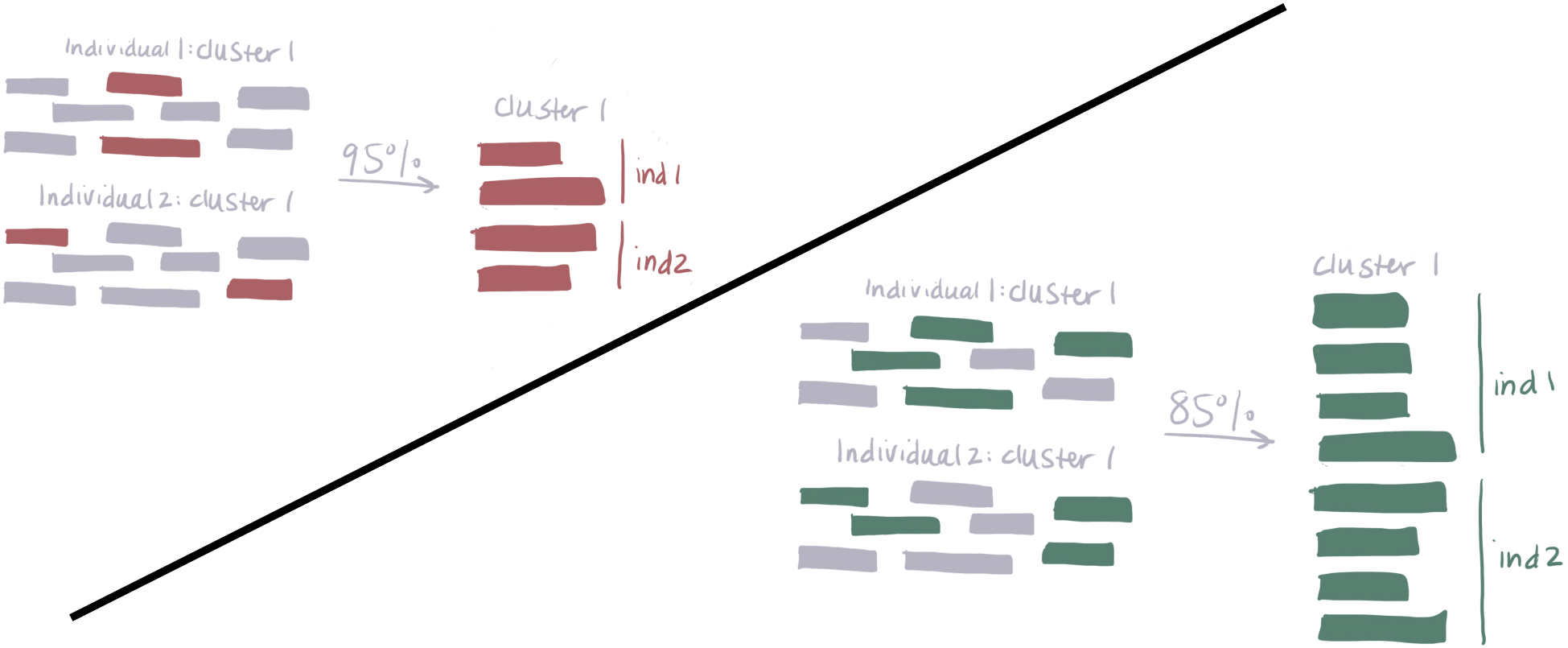
RADseq bioinformatics overview



Common RADseq pipelines

	Modularity	Installation	Documentation	Output	Bonus	Anti-bonus
iPyrad	Modular with wrapper	Easy (conda)	Excellent	Most	.phy output	Crashes with large datasets
stacks	Modular (uncurated wrappers)	Medium (zip, make)	Good	Most (except .phy)	Very flexible	No indels
dDocent	Not modular	Difficult (conda with issues)	Good	Vcf	Interactive parameter entry; SNP filtering tutorial	Separate pre- and post-filtering; cannot pause
2bRAD (native)	Modular	Medium (git, individual programs)	Fair	Most (except .phy)	Can incorporate replicates	Unclear documentation
AftrRAD	Modular	Easy (zip)	Good	Most	Reduced run times; .phy output	Crashes with large datasets; only SE reads

Rebecca Tarvin, https://rdtarvin.github.io/RADseq_Quito_2017/

Errors to watch for

- PCR duplicates and genotyping errors
  - By chance some alleles will amplify more than others
  - Can cause some individuals to appear as homozygotes
  - Errors can be interpreted as true diversity

Rebecca Tarvin, https://rdtarvin.github.io/RADseq_Quito_2017/

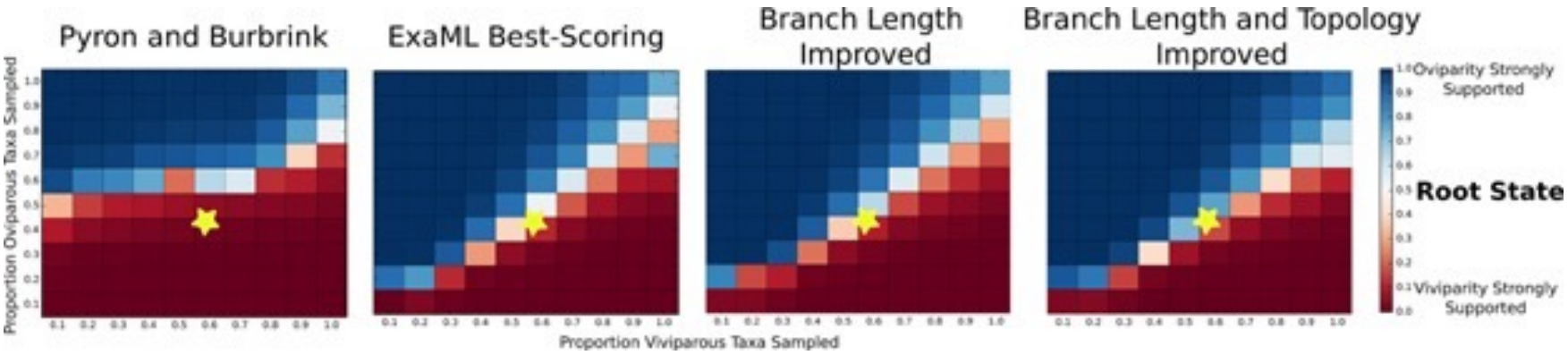
Errors to watch for

- PCR duplicates and genotyping errors
  - By chance some alleles will amplify more than others
  - Can cause some individuals to appear as homozygotes
  - Errors can be interpreted as true diversity
- Variance in depth and coverage among loci
  - Shorter loci are sequenced more often
  - High GC = more PCR

Rebecca Tarvin, https://rdtarvin.github.io/RADseq_Quito_2017/

Things to keep in mind while processing data

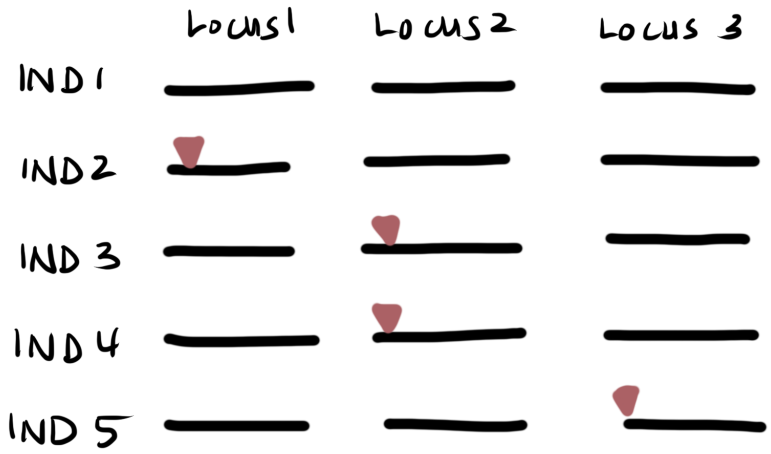
- **Robustness** of your data is a crucial first step to understanding and interpreting it
  - Allelic dropout & missing data
  - Clustering threshold



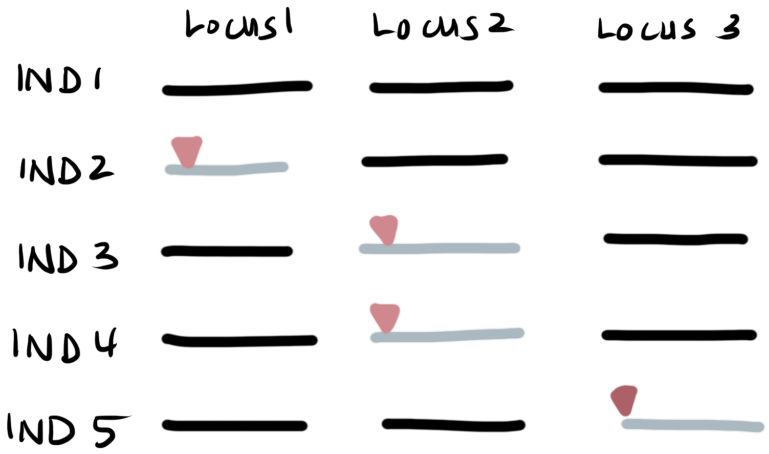
Wright et al. (2015) J. Exp. Biol. 324(6):504–516

Allelic dropout is a source of missing data

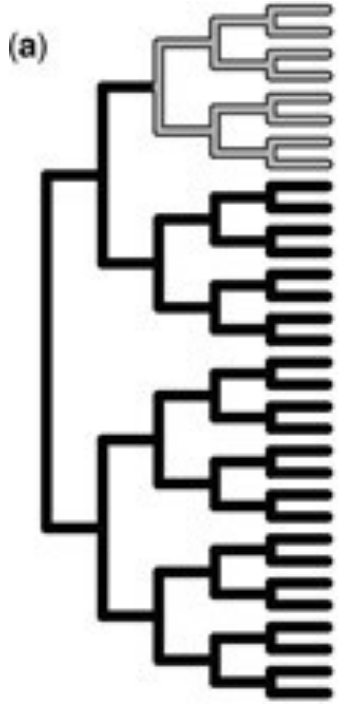
Allelic dropout is a source of missing data



Allelic dropout is a source of missing data



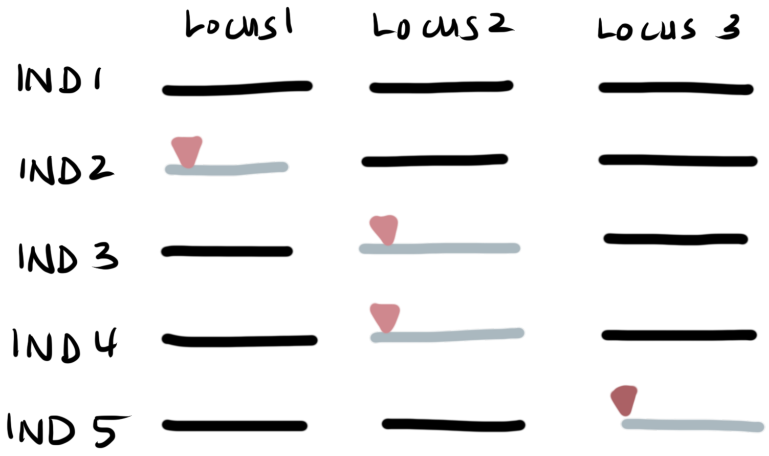
Mutation disruption vs. mutation generation



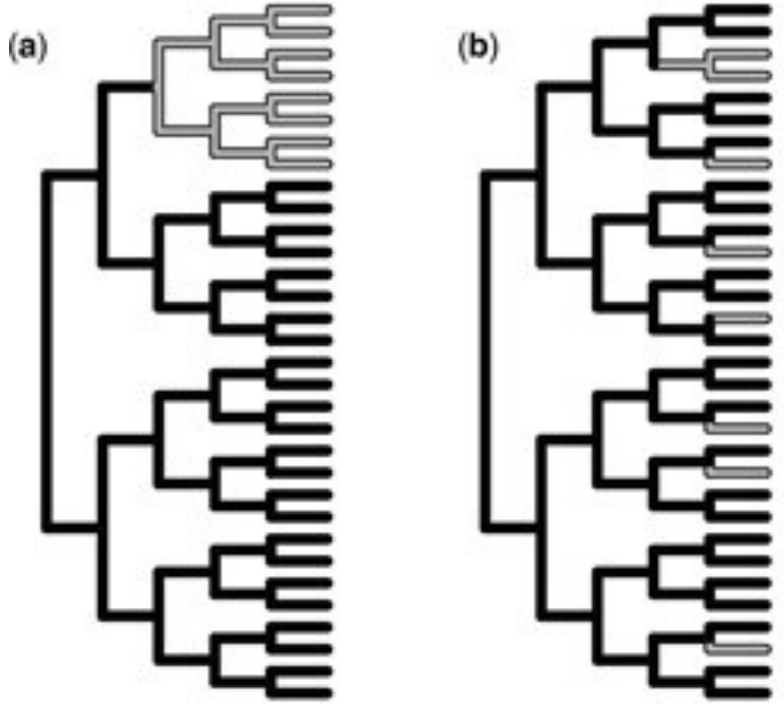
Missing data is a source of information (in a way) for building trees

Eaton et al. (2017) Syst. Biol. 66(3):399–412

Allelic dropout is a source of missing data



Mutation disruption vs. mutation generation

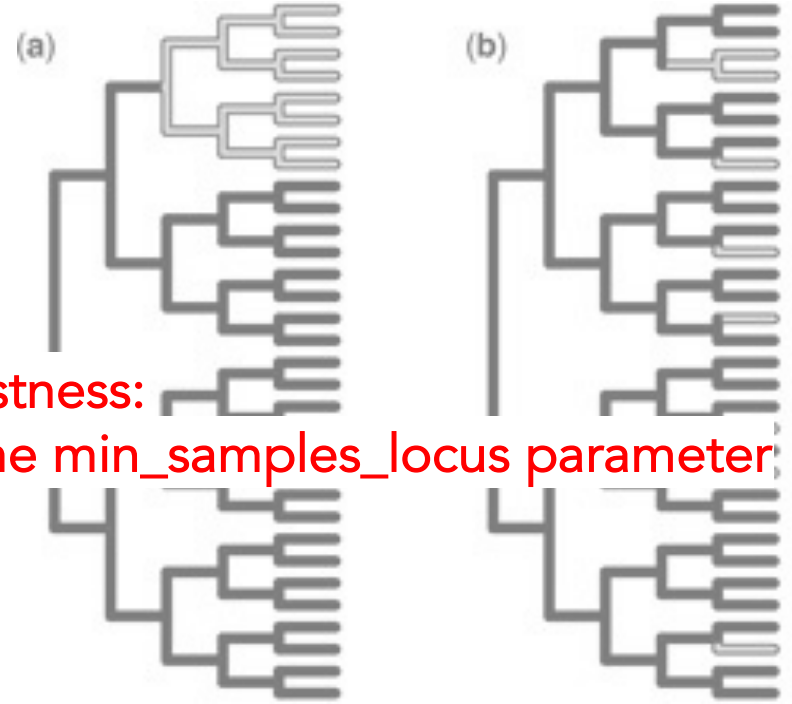


Missing data is a source of information (in a way) for building trees

Eaton et al. (2017) Syst. Biol. 66(3):399–412

Allelic dropout is a source of missing data

Mutation disruption vs. mutation generation

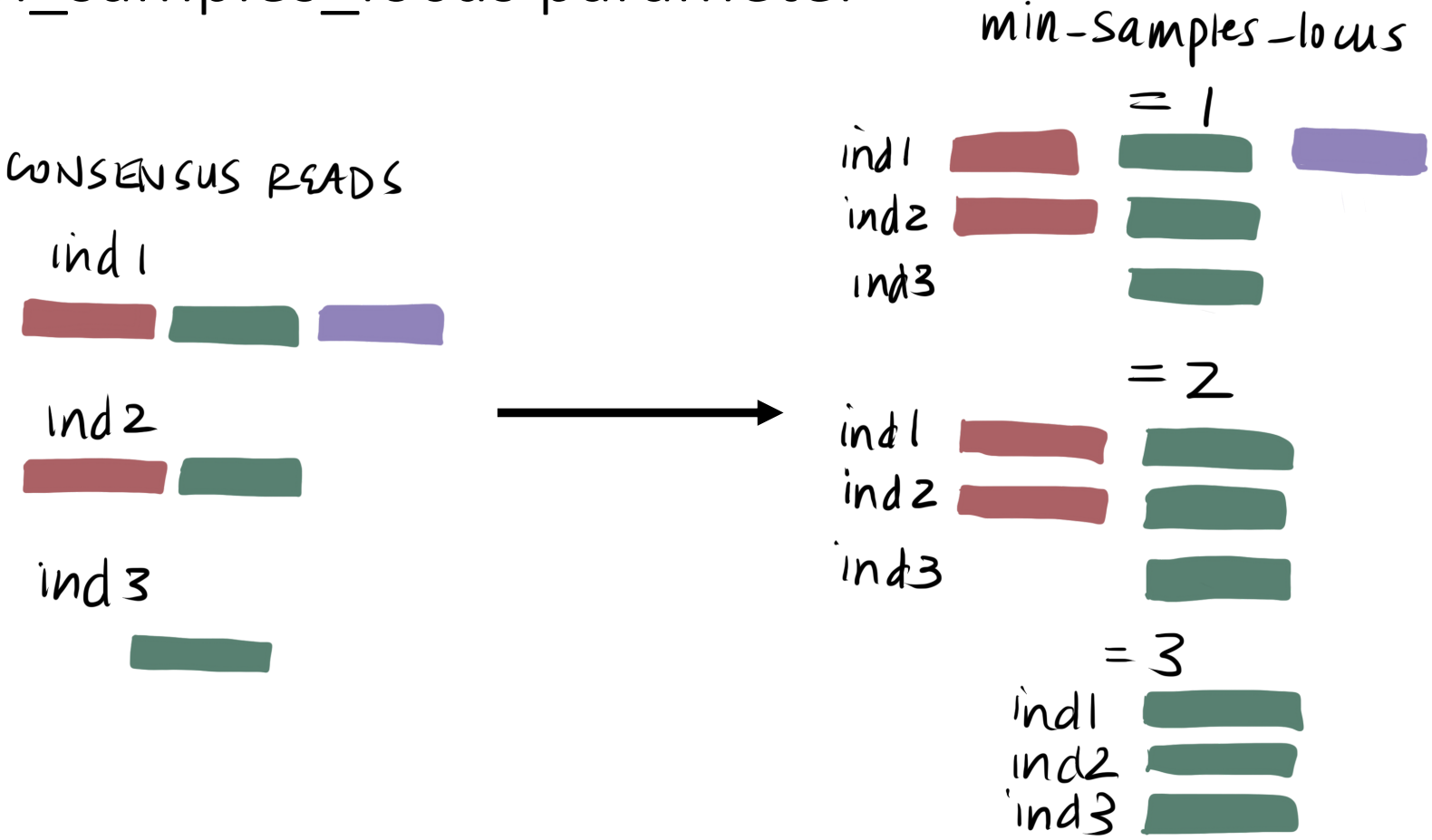


Test for robustness:
adjust levels of missing data through the min_samples_locus parameter

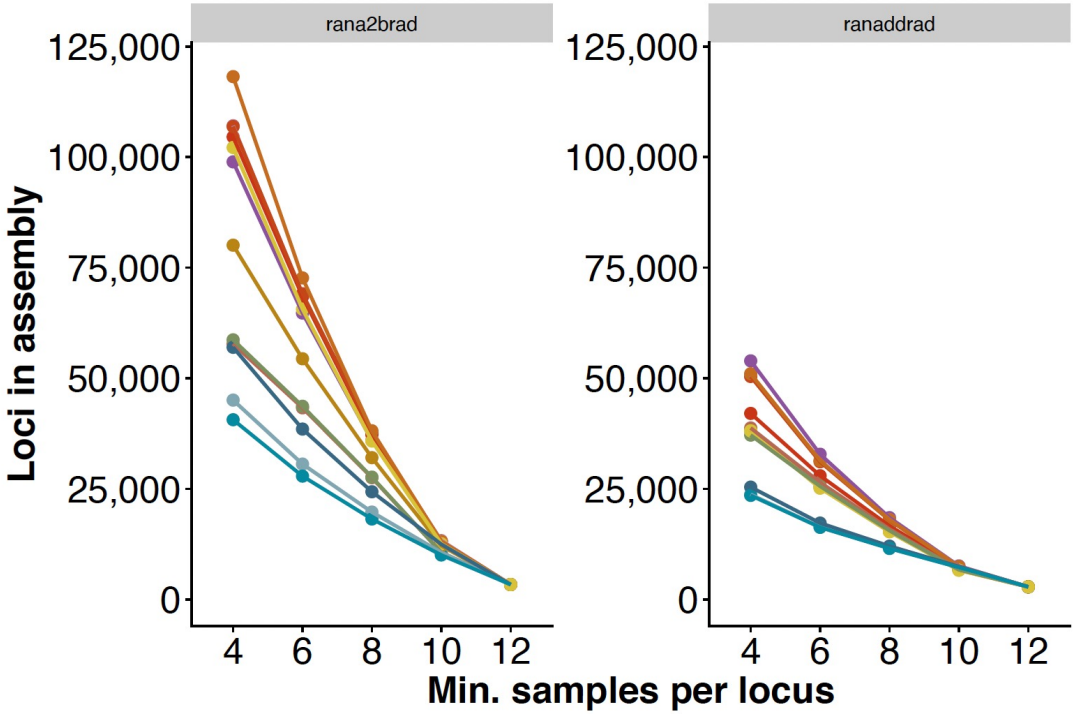
Missing data is a source of information (in a way) for building trees

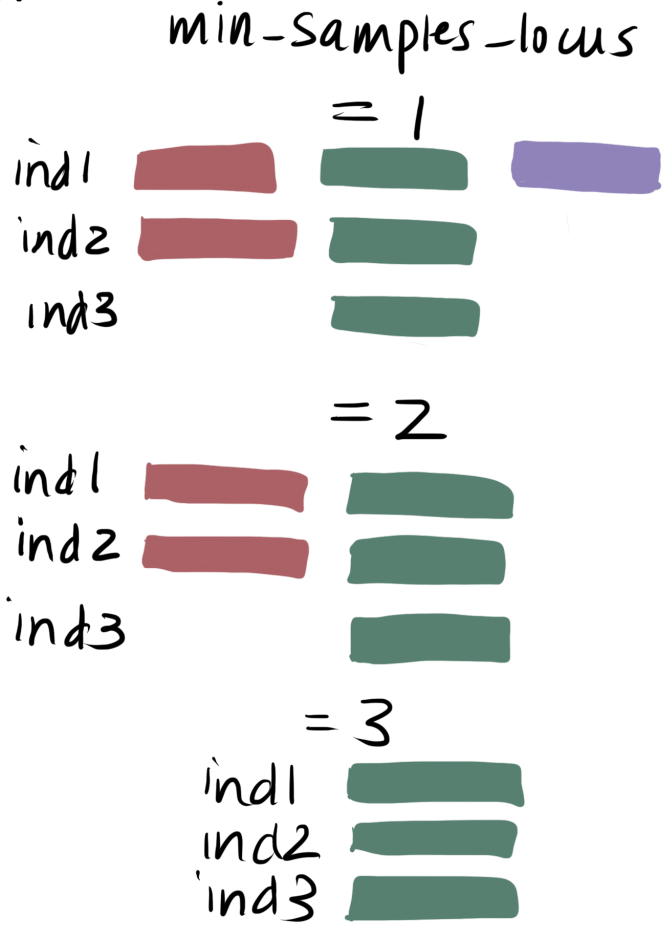
Eaton et al. (2017) Syst. Biol. 66(3):399–412

Min_samples_locus parameter

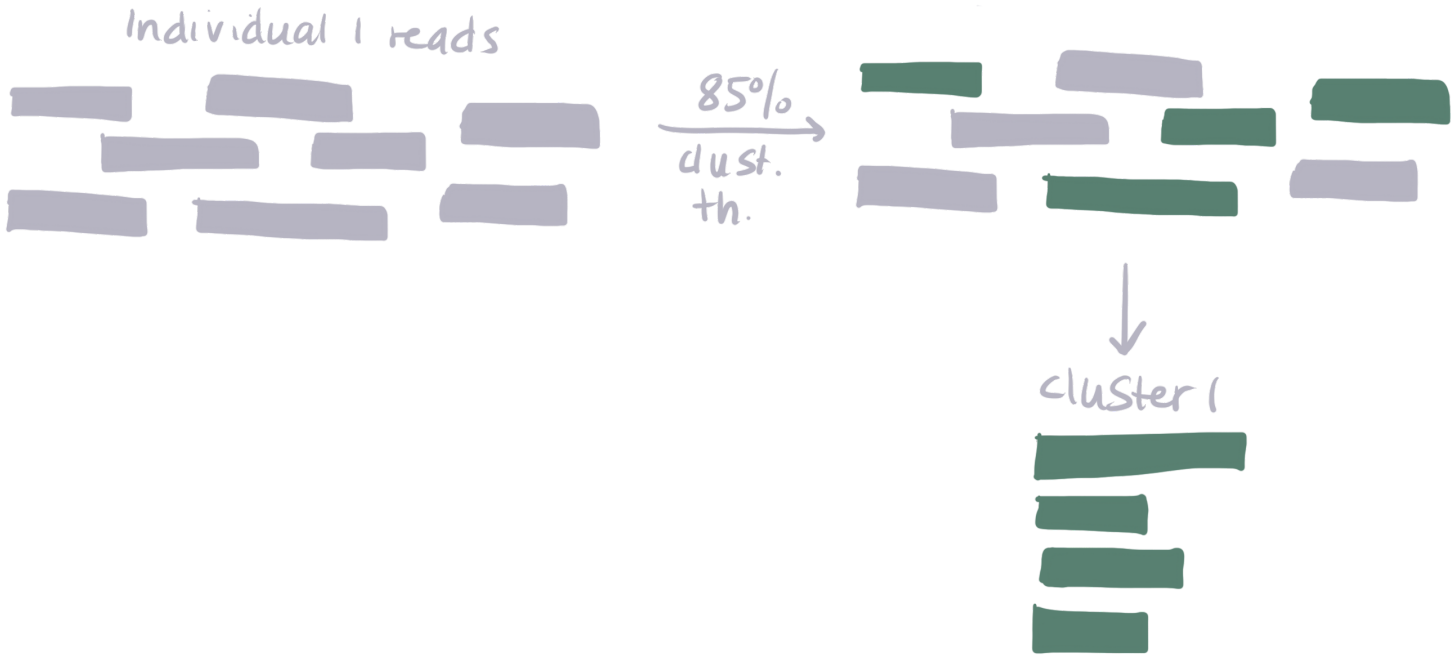


Min_samples_locus parameter

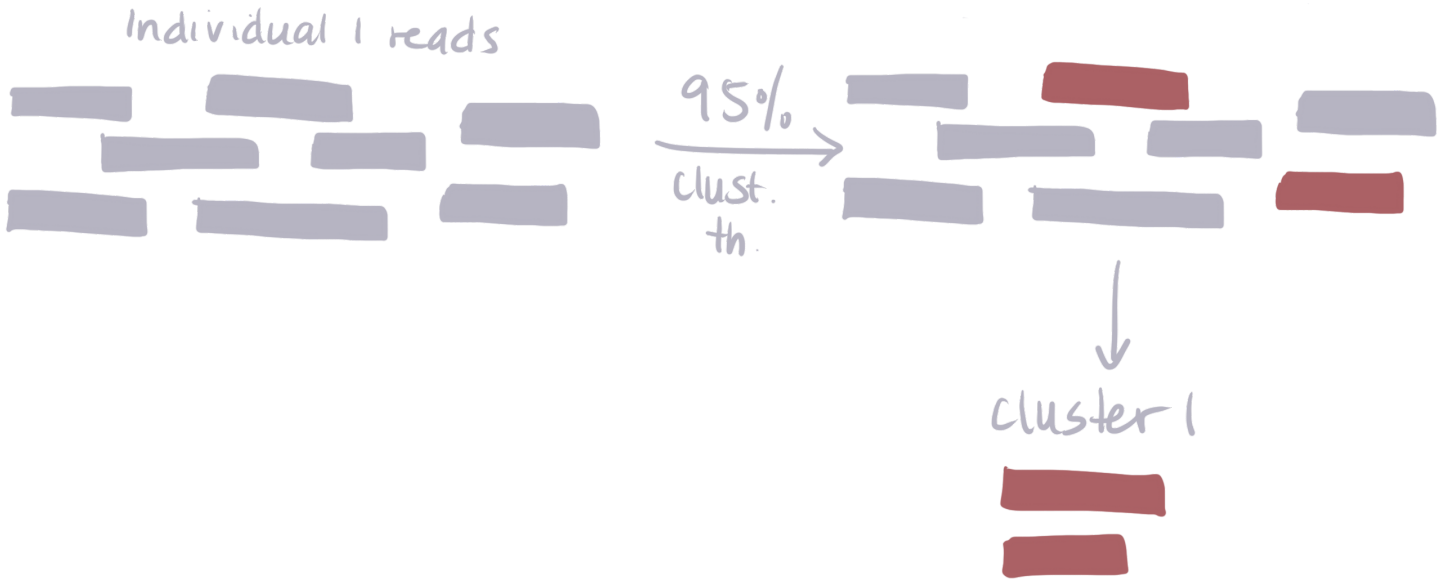




Clustering threshold is another way to check robustness



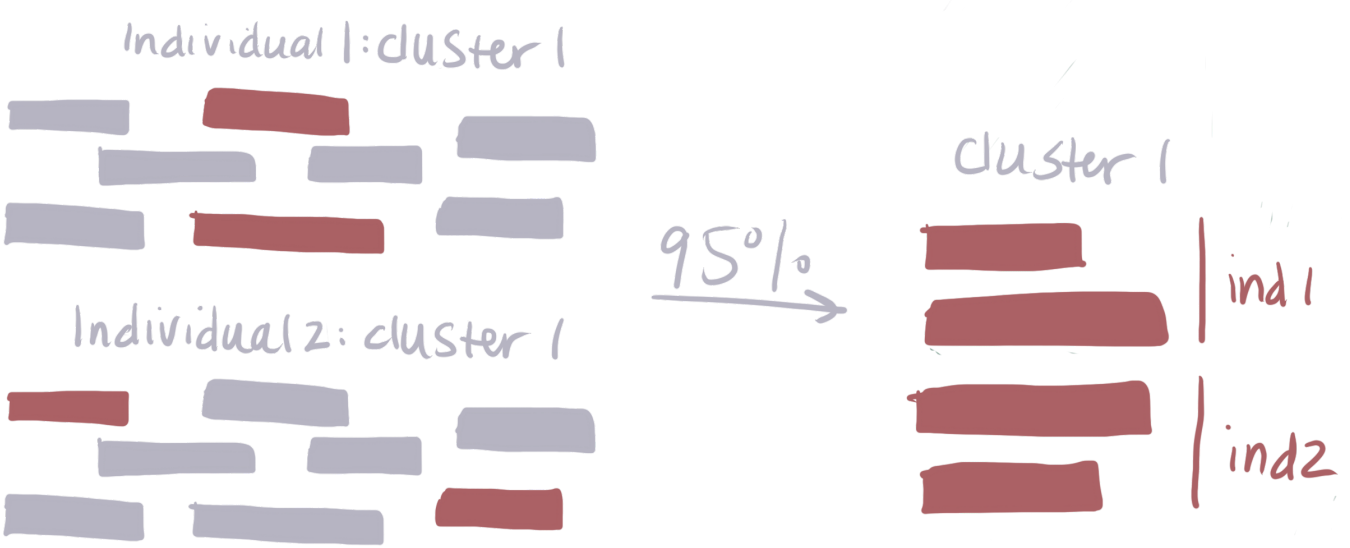
Clustering threshold



Clustering threshold: among individuals

Clustering threshold: among individuals

Clustering threshold: among individuals

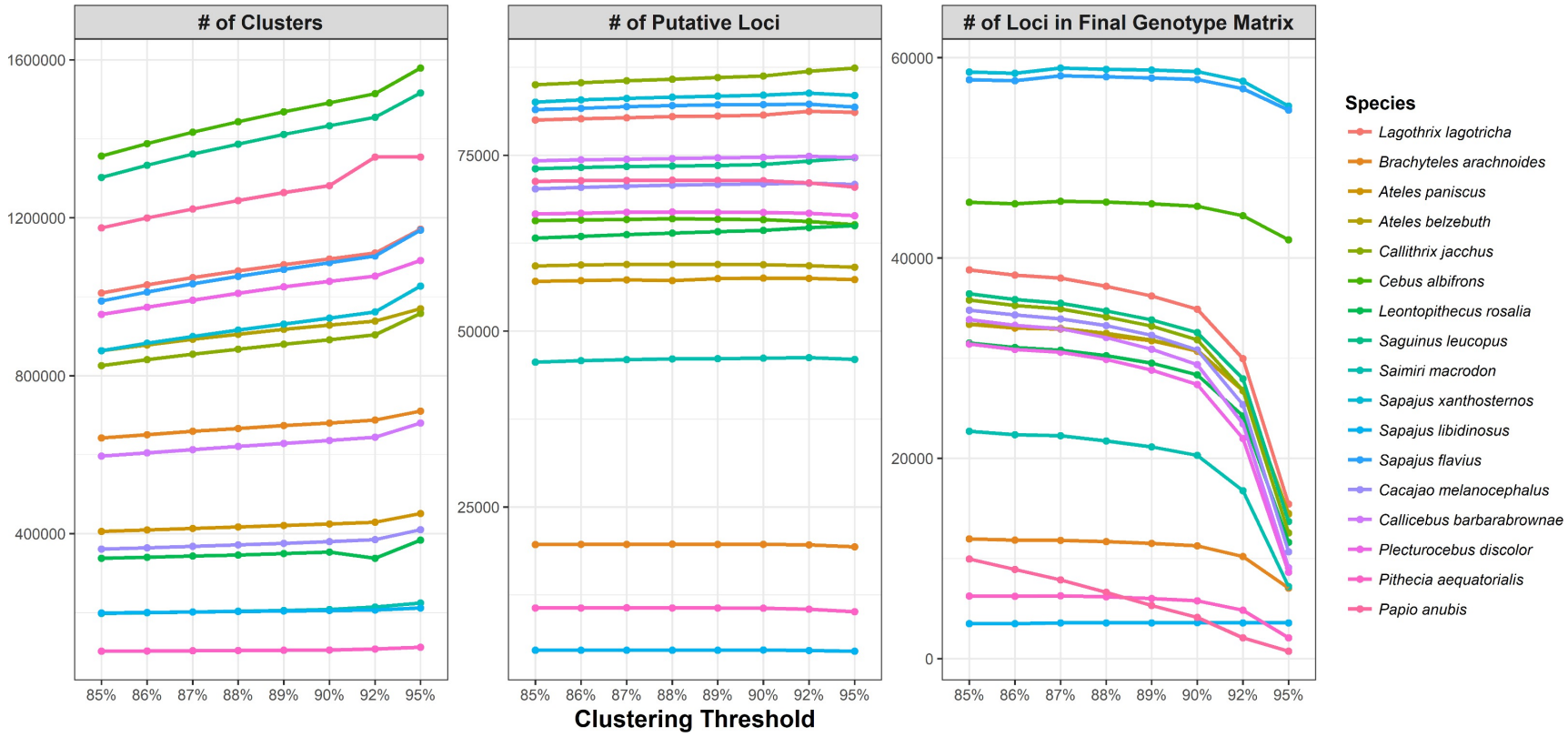


Clustering threshold

- What happens if we make the clustering threshold more relaxed (i.e., a lower %)?
  - We get more loci per cluster
  - Higher likelihood of clustering paralogs

- What happens if we make the clustering threshold more stringent (i.e., a higher %)?
  - We get fewer loci per cluster
  - Higher likelihood of losing homologs

Clustering threshold



Valencia et al. (2018) PLoS ONE 13(8):e0201254

What to expect

- You can’t learn programming overnight!
- Programming is problem-solving and all about fixing errors
  - Good judgment comes from experience
  - Experience comes from bad judgment
  - Make mistakes!
- Take breaks, avoid frustration

Rebecca Tarvin, https://rdtarvin.github.io/RADseq_Quito_2017/

Best practices

- Keep a “lab” notebook
- Find a supercomputer
- Adequate storage space (1-5 TB)
- Save multiple copies of unedited, raw data (processed data can be recreated)

```
############################## SUBSAMPLING THE DATA [.fq.tx] ##############################
# subsampling raw reads, placing into separate directories:
# EPIPEDOBATES
# Easiest to run these lines on their own (and not in a job); be sure to enter idev node
# copy the following; make sure you're within the seqtk folder to run this
# also be sure to have copied the epi.fq file into the seqtk folder, otherwise change
./seqtk sample epi.fq 27199735 > epi.fq.t1
./seqtk sample epi.fq 78841547 > epi.fq.t2
./seqtk sample epi.fq 133201348 > epi.fq.t3

# RANA
nano ranasampl
# copy the following
./seqtk sample rana.fq 24115588 >rana.fq.t1
./seqtk sample rana.fq 49309082 >rana.fq.t2
./seqtk sample rana.fq 99696069 >rana.fq.t3

# Move all .fq samples back into your main project folder


############################## TRIMMING BARCODES [.tr0] ##############################
# trim barcodes off sequences; creates trims (tr0) files
# need to make sure you do this SEPARATELY for each taxon since barcodes are shared

for F in epi.fq*; do echo "trim2bRAD_2barcodes_dedup.pl input=$F sampleID=1" >> epitrims;done
ls5_launcher_creator.py -j epitrims -n epitrims -t 0:30:00 -a Dendrobatidae -e eachambers@utexas.edu -w 2
# manually change CONTROL_FILE to LAUNCHER_JOB_FILE (see ***)
sbatch epitrims.slurm

for F in rana.fq*; do echo "trim2bRAD_2barcodes_dedup.pl input=$F sampleID=1" >>ranatrims;done
ls5_launcher_creator.py -j ranatrims -n ranatrims -t 0:30:00 -a Dendrobatidae -e eachambers@utexas.edu -w 2
# manually change CONTROL_FILE to LAUNCHER_JOB_FILE (see ***)
sbatch ranatrims.slurm

# *** Output from job: NOTICE: CONTROL_FILE variable depricated. Use LAUNCHER_JOB_FILE in the future.

# Move each of the 12 .tr0 files into their corresponding folder for each seq depth

#--------- duplication rates; get these from the output files
```

Rebecca Tarvin, https://rdtarvin.github.io/RADseq_Quito_2017/