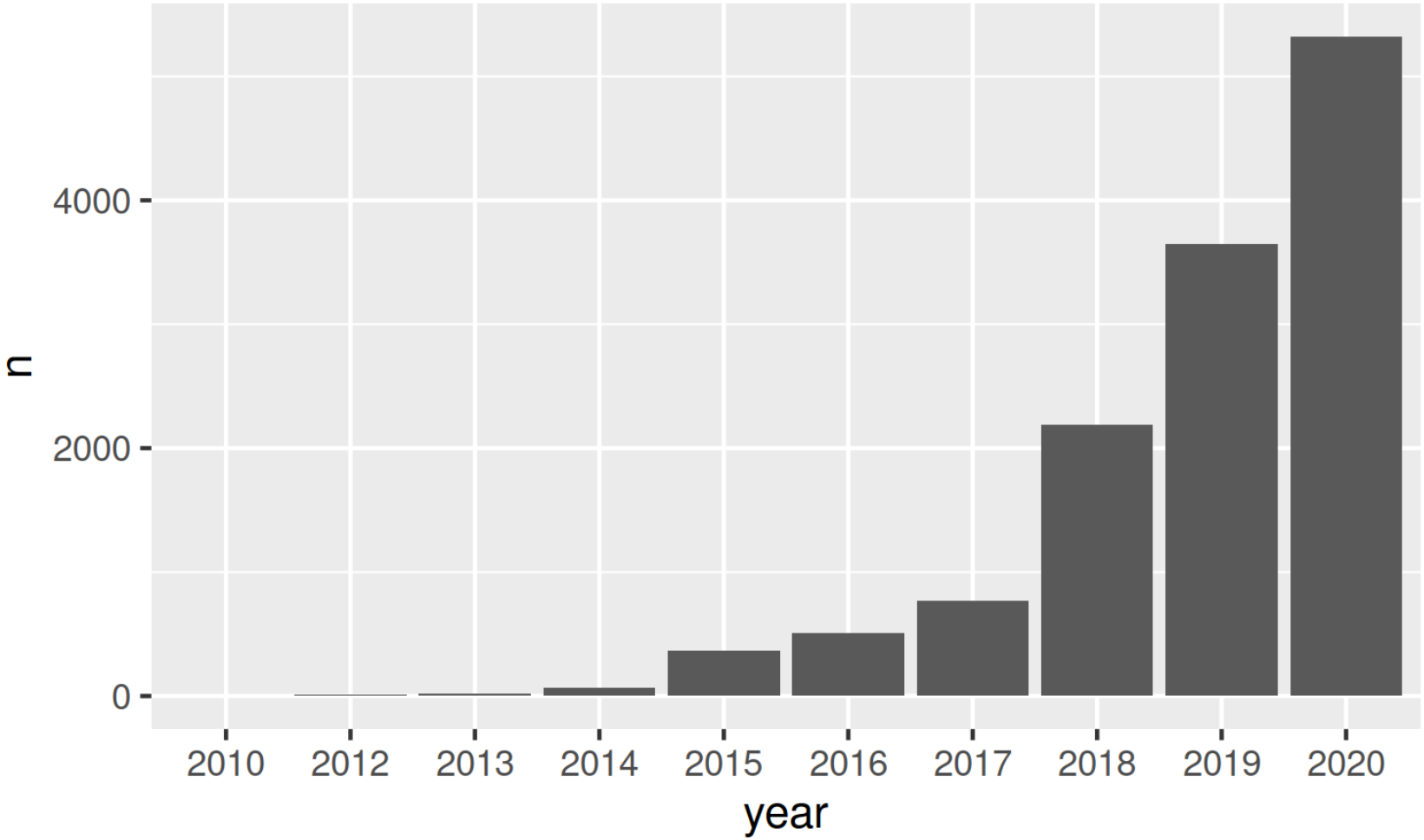
F-statistics and PCA

Benjamin Peter

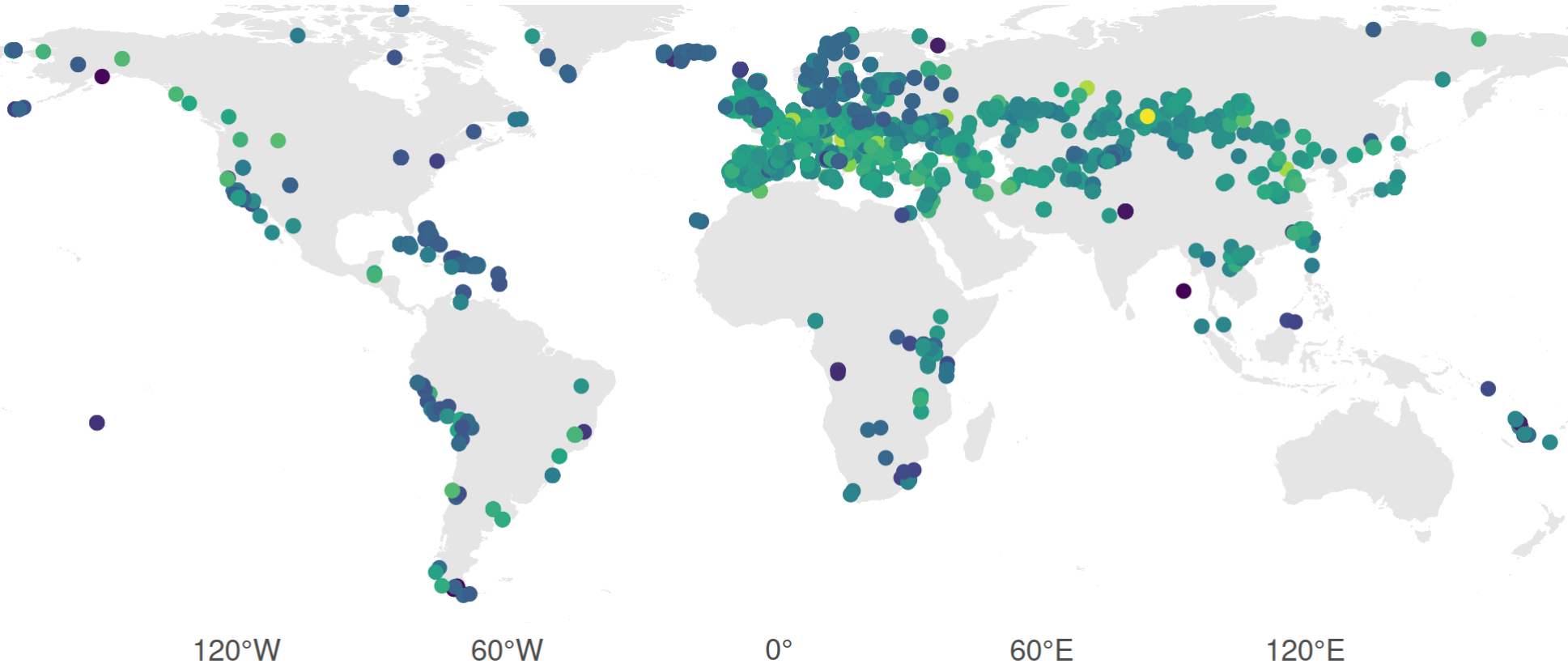
April 22, 2021

Population structure and ancient DNA



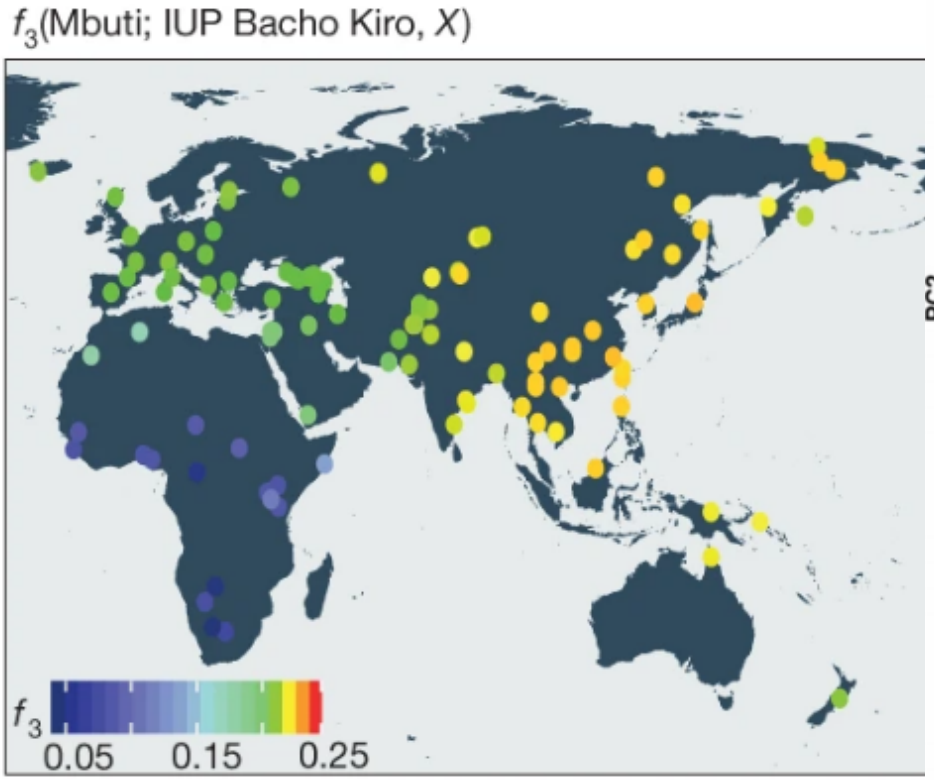
https://reich.hms.harvard.edu/

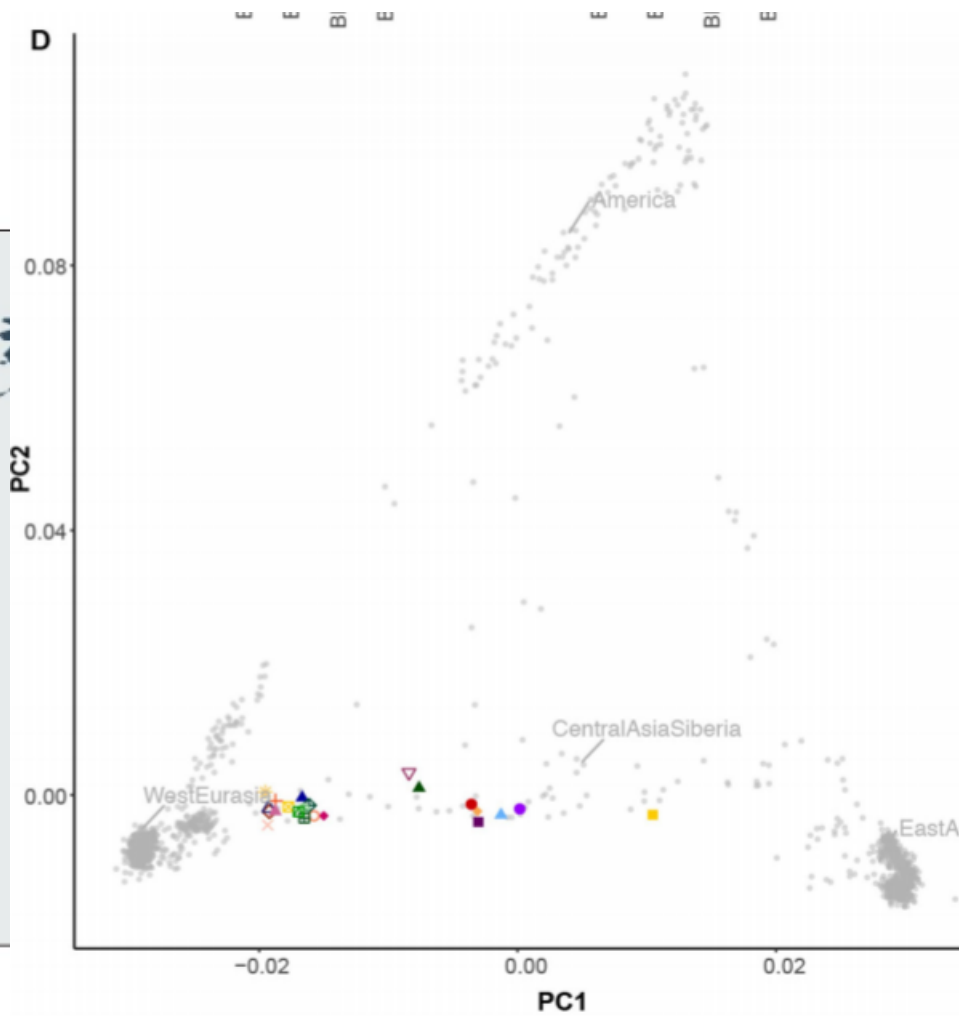
Population structure and ancient DNA



https://reich.hms.harvard.edu/

PCA and F-statistics





Hajdinjak et al. 2021

Goals of this talk

- Technical & Conceptual Background
- Establish conceptual links between frameworks
  1. How can we interpret PCA in context of F-stats?
  2. How can we interpret F-stats in the context of PCA?
- (Use established links to improve data interpretation)

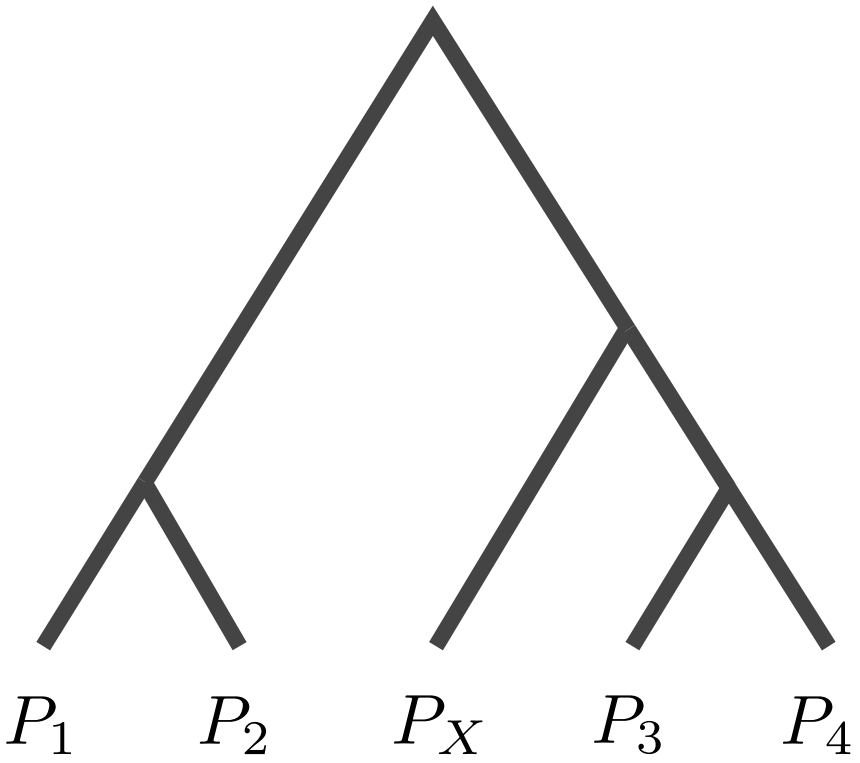
Goals of this talk

- Technical & Conceptual Background
- Establish conceptual links between frameworks
  1. How can we interpret PCA in context of F-stats?
  2. How can we interpret F-stats in the context of PCA?
- (Use established links to improve data interpretation)

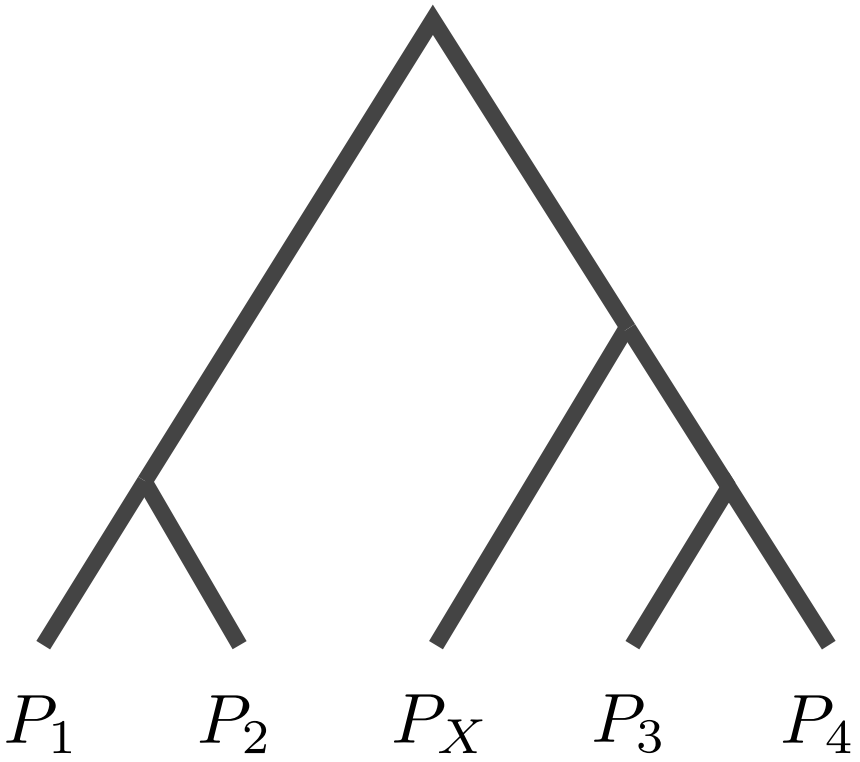
Focus on intuition

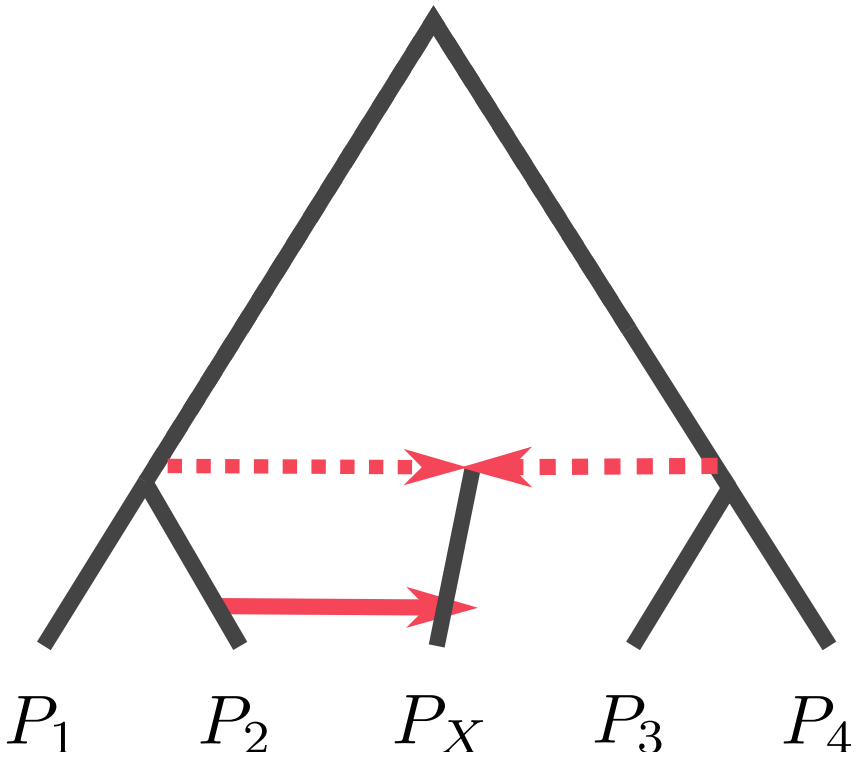
Some details in terms of estimation, normalization, missing data will be glossed over

Trees vs. Admixture Graphs

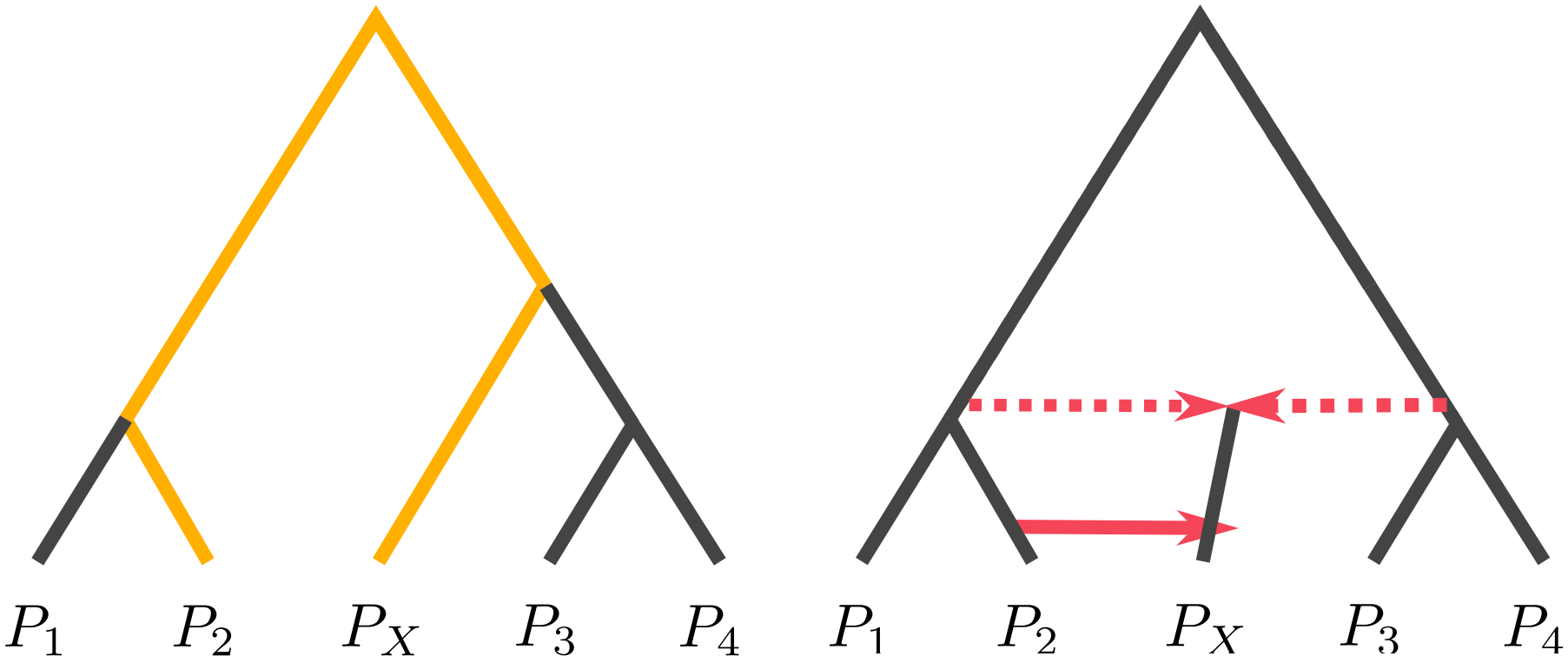


Trees vs. Admixture Graphs

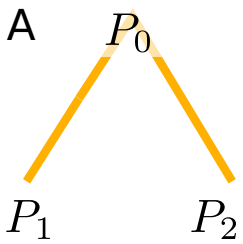




Trees vs. Admixture Graphs

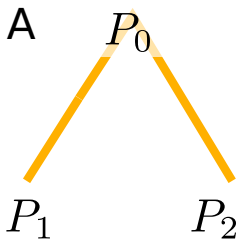
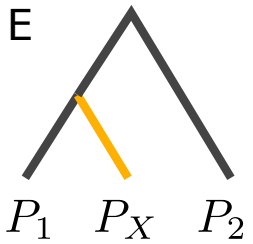


F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	

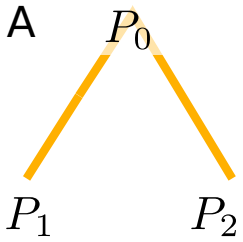
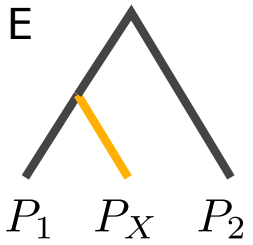
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$	E 

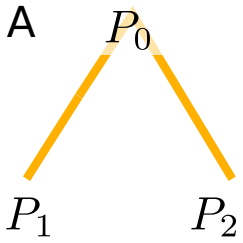
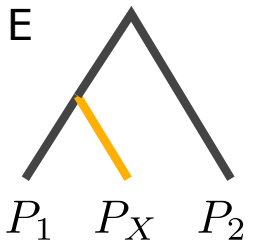
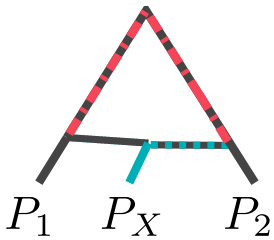
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$ $F_3(X_x; X_1, X_2) = \sum_l (X_{xl} - X_{1l})(X_{xl} - X_{2l}) - H_X$	E 

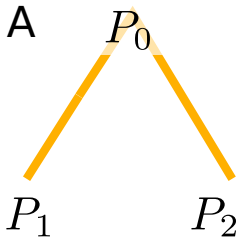
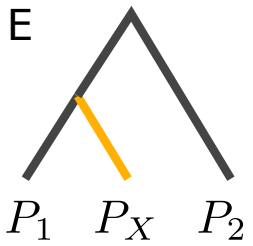
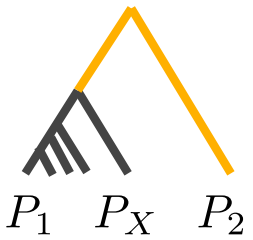
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$ $F_3(X_x; X_1, X_2) = \sum_l (X_{xl} - X_{1l})(X_{xl} - X_{2l}) - H_X$	E 
"Admixture"-F_3-statistic: If data is generated by a tree-like relationship, $F_3(P_X; P_1, P_2) \geq 0$	

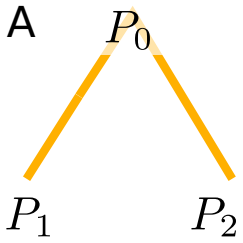
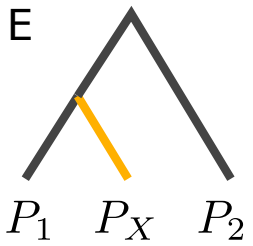
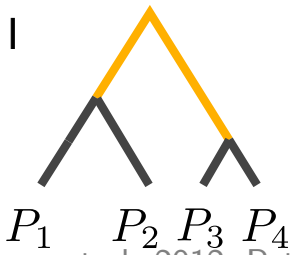
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$ $F_3(X_x; X_1, X_2) = \sum_l (X_{xl} - X_{1l})(X_{xl} - X_{2l}) - H_X$	E 
"Outgroup"-F_3-statistic: Most similar pops have highest $F_3(P_2; P_X, P_1)$	

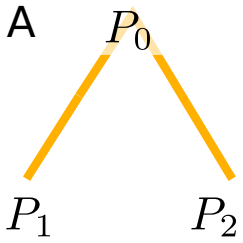
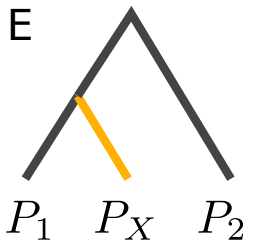
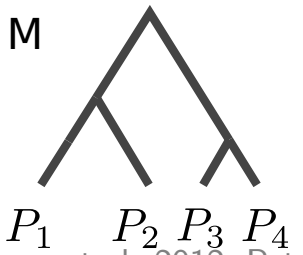
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$ $F_3(X_x; X_1, X_2) = \sum_l (X_{xl} - X_{1l})(X_{xl} - X_{2l}) - H_X$	E 
$F_4^{(B)}(X_1; X_2; X_3, X_4) = \sum_l (X_{1l} - X_{3l})(X_{2l} - X_{4l})$	I 

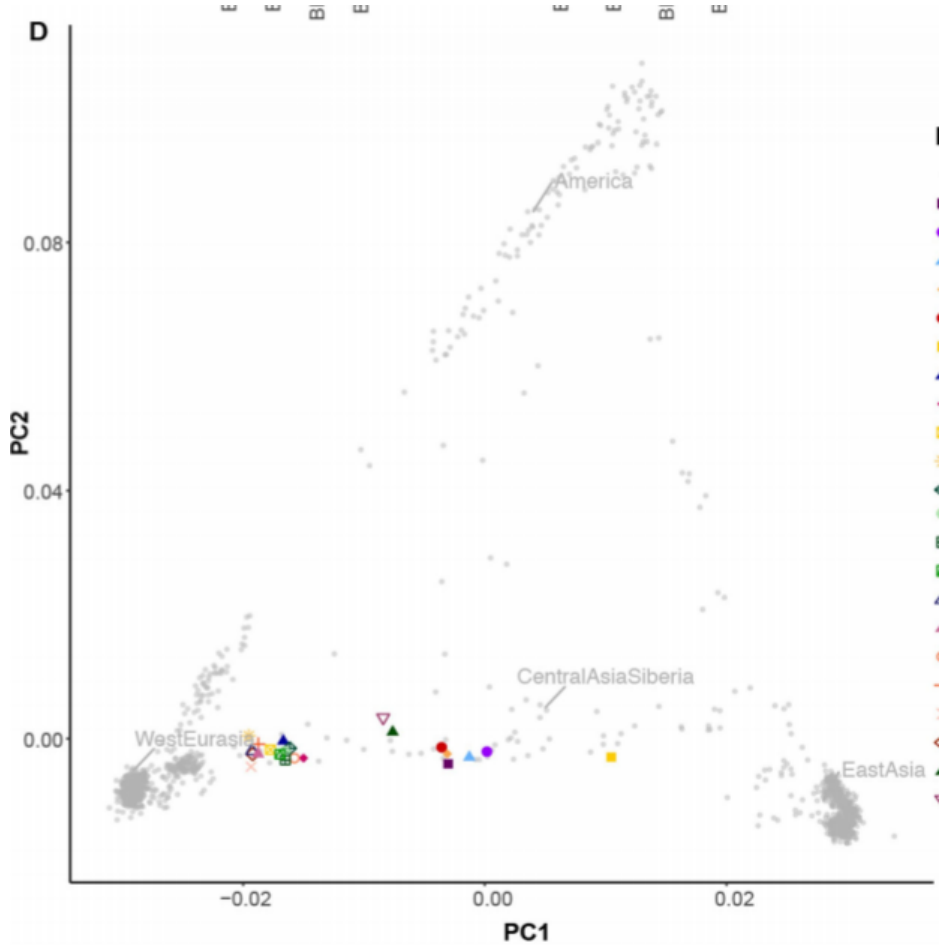
Patterson et al. 2012; Peter 2016

F-statistics

Definition	Branch length
$F_2(X_1, X_2) = \sum_l (X_{il} - X_{jl})^2 - H_1 - H_2$	A 
$2F_3(X_x; X_1, X_2) = F_2(X_x, X_1) + F_2(X_x, X_2) - F_2(X_1, X_2)$ $F_3(X_x; X_1, X_2) = \sum_l (X_{xl} - X_{1l})(X_{xl} - X_{2l}) - H_X$	E 
$F_4^{(T)}(X_1; X_2; X_3, X_4) == \sum_l (X_{1l} - X_{2l})(X_{3l} - X_{4l})$	M 

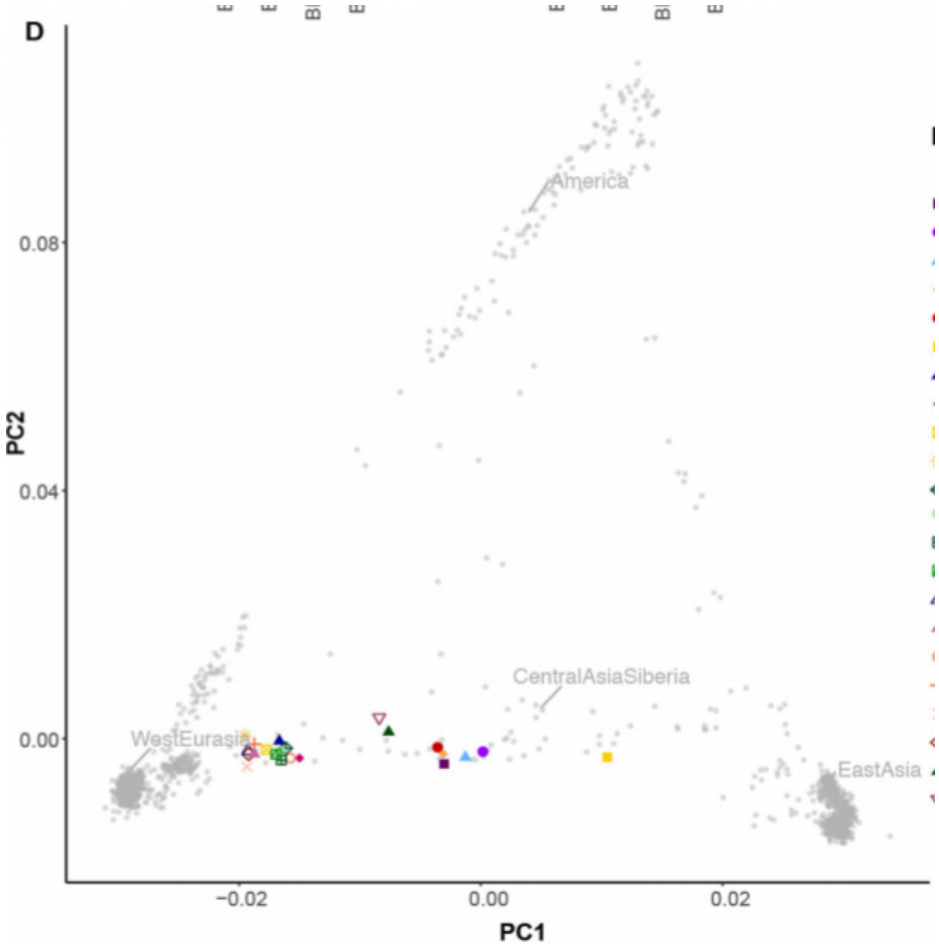
Patterson et al. 2012; Peter 2016

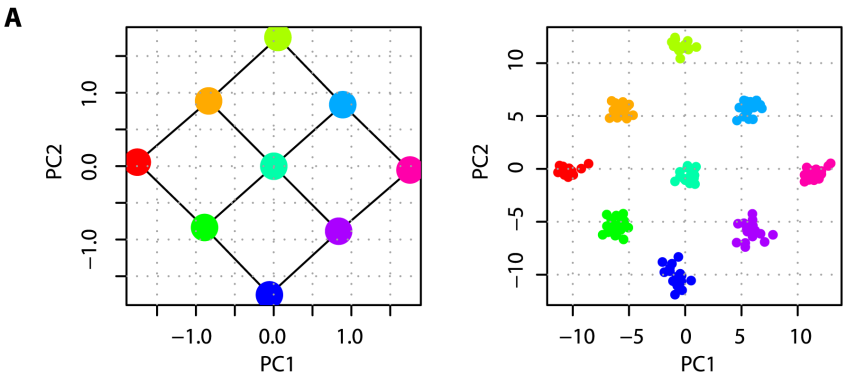
Principal Component Analysis



McVean, 2009

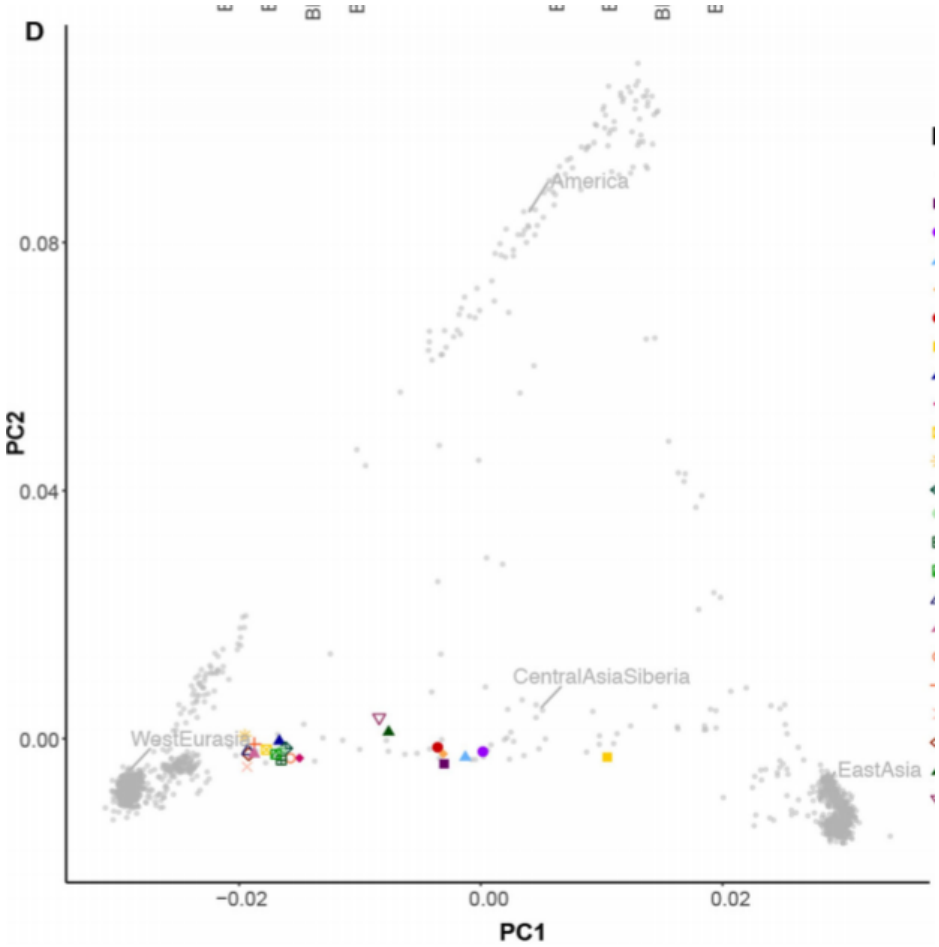
Principal Component Analysis

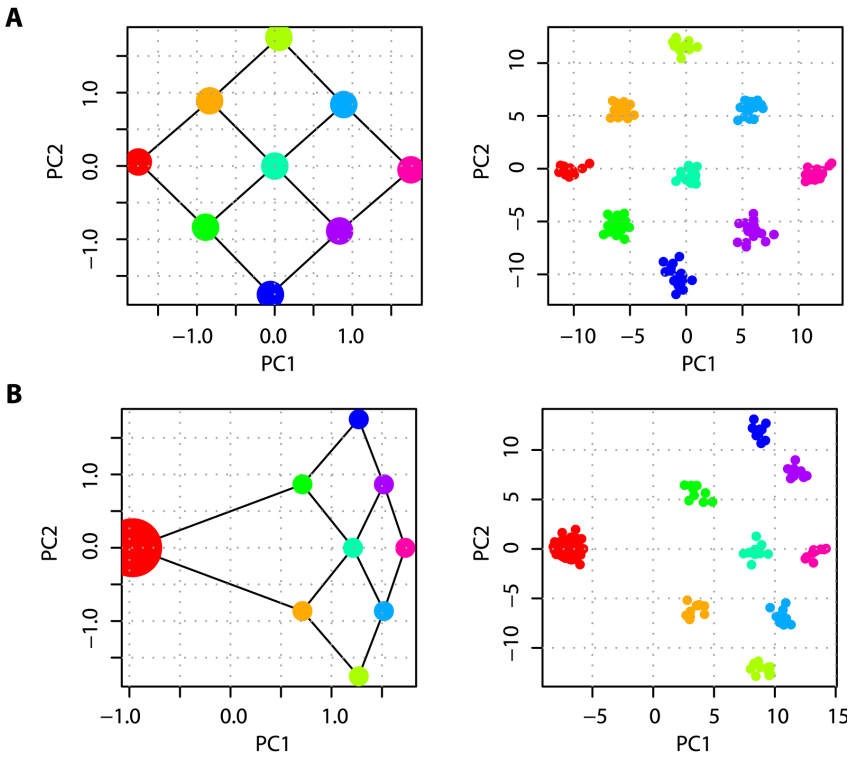




McVean, 2009

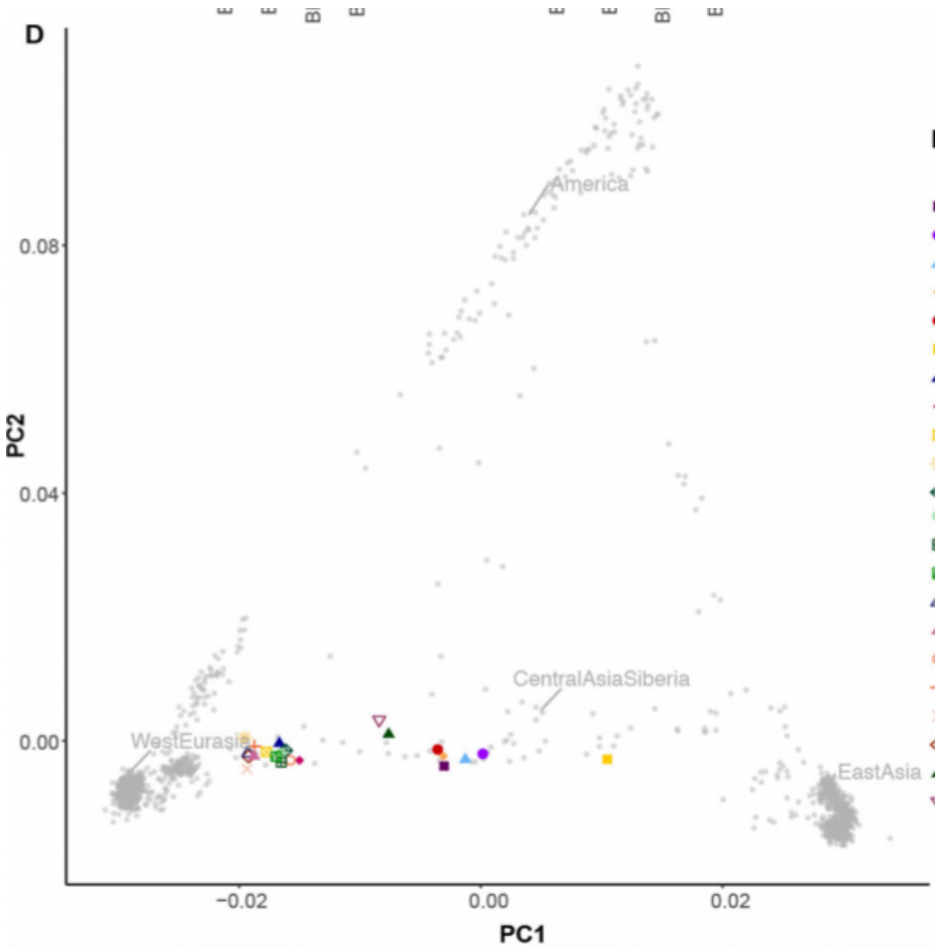
Principal Component Analysis

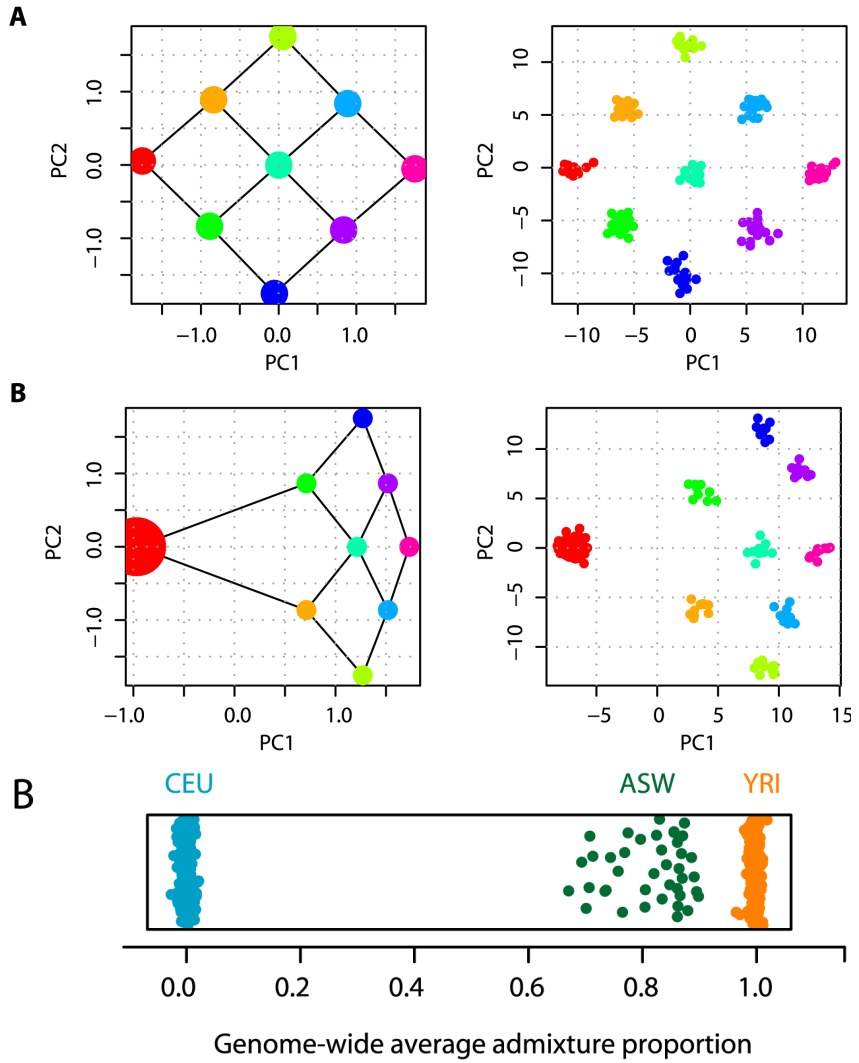




McVean, 2009

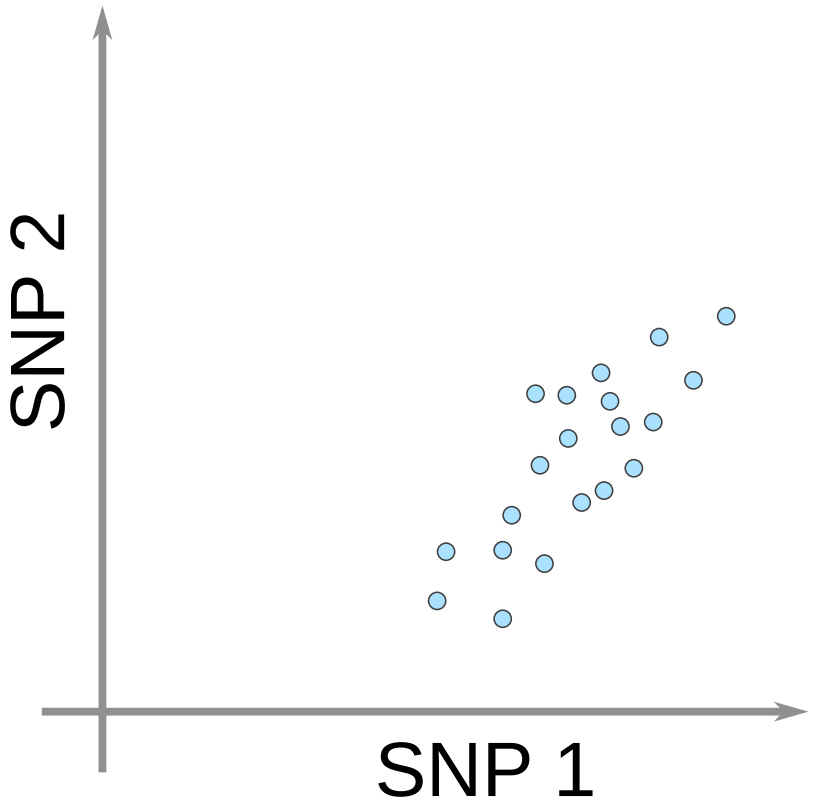
Principal Component Analysis





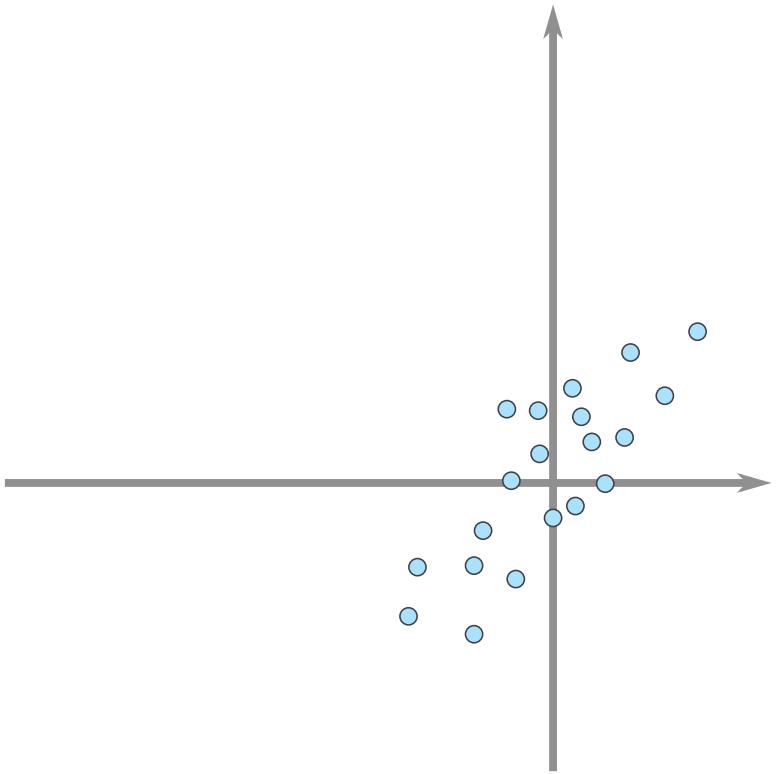
McVean, 2009

Principal Component Analysis



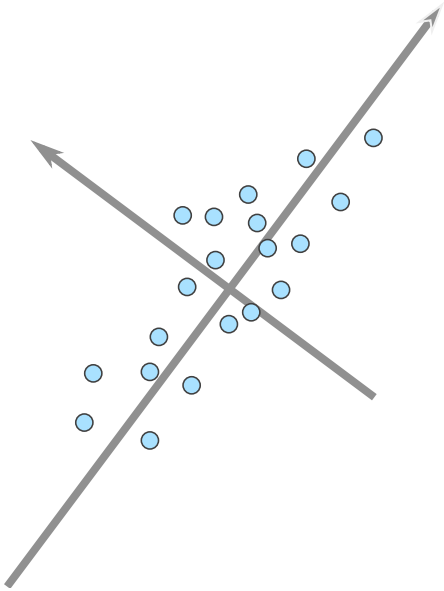
- Raw SNP data $\mathbf{X}$; x_{ij}

Principal Component Analysis



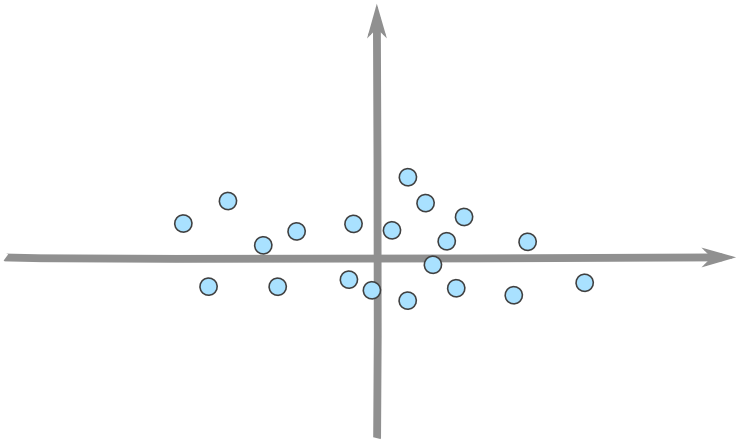
- Raw SNP data $\mathbf{X}; x_{ij}$
- Centering
  $\mathbf{Y} = \mathbf{CX}; y_{ij} = x_{ij} - \mu_j$

Principal Component Analysis



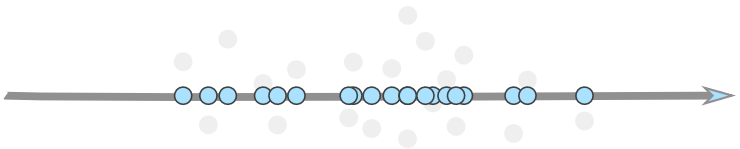
- Raw SNP data $\mathbf{X}; x_{ij}$
- Centering
  $\mathbf{Y} = \mathbf{CX}; y_{ij} = x_{ij} - \mu_j$
- Rotation $\mathbf{Y} = \underbrace{\mathbf{P}}_{\text{PCs}} \underbrace{\mathbf{L}}_{\text{Rotation}}$

Principal Component Analysis



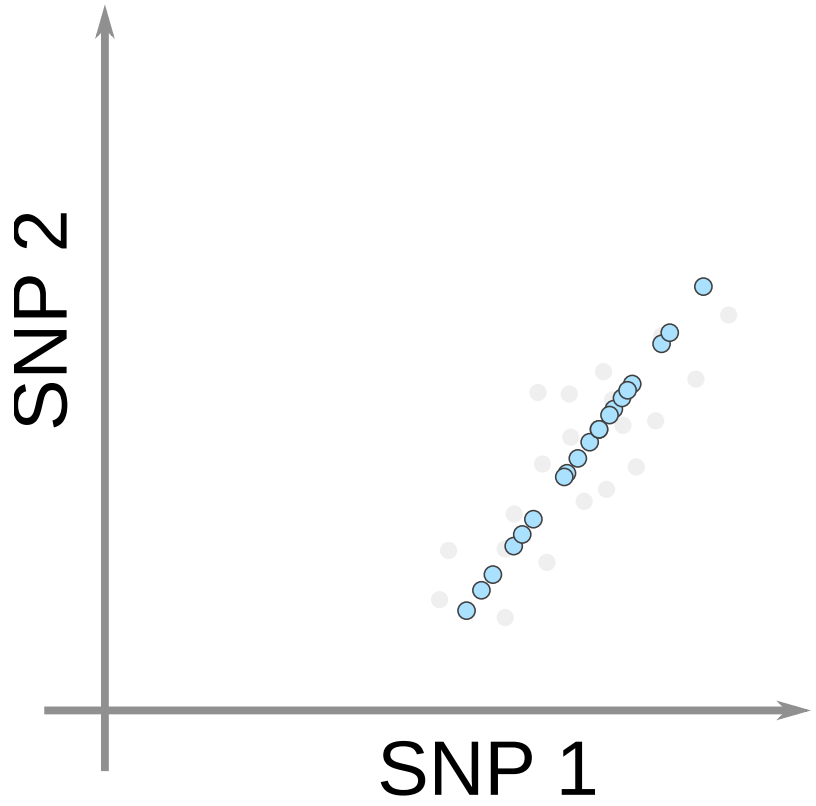
- Raw SNP data $\mathbf{X}; x_{ij}$
- Centering
  $\mathbf{Y} = \mathbf{CX}; y_{ij} = x_{ij} - \mu_j$
- Rotation $\mathbf{Y} = \mathbf{PL}$

Principal Component Analysis



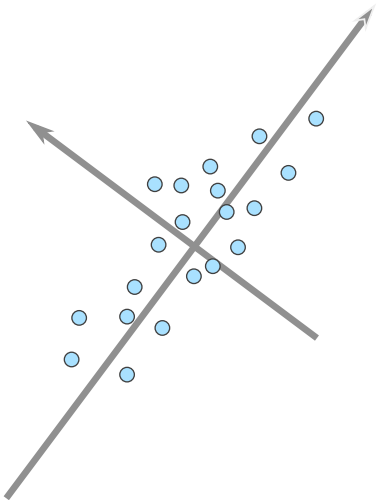
- Raw SNP data $\mathbf{X}; x_{ij}$
- Centering
  $\mathbf{Y} = \mathbf{CX}; y_{ij} = x_{ij} - \mu_j$
- Rotation $\mathbf{Y} = \mathbf{PL}$
- Truncation $\hat{\mathbf{P}} = \left(\begin{array}{c} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \hline \\ \vdots \\ \\ \mathbf{p}_n \end{array}\right)$

Principal Component Analysis



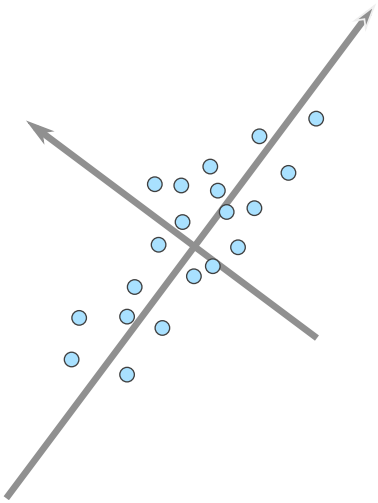
- Raw SNP data $\mathbf{X}; x_{ij}$
- Centering
  $\mathbf{Y} = \mathbf{CX}; y_{ij} = x_{ij} - \mu_j$
- Rotation $\mathbf{Y} = \mathbf{PL}$
- Truncation $\hat{\mathbf{P}} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \hline \\ \vdots \\ \mathbf{p}_n \end{pmatrix}$
- Approximation $\hat{\mathbf{Y}} = \hat{\mathbf{P}}\hat{\mathbf{L}}$

How to find PCs



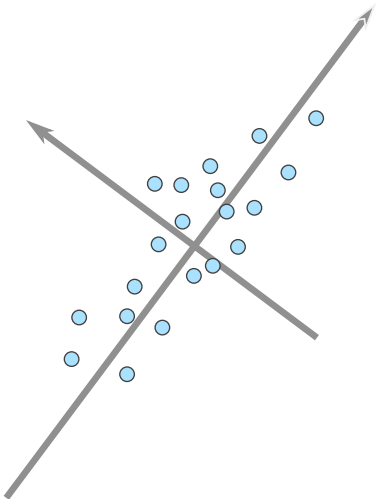
- Singular Value Decomposition:
  $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$

How to find PCs



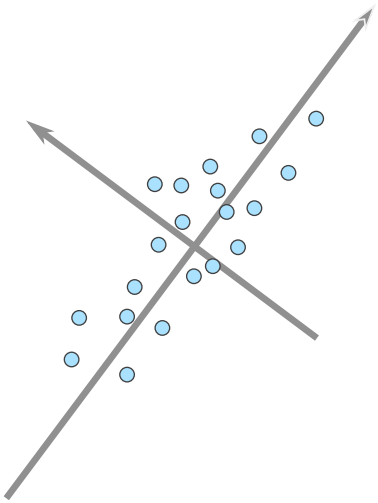
- Singular Value Decomposition:
  $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$
- Eigendecomposition of $\mathbf{YY}^T$:
  $\mathbf{YY}^T = \mathbf{UD}^2\mathbf{U}^T = \mathbf{PP}^T$

How to find PCs



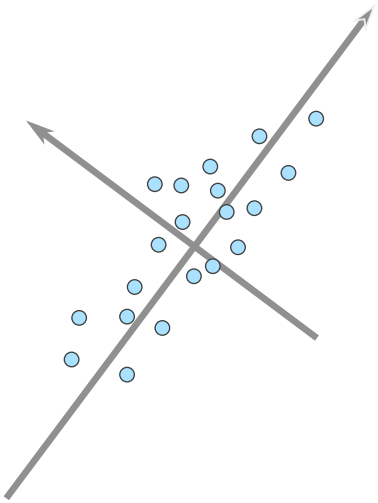
- Singular Value Decomposition: $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$
- Eigendecomposition of $\mathbf{YY}^T$: $\mathbf{YY}^T = \mathbf{UD}^2\mathbf{U}^T = \mathbf{PP}^T$
- $(\mathbf{YY}^T)_{ij}$

How to find PCs



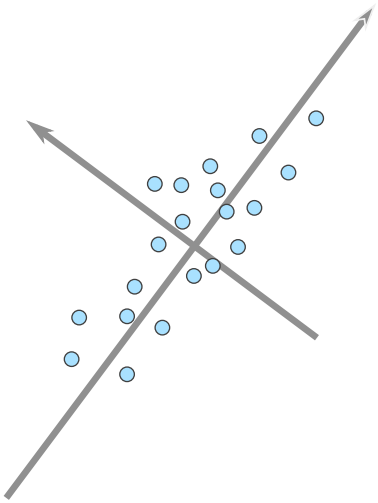
- Singular Value Decomposition:
  $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$
- Eigendecomposition of $\mathbf{YY}^T$:
  $\mathbf{YY}^T = \mathbf{UD}^2\mathbf{U}^T = \mathbf{PP}^T$
- $(\mathbf{YY}^T)_{ij} = \sum_l (x_{il} - \mu_l)(x_{jl} - \mu_l)$

How to find PCs



- Singular Value Decomposition:
  $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$
- Eigendecomposition of $\mathbf{YY}^T$:
  $\mathbf{YY}^T = \mathbf{UD}^2\mathbf{U}^T = \mathbf{PP}^T$
- $(\mathbf{YY}^T)_{ij} = \sum_l (x_{il} - \mu_l)(x_{jl} - \mu_l)$
- $F_3(X_x; X_1, X_2) = \sum_l (X_{1l} - X_{xl})(X_{2l} - X_{xl})$

How to find PCs

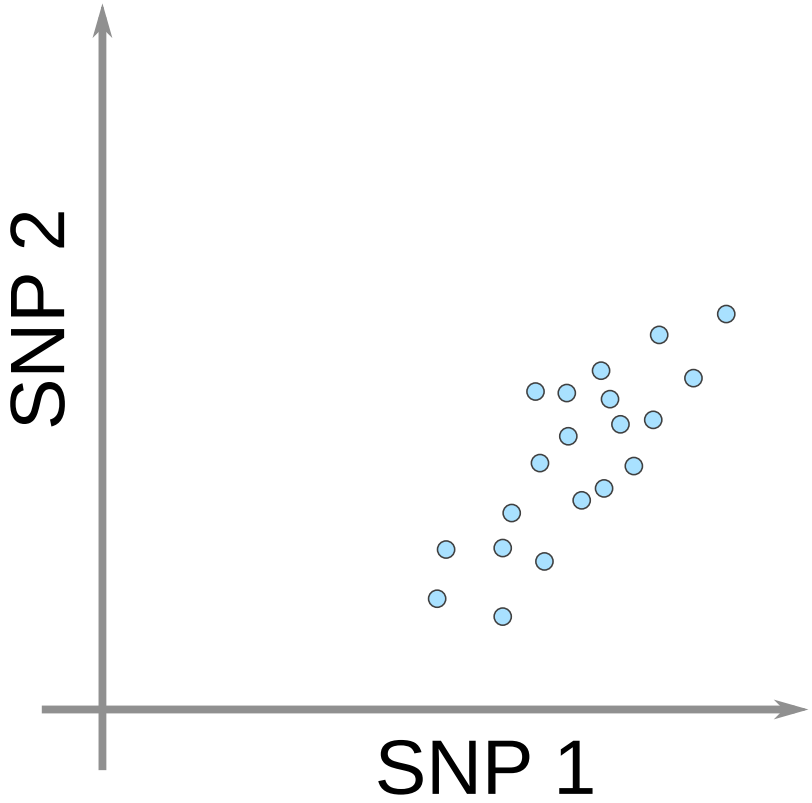


- Singular Value Decomposition:
  $\mathbf{Y} = (\mathbf{UD})\mathbf{L} = \mathbf{PL}$
- Eigendecomposition of $\mathbf{YY}^T$:
  $\mathbf{YY}^T = \mathbf{UD}^2\mathbf{U}^T = \mathbf{PP}^T$
- $(\mathbf{YY}^T)_{ij} = \sum_l (x_{il} - \mu_l)(x_{jl} - \mu_l)$
- $F_3(X_x; X_1, X_2) = \sum_l (X_{1l} - X_{xl})(X_{2l} - X_{xl})$
- $(\mathbf{YY}^T)_{ij} = F_3(\boldsymbol{\mu}; \mathbf{X}_i, \mathbf{X}_j)$

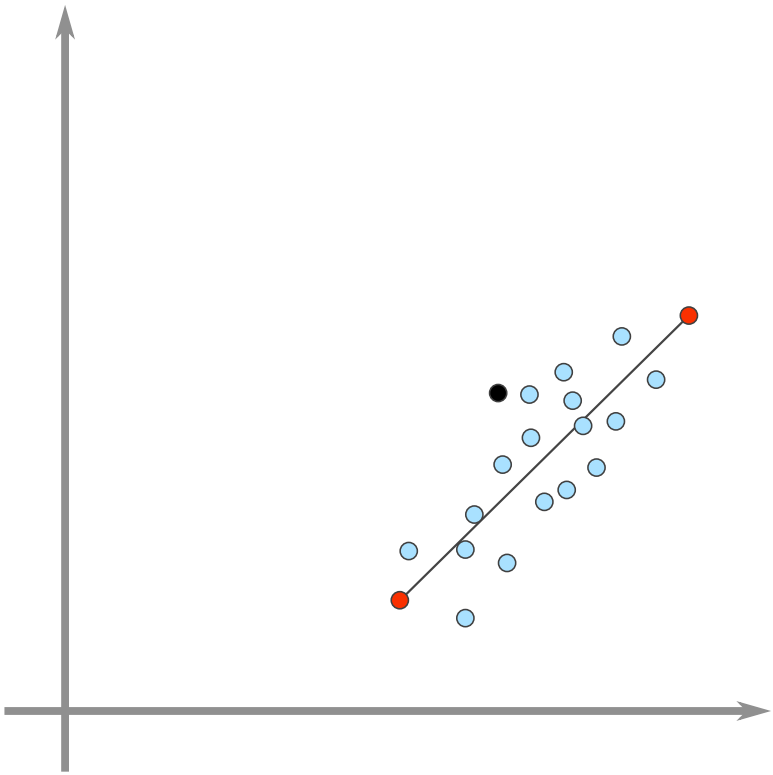
Observation

PCA is equivalent to outgroup-F_3-analysis with sample mean as outgroup

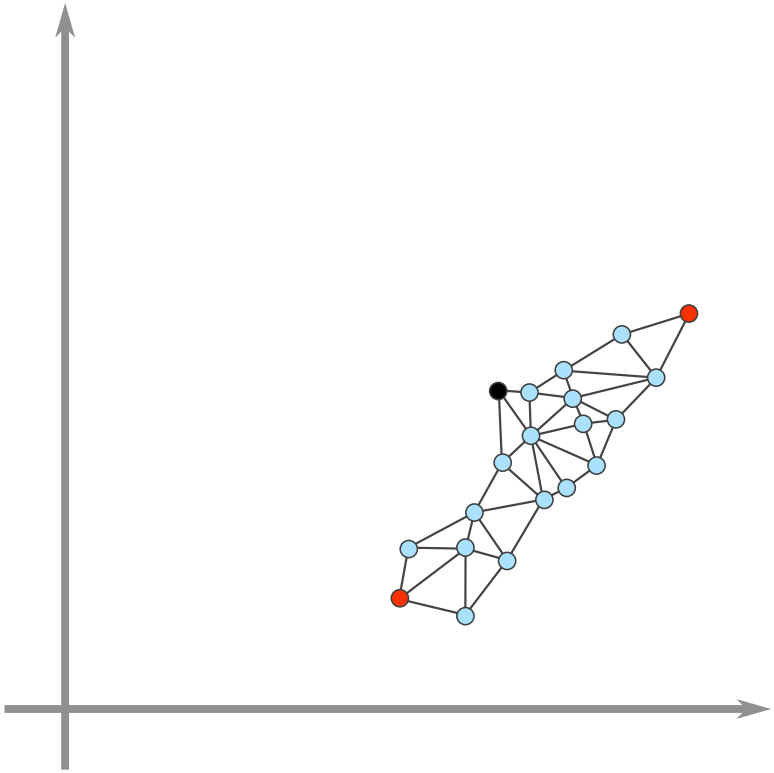
(metric) Multi-Dimensional Scaling (MDS)



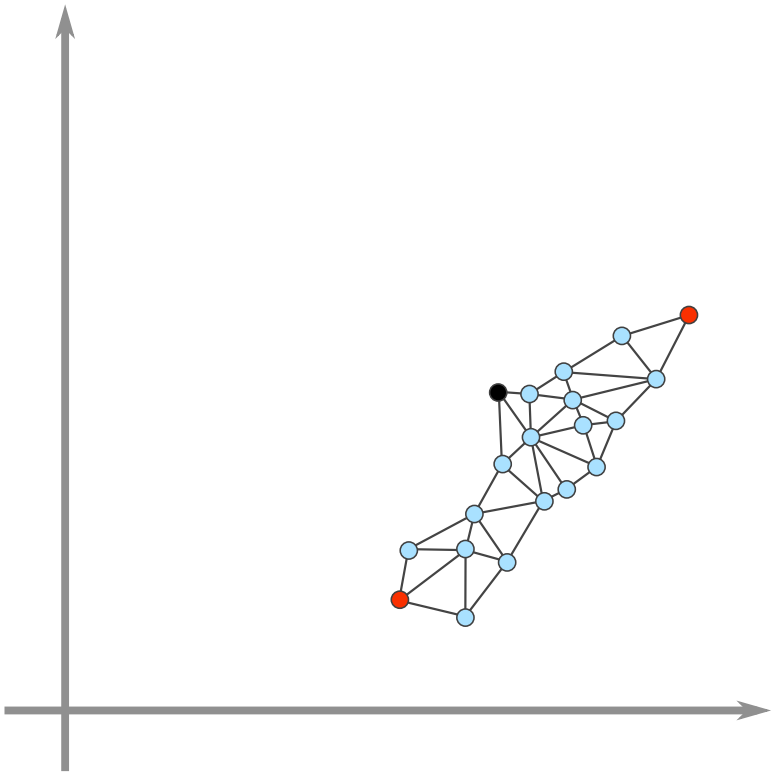
(metric) Multi-Dimensional Scaling (MDS)

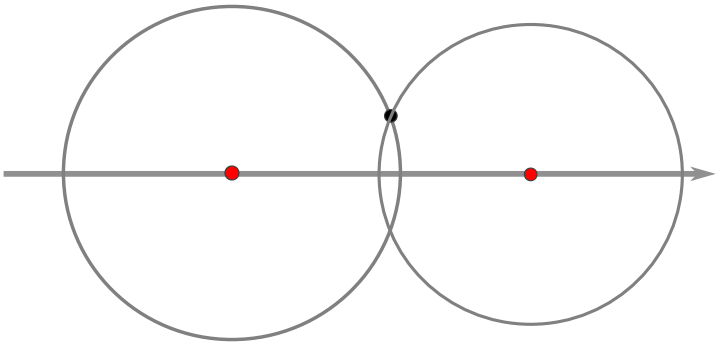


(metric) Multi-Dimensional Scaling (MDS)

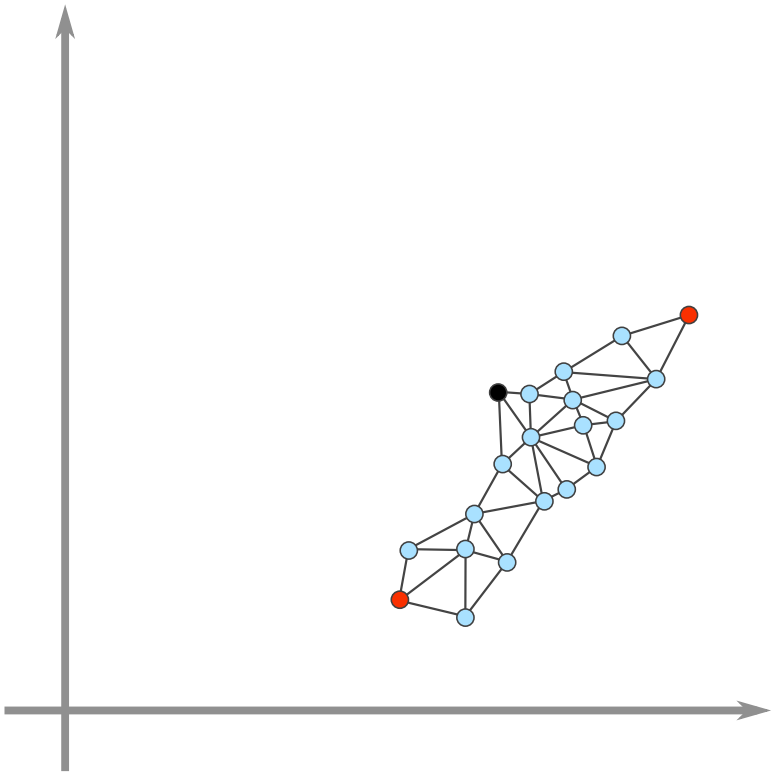


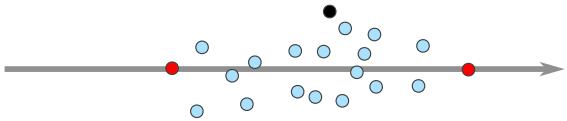
(metric) Multi-Dimensional Scaling (MDS)



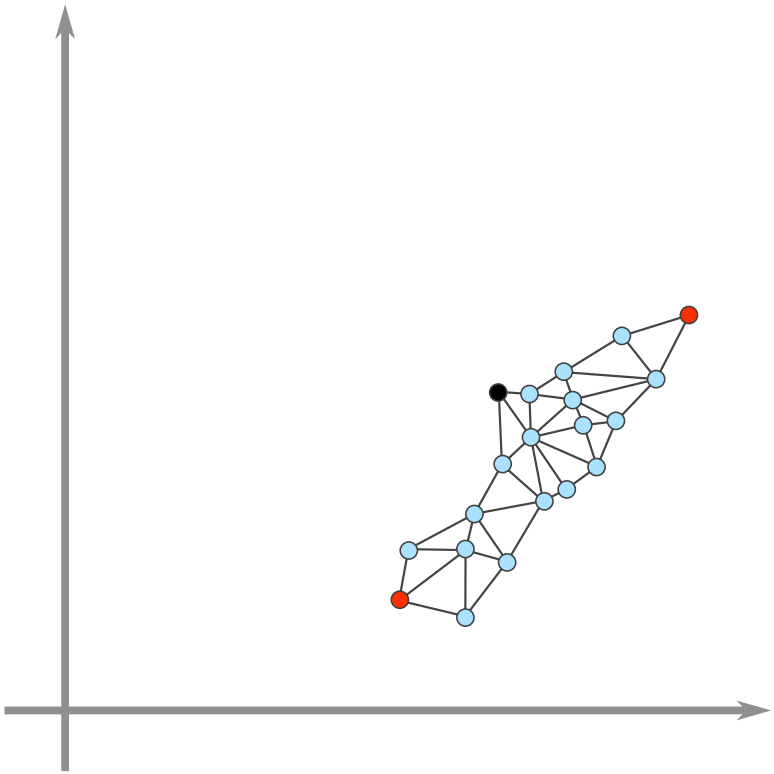


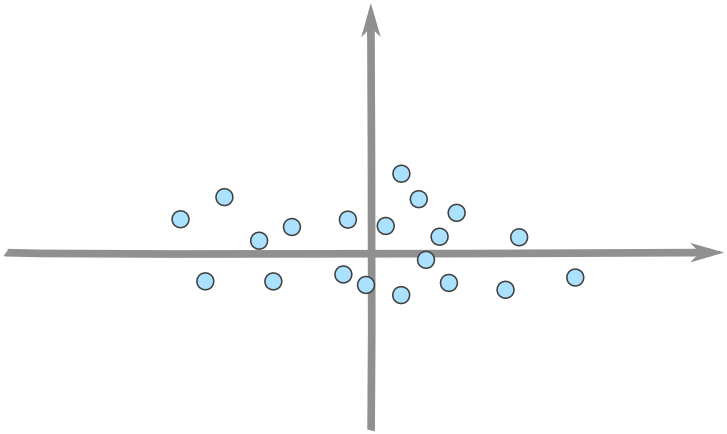
(metric) Multi-Dimensional Scaling (MDS)



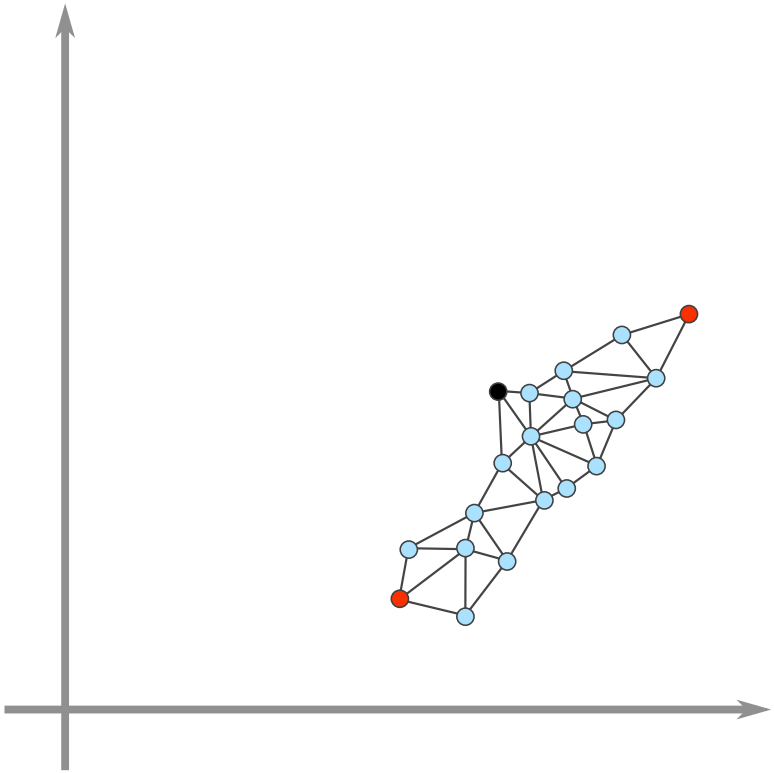


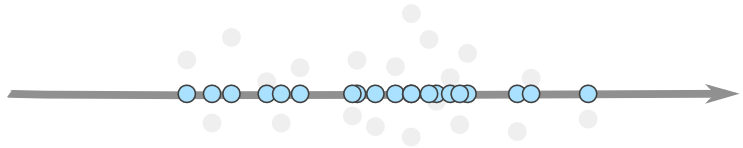
(metric) Multi-Dimensional Scaling (MDS)





(metric) Multi-Dimensional Scaling (MDS)





PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$
- Consider $\mathbf{F}_2$; $f_{ij} = \sum_l X_{li}^2 + \sum_l X_{lj}^2 - 2\sum_l X_{li}X_{lj}$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$
- Consider $\mathbf{F}_2$; $f_{ij} = \sum_l X_{li}^2 + \sum_l X_{lj}^2 - 2\sum_l X_{li}X_{lj}$
- MDS is Eigendecomposition of $-\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$
- Consider $\mathbf{F}_2$; $f_{ij} = \sum_l X_{li}^2 + \sum_l X_{lj}^2 - 2\sum_l X_{li}X_{lj}$
- MDS is Eigendecomposition of $-\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}$
- $\mathbf{C}\mathbf{F}_2\mathbf{C} = \mathbf{C}\mathbf{X}_i^2\mathbf{C} + \mathbf{C}\mathbf{X}_j^2\mathbf{C} - 2\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$
- Consider $\mathbf{F}_2$; $f_{ij} = \sum_l X_{li}^2 + \sum_l X_{lj}^2 - 2\sum_l X_{li}X_{lj}$
- MDS is Eigendecomposition of $-\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}$
- $\mathbf{C}\mathbf{F}_2\mathbf{C} = \underbrace{\mathbf{C}\mathbf{X}_i^2\mathbf{C}}_{0} + \underbrace{\mathbf{C}\mathbf{X}_j^2\mathbf{C}}_{0} - 2\underbrace{\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}}_{\mathbf{Y}\mathbf{Y}^T}$

PCA is MDS on $\mathbf{F}_2$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- Consider $\mathbf{F}_2$; $f_{ij} = F_2(X_i, X_j) = \sum_l (X_{li}^2 - X_{lj})^2$
- Consider $\mathbf{F}_2$; $f_{ij} = \sum_l X_{li}^2 + \sum_l X_{lj}^2 - 2\sum_l X_{li}X_{lj}$
- MDS is Eigendecomposition of $-\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}$
- $\mathbf{C}\mathbf{F}_2\mathbf{C} = \underbrace{\mathbf{C}\mathbf{X}_i^2\mathbf{C}}_{0} + \underbrace{\mathbf{C}\mathbf{X}_j^2\mathbf{C}}_{0} - 2\underbrace{\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}}_{\mathbf{Y}\mathbf{Y}^T}$

Observation

PCA is equivalent to MDS on $\mathbf{F}_2$

PCA is MDS on Outgroup$\mathbf{F}_3$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$

PCA is MDS on Outgroup$\mathbf{F}_3$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- $\mathbf{F}_3(O); f_{ij} = F_3(O; X_i, X_j)$
- $f_ij = \sum_l \left[o_l^2 - o_l X_{li} - o_l X_{li} + X_{li} X_{lj}\right]$

PCA is MDS on Outgroup$\mathbf{F}_3$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- $\mathbf{F}_3(O); f_{ij} = F_3(O; X_i, X_j)$
- $f_ij = \sum_l \left[o_l^2 - o_l X_{li} - o_l X_{li} + X_{li} X_{lj}\right]$
- $\mathbf{C}\mathbf{F}_3\mathbf{C} = \underbrace{\mathbf{C}O^2\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_i\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_j\mathbf{C}}_{0} + \underbrace{\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}}_{\mathbf{Y}\mathbf{Y}^T}$

PCA is MDS on Outgroup$\mathbf{F}_3$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- $\mathbf{F}_3(O); f_{ij} = F_3(O; X_i, X_j)$
- $f_ij = \sum_l \left[o_l^2 - o_l X_{li} - o_l X_{li} + X_{li} X_{lj}\right]$
- $\mathbf{C}\mathbf{F}_3\mathbf{C} = \underbrace{\mathbf{C}O^2\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_i\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_j\mathbf{C}}_{0} + \underbrace{\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}}_{\mathbf{Y}\mathbf{Y}^T}$

PCA is MDS on Outgroup$\mathbf{F}_3$

- PCA is decomposition of Covariance matrix: $\mathbf{Y}\mathbf{Y}^T$
- $\mathbf{F}_3(O); f_{ij} = F_3(O; X_i, X_j)$
- $f_ij = \sum_l \left[o_l^2 - o_l X_{li} - o_l X_{li} + X_{li} X_{lj}\right]$
- $\mathbf{C}\mathbf{F}_3\mathbf{C} = \underbrace{\mathbf{C}O^2\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_i\mathbf{C}}_{0} - \underbrace{\mathbf{C}O\mathbf{X}_j\mathbf{C}}_{0} + \underbrace{\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}}_{\mathbf{Y}\mathbf{Y}^T}$

Observation

Decomposition of *any* centered F_3-matrix is equivalent to PCA.

0-diagonal MDS

In aDNA applications (e.g. Fu et al 2016), MDS is performed on

$$\mathbf{M} = \begin{pmatrix} 0 & 1 - f_3(O; X_1, X_2) & \dots & 1 - f_3(O; X_1, X_n) \\ 1 - f_3(O; X_1, X_2) & 0 & \dots & 1 - f_3(O; X_2, X_n) \\ \vdots & \vdots & \ddots & 1 - f_3(O; X_i, X_n) \\ 1 - f_3(O; X_1, X_n) & 1 - f_3(O; X_2, X_n) & \dots & 0 \end{pmatrix}$$

It can be shown that

$$\mathbf{CMC} = \mathbf{CF_2C} + \mathbf{COC}$$

with $o_{ii} = f_2(O, X_i) - 1; o_{ij} = 0$

Fu et al. 2016

How can we interpret PCA in context of F-stats?

- PCA is decomposition of $\mathbf{F}_3(\boldsymbol{\mu}, X_1, X_2)$-matrix
- PCA is a function of (unnormalized)-$\mathbf{F}_2$-matrix
  - same information content
  - differences in analysis choices (individual/population, normalization, estimate) need to be justified

How can we interpret PCA in context of F-stats?

- PCA is decomposition of $\mathbf{F}_3(\boldsymbol{\mu}, X_1, X_2)$-matrix
- PCA is a function of (unnormalized)-$\mathbf{F}_2$-matrix
  - same information content
  - differences in analysis choices (individual/population, normalization, estimate) need to be justified
- setting diagonal allows some leeway

How can we interpret PCA in context of F-stats?

- PCA is decomposition of $\mathbf{F}_3(\boldsymbol{\mu}, X_1, X_2)$-matrix
- PCA is a function of (unnormalized)-$\mathbf{F}_2$-matrix
  - same information content
  - differences in analysis choices (individual/population, normalization, estimate) need to be justified
- setting diagonal allows some leeway
- *Conjecture*: PCA is analogous to midpoint-rooted phylogenetic analyses; Outgroup-F_3 allow other rootings.

F-statistics on PCA-plot

- Recall that PCA is just translation + rotation

F-statistics on PCA-plot

- Recall that PCA is just translation + rotation
- Distances (such as F_2) are invariant to translation + rotation

F-statistics on PCA-plot

- Recall that PCA is just translation + rotation
- Distances (such as F_2) are invariant to translation + rotation
- $$F_2(X_1, X_2) = \sum_{\text{loci}} (x_{1l} - x_{2l})^2$$

F-statistics on PCA-plot

- Recall that PCA is just translation + rotation
- Distances (such as F_2) are invariant to translation + rotation
- $$F_2(X_1, X_2) = \sum_{\text{loci}} (x_{1l} - x_{2l})^2$$
- $$F_2(X_1, X_2) = \sum_{\text{PCs}} (x_{1p} - x_{2p})^2$$

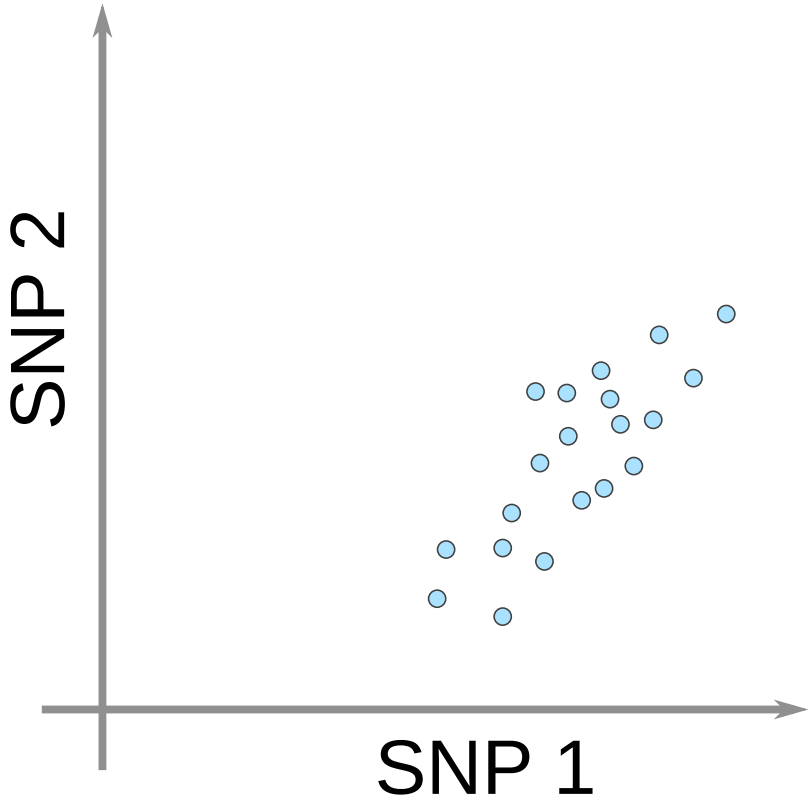
F-statistics on PCA-plot

- Recall that PCA is just translation + rotation
- Distances (such as F_2) are invariant to translation + rotation
- $$F_2(X_1, X_2) = \sum_{\text{loci}} (x_{1l} - x_{2l})^2$$
- $$F_2(X_1, X_2) = \sum_{\text{PCs}} (x_{1p} - x_{2p})^2$$

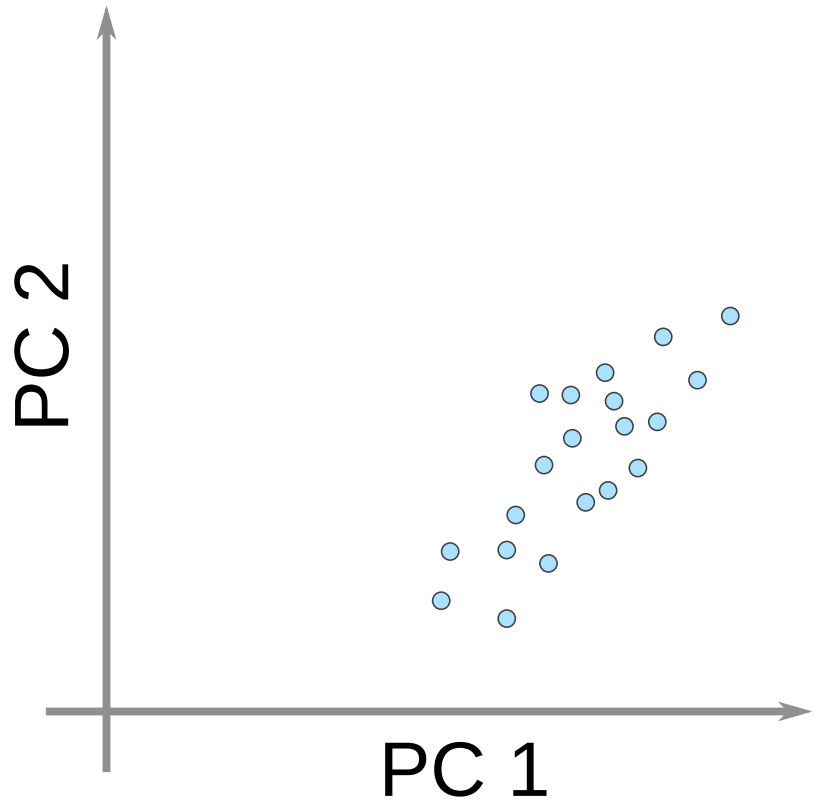
Observation

F_2 can be decomposed in contributions of different principal components

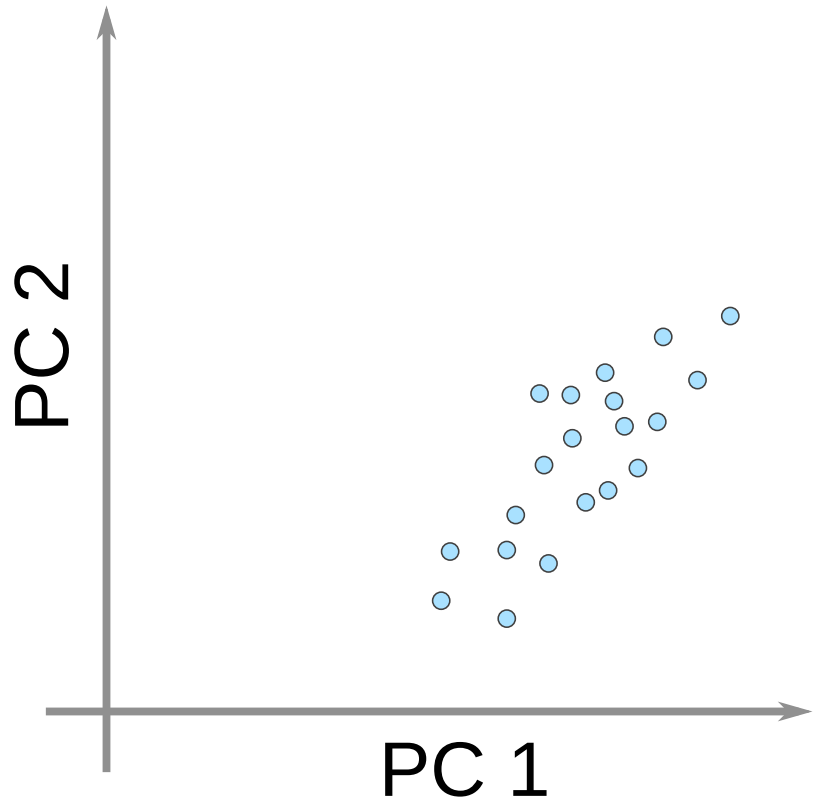
F-statistics on PCA-plot



F-statistics on PCA-plot

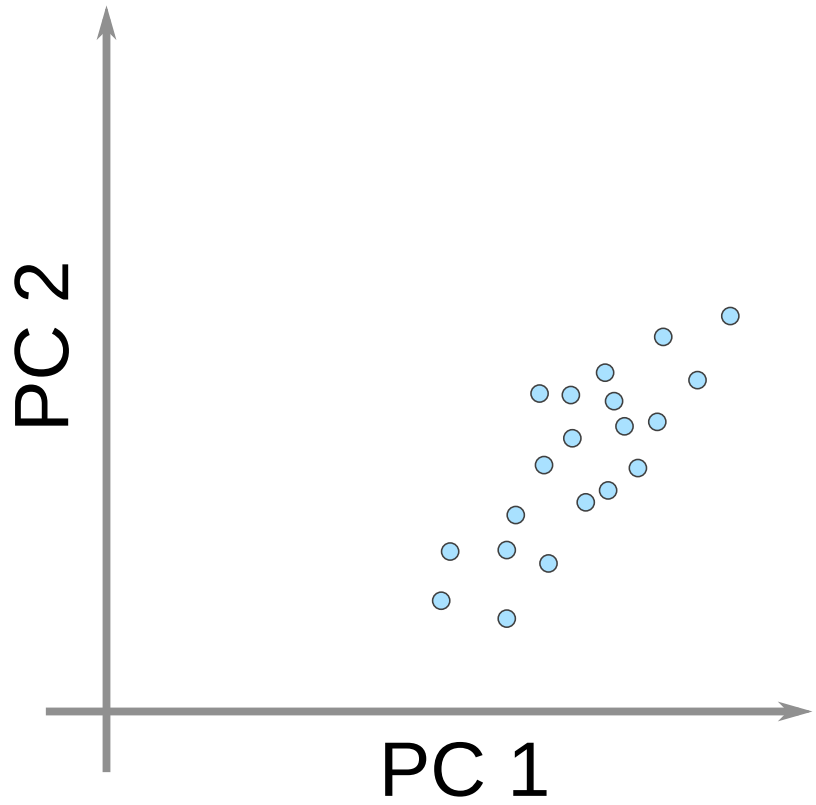


F-statistics on PCA-plot



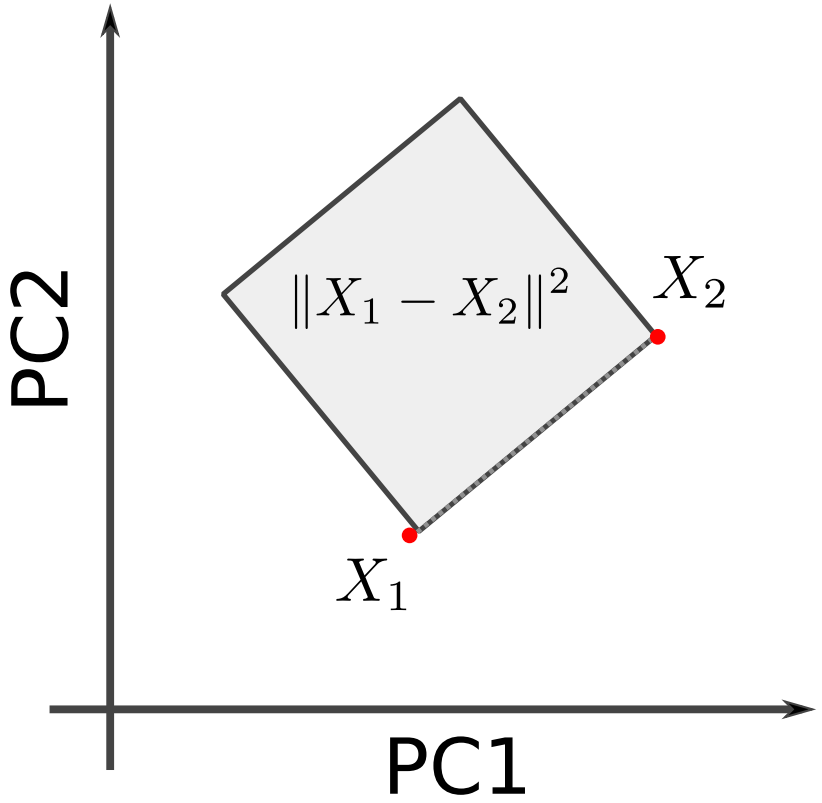
- F-statistics have a geometrical representation on PCA-plot
- Exact only if we use *all* PCs

F-statistics on PCA-plot



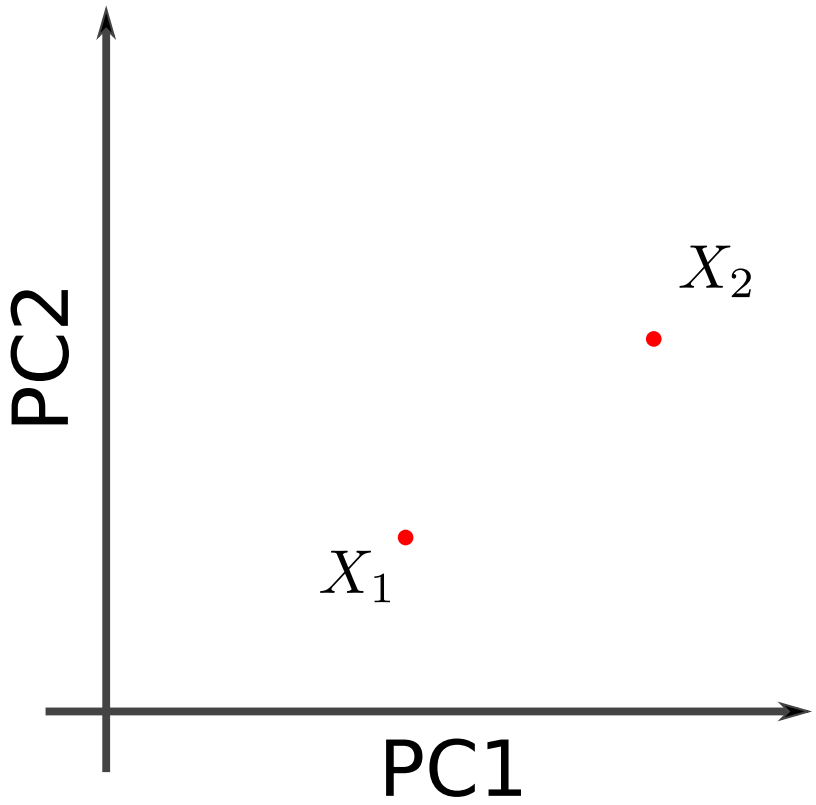
- F-statistics have a geometrical representation on PCA-plot
- Exact only if we use *all* PCs
- Good approximation for 2D-plot if first 2 PCs capture relevant population structure

F_2-statistic on PCA-plot



- $F_2(X_1, X_2) = \sum_l (X_{1l} - X_{2l})^2$
- $F_2(X_1, X_2) = \|X_1, X_2\|^2$

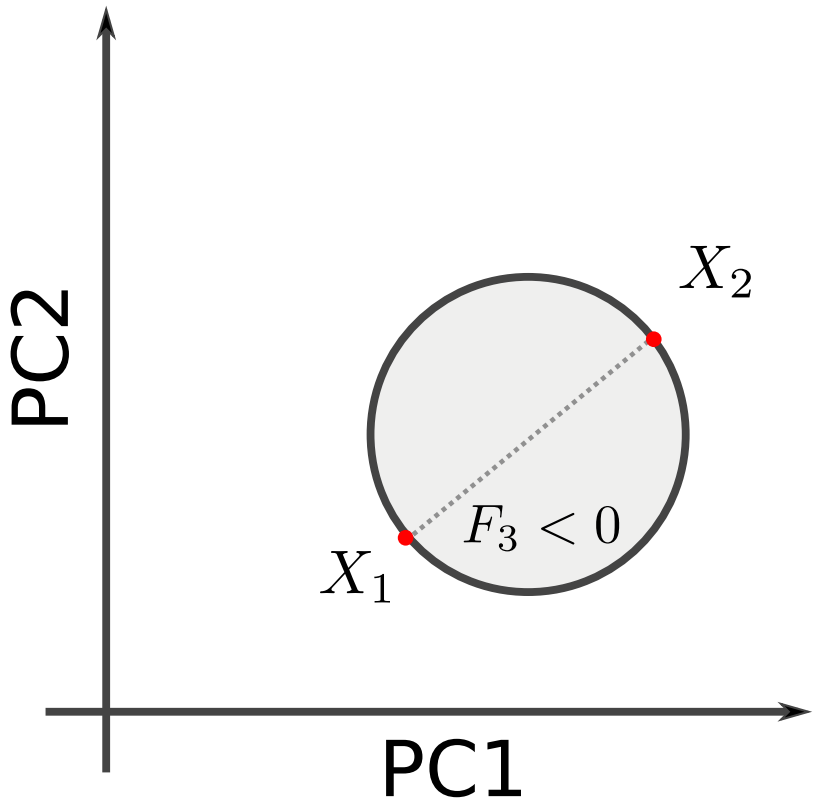
Admixed populations (F_3) on PCA-plot



- Given X_1, X_2, which pops have $F_3 < 0$?

McVean 2009

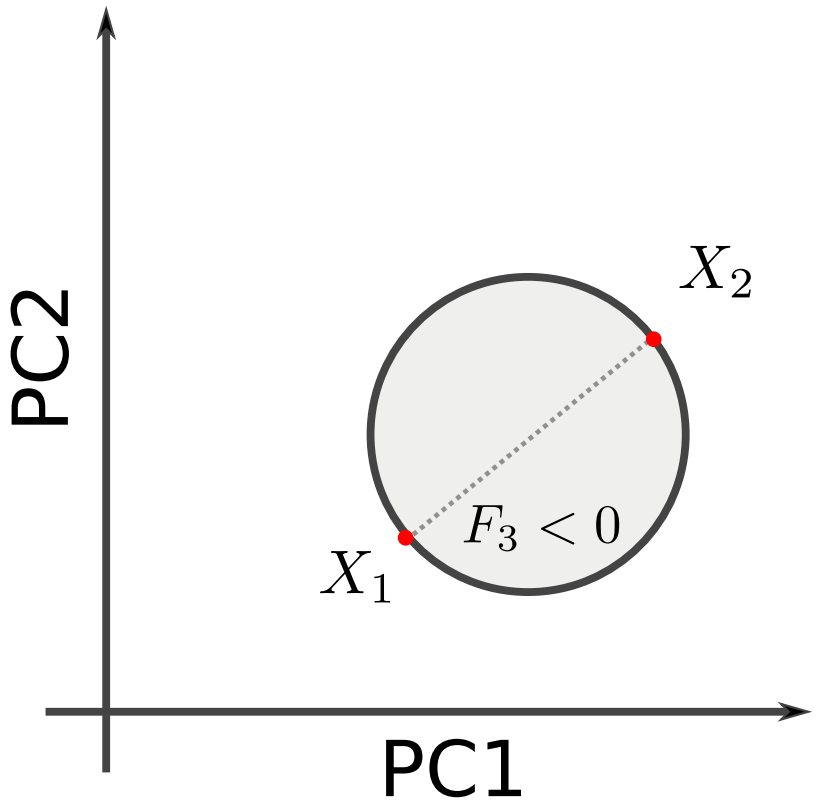
Admixed populations (F_3) on PCA-plot



- Given X_1, X_2, which pops have $F_3 < 0$?
- $F_3(Y; X_1, X_2) = 0$ is a circle!

McVean 2009

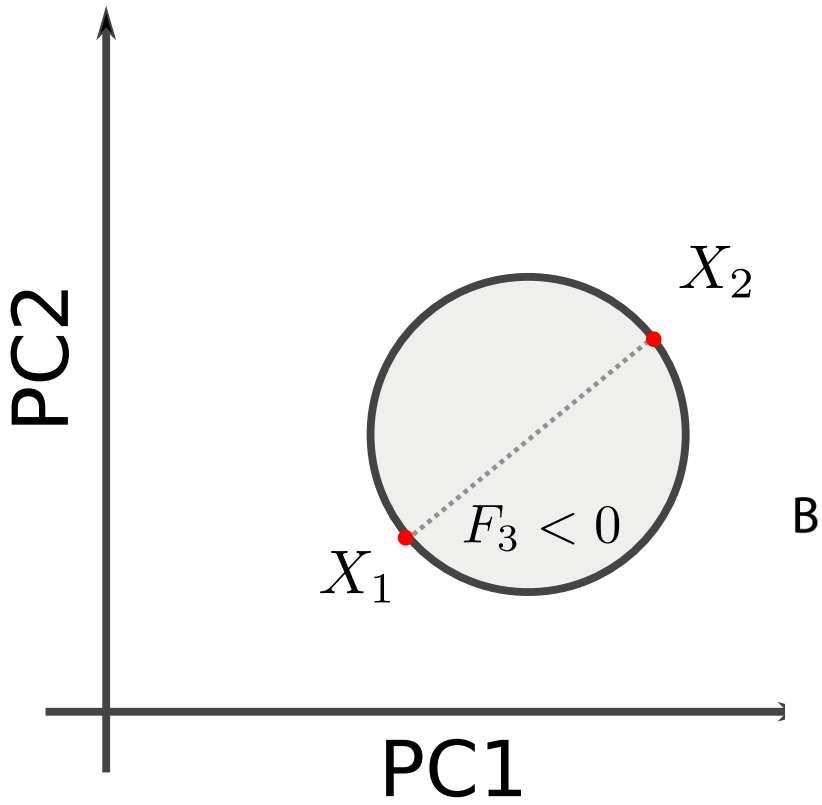
Admixed populations (F_3) on PCA-plot



- Given X_1, X_2, which pops have $F_3 < 0$?
- $F_3(Y; X_1, X_2) = 0$ is a circle!

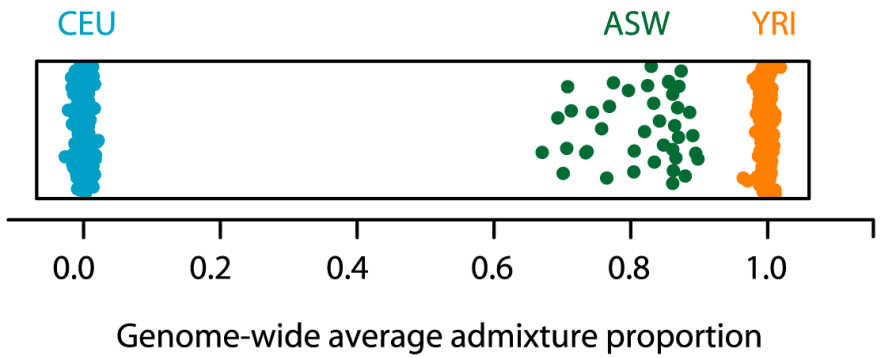
McVean 2009

Admixed populations (F_3) on PCA-plot



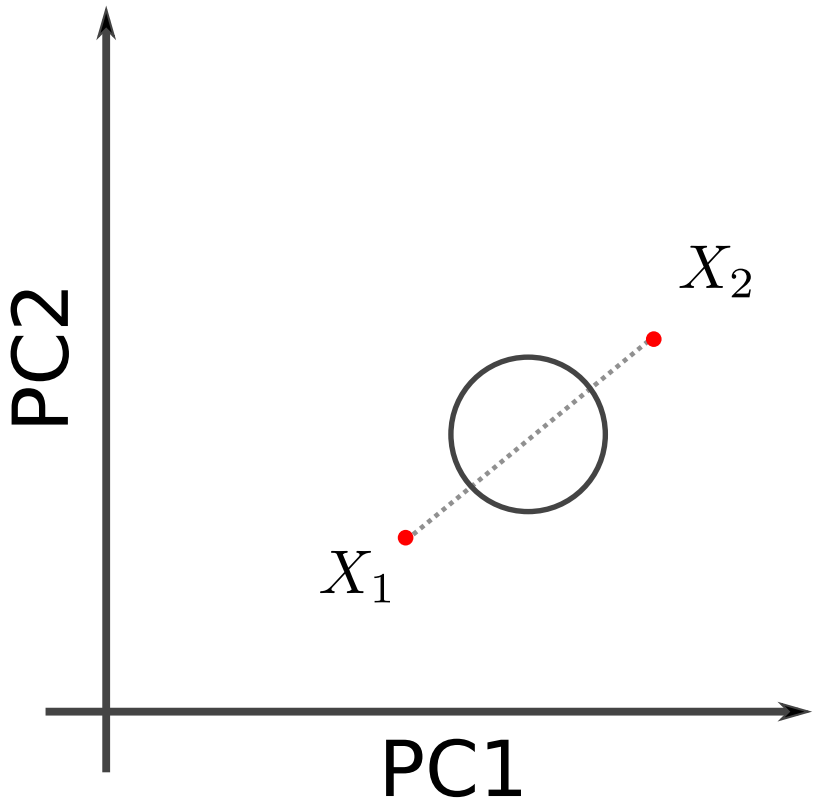
- Given X_1, X_2, which pops have $F_3 < 0$?
- $F_3(Y; X_1, X_2) = 0$ is a circle!

B



McVean 2009

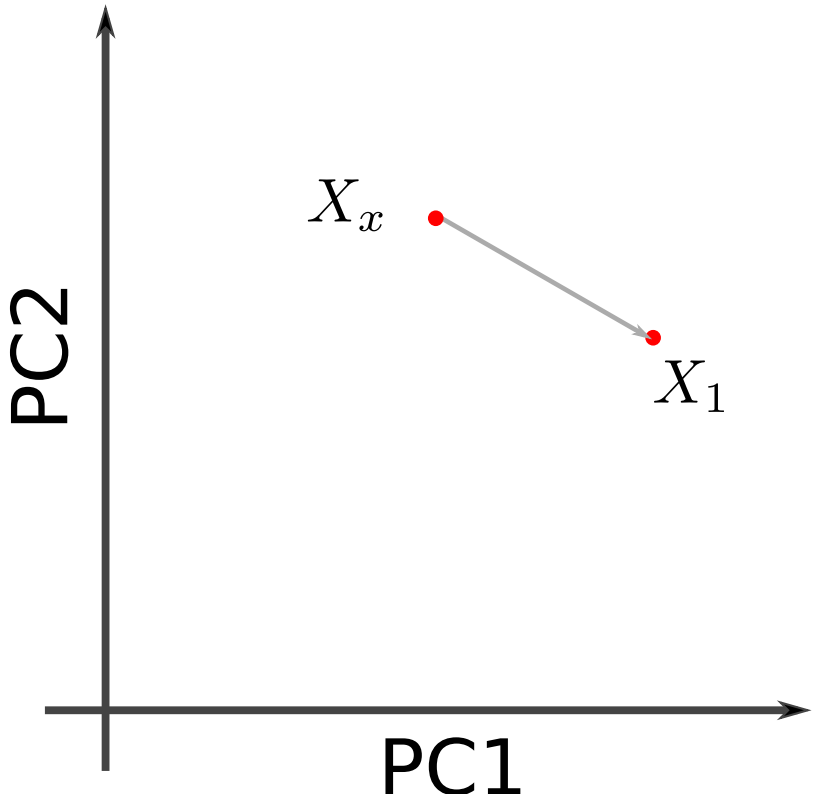
Admixed populations (F_3) on PCA-plot



- Given X_1, X_2, which pops have $F_3 < 0$?
- $F_3(Y; X_1, X_2) = 0$ is a circle!
- $F_3(Y; X_1, X_2) = k < 0$ is smaller circle

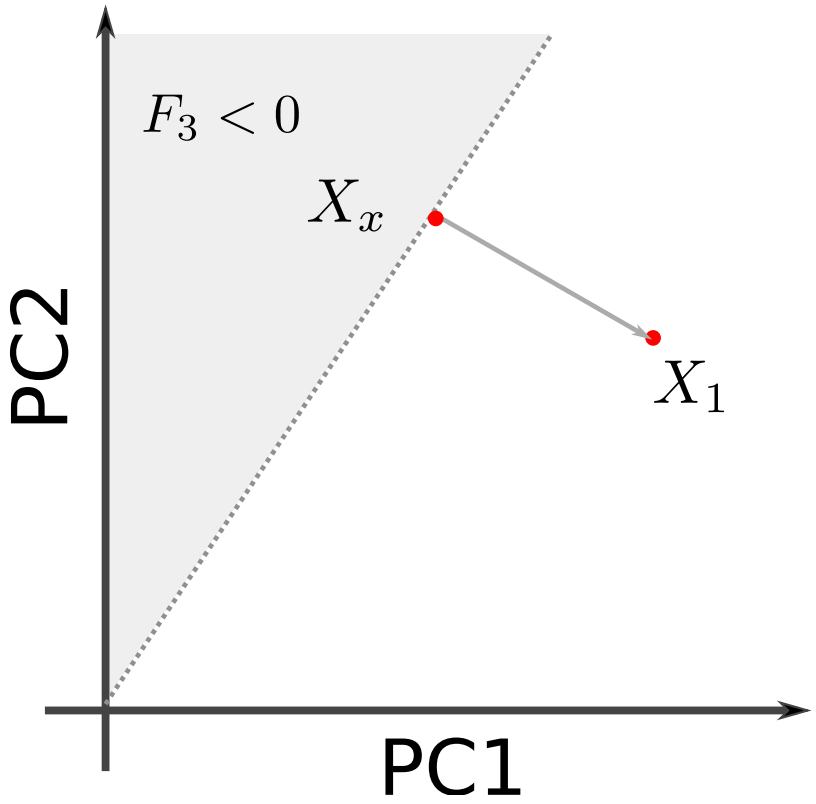
McVean 2009

Admixture F_3-stats on PCA-plot



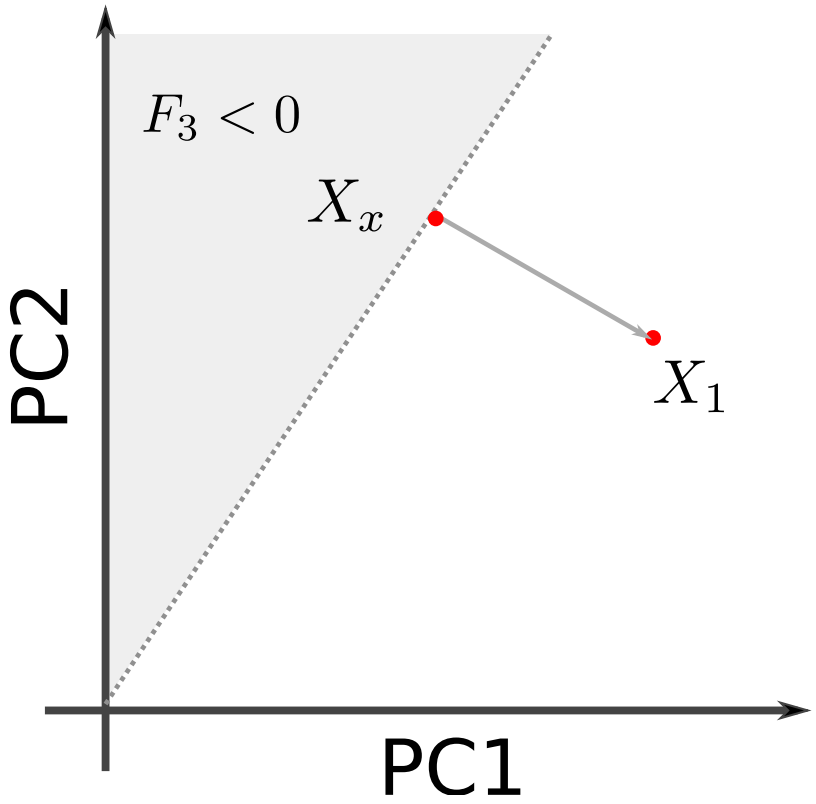
- Given X_1, X_x, which pops X_2 have $F_3 < 0$?

Admixture F_3-stats on PCA-plot



- Given X_1, X_x, which pops X_2 have $F_3 < 0$?
- F_3 is 0 if $(X_x; X_1), (X_x; X_2)$ form a right angle!

Admixture F_3-stats on PCA-plot

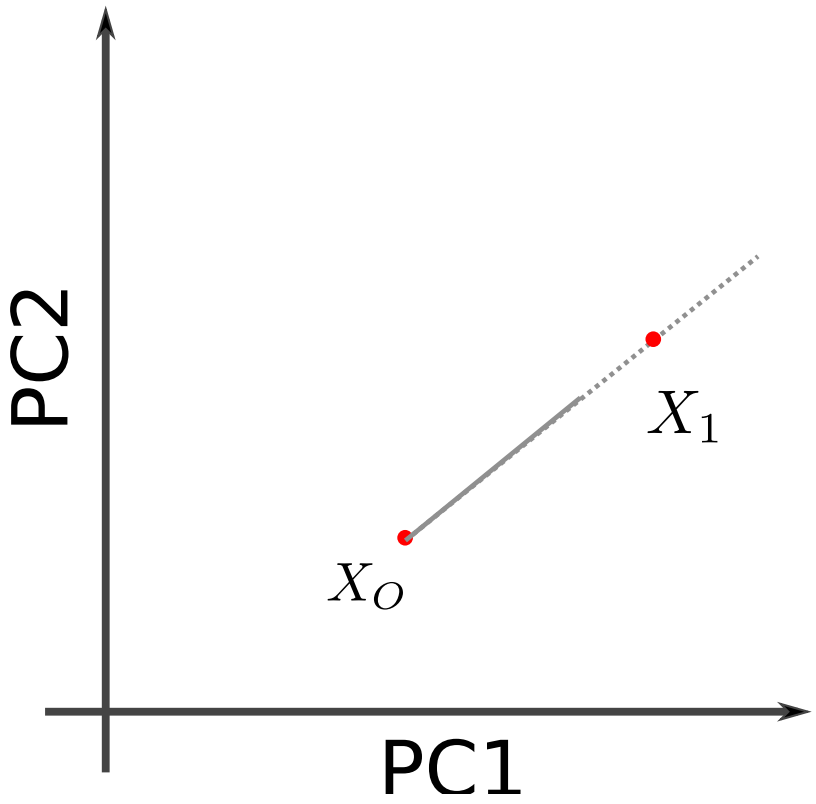


- Given X_1, X_x, which pops X_2 have $F_3 < 0$?
- F_3 is 0 if $(X_x; X_1), (X_x; X_2)$ form a right angle!
- Inner (dot) product: $F_3(X_x; X_1, X_2) = \langle X_x - X_1, X_x - X_2 \rangle$

Outgroup F_3-stats on PCA-plot

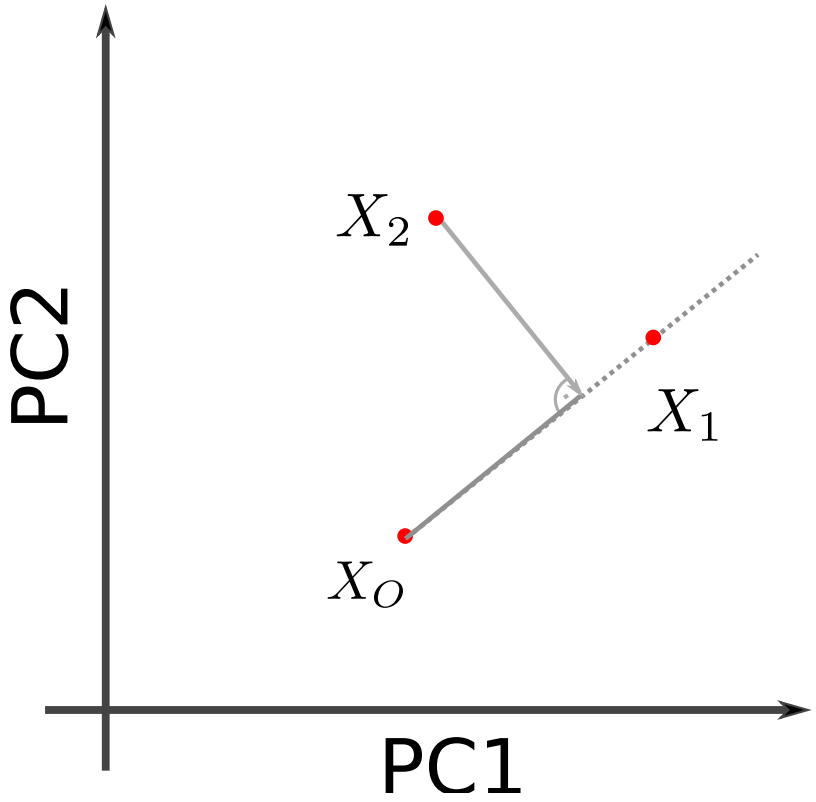
- Given X_O, X_1, which pop X_2 has highest F_3?

Outgroup F_3-stats on PCA-plot



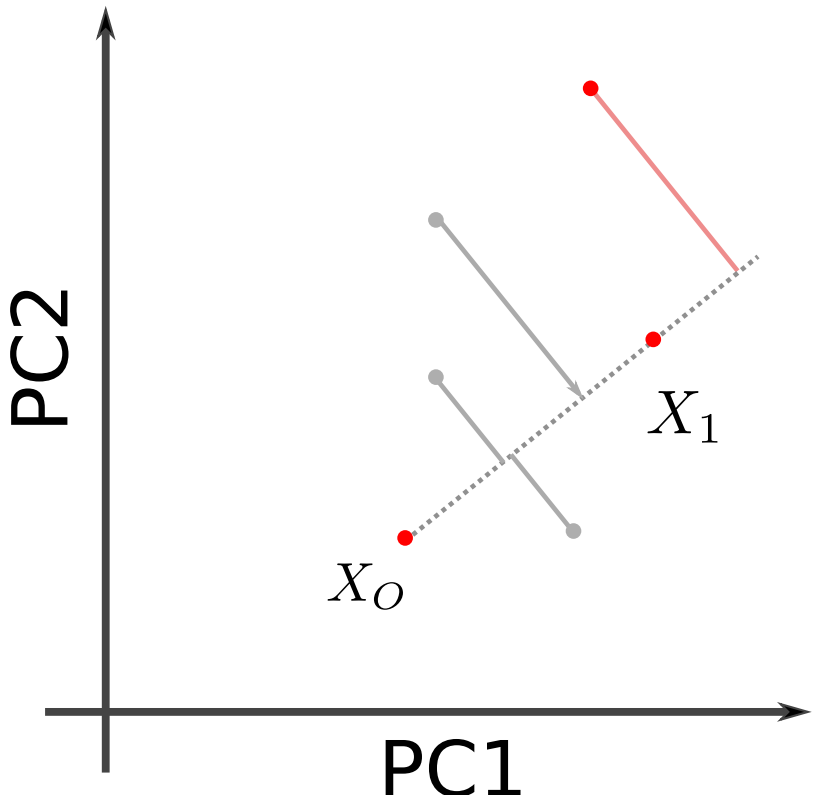
- Given X_O, X_1, which pop X_2 has highest F_3?
- As $F_3 = \langle X_O - X_1, X_O - X_2 \rangle$, project $\overline{X_O X_2}$ on $\overline{X_O X_1}$

Outgroup F_3-stats on PCA-plot



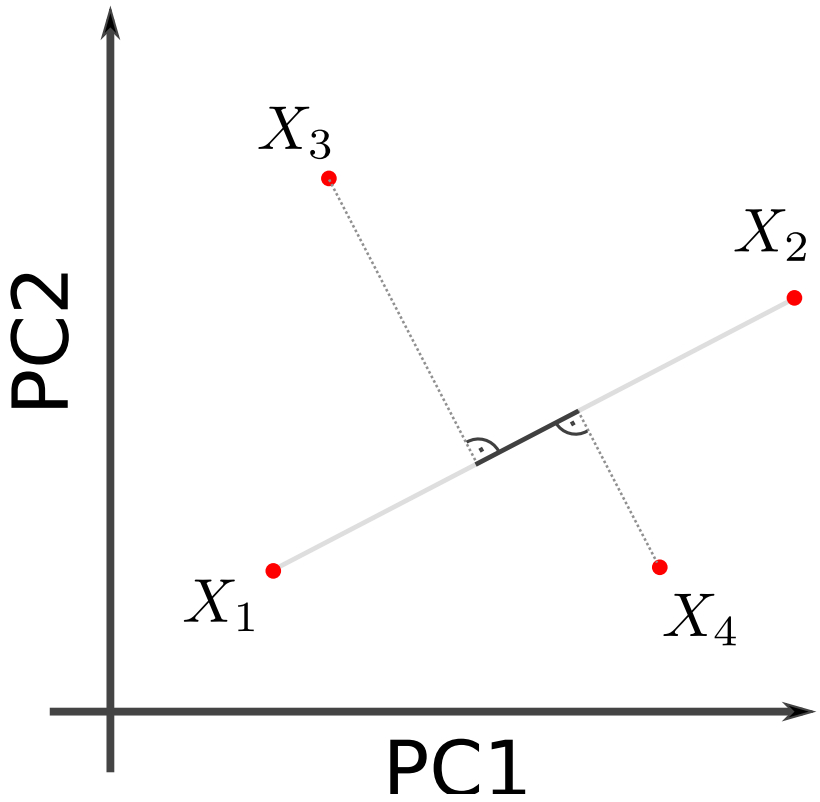
- Given X_O, X_1, which pop X_2 has highest F_3?
- As $F_3 = \langle X_O - X_1, X_O - X_2 \rangle$, project $\overline{X_O X_2}$ on $\overline{X_O X_1}$
- F_3 is proportional to segment in dir $X_0 \to X_1$

Outgroup F_3-stats on PCA-plot



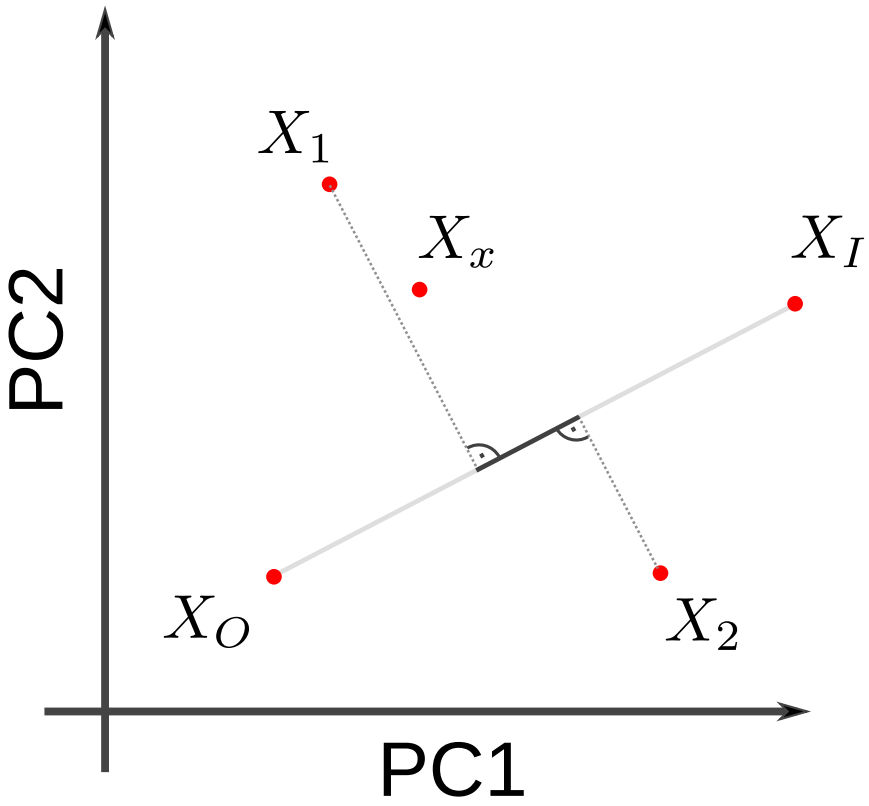
- Given X_O, X_1, which pop X_2 has highest F_3?
- As $F_3 = \langle X_O - X_1, X_O - X_2 \rangle$, project $\overline{X_O X_2}$ on $\overline{X_O X_1}$
- F_3 is proportional to segment in dir $X_0 \to X_1$
- Pop furthest on axis has highest F_3.

F_4-stats on PCA-plot



- F_4 is projection of $\overline{X_3X_4}$ on $\overline{X_1X_2}$

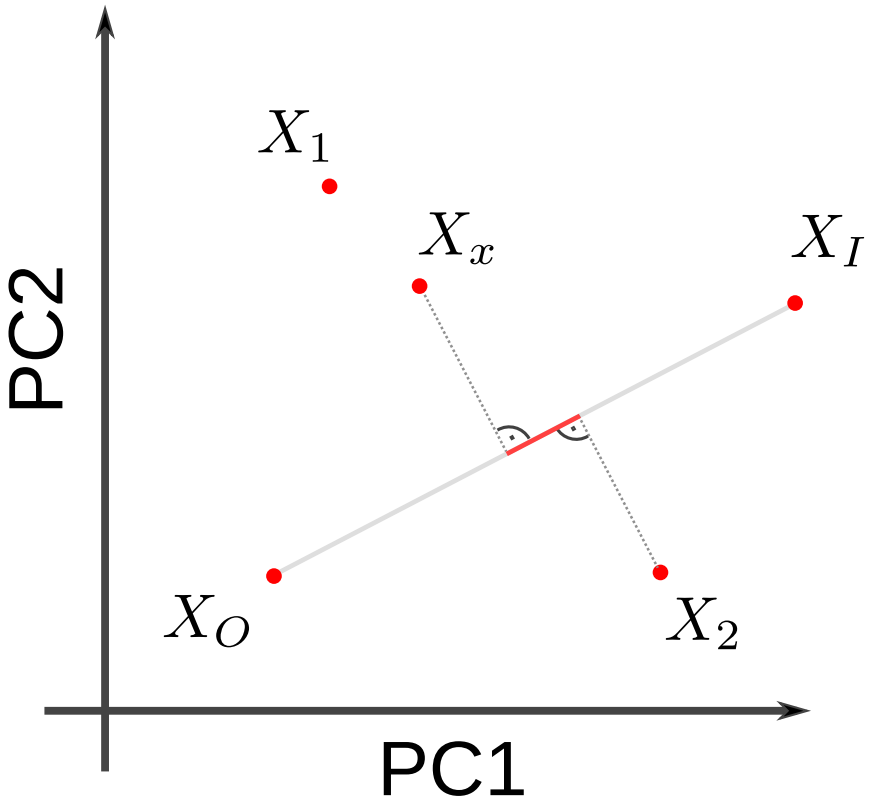
F_4-ratio on PCA plot



- Estimate admixture proportions using

$$\alpha = \frac{F_4(X_X, X_2; X_I, X_O)}{F_4(X_1, X_2; X_I, X_O)}$$

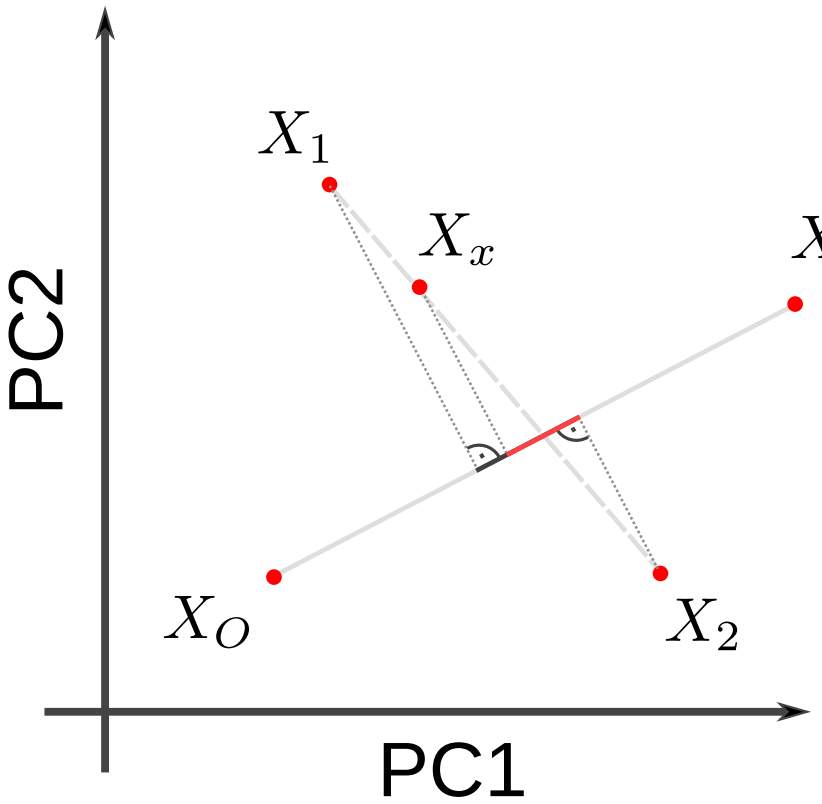
F_4-ratio on PCA plot



- Estimate admixture proportions using

$$\alpha = \frac{F_4(X_X, X_2; X_I, X_O)}{F_4(X_1, X_2; X_I, X_O)}$$

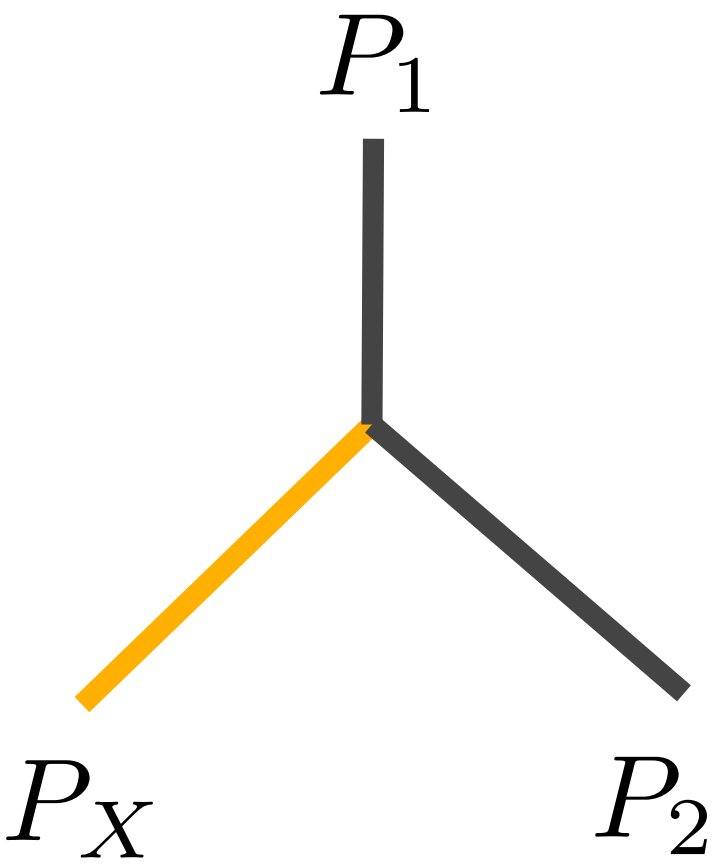
F_4-ratio on PCA plot



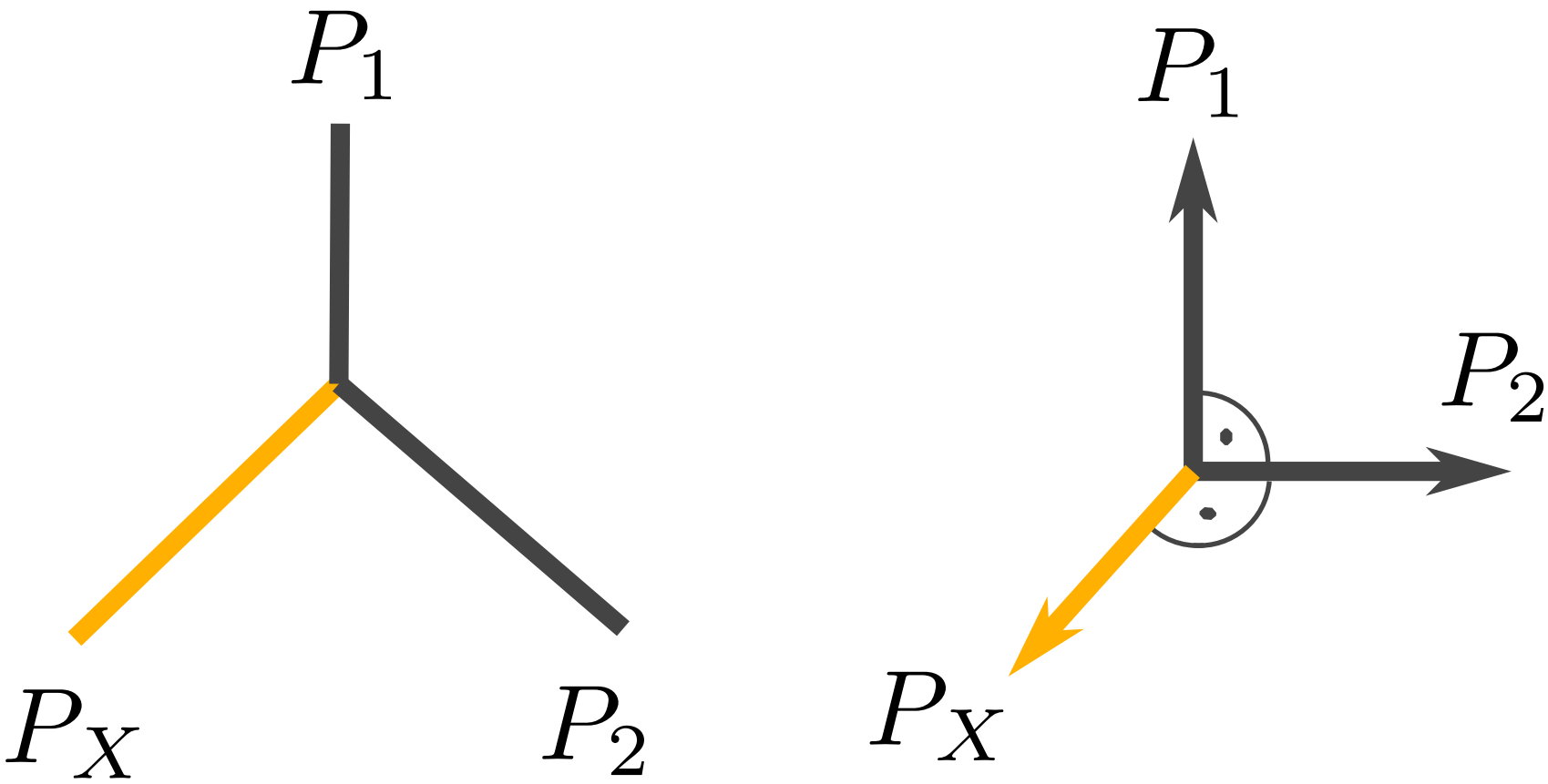
- Estimate admixture proportions using

$$\alpha = \frac{F_4(X_X, X_2; X_I, X_O)}{F_4(X_1, X_2; X_I, X_O)}$$

Where does Orthogonality come from?



Where does Orthogonality come from?



How can we interpret F-stats in the context of PCA?

- F-stats have geometric interpretation

How can we interpret F-stats in the context of PCA?

- F-stats have geometric interpretation
- exact in n-dimensional, approximate in 2D

How can we interpret F-stats in the context of PCA?

- F-stats have geometric interpretation
- exact in n-dimensional, approximate in 2D
- F_3 is a (hyper)-circle

How can we interpret F-stats in the context of PCA?

- F-stats have geometric interpretation
- exact in n-dimensional, approximate in 2D
- F_3 is a (hyper)-circle
- F_4 established orthogonality

How can we interpret F-stats in the context of PCA?

- F-stats have geometric interpretation
- exact in n-dimensional, approximate in 2D
- F_3 is a (hyper)-circle
- F_4 established orthogonality
- Orthogonality stems from independence of drift in distinct parts of history

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events
   - same F_3 value may arise from distinct admixture events, PCs may point to differences

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events
   - same F_3 value may arise from distinct admixture events, PCs may point to differences
4. Understand discrepancies

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events
   - same F_3 value may arise from distinct admixture events, PCs may point to differences
4. Understand discrepancies
   - most likely due to data artifacts / higher PCs

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events
   - same F_3 value may arise from distinct admixture events, PCs may point to differences
4. Understand discrepancies
   - most likely due to data artifacts / higher PCs
5. Standardize normalization
   - $F_2^{(\text{PCA})} = \frac{1}{\hat{\sigma}} \sum (X_i - X_j)^2$
   - $F_2^{(\text{F-stats})} = \sum (X_i - X_j)^2$

Applications

1. Better link F-stats and PCA results
   - use Dimensions / Orthogonality for useful data representations
2. Diagnostic plots
   - Easy checks whether assumptions for f_4-ratio, `qpadm`, etc. are satisfied
3. Distinguish admixture events
   - same F_3 value may arise from distinct admixture events, PCs may point to differences
4. Understand discrepancies
   - most likely due to data artifacts / higher PCs
5. Standardize normalization
   - $F_2^{(\text{PCA})} = \frac{1}{\hat{\sigma}} \sum (X_i - X_j)^2$
   - $F_2^{(\text{F-stats})} = \sum (X_i - X_j)^2$
6. Better out-of-sample predictions
   - `qpGraph` and other tools fail with large samples

Thank You!