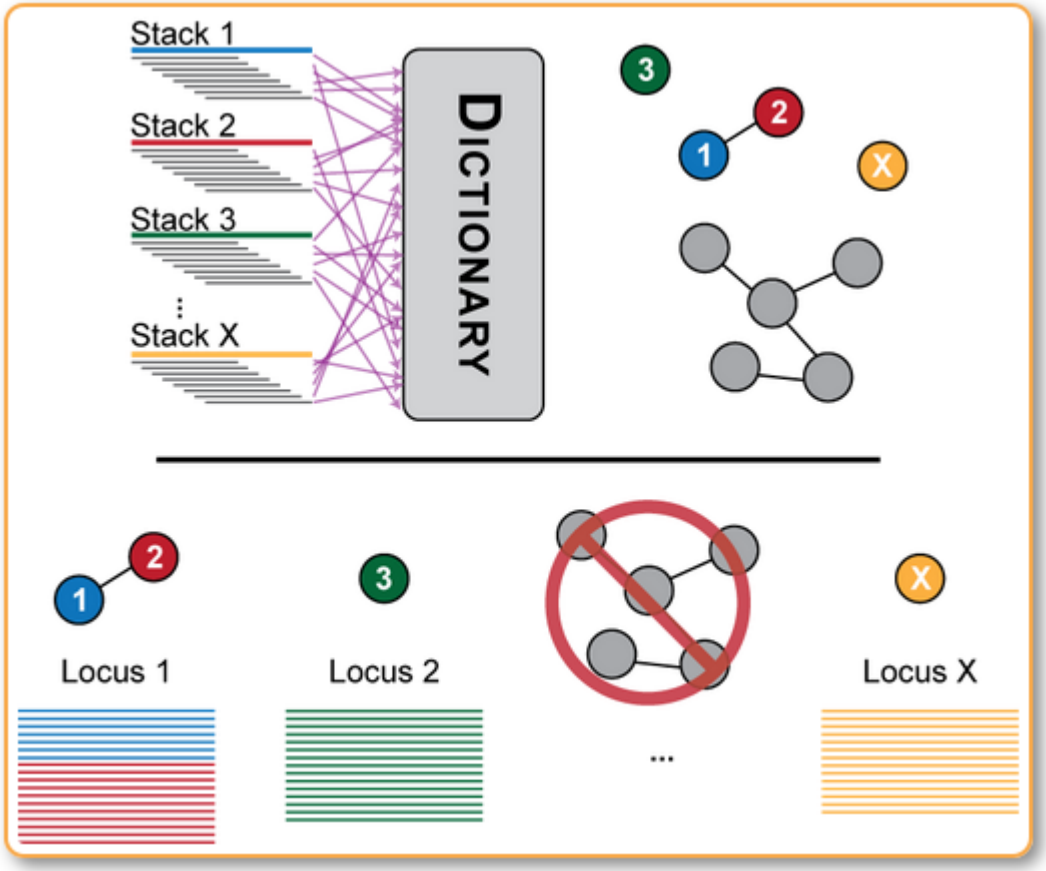


Figure 1. The initial stage of the **ustacks** *de novo* assembly algorithm forms exactly matching stacks from raw short-reads.

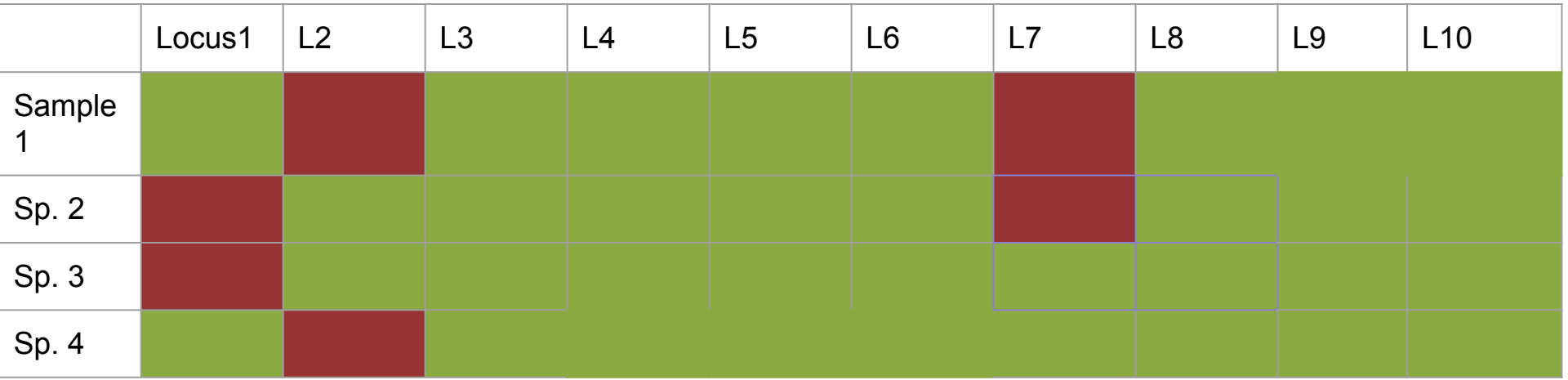
Locus-building: across individuals



	Locus1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Sample 1										
Sp. 2										
Sp. 3										
Sp. 4										

Green = present
Red = absent
Blue = fragment

	Locus1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Sample 1										
Sp. 2										
Sp. 3										
Sp. 4										



Green = present
Red = absent
Blue = fragment

Making the matrix

- How many populations must a loci be present in in order to be output?
- What kind of output do you want?

Making the Matrix

- Phylogenetic trees
  - Often take all SNPs at a locus

- STRUCTURE
  - Don’t

Making the Matrix

- Phylogenetic trees
  - Often take all SNPs at a locus
  - Lots of literature on missing data

- STRUCTURE
  - Don’t
  - Not nearly as much

Making the Matrix

- Phylogenetic trees
  - Often take all SNPs at a locus
  - Lots of literature on missing data
  - Do you need all individuals?
- STRUCTURE
  - Don’t
  - Not nearly as much
  - Specific assumptions about Hardy-Weinberg equilibrium

Structure

- Data can generally be run as-is from either Stacks or pyRAD
  - Note: Stacks adds a footer to Phylip files and a header to structure: remove these

Special challenges for phylogenetic trees

- We don’t know where our data came from

Special challenges for phylogenetic trees

- We don’t know where our data came from
  - Partitions?
  - http://www.robertlanfear.com/partitionfinder/
- Huge amounts of missing data
- Acquisition bias
  - Parsimony
  - Mk
  - https://github.com/stamatak/standard-RAxML

Here at UT

- GSAF manages this data type

Here at UT

- GSAF manages this data type
  - Sample prep managed in-lab, sequencing performed at GSAF

Analyses on TACC

- TACC has STACKS on both Lonestar and Stampede

Analyses on TACC

- TACC has STACKS on both Lonestar and Stampede
  - Do not have pyRAD, though the SciPy Stack is present to run pyRAD
  - http://wrightaprilm.github.io/posts/pyrad-and-tacc.html

Questions?