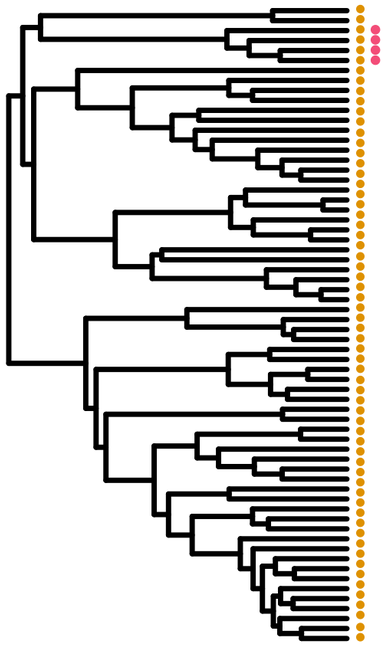
Phylogenomic perspectives on reproductive isolation and introgression

Botany Conference 2019, Tucson

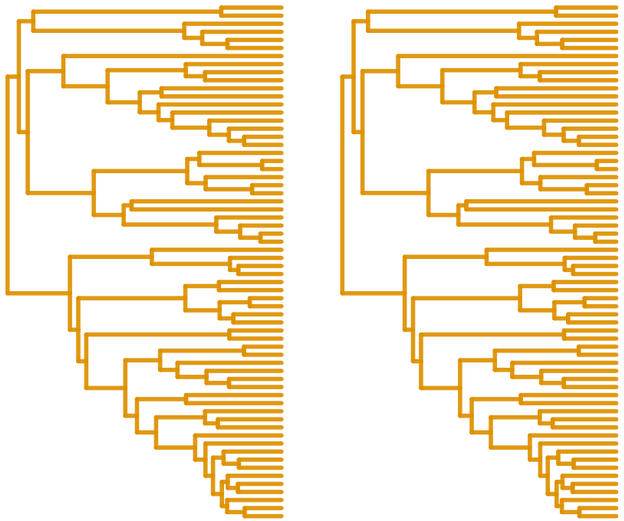
Deren Eaton, Columbia University

The goal of phylogenomics

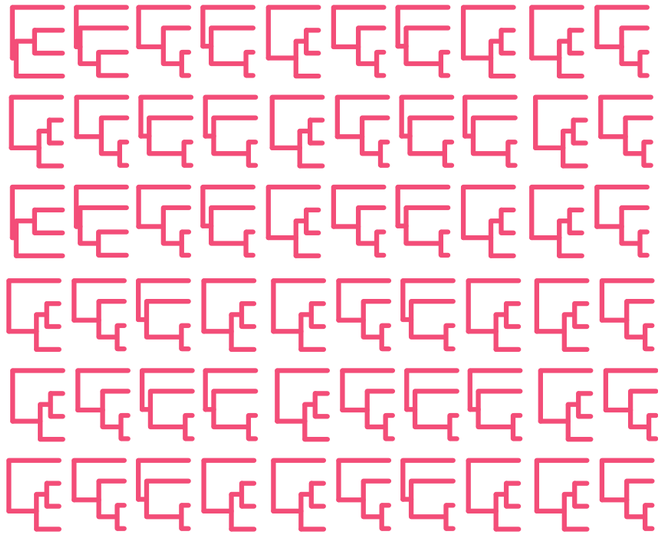
Characterize evolutionary relationships from a subset of sampled genomes.



few genes across many taxa

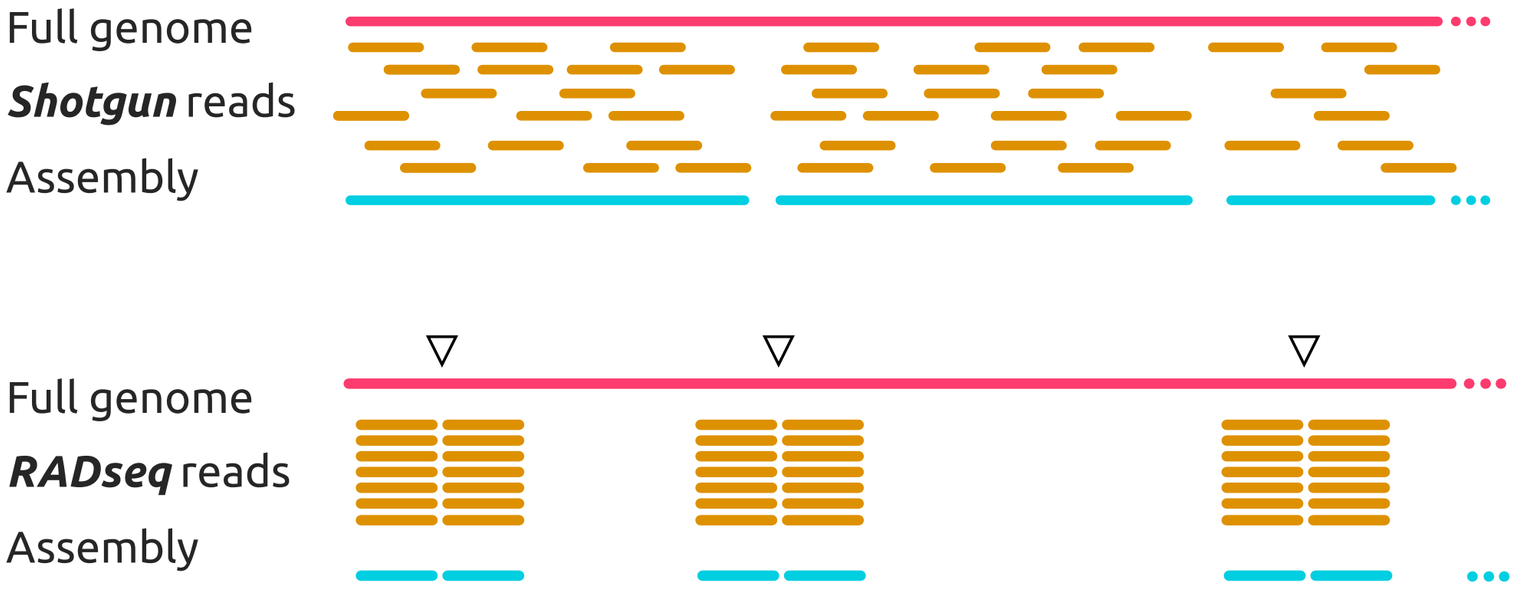


many genes across few taxa



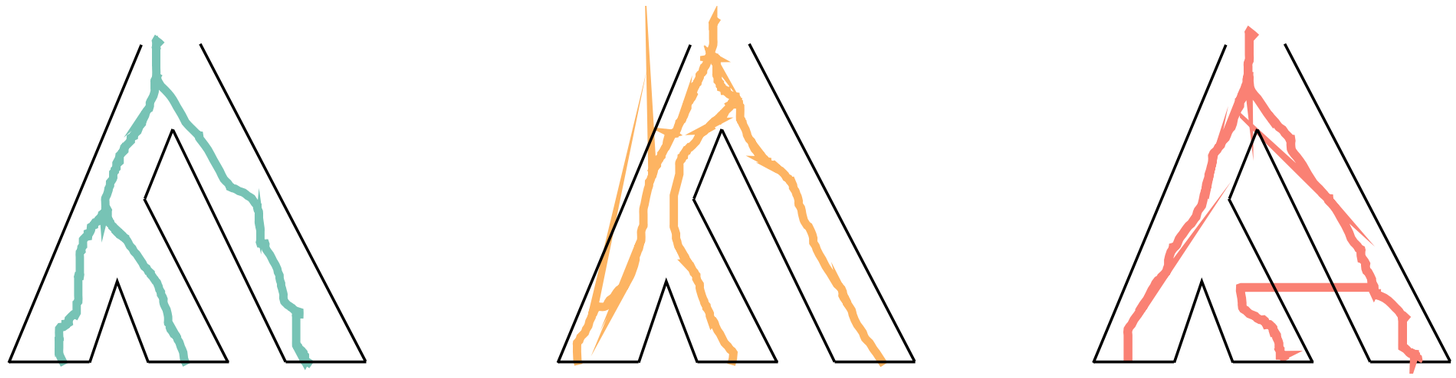
WGS vs. RAD-seq genomic sampling

Characterize whole genomes from a subset of sequenced markers.



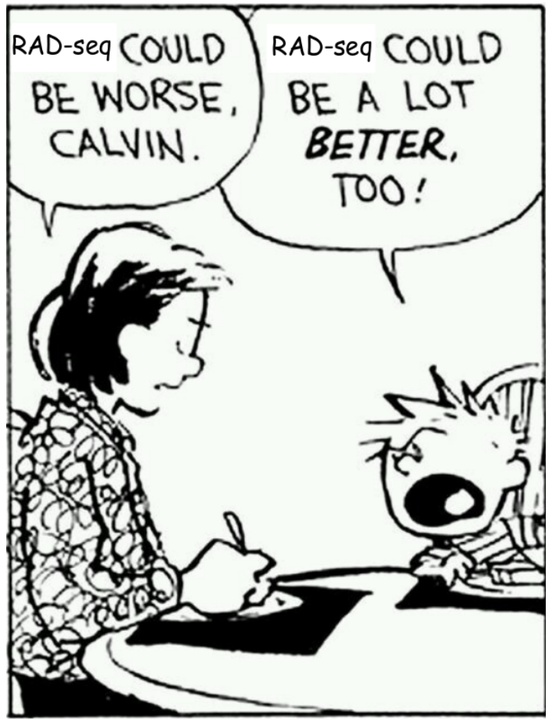
Coalescent variation

Different genomic regions have different genealocial histories.



Can sparse SNP sampling reconstruct genome-wide patterns?

Filtering and formatting to deal with missing data…

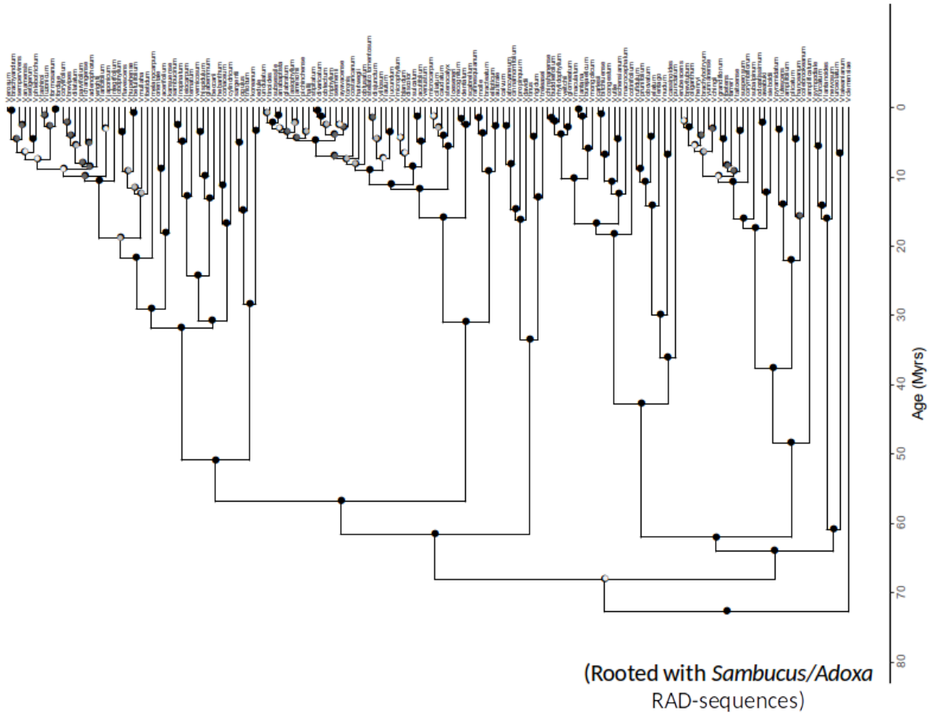


Viburnum Phylogeny

Species-level phylogenetic sampling
Published: 65 species; Eaton et al. (2015)
Current: 127 species; In Prep.

Assembled in *ipyrad* (Eaton 2014; Eaton & Overcast)
290K RAD loci (75% missing)
3.1M SNPs across 127 species

Species tree inferred with *tetrad* (Eaton et al. 2015)
Uses all SNP information for each quartet
(average ~30K SNPs per quartet)



(Rooted with *Sambucus/Adoxa*
RAD-sequences)

Viburn'ers

Michael Donoghue

Ivalu Cacho

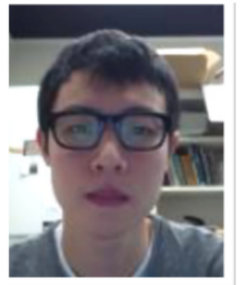
Erika Edwards

Beth Spriggs

Patrick Sweeney

Mark Olson

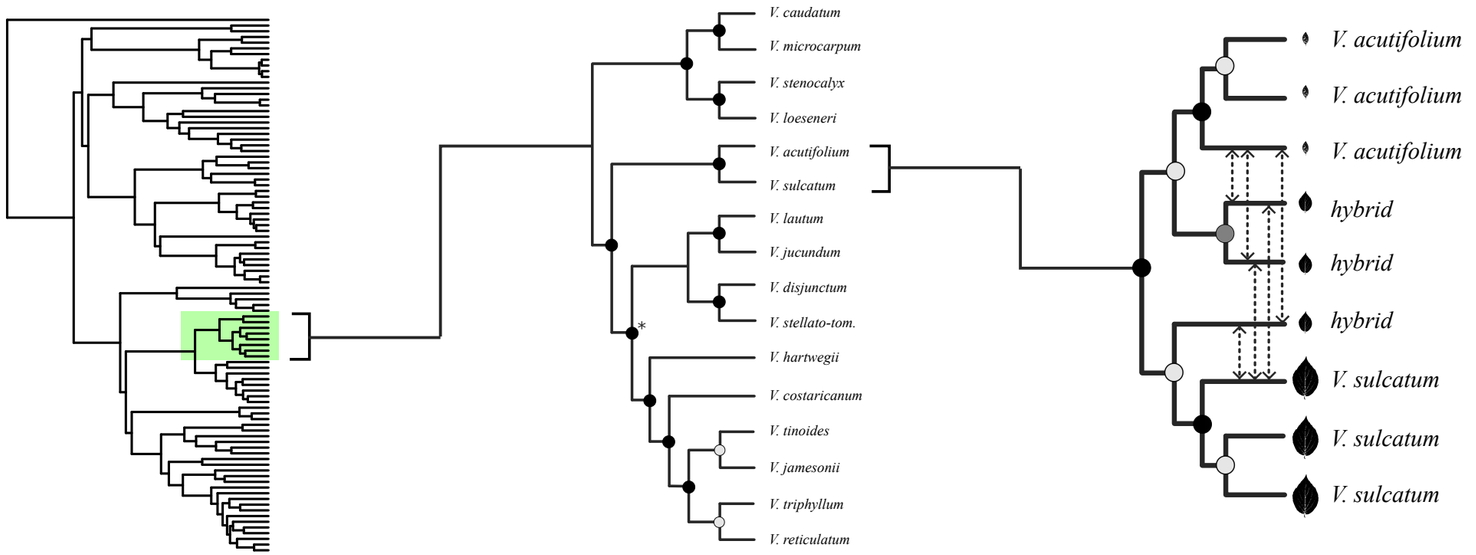
Morgan Moeglin



Brian Park

Viburnum global RAD sampling

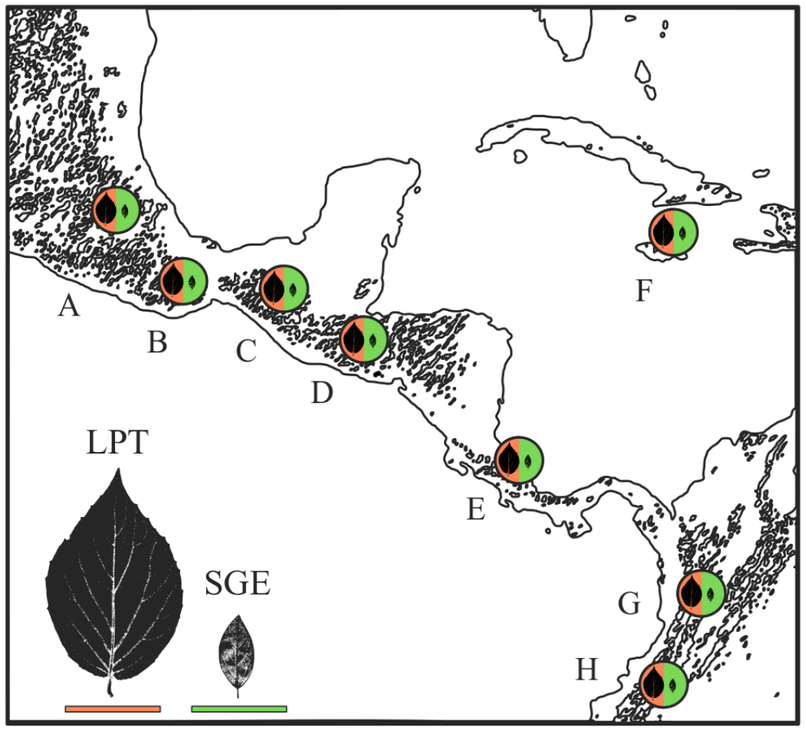
From global to population-level variation.

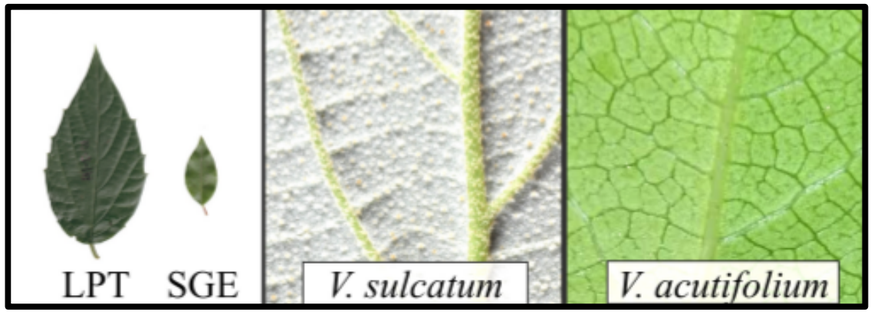


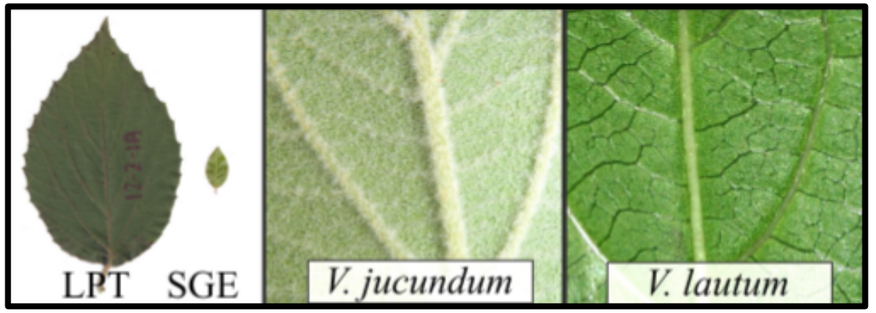
global *sub-clade* *populations*

Viburnum Orienotinus rapid radiation

~35 species from Mexico to Bolivia over ~10Ma

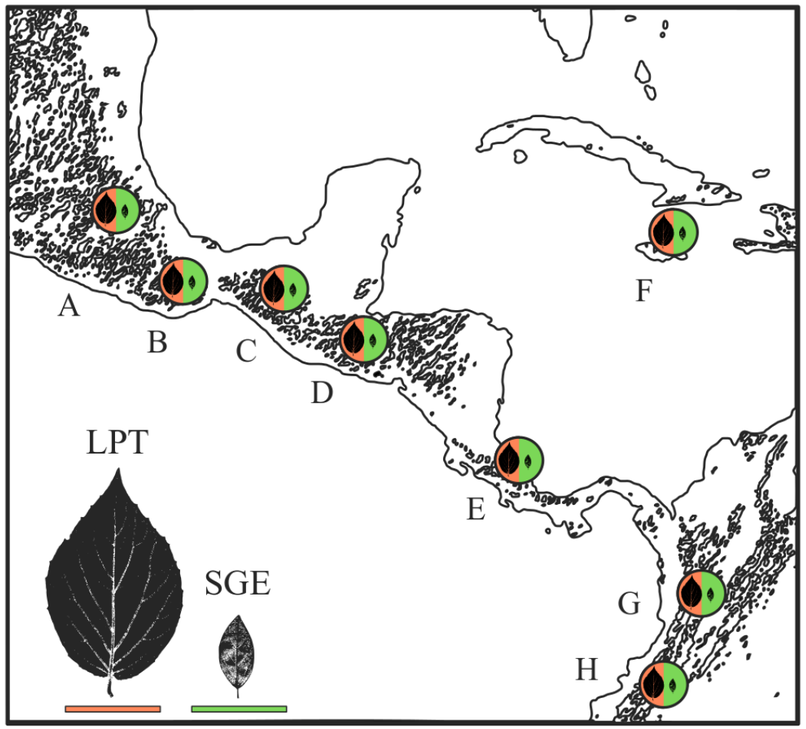


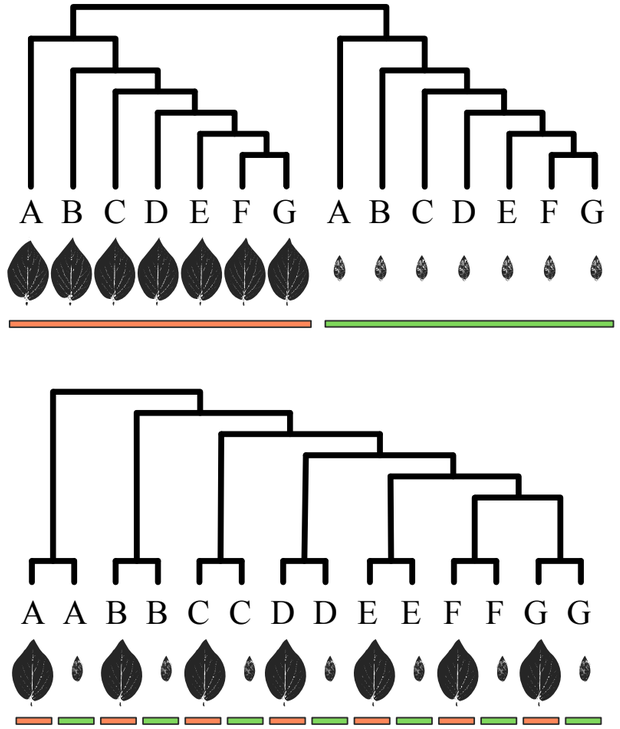




Viburnum Orienotinus rapid radiation

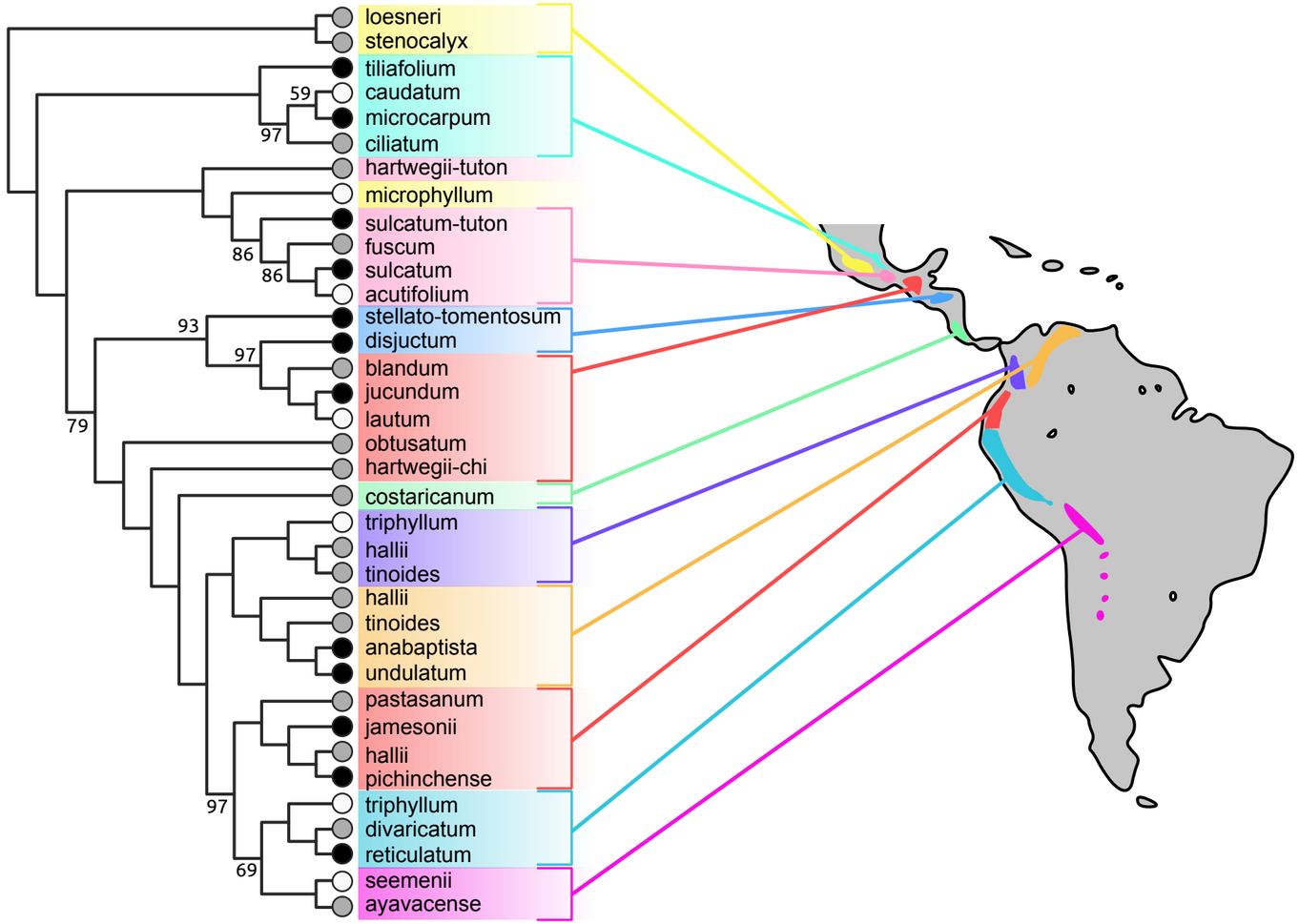
~35 species from Mexico to Bolivia over ~10Ma





Viburnum Orienotinus rapid radiation

~35 species from Mexico to Bolivia over ~10Ma


loesneri
stenocalyx
tiliafolium
caudatum
microcarpum
ciliatum
hartwegii-tuton
microphyllum
sulcatum-tuton
fuscum
sulcatum
acutifolium
stellato-tomentosum
disjuctum
blandum
jucundum
lautum
obtusatum
hartwegii-chi
costaricanum
triphyllum
hallii
tinoides
hallii
tinoides
anabaptista
undulatum
pastasanum
jamesonii
hallii
pichinchense
triphyllum
divaricatum
reticulatum
seemenii
ayavacense
59
97
86
86
93
97
79
97
69

Outline: RAD-seq phylogenomics in *ipyrad*

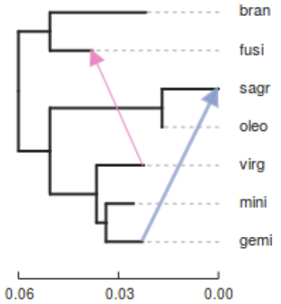
1. *ipyrad-analysis* toolkit.
2. Gene tree extraction: concatenation.
3. Gene tree distributions: sliding window consensus.
4. Sticking with SNPs: genome-wide inference.

ipyrad-analysis toolkit (and toytree) and jupyter

In [6]:
```
# init a treemix analysis object with some param arguments
tmx = ipa.treemix(
    data=data,
    imap=imap,
    minmap=minmap,
    seed=1234,
    root="bran,fusi",
    m=2,
)
```

```
Samples: 29
Sites before filtering: 349914
Filtered (indels): 0
Filtered (bi-allel): 13379
Filtered (mincov): 0
Filtered (minmap): 99517
Filtered (combined): 108292
Sites after filtering: 241622
Sites containing missing values: 231436 (95.78%)
Missing values in SNP matrix: 905662 (12.93%)
subsampled 30621 unlinked SNPs
```

In [7]:
```
# draw the best scoring admixture graph
tmx.run()
tmx.draw_tree();
```



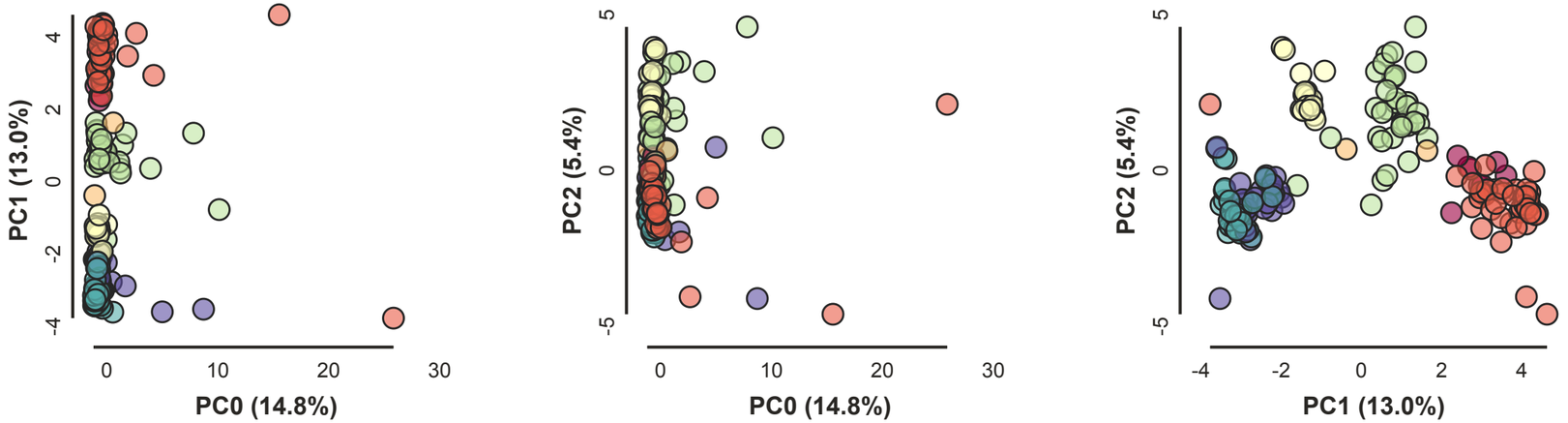
ipyrad-analysis toolkit

Filter or impute missing data; easily distribute massively parallel jobs.

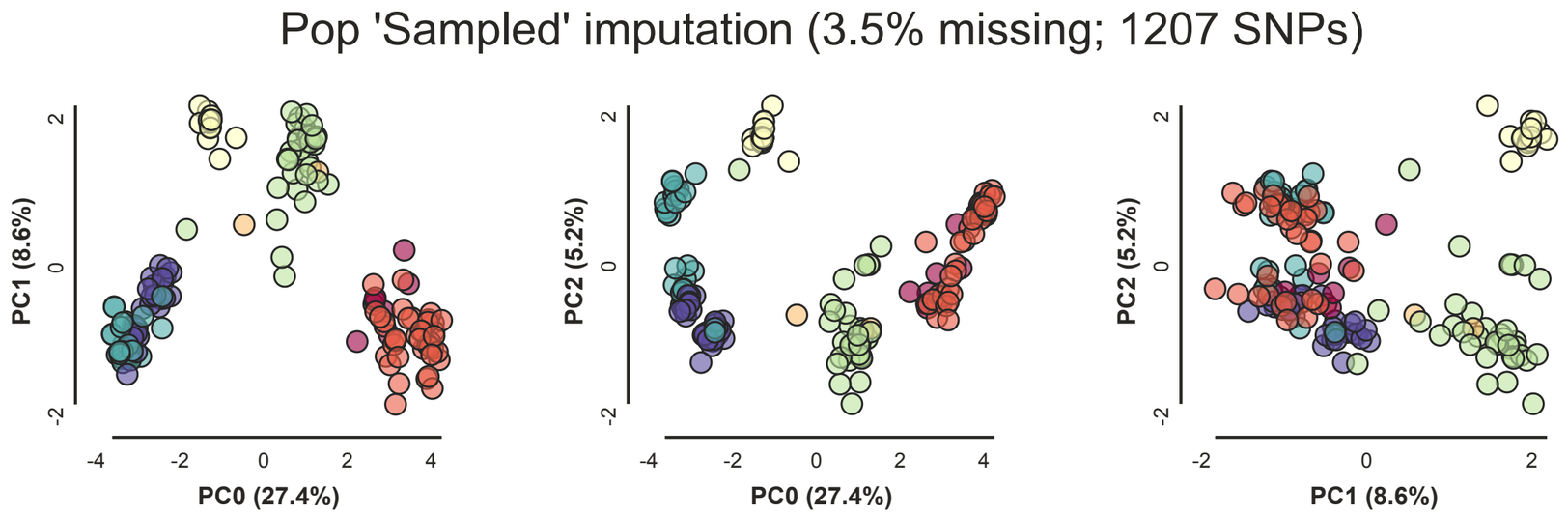
```
import ipyrad.analysis as ipa

# initiate an analysis tool with arguments
tool = ipa.pca(data=data, ...)

# run job (distribute in parallel)
tool.run()

# examine results
...
```

PCA: very sensitive to missing data

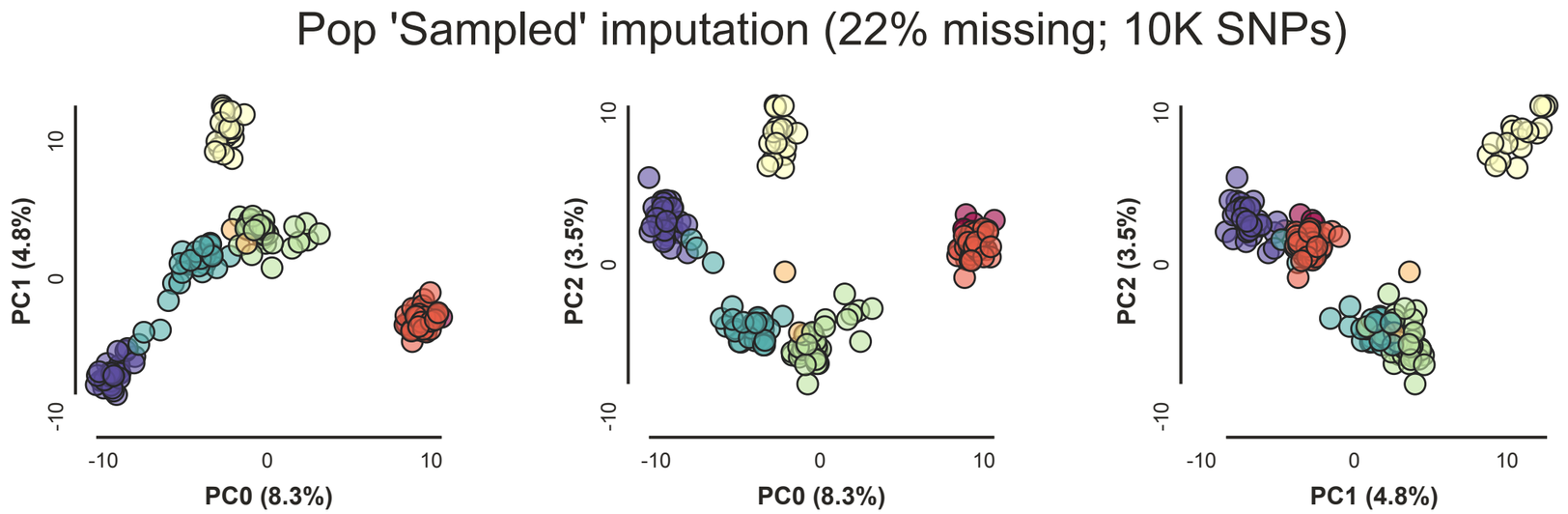
No imputation (3% missing; 1250 SNPs)



PCA: missing data imputed

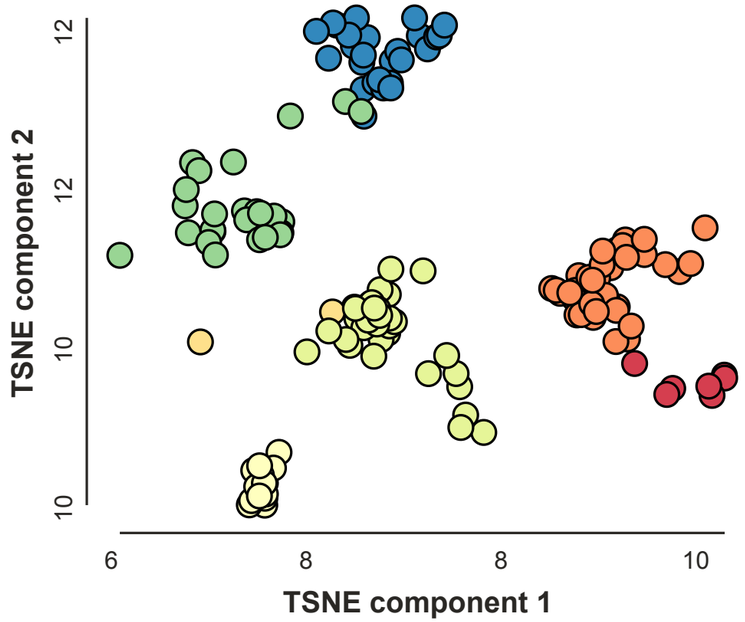


PCA: missing data imputed



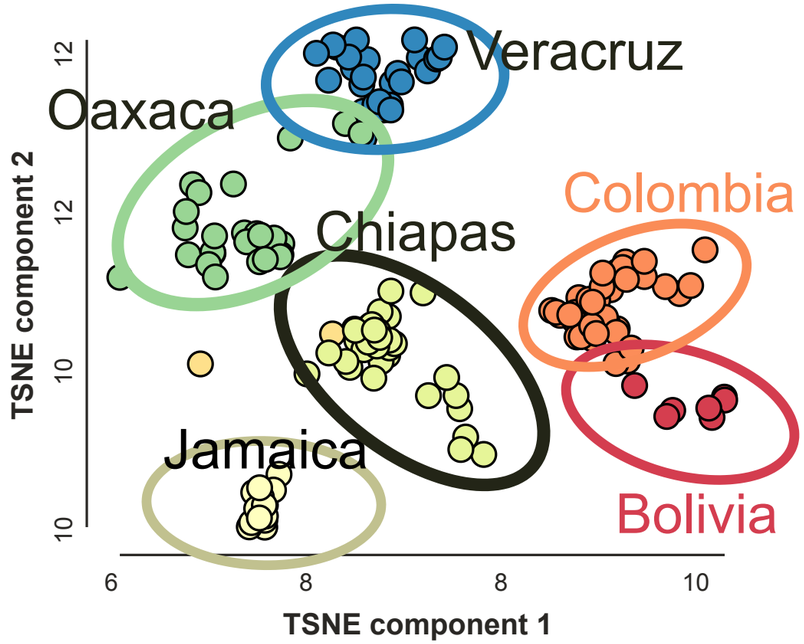
PCA + T-SNE: missing data imputed

TSNE manifold projection method (sckit-learn)



PCA + T-SNE: missing data imputed

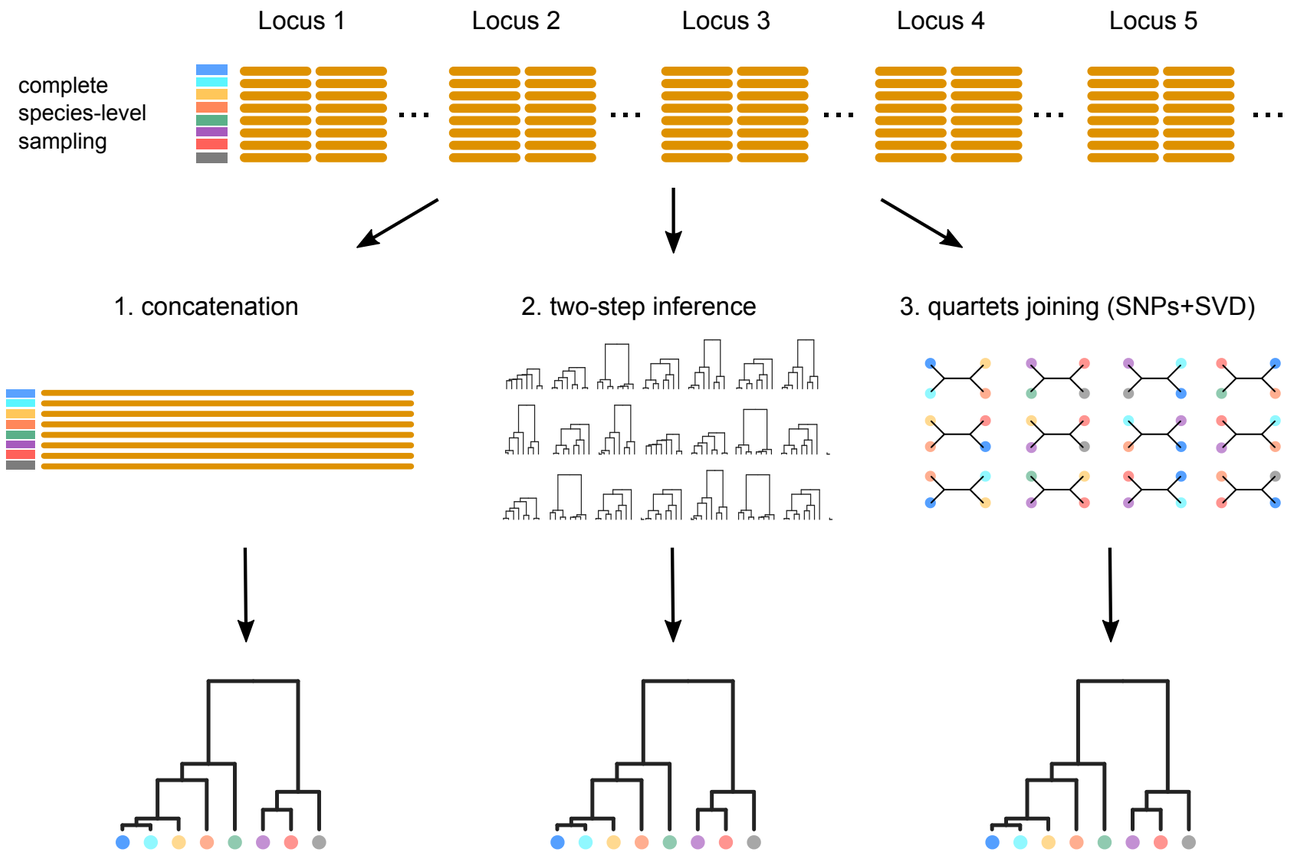
TSNE manifold projection method (sckit-learn)



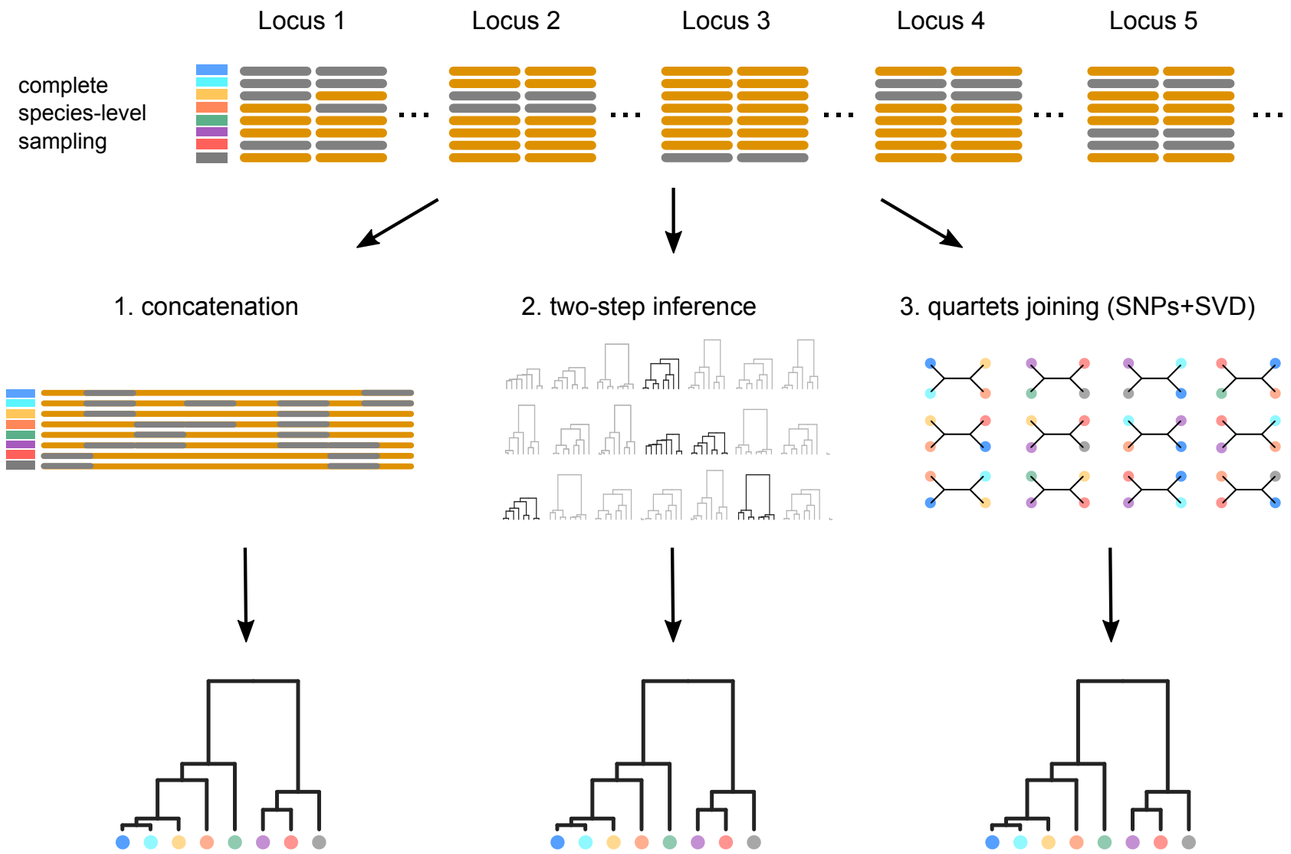
Outline: RAD-seq phylogenomics in *ipyrad*

1. *ipyrad-analysis* toolkit.
2. Gene tree extraction: concatenation.
3. Gene tree distributions: sliding window consensus.
4. Sticking with SNPs: genome-wide inference.

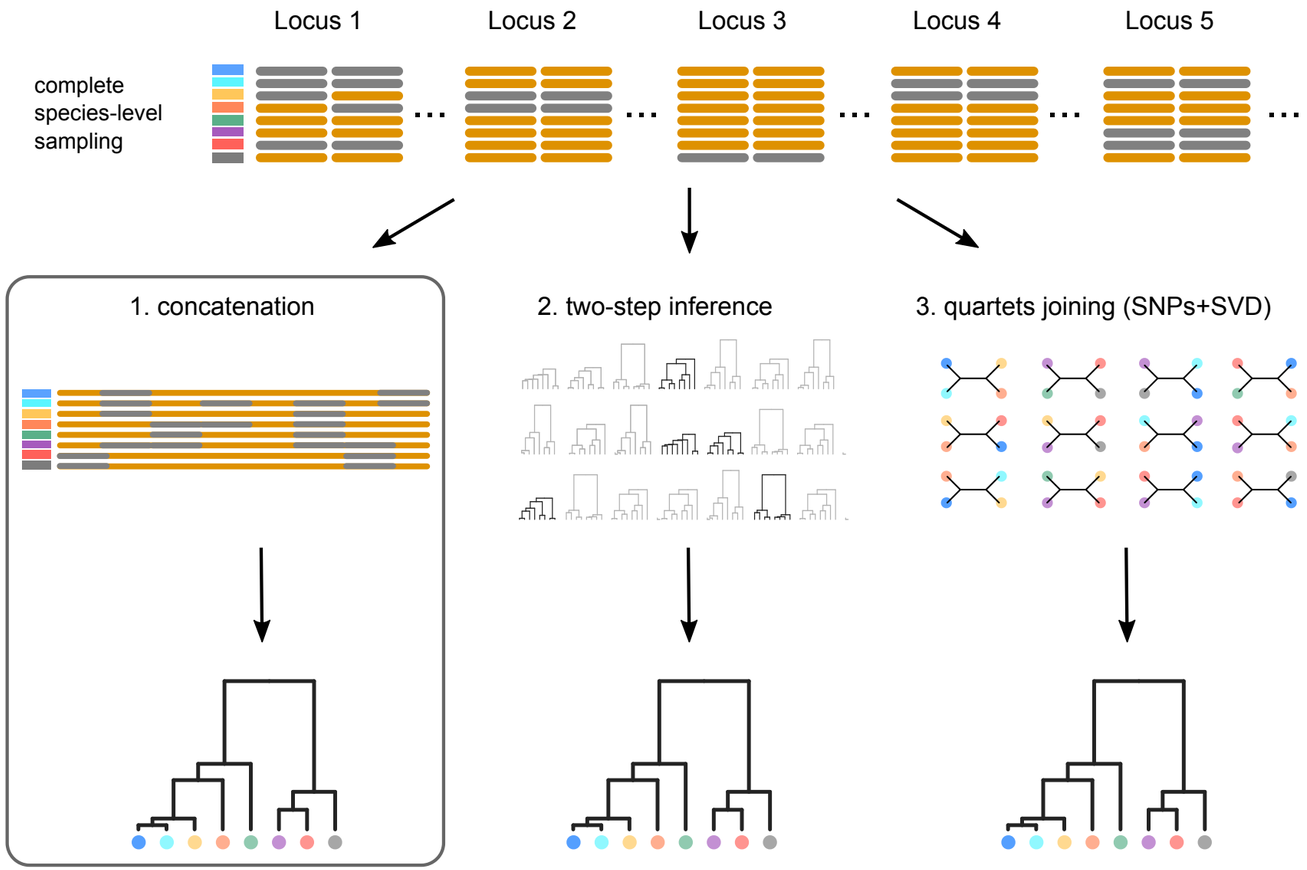
Missing data in phylogenetics



Missing data in phylogenetics



Missing data in phylogenetics



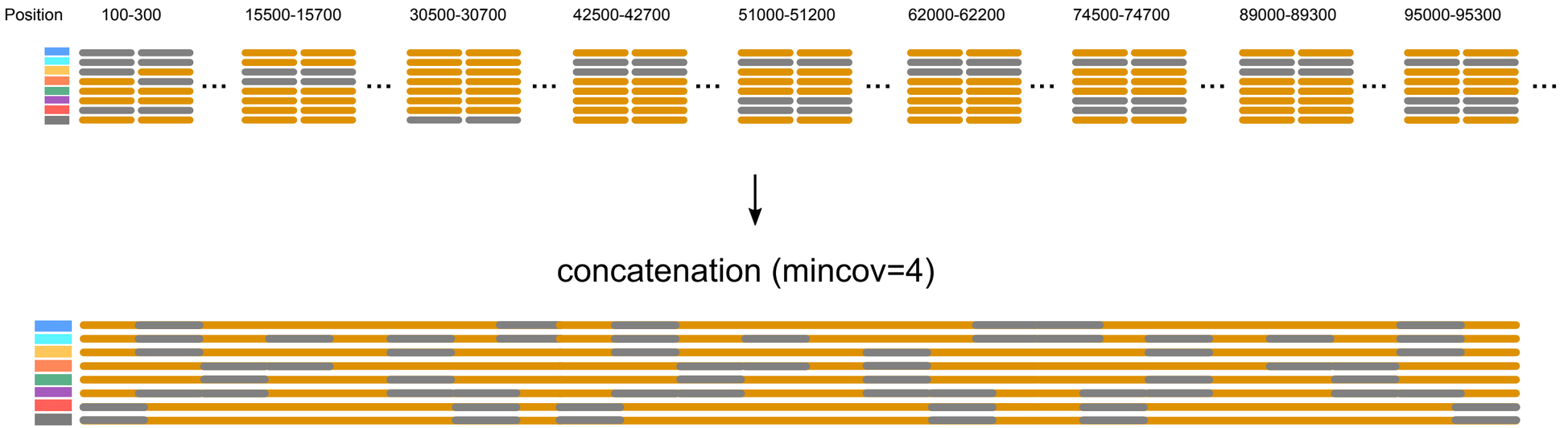
Window_extracter: extract, filter, format.

Reference mapped RAD loci can be *"spatially binned"* to form larger loci.

```
import ipyrad.analysis as ipa

# initiate an analysis tool with arguments
tool = ipa.window_extacter(
    data=data,
    scaffold_idx=0,
    start=0,
    end=1000000,
)

# writes a phylip file
tool.run()
```

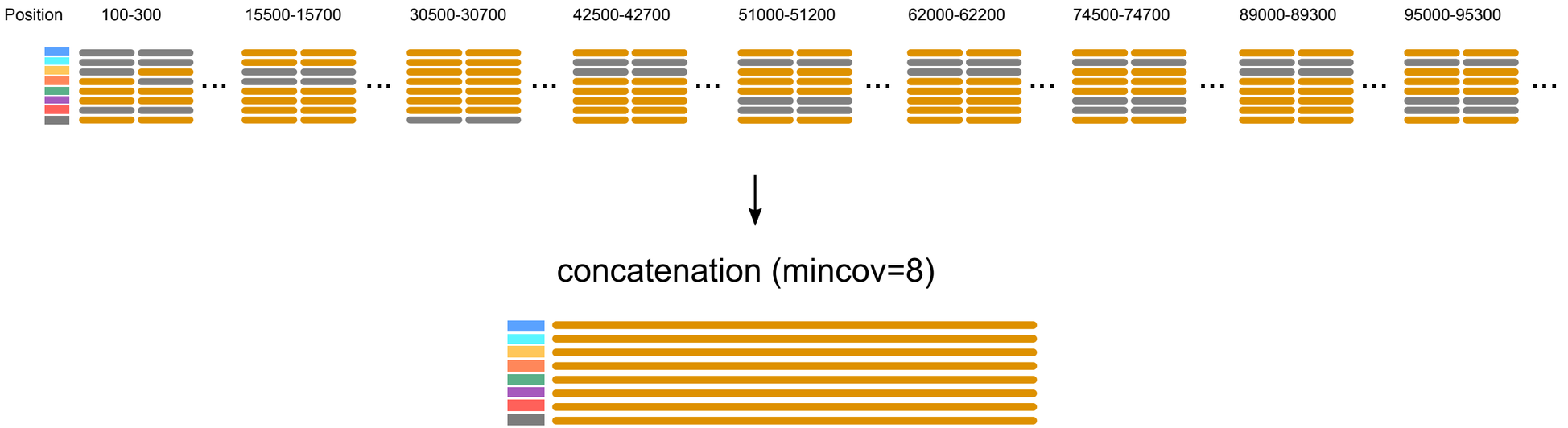
Window_extracter: extract, filter, format.

Reference mapped RAD loci can be *"spatially binned"* to form larger loci.



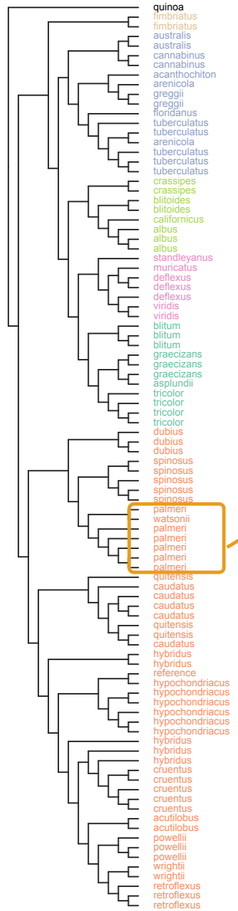
Window_extracter: extract, filter, format.

Reference mapped RAD loci can be *"spatially binned"* to form larger loci.

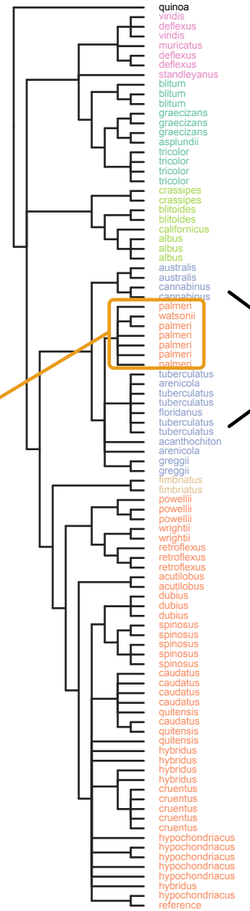


Herbicide resistance among *Amaranthus* species.

Chromosome 1
concatenation tree



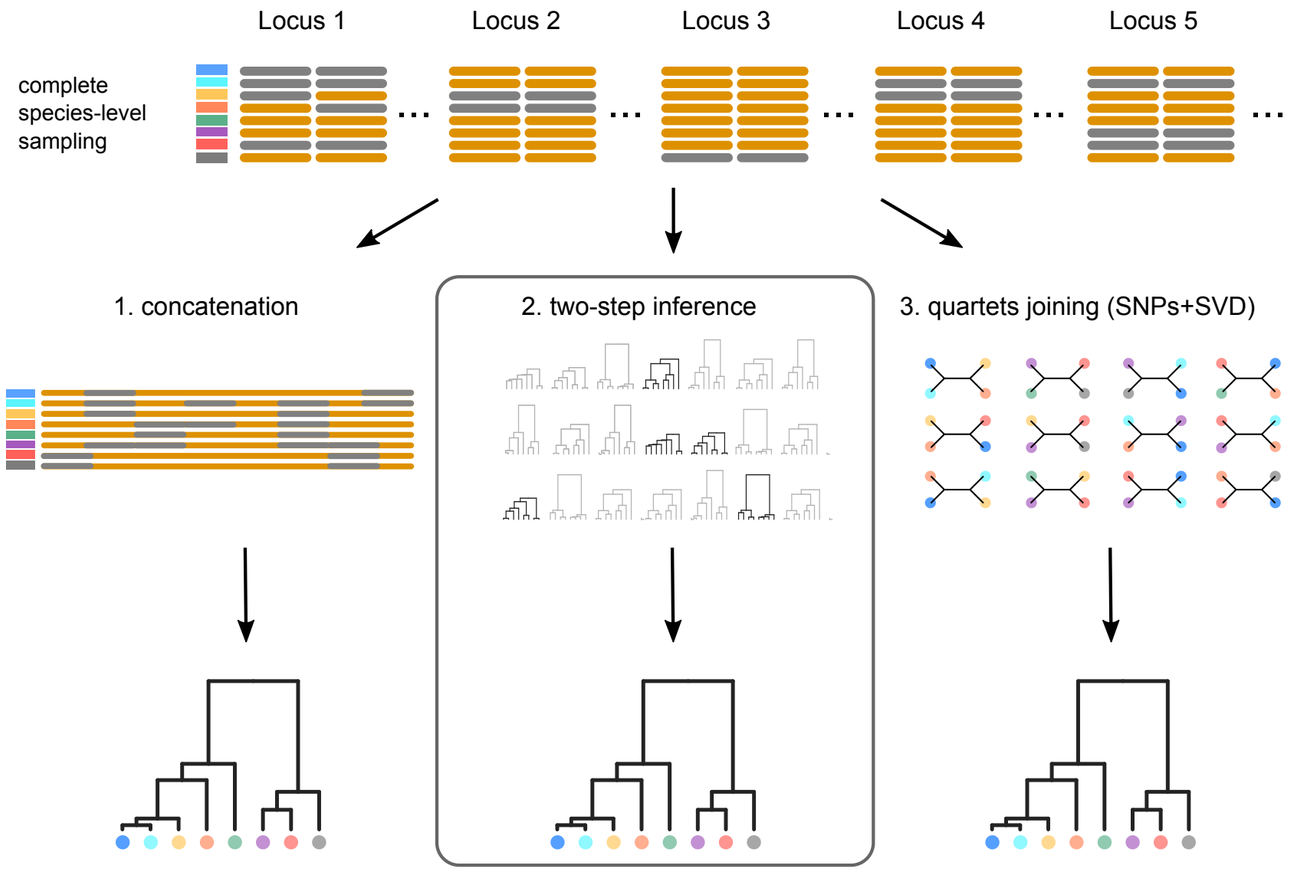
1Mb window at known
herbicide resistance gene



Sandra Hoffberg
Eaton Lab Postdoc

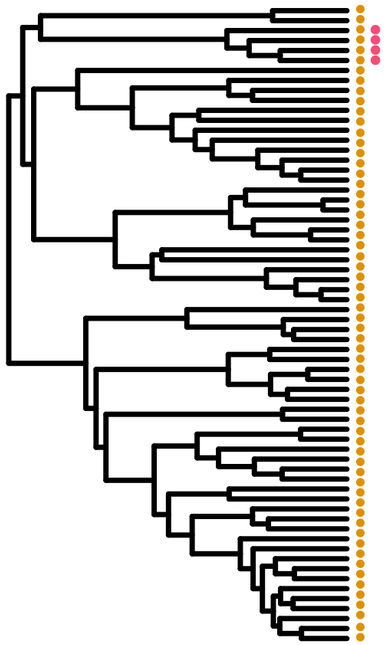
Introgression among
the two most notorious weeds:
A. palmeri (pigweed)
A. tuberculatus (waterhemp)

Missing data in phylogenetics

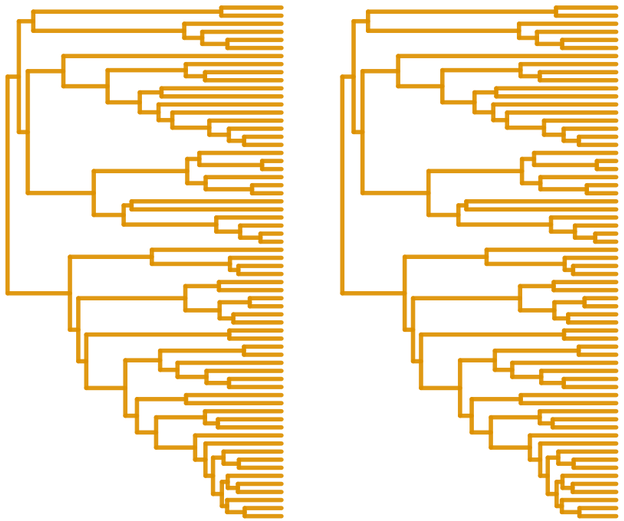


Missing data in phylogenetics

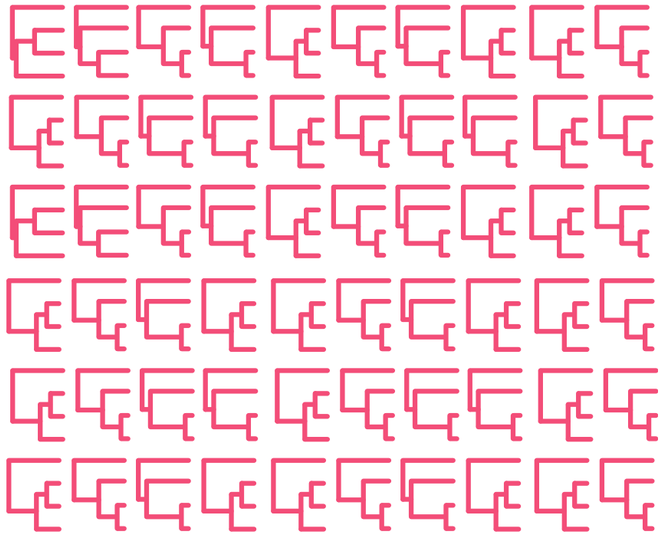
Goal: A distribution of gene trees representing every species.



few genes across many taxa

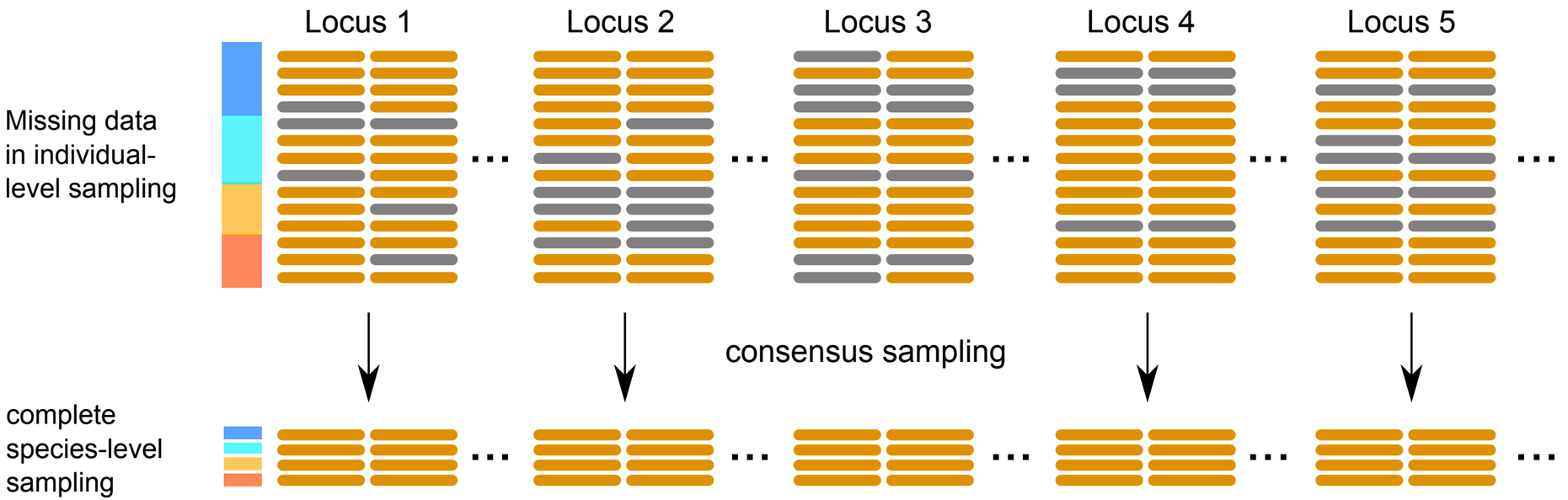


many genes across few taxa



Missing data: Consensus sampling

Represent species by the consensus genotype across sampled individuals



treeslider: extract windows across chromosomes.

Runs raxml on windows and parses results into a "tree_table"

```
# define population groups
imap = {
    "sp1": ["a0", "a1", "a2", "a3"],
    "sp2": ["b0", "b1", "b2", "b3"],
    "sp3": ["c0", "c1", "c2", "c3"],
    "sp4": ["d0", "d1", "d2", "d3"],
}

# initiate an analysis tool with arguments
tool = ipa.treeslider(
    data=data,
    window_size=1e6,
    slide_size=1e6,
    imap=imap,
)

# distributes raxml jobs across all 1M windows in data set
tool.run()
```

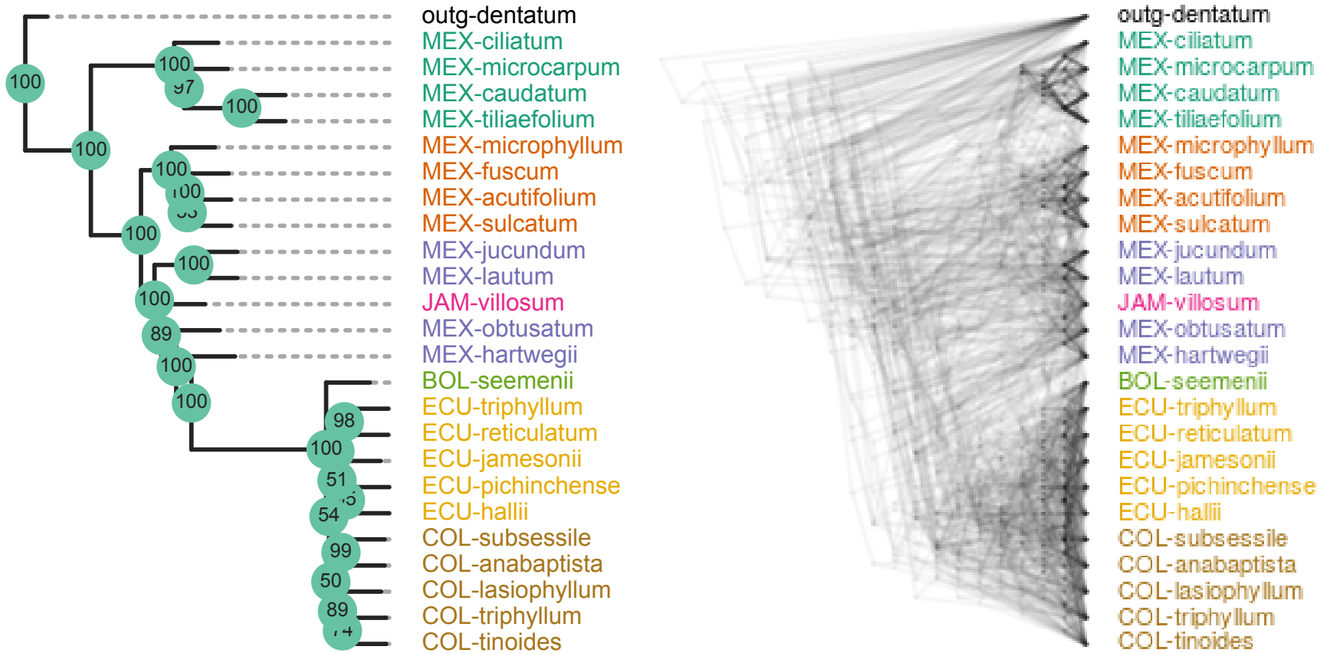
treeslider: extract windows across chromosomes.

	index	scaffold	start	end	sites	snps	samples	missing	tree
0	21	21	0	672631	4534	64	25	0.12	(BOL-seemenii:0.000722696,(ECU-pichinchense:0....
1	46	46	0	134933	1643	20	25	0.16	(BOL-seemenii:1e-06,(COL-subsessile:0.00106464...
2	56	56	0	537492	2223	33	25	0.13	(BOL-seemenii:1e-06,(COL-subsessile:0.00106464...
3	111	111	0	127917	524	11	25	0.25	(MEX-obtusatum:0.00153093,BOL-seemenii:1e-06,(...
4	127	127	0	259125	2006	20	25	0.13	(MEX-obtusatum:0.00153093,BOL-seemenii:1e-06,(...
5	182	182	0	1881	167	10	15	0.15	(COL-subsessile:1e-06,BOL-seemenii:0.00601782,...
6	320	320	0	321153	445	11	25	0.15	(BOL-seemenii:1e-06,(COL-triphyllum:1e-06,COL-...
7	340	340	0	429078	4348	48	25	0.16	(BOL-seemenii:0.00150841,((COL-anabaptista:0.0...
8	342	342	0	117234	981	10	25	0.13	(BOL-seemenii:0.0015864,(ECU-jamesonii:1e-06,E...
9	384	384	0	108573	892	14	25	0.20	(BOL-seemenii:1e-06,(COL-subsessile:1e-06,(COL...
10	395	395	0	851940	2042	28	25	0.20	(ECU-jamesonii:0.00398544,BOL-seemenii:1e-06,(...
11	401	401	0	225291	3718	57	25	0.15	(BOL-seemenii:0.000999779,(COL-anabaptista:0.0...
12	454	454	0	367849	537	10	25	0.36	(COL-subsessile:1e-06,BOL-seemenii:1e-06,((COL...
13	478	478	0	88302	360	13	25	0.30	(COL-lasiophyllum:1e-06,BOL-seemenii:1e-06,(EC...
14	515	515	0	235196	1111	11	25	0.18	(ECU-hallii:1e-06,BOL-seemenii:0.000984764,((C...
15	548	548	0	210332	2825	32	25	0.20	(BOL-seemenii:0.000448699,(COL-tinoides:0.0025...
16	554	554	0	126796	712	13	25	0.25	(ECU-hallii:1e-06,BOL-seemenii:0.001867,((COL-...
17	569	569	0	49074	360	11	25	0.27	(COL-subsessile:0.00379025,BOL-seemenii:0.0055...
18	627	627	0	198453	1597	16	25	0.12	(BOL-seemenii:0.000717776,(ECU-hallii:1e-06,(E...
19	654	654	0	135088	625	12	25	0.13	(ECU-pichinchense:1e-06,BOL-seemenii:0.0018493...
20	656	656	0	499297	5311	75	25	0.12	(BOL-seemenii:0.00293187,(COL-triphyllum:0.001...

Consensus sampling

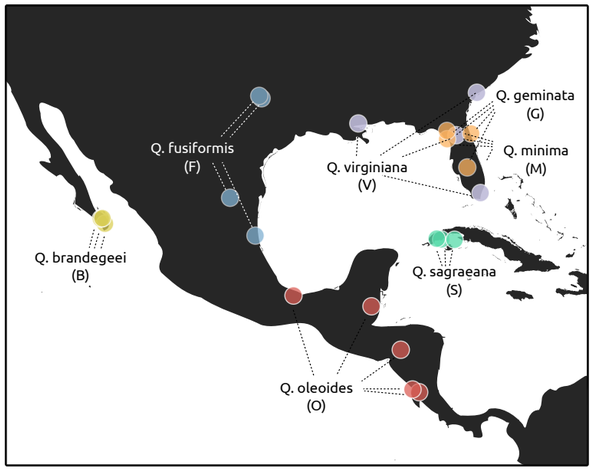
Recovers 5,500 informative gene trees (>50 SNPs) with no missing data across 25 taxa.

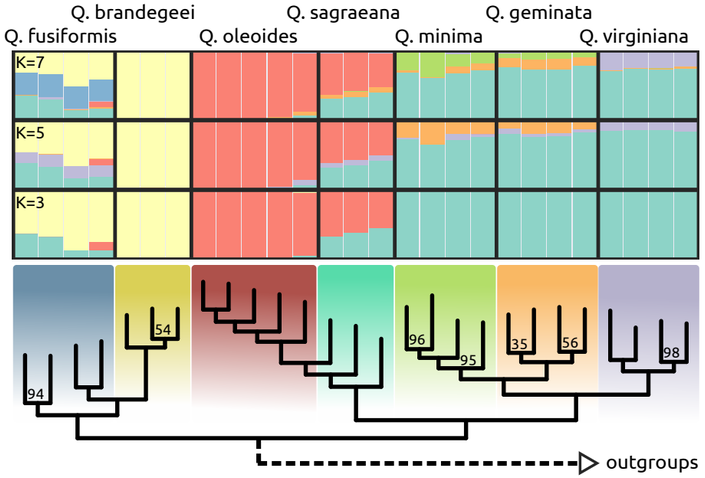
ASTRAL species tree and cloud tree of RAxML gene trees

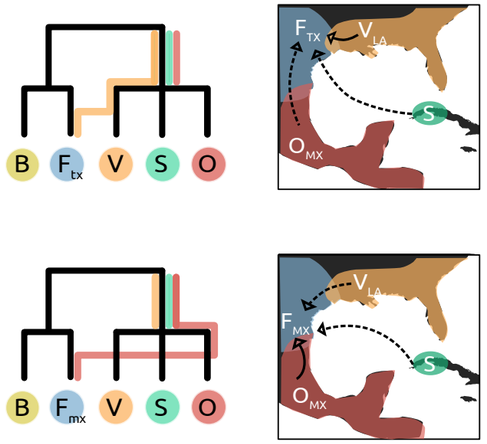


Another data set: *Quercus* section Virentes

Consensus sampling yields 3X as many fully sampled loci (>30K)





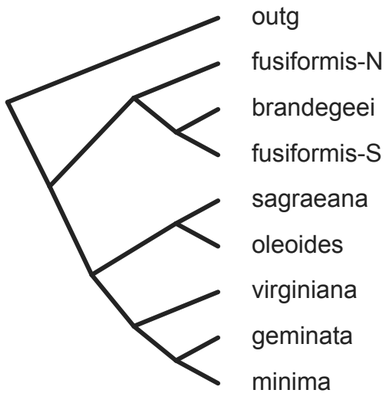


Hipp et al. (2014); Eaton et al. (2015); Cavender-Bares et al. (2015)

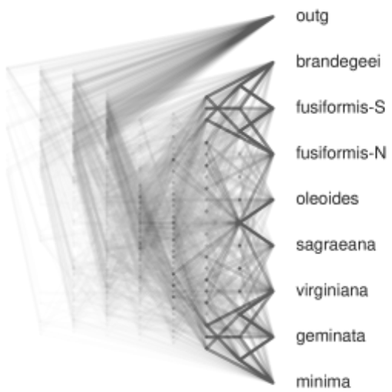
Sliding windows

How well do concatenated RAD windows represent gene tree variation?

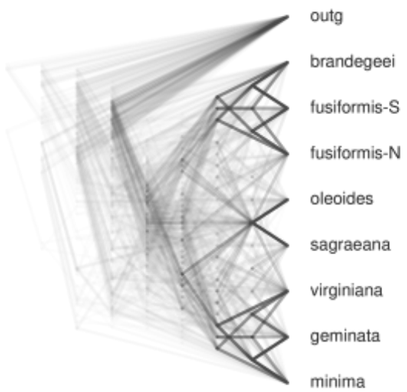
1 tree
chromosome 2
~1.1M sites
10,139 SNPs



711 trees
1Mb windows
mean 10K sites
mean 96 SNPs



352 trees
2Mb windows
mean 20K sites
mean 193 SNPs



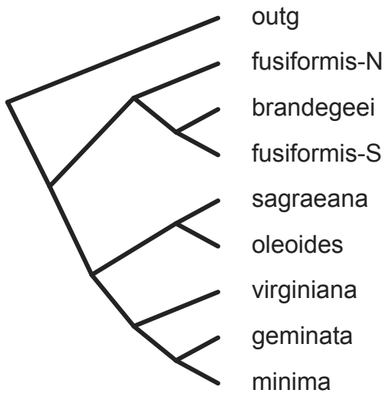
139 trees
5Mb windows
mean 50K sites
mean 480 SNPs

RAxML gene trees.

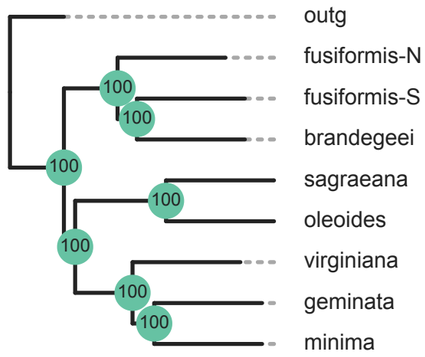
Sliding windows

How well do concatenated RAD windows represent gene tree variation?

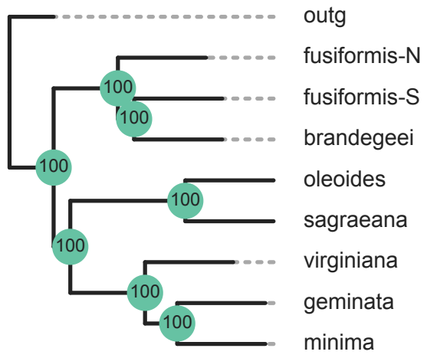
1 tree
chromosome 2
~1.1M sites
10,139 SNPs



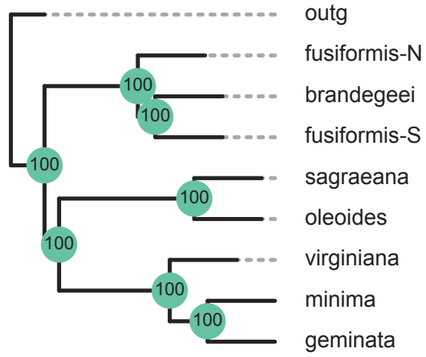
711 trees
1Mb windows
mean 10K sites
mean 96 SNPs



352 trees
2Mb windows
mean 20K sites
mean 193 SNPs



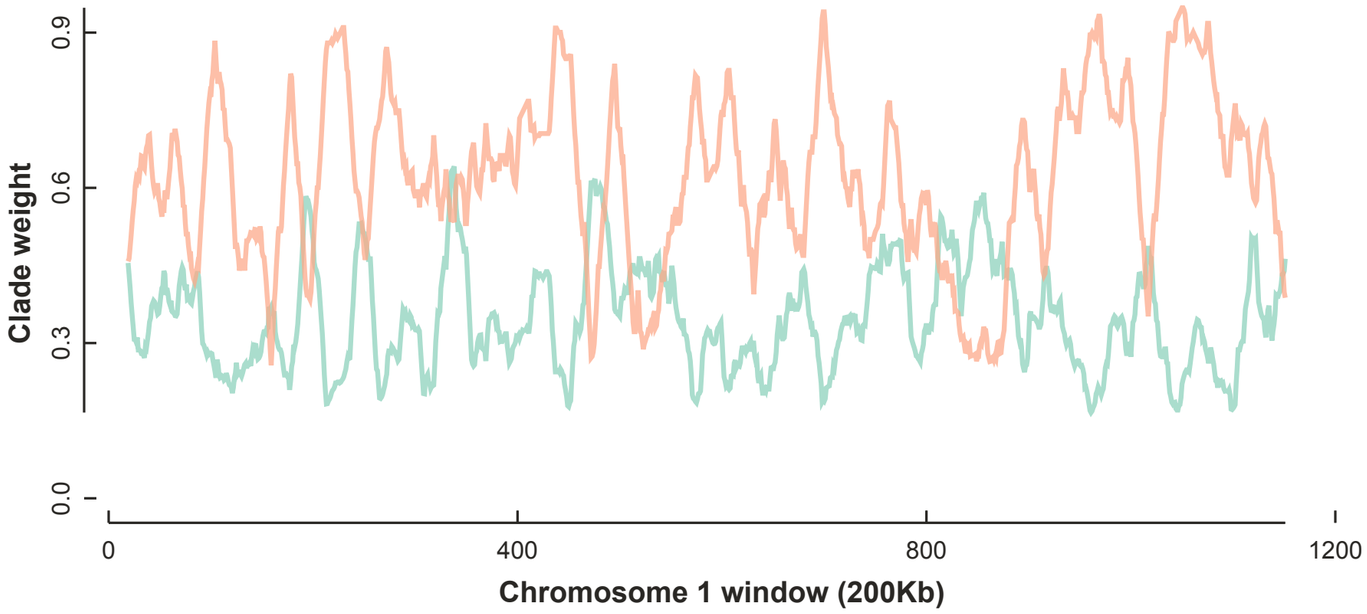
139 trees
5Mb windows
mean 50K sites
mean 480 SNPs



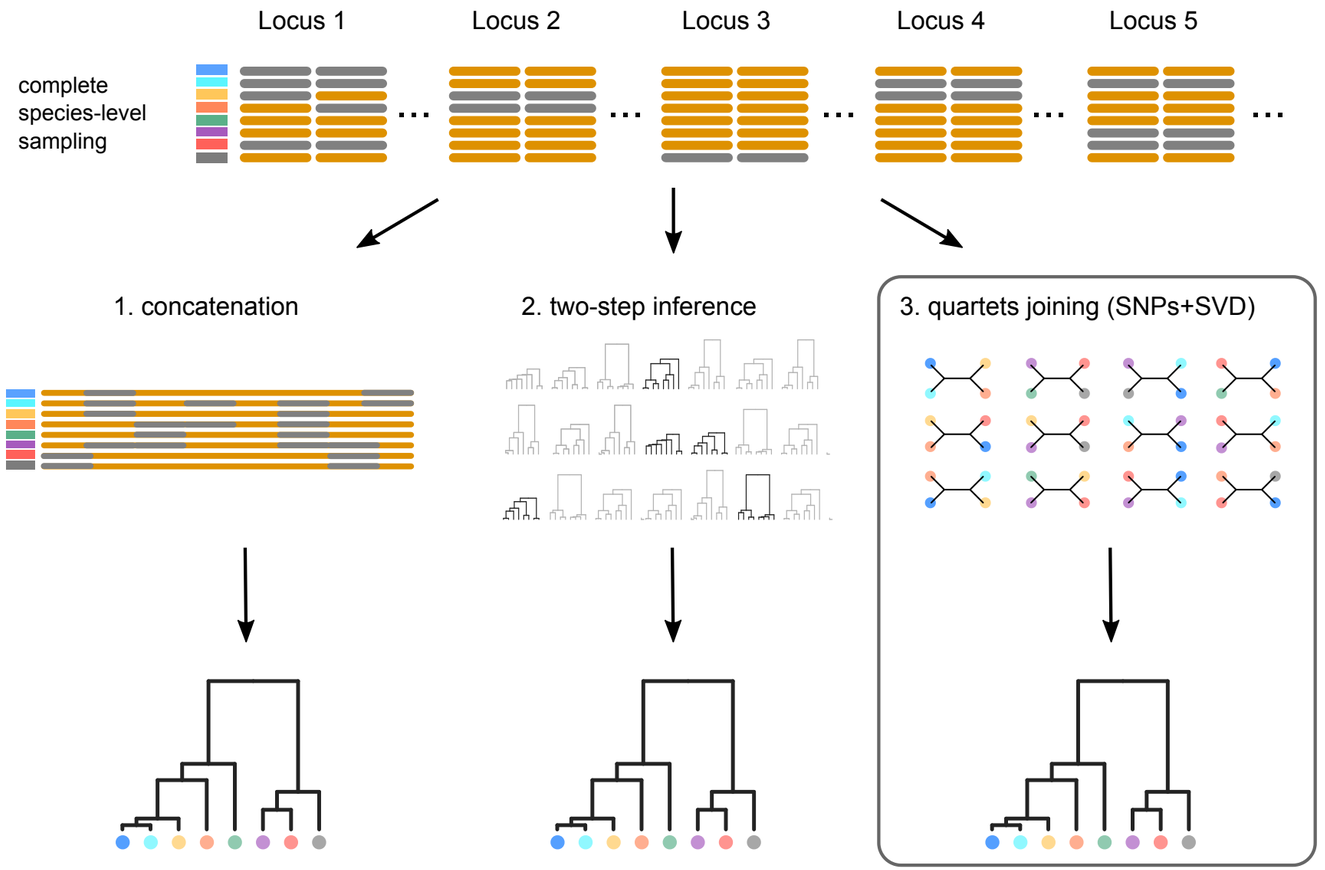
Astral species trees inferred from gene trees.

Clade weights (*sensu* Martin et al. 2017)

Chrom 1 weighted support for a (Cuba, Florida) vs (Cuba, Mexico)



Missing data in phylogenetics



SNP-based species trees

SNAPP: joint inference of gene trees and species trees.

(Bryant et al. 2012)

SVDquartets: infers quartet trees from SNPs and joins these into a species tree.

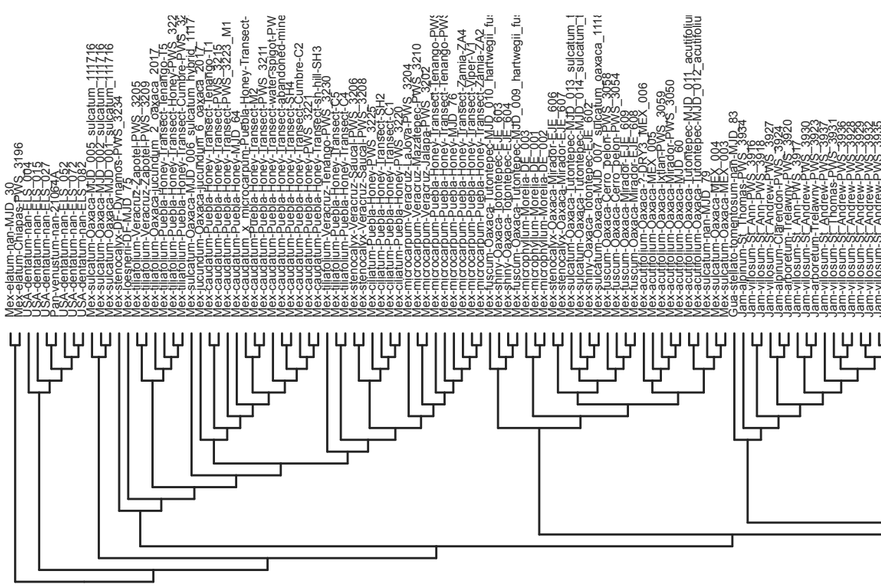
(Chifman and Kubatko 2014)

Advantages to SVDquartets-based methods

Each quartet is inferred independently: *missing data has almost no effect.*

Huge sptrees: 204 *Viburnum* taxa; 2.4M SNPs; 176K SNPs/quartet; >70M quartets.

It is fast! *tetrad*: 40 cores, one bootstrap ~24 hours.



Conclusions

1. With *ipyrad-analysis* it is easy to run dozens of analyses optimized for RAD missing-ness with a few lines of code.

2. Concatenating RAD loci in scaffold windows, and consensus or imputation sampling, dramatically improve the utility of RAD.

3. SNP based methods are in their infancy, but work well with RAD data.

Announcement

RADcamp wetlab AND bioinformatics workshop in New York City Oct. 2019

Bring **your DNA samples.** Library preparation and sequencing will be **free.**

(sponsored by SSB, SSE, Columbia, CCNY).

https://radcamp.github.io/NYC2019/

Acknowledgements

Viburn'ers: Donoghue-lab, Edwards-lab, M. Olson, I. Cacho.

ipyrad development: Isaac Overcast

Eaton lab members

Funding: NSF DEB 1557059; Columbia University

Questions?