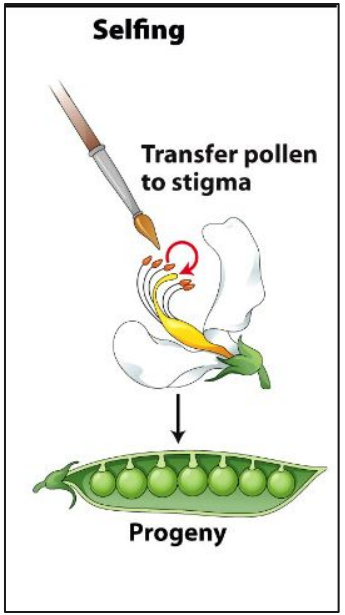
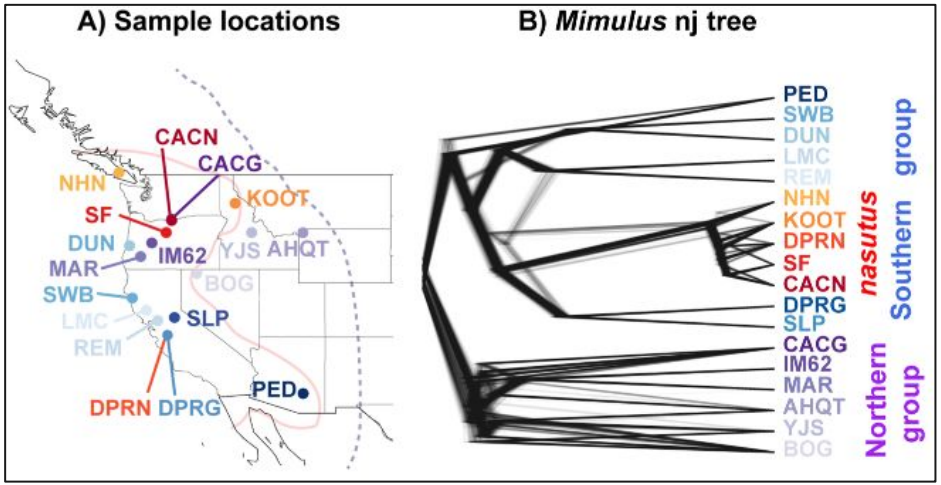
PCA and D-Statistics on Mimulus nasutus and Mimulus guttatus

Context

- *Mimulus guttatus* and *M. nasutus* are sister species pair
- They differ in that *M. nasutus* are selfing
    - where ovules are pollinated by the plant's own pollen
- Unidirectional introgression is heavily present with some bidirectional
- There appears to be a negative relationship between the recombination rate and divergence between these two species
    - Selection acts against *M. nasutus*
- *Mimulus guttatus* are further differentiated through geographical location (i.e Northern and Southern)



Context (cont.)



Key Idea:

- The evolutionary history is depicted where M. Guttatus first split off into Northern and Southern groups. And then M. Nasutus buds off from the Southern group.
- The geographical location of these three groups differ in that more M. Nasutus are closer to the Northern group.

Inferences:

- Which aspect plays a bigger role in introgression, the evolutionary history or geographical location of the species?
- In geography does allopatric or sympatric play a role in genetic sharing.

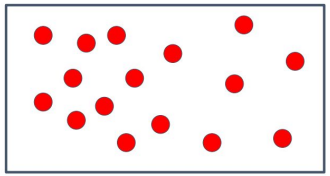
Motivation

- The genealogy history between these two flowering plants, Mimulus nasutus and Mimulus guttatus, are very complex.
- Although past studies have detected gene flow between these closely related species, the exact scope and direction of introgression events remain unclear.
- Through computational methods and analysis we can get obtain a more consistent story and history of a population.
- Additionally, providing detailed information on the experimental procedures ensures that science is reproducible, which helps to confirm and reinforce the results of prior studies.
- PCA and D-statistic can provide insights into the extent and direction of gene flow between these species, which can inform future research on their genetic diversity and adaptation. Moreover, these methods can be applied to other species to study their evolutionary history and population dynamics

Principal Component Analysis: PCA

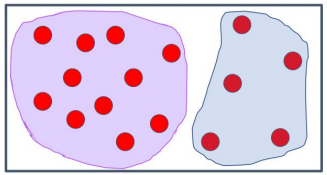
Unlabeled

Data:

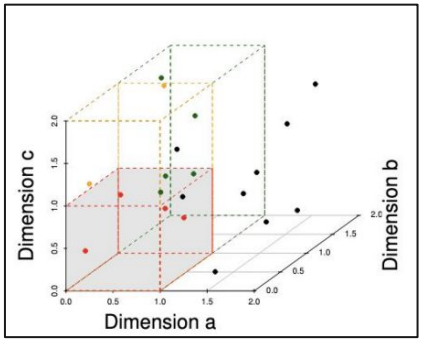


Clustered

Data:



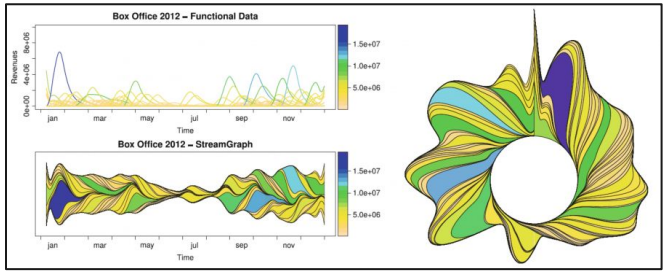
Dimensions



Why add dimensions?

- More Information
- Flexibility (Manipulating or Analyzing Data)
- Detailed Visualization

Principal Component Analysis: PCA (cont.)

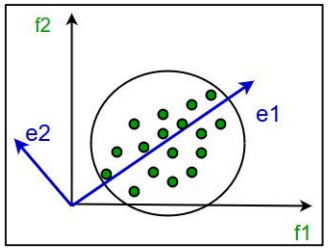


Flaws In Adding Dimensions

- Unnecessarily Complex
- Outperforms The Human Eye

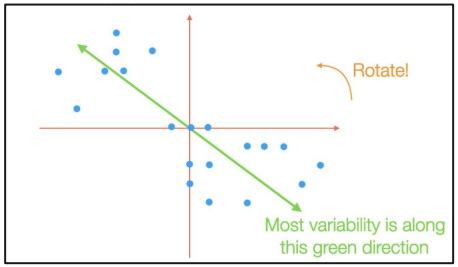
Solutions

- Simply Reduce Dimension Through PCA.
- Preserves Interpretability and Visualization

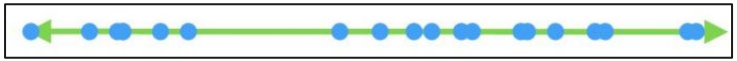


Principal Component Analysis: PCA (cont.)

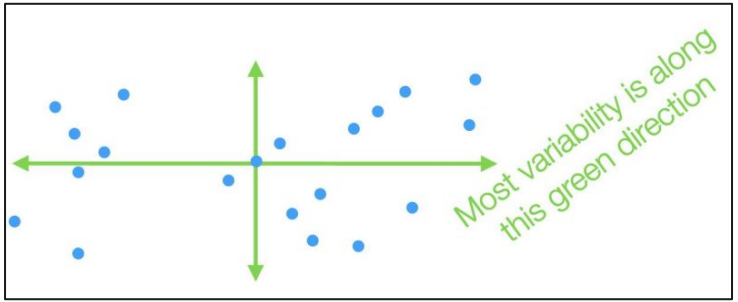
Find Variation



Rotate and Flatten To 1D



Orthogonalize For More Interpretability

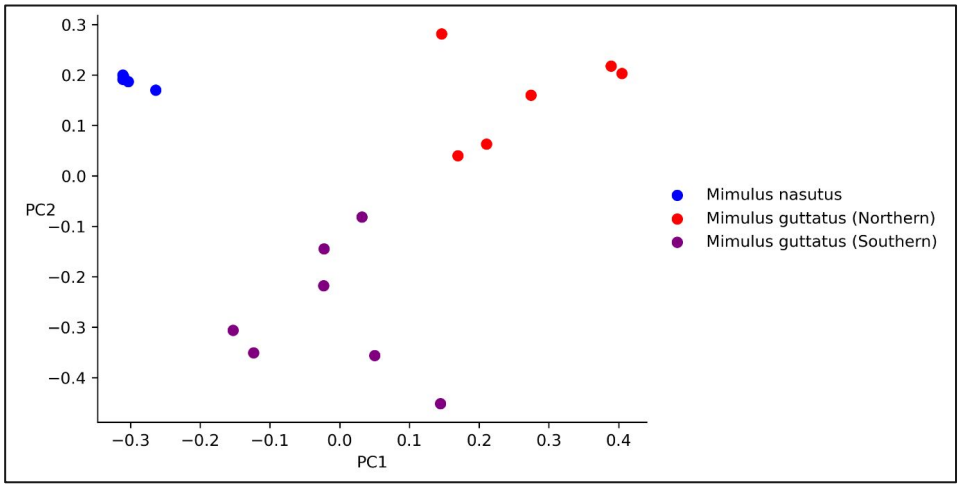


Principal Component Analysis: PCA (cont.)

Steps To PCA From A Data Set:

1. Normalize and Simplify To Count Matrix (Individuals x Sites)
2. Find Relevant Data by Removing NAN and Setting Up Conditions
3. Find the Covariance Between Them
4. Perform the Eigendecomposition to Compute the Eigenvalues and Eigenvectors from the Covariance Matrix.
5. Visualize!

Principal Component Analysis: PCA (cont.)

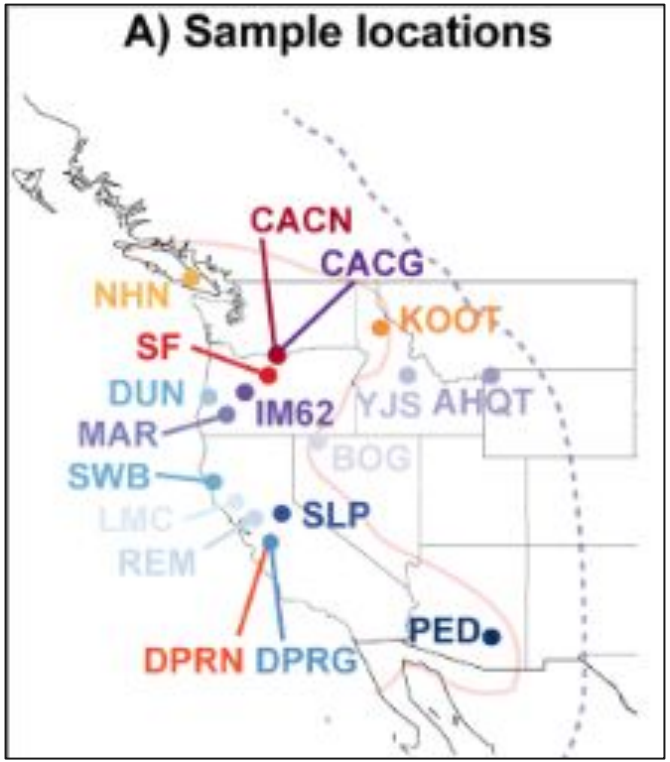


Result: The PCA analysis showed that the biggest variation was the split between M. Nasutus and M. Guttatus.

And the second principle highlights the second biggest variation which seems to be influenced by geography.

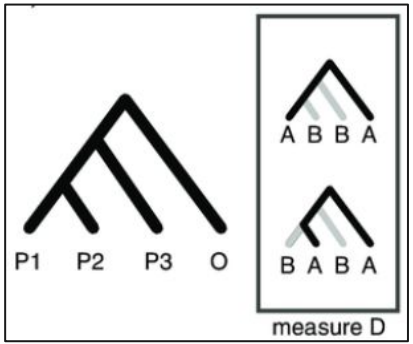
Northern M. Guttatus and M. Nasutus are closer as shown in the second principle component.

This is consistent with the map given.


A) Sample locations
CACN
CACG
NHN
KOOT
SF
DUN
IM62
YJS
AHQT
MAR
BOG
SWB
LMC
SLP
REM
PED
DPRN
DPRG

D - Statistics: Patterson's D

Visualization



Derivation

$$\mathbb{E}(D) = \frac{\mathbb{E}(\tau_{ABBA}) - \mathbb{E}(\tau_{BABA})}{\mathbb{E}(\tau_{ABBA}) + \mathbb{E}(\tau_{BABA})}$$

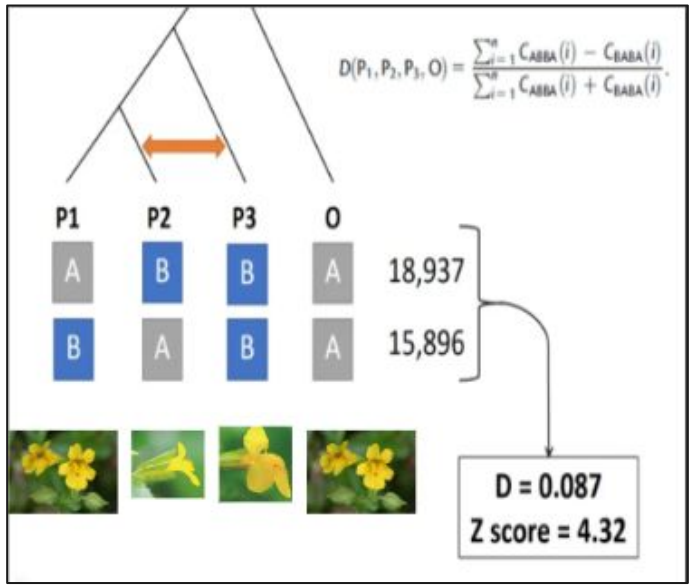
Analysis

- D-Statistic reveals whether there is hybridization or introgression.
- Can reveal the direction of gene flow and population (i.e sympatric vs allopatric)

Interpretation

- To measure the degree of genetic differentiation between two populations based on their allele frequencies.
- Ranges from -1 to 1 where a positive value indicates introgression

D - Statistics: Patterson's D (cont.)



Setup

P1: Sample that is related to P2 but not P3. This ensures consistency where P1 and P3 should not be related to compute the Patterson's D.

$$\mathbb{E}(D) = \frac{\mathbb{E}(\tau_{ABBA}) - \mathbb{E}(\tau_{BABA})}{\mathbb{E}(\tau_{ABBA}) + \mathbb{E}(\tau_{BABA})}$$

P2: Sample that is used to compare between itself and P3.

P3: Sample that is used to compare between itself and P2.

P4: Serves as a normalization and reference for the other samples

D - Statistics: Patterson's D (cont.)

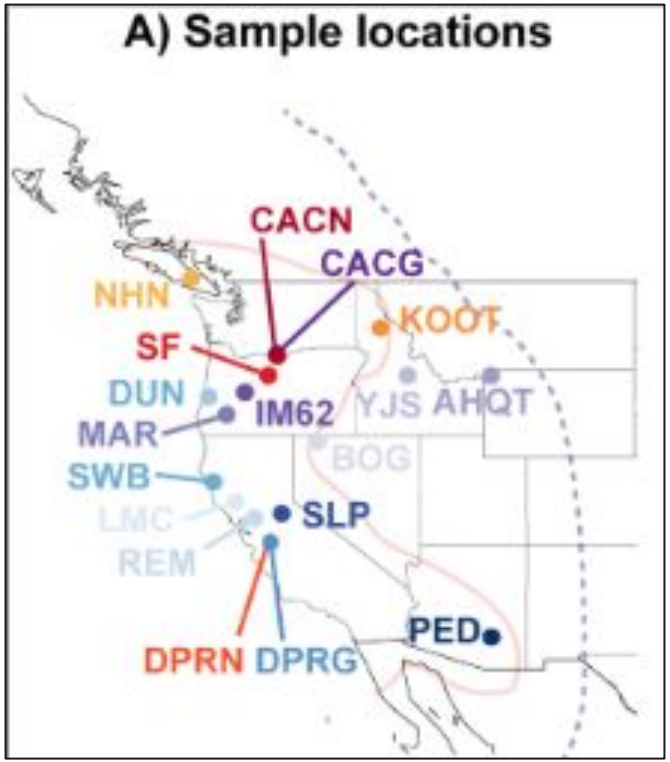
Steps To D-Statistics From A Data Set:

1. Extract the Allele Count Matrix from the Data Set
2. Polarize the Data Set
3. Compute Both ABBA and BABA
4. Compute the Patterson's D Test
5. Visualize and Infer!

D - Statistics: Patterson's D (cont.)

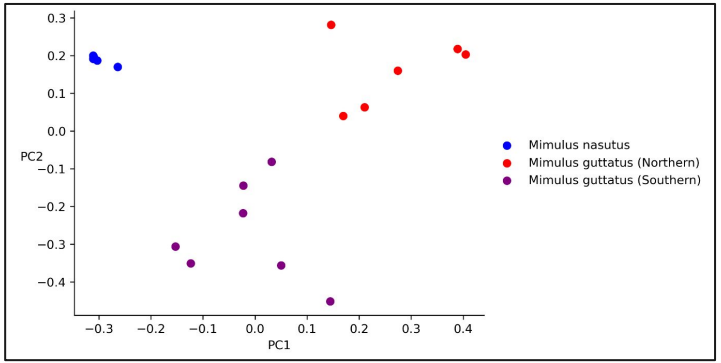
	P1	P2	P3	OUTGROUP (P4)	ABBA	BABA	PATTERSON'S D
Northern	AHQT1G	CAC6G	DPRN	SLP9G	1032	567	0.2908
Southern	SLP9G	DPRG	DPRN	AHQT1G	966	832	0.0745

Result: We know that a positive D-Statistic reveals that there is a genetic share of material among two species and even indicates gene flow. And given the previous information where one population is selfing, the direction of genetic share should be unidirectional where Mimulus Nasutus contributes to the Mimulus Guttatus. The results also reinforces that sympatry plays an essential role as both population are close in proximity which permits the event of introgression. Lastly, the greater ABBA count in the derivation suggests that there is introgression between P2 and P3


A) Sample locations
CACN
CACG
NHN
KOOT
SF
DUN
IM62
YJS
AHQT
MAR
BOG
SWB
LMC
SLP
REM
PED
DPRN
DPRG

Further Interpretation

PCA: The clustering is consistent in that the M. Nasutus does not have that many variation because they are selfing. In addition, there is a closer relationship between the Northern group and M. Nasutus because they are geographically closer.



D-Statistic: The D Statistics also supports the idea that the M. Nasutus is closer related to the Northern group in that there is a higher D result compared to the Southern group. In addition, the sympatry and allopatry of populations contribute to introgression.

Northern

AHQT (Allopatric Nothern Mimulus Guttatus)

CACG (Sympatric Northern Mimulus Guttatus)

DPRN (Sympatric Mimulus Nasutus)

SLP (Allopatric Southern Mimulus Guttatus)

Southern

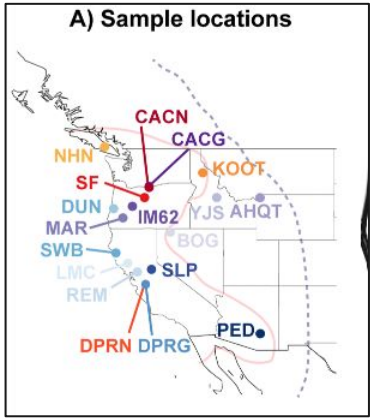
SLP (Allopatric Southern Mimulus Guttatus)

DPRG (Sympatric Southern Mimulus Guttatus)

DPRN (Sympatric Mimulus Nasutus)

AHQT (Allopatric Nothern Mimulus Guttatus)

	P1	P2	P3	OUTGROUP (P4)	ABBA	BABA	PATTERSON'S D
Northern	AHQT1G	CAC6G	DPRN	SLP9G	1032	567	0.2908
Southern	SLP9G	DPRG	DPRN	AHQT1G	966	832	0.0745



Looking Ahead

- Even with PCA and D-Statistic, the evolution history and population structure of both these species are still very complex, so maybe we can more use more computational methods
  - Coalescent Simulation
  - F-Statistic
  - Linkage Disequilibrium Analysis:

References

- https://biol1435.jupyter.brown.edu/user/john_zou/files/23feb23_exercise%20(1).ipynb?_xsrf=2%7Cc8f6ec77%7Cea9c8a17f71e576ca009325a2649d8f4%7C1677783608
- Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2(12): e190. https://doi.org/10.1371/journal.pgen.0020190
- Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet 10(6): e1004410. https://doi.org/10.1371/journal.pgen.1004410
- *Durand E., Patterson N., Reich D., Slatkin M., Molecular Biology and Evolution*, Volume 28, Issue 8, August 2011, Pages 2239–2252, https://doi.org/10.1093/molbev/msr048