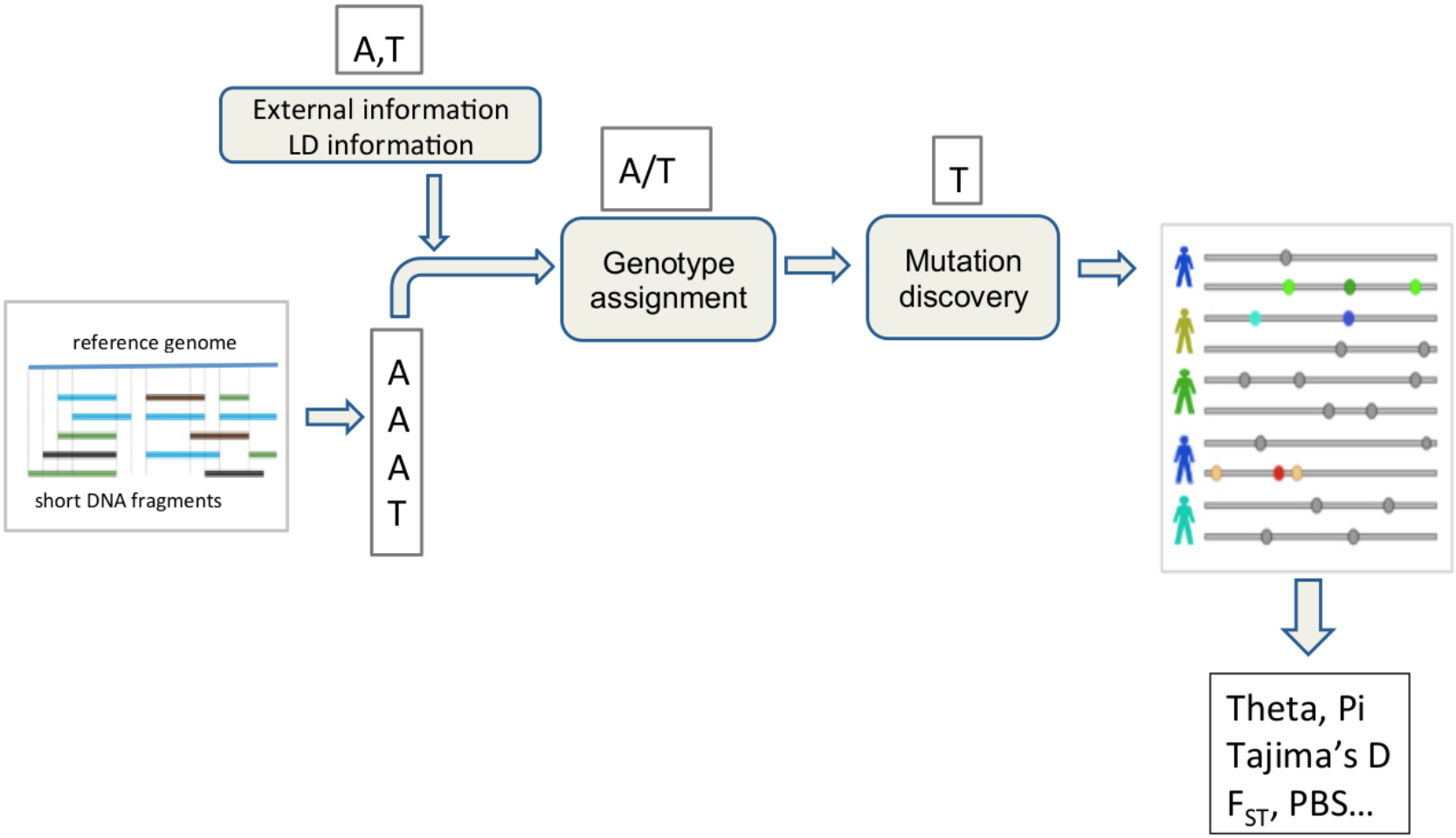
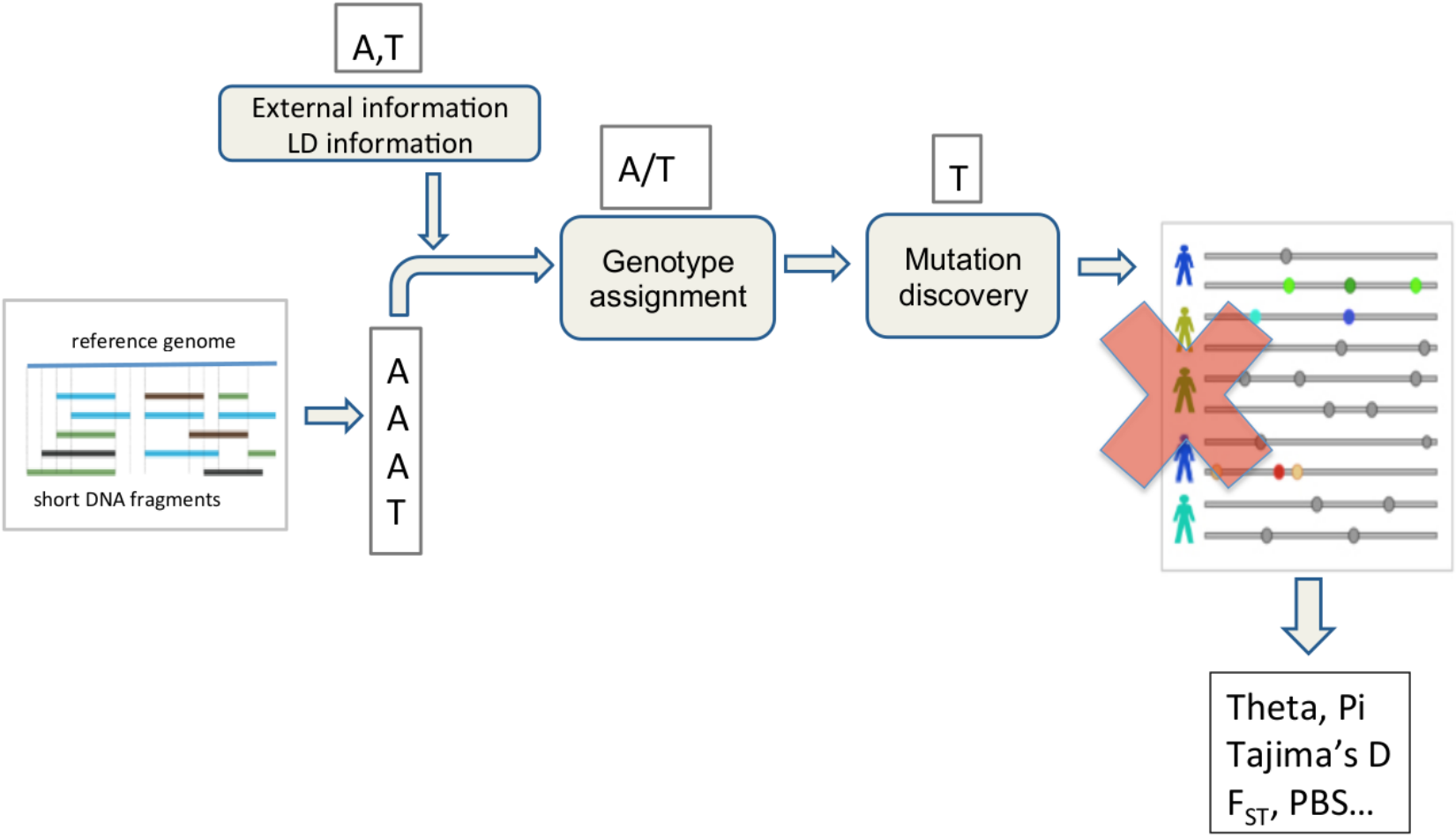
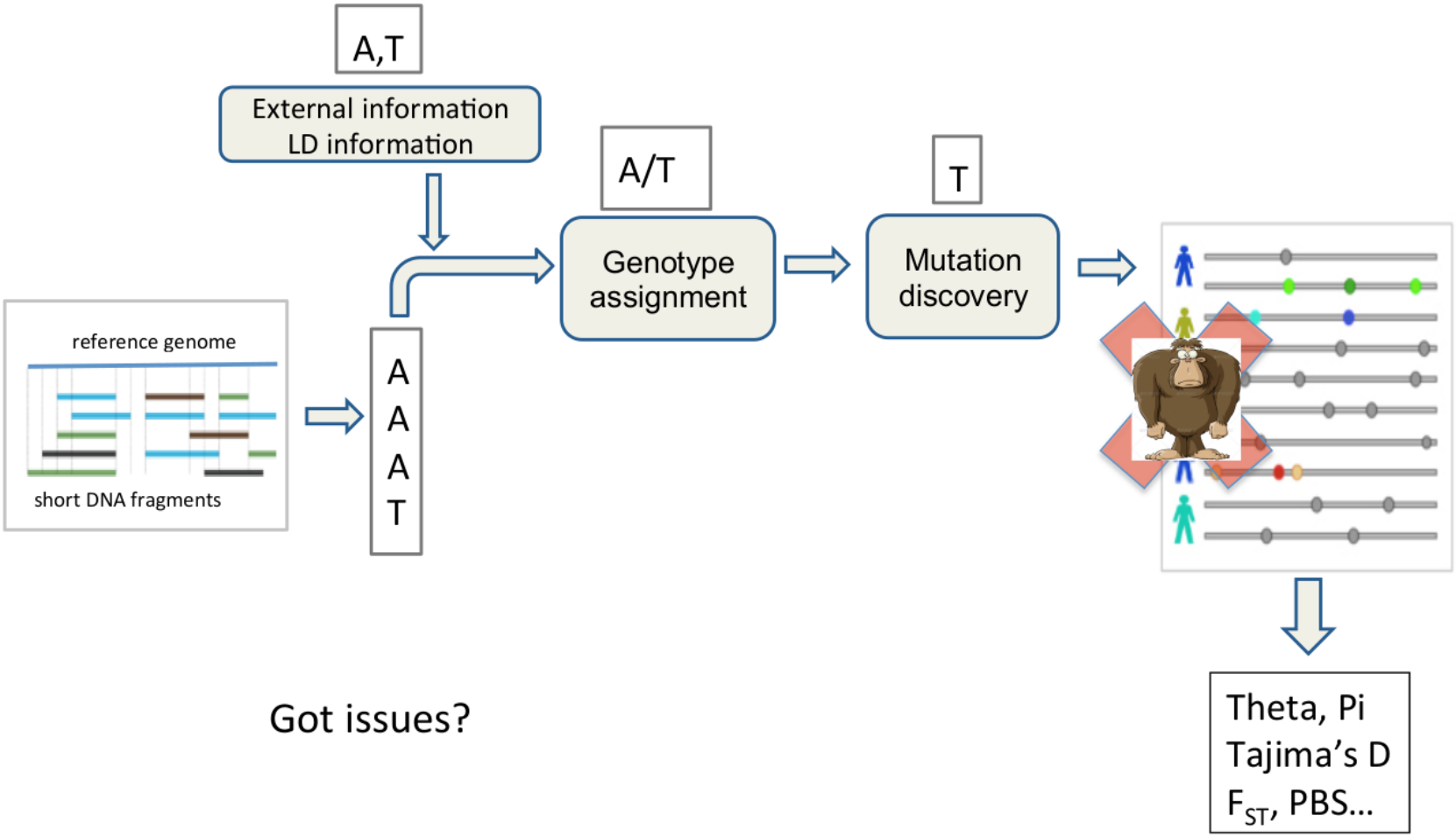
Imperial College London

Population structure and admixture analysis

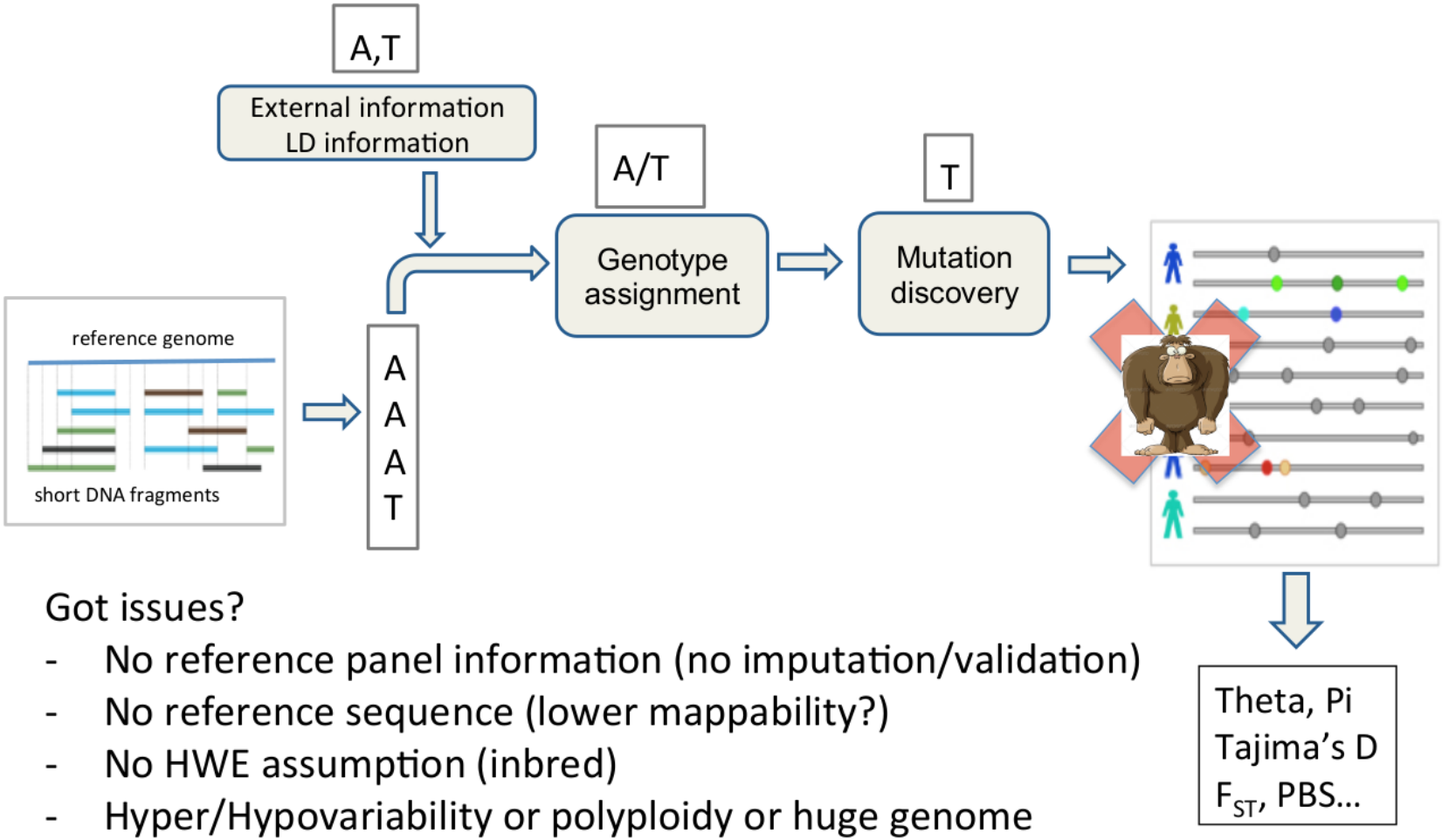
Matteo Fumagalli


A,T
External information
LD information
A/T
Genotype
assignment
T
Mutation
discovery
reference genome
short DNA fragments
A
A
A
T
Theta, Pi
Tajima's D
F_{ST}, PBS...


A,T
External information
LD information
A/T
Genotype
assignment
T
Mutation
discovery
reference genome
short DNA fragments
A
A
A
T
Theta, Pi
Tajima's D
F_{ST}, PBS...

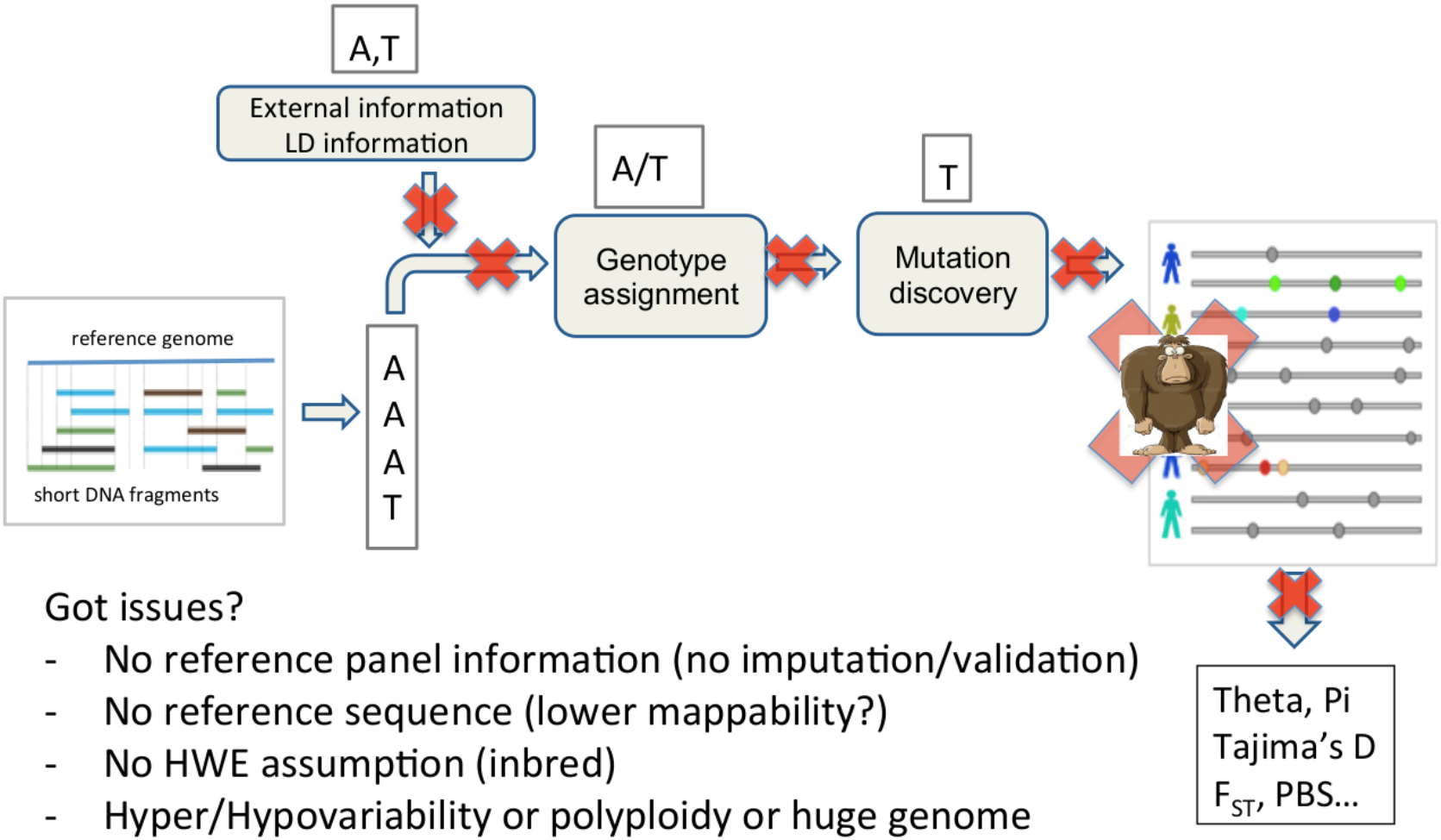

A,T
External information
LD information
A/T
Genotype assignment
T
Mutation discovery
reference genome
short DNA fragments
A
A
A
T
Theta, Pi
Tajima's D
F_{ST}, PBS...

Got issues?



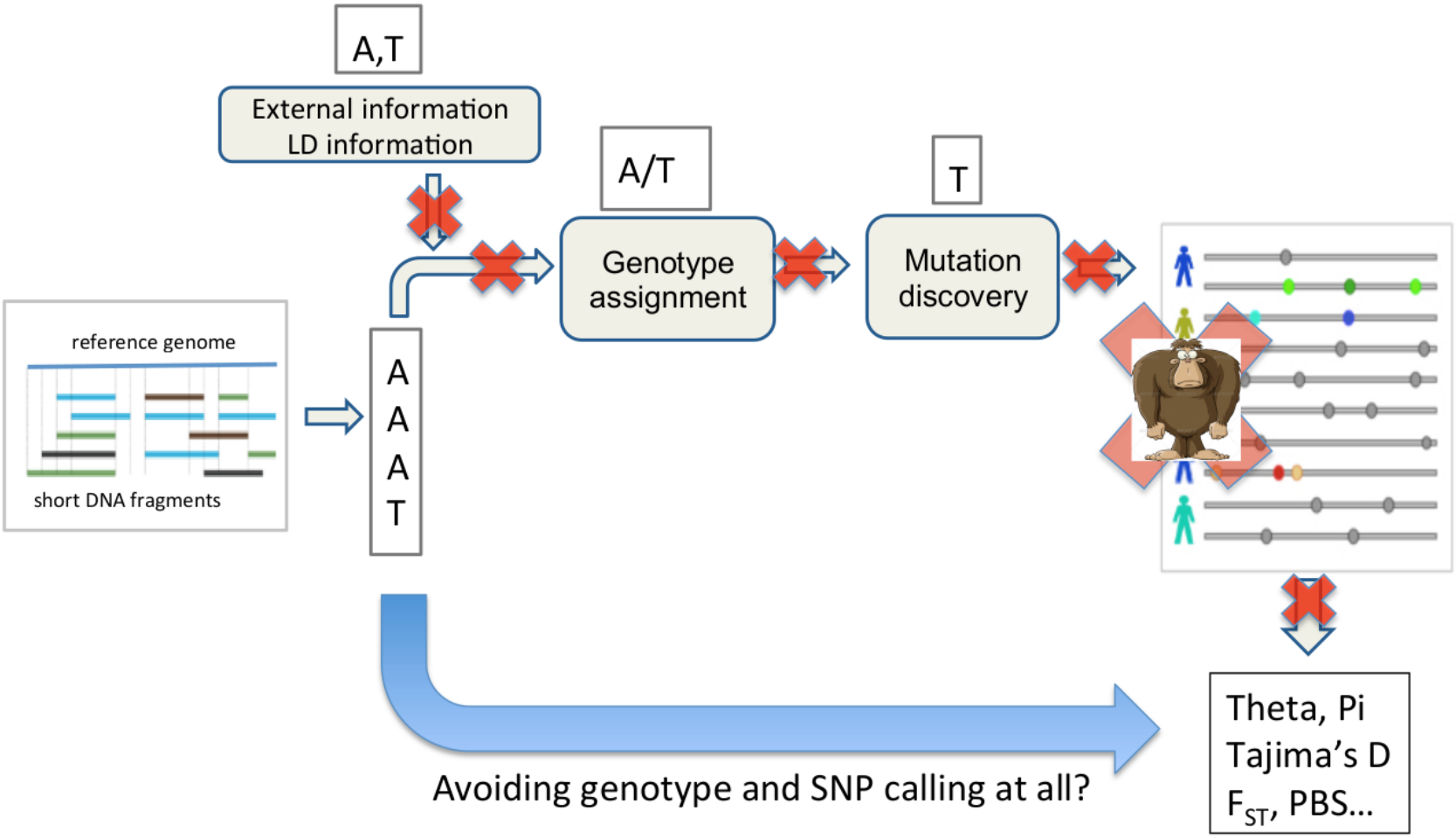
Got issues?

- No reference panel information (no imputation/validation)
- No reference sequence (lower mappability?)
- No HWE assumption (inbred)
- Hyper/Hypovariability or polyploidy or huge genome
- No money (?)
- …



Got issues?

- No reference panel information (no imputation/validation)
- No reference sequence (lower mappability?)
- No HWE assumption (inbred)
- Hyper/Hypovariability or polyploidy or huge genome
- No money (?)
- **Your inferences will be wrong!**


A,T
External information
LD information
A/T
Genotype assignment
T
Mutation discovery
reference genome
short DNA fragments
A
A
A
T
Avoiding genotype and SNP calling at all?
Theta, Pi
Tajima's D
F_{ST}, PBS...

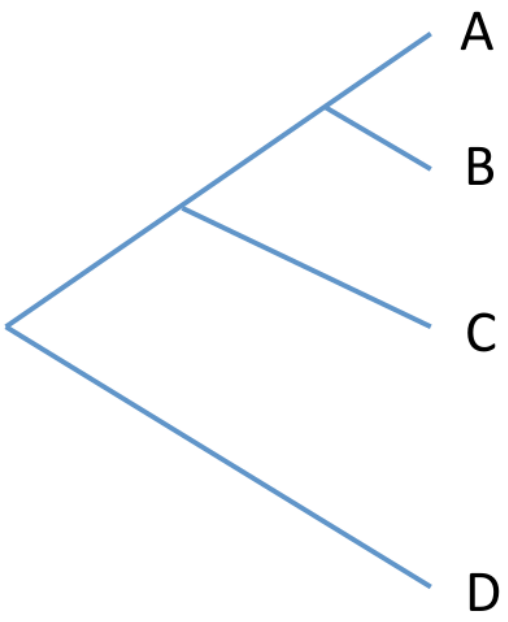
Intended Learning Outcomes

By the end of this session you will be able to

- understand the theory underlying distance and covariance matrices
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring population structure and admixture from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses

Imperial College London

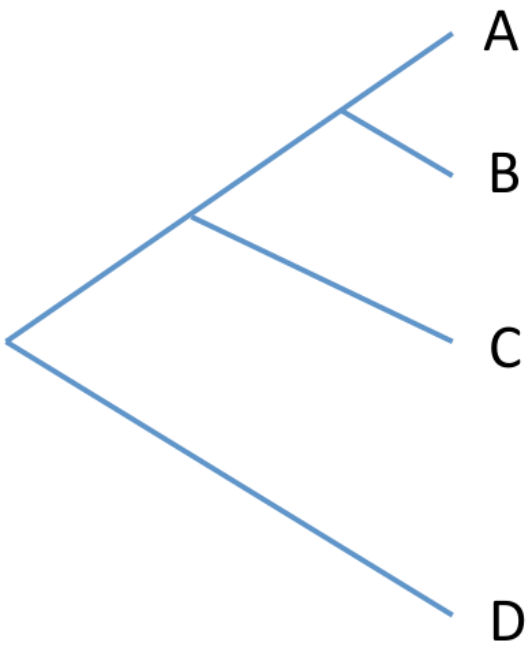
Genetic distances



Genotype 1	Genotype 2	Distance
aa	aa	0
aa	aA	1
aa	AA	2
aA	aa	1
aA	aA	0
aA	AA	2
…	…	…

Imperial College London

Genetic distances



Genotypes are {aa, aA, AA} as {0, 1 ,2}

For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N} \frac{|g(i,s) - g(j,s)|}{2}\right)$$

genotype of *i* at site *s*

e.g. G(i=A,s=1)=0 and G(j=B,s=1)=1 then d(i,j)=1

Imperial College London

Genetic distances from known genotypes

Genotypes are {aa, aA, AA} as {0, 1 ,2}
For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N} \frac{|g(i,s) - g(j,s)|}{2}\right)$$

d(i,j) = 1*1.00 = 1.00/2

B

A

	0	1	2
0	0	1	0
1	0	0	0
2	0	0	0

Expected genotype

Expected value

The expected value of a discrete random variable is the probability-weighted average of all possibles outcomes of the experiment.

It is equivalent to the average value when the experiment is performed many times.
Let's go back to the whiteboard!

Expected genotype

Expected value

The expected value of a discrete random variable is the probability-weighted average of all possibles outcomes of the experiment.

It is equivalent to the average value when the experiment is performed many times.

$$E[X|D] = \sum_{i=1}^{N} x_i p(X = x_i|D)$$

Imperial College London

Genetic distances from (un)known genotypes

Genotypes are {aa, aA, AA} as {0, 1 ,2}
For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

E[d(i,j)] = 0*0.30 + 1*0.50 + 2*0.10 + 1*0.10 + … = 0.80/2

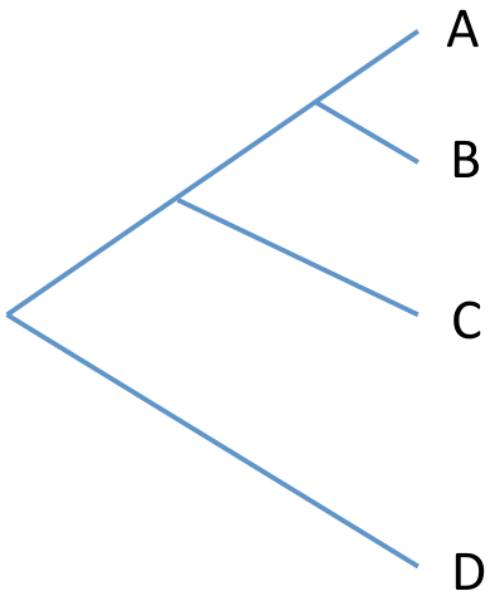
B

A

	0	1	2
0	0.30	0.50	0.10
1	0.10	0	0
2	0	0	0

Imperial College London

Genetic distances from unknown genotypes



Genotypes are {aa, aA, AA} as {0, 1 ,2}

For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N} \frac{|g(i,s) - g(j,s)|}{2}\right)$$

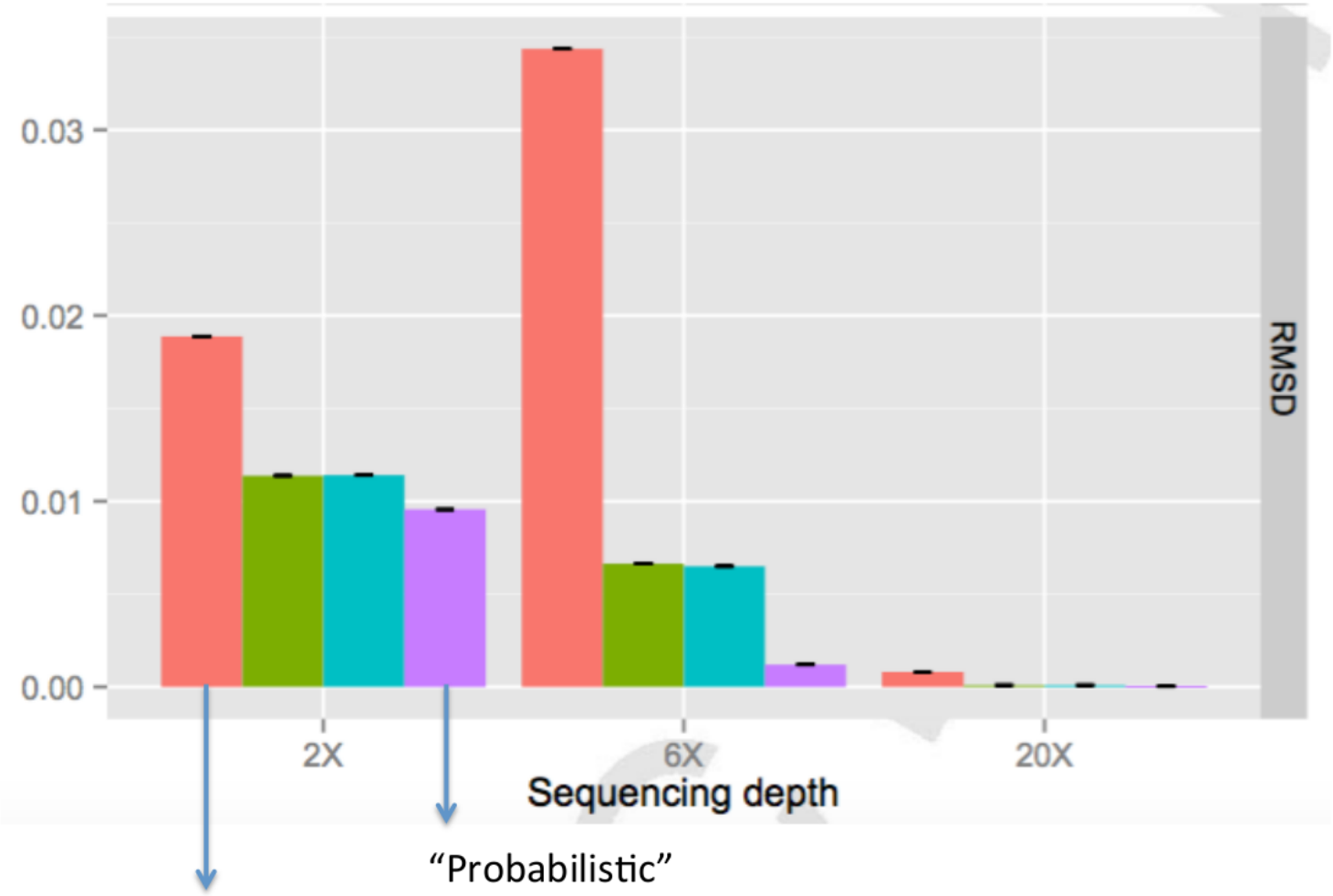
Iterate across all possible genotypes

Genotypes probability

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N} \sum_{g(i,s)=0}^{2} \sum_{g(j,s)=0}^{2} \frac{|g(i,s) - g(j,s)|}{2} * P(g(i,s), g(j,s))\right)$$

Imperial College London

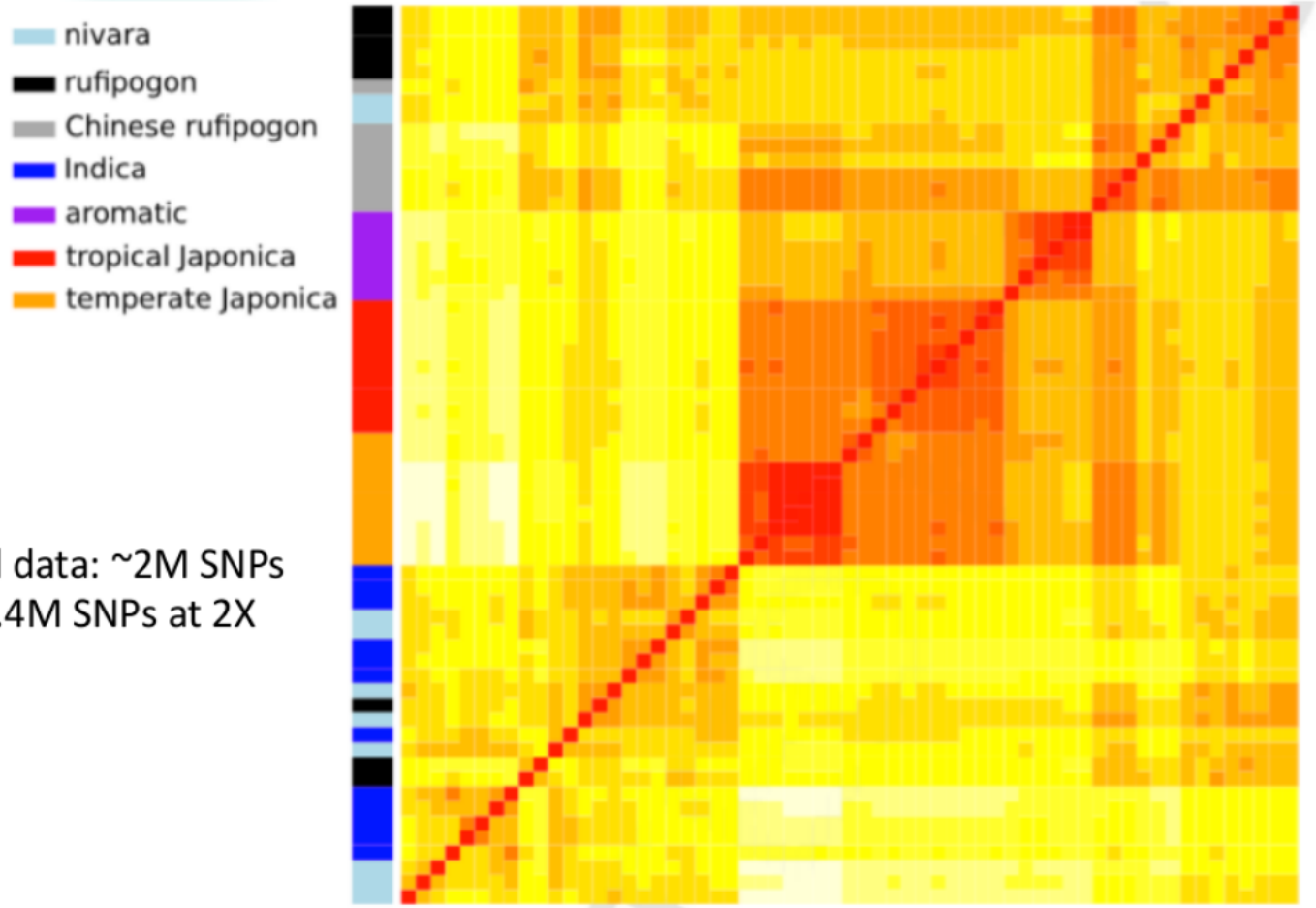
Genetic distances from unknown genotypes



Vieira et al. BJLS 2016

Imperial College London

Clustering from unknown genotypes

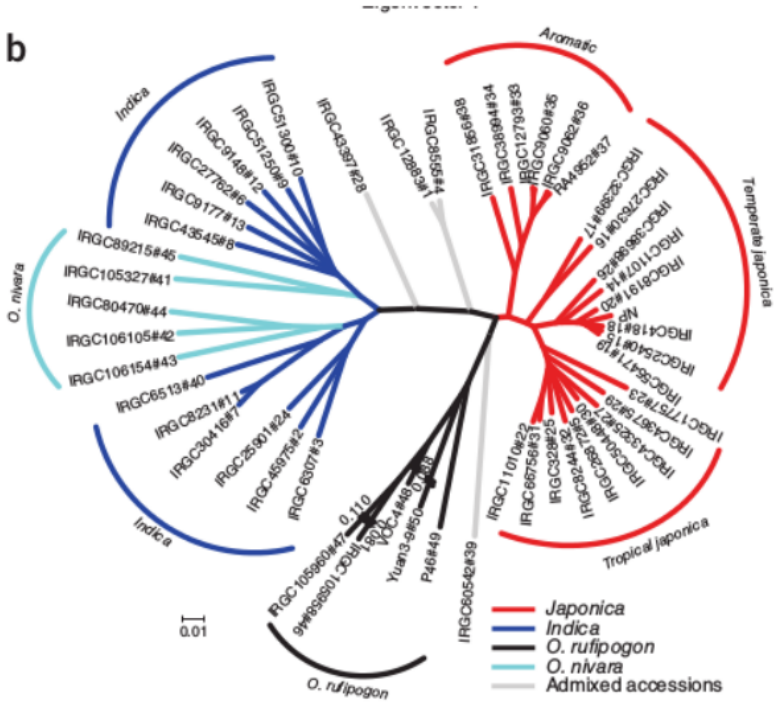


Original data: ~2M SNPs
Here: 5.4M SNPs at 2X

Imperial College London

Population structure

Graphical representation of genetic distance in form of a tree*.

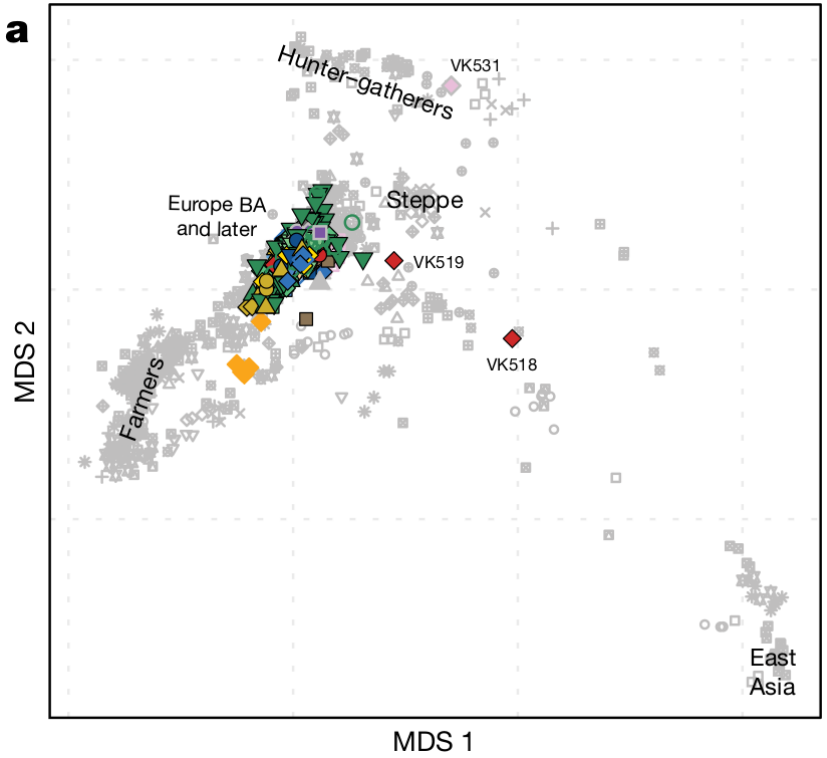


* this should not be considered as a proper phylogenetic tree

Xu et al. Nature Biotech. 2012

Population genomics of the Viking world

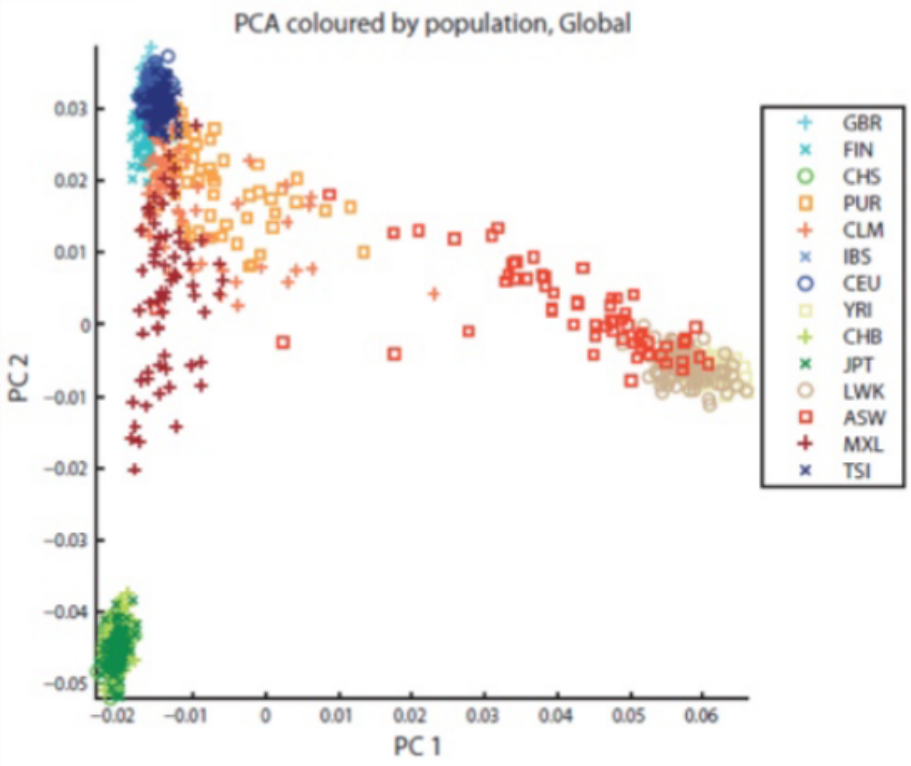
https://doi.org/10.1038/s41586-020-2688-8 Ashot Margaryan[1,2,3,71], Daniel J. Lawson[4,5,71], Martin Sikora[1,71], Fernando Racimo[1,71],



Nature 2020

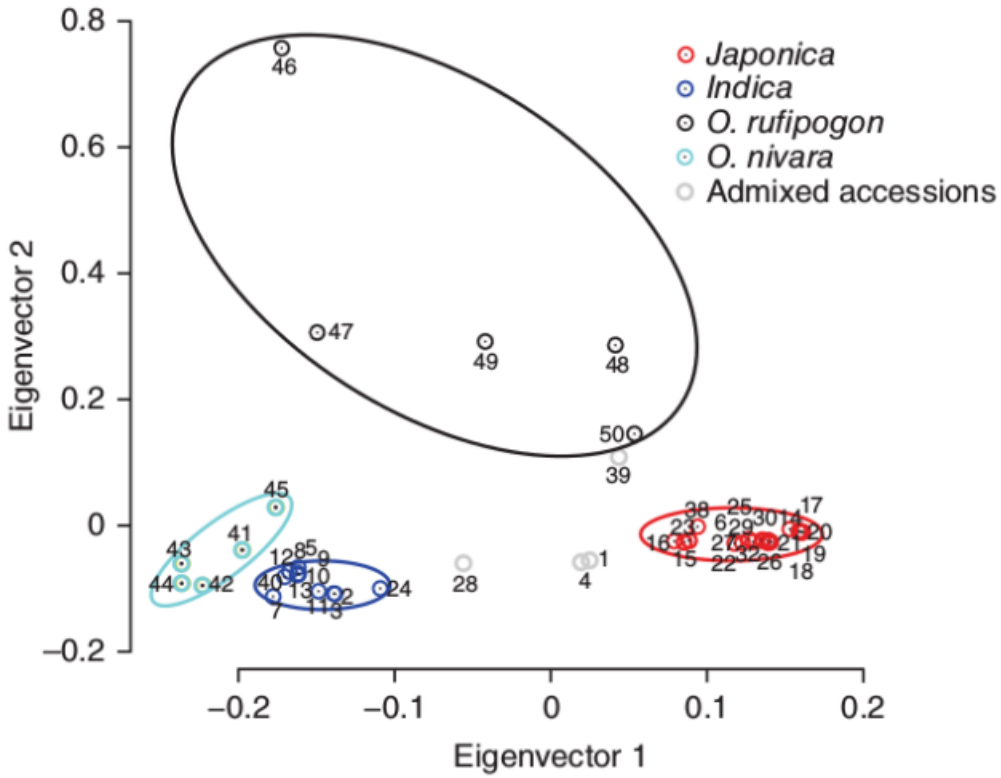
Principal Compoment Analysis (PCA) is a data reduction method for:

- visualisation
- correction for population stratification
- information on population history and differentiation?



- An eigenvector decomposition of the covariance matrix is computed. Eigenvectors are then plotted.

- PCA is a **descriptive** analysis of your dataset, for clustering individuals (and identify population assignment).



p mutations

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$ *n* samples

where X is standardise to mean=0 and sd=1 for each column

V. SOLVING PCA USING EIGENVECTOR DECOMPOSITION

We derive our first algebraic solution to PCA based on an important property of eigenvector decomposition. Once again, the data set is $\mathbf{X}$, an $m \times n$ matrix, where m is the number of measurement types and n is the number of samples. The goal is summarized as follows.

> Find some orthonormal matrix $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$ such that $\mathbf{C_Y} \equiv \frac{1}{n}\mathbf{YY}^T$ is a diagonal matrix. The rows of $\mathbf{P}$ are the *principal components* of $\mathbf{X}$.

$$\mathbf{C_Y} = \mathbf{PC_XP}^T$$

One interpretation of $\mathbf{X}$ is the following. Each *row* of $\mathbf{X}$ corresponds to all measurements of a particular type. Each *column* of $\mathbf{X}$ corresponds to a set of measurements from one particular trial (this is $\vec{X}$ from section 3.1). We now arrive at a definition for the *covariance matrix* $\mathbf{C_X}$.

$$\mathbf{C_X} \equiv \frac{1}{n}\mathbf{XX}^T.$$

https://arxiv.org/pdf/1404.1100.pdf

Imperial College London

Covariance matrix

Genotype (0,1,2)

Allele frequency

$$cov(i,j) = \frac{1}{(m-1)} \frac{\sum_{s=1}^{m} (G_s^{(i)} - 2\hat{p}_s)(G_s^{(j)} - 2\hat{p}_s)}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

Imperial College London

Covariance matrix

Genotype (0,1,2)

Allele frequency

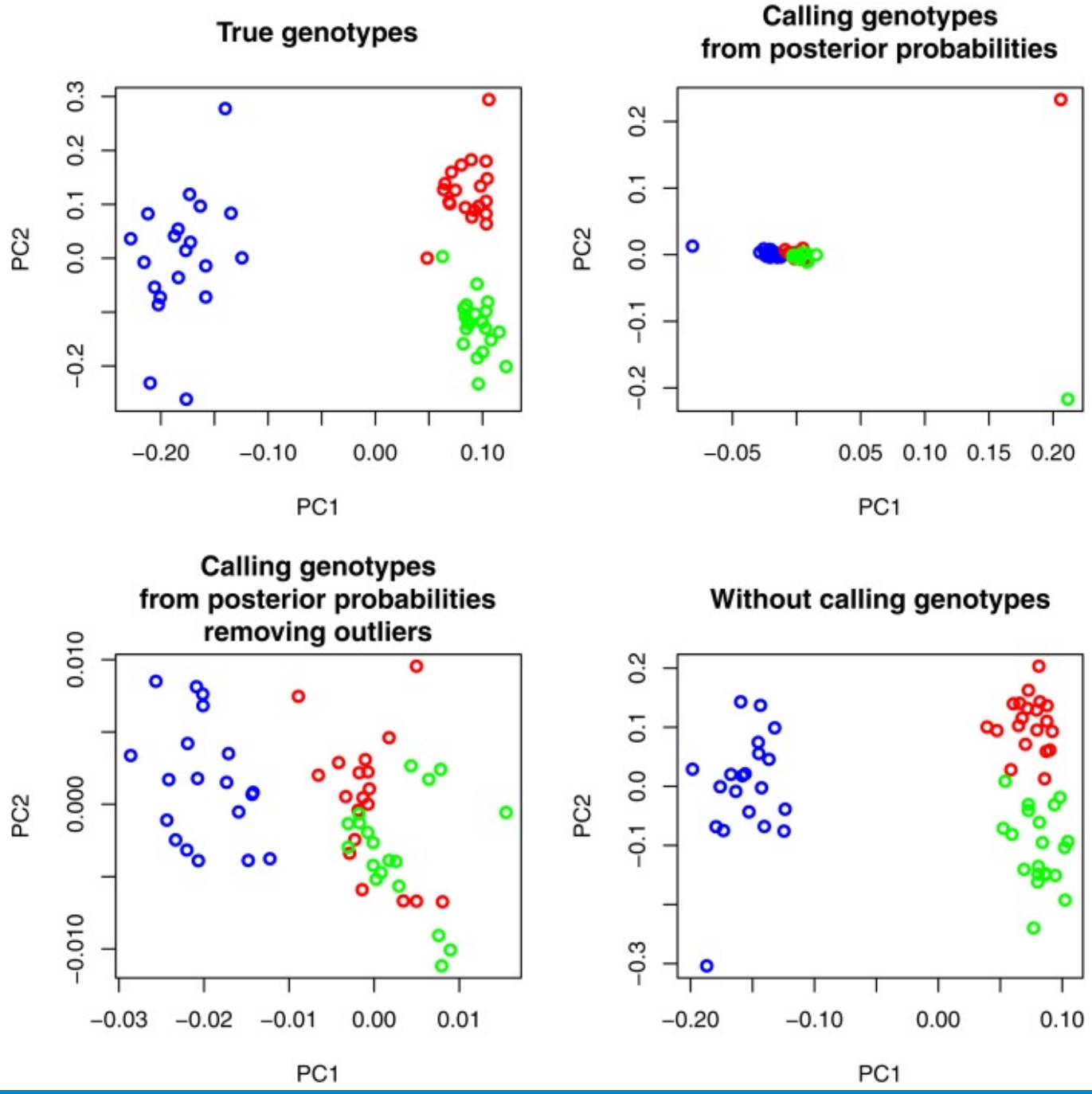
$$cov(i,j)=\frac{1}{(m-1)}\frac{\sum_{s=1}^{m}(G_s^{(i)}-2\hat{p}_s)(G_s^{(j)}-2\hat{p}_s)}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

Iterate across all genotypes

Weight by their probability

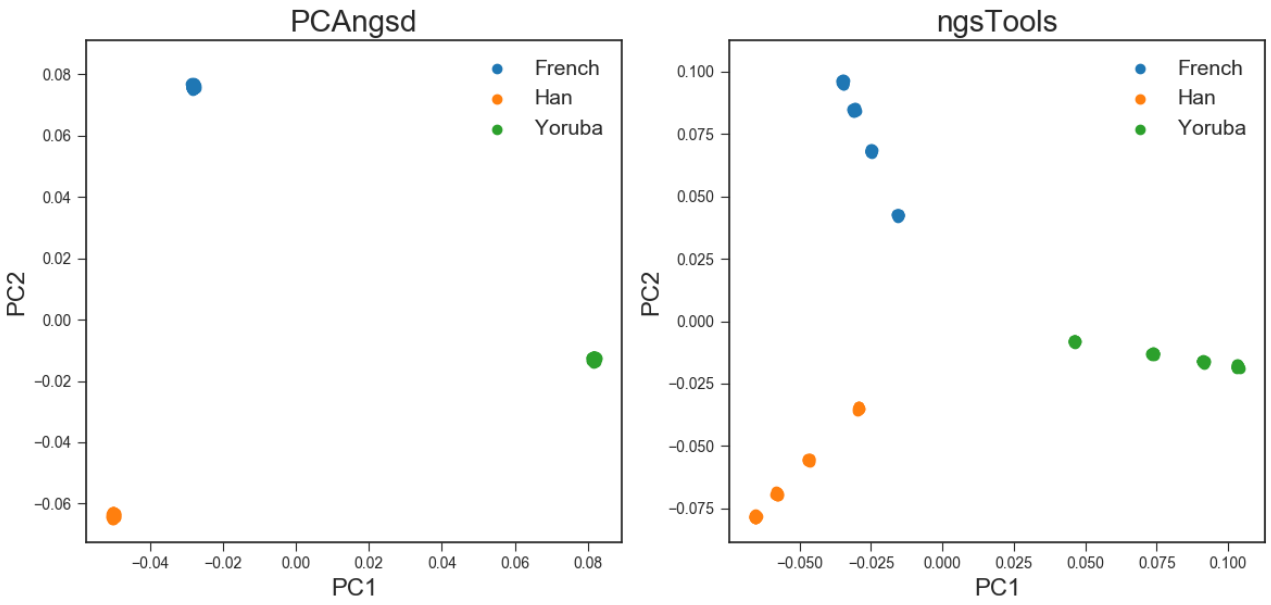
$$\hat{cov}_{(i,j)}:=\frac{1}{(\sum_{s=1}^{m}P_{var,s})-1}\frac{\sum_{s=1}^{m}\left(\sum_{G_s^{(i)}=0}^{2}\sum_{G_s^{(j)}=0}^{2}(G_s^{(i)}-2\hat{p}_s)(G_s^{(j)}-2\hat{p}_s)P(G_s^{(i)}|X_s^{(i)})P(G_s^{(j)}|X_s^{(j)})\right)P_{var,s}}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

Probability of the site being variable (to avoid SNP calling)


True genotypes
PC2
PC1
0.3
0.2
0.1
0.0
−0.2
−0.20
−0.10
0.00
0.10
Calling genotypes
from posterior probabilities
PC2
PC1
0.2
0.1
0.0
−0.1
−0.2
−0.05
0.05
0.10
0.15
0.20
Calling genotypes
from posterior probabilities
removing outliers
PC2
PC1
0.010
0.000
−0.010
−0.03
−0.02
−0.01
0.00
0.01
Without calling genotypes
PC2
PC1
0.2
0.1
0.0
−0.1
−0.3
−0.20
−0.10
0.00
0.10

Imperial College London

???



Jonas Meisner

$$c_{ij} = \frac{1}{m} \sum_{s=1}^{m} \frac{\sum_{g_i=0}^{2} \sum_{g_j=0}^{2} (g_i - 2\hat{p}_s)(g_j - 2\hat{p}_s) P(G_{is} = g_i, G_{js} = g_j \mid X_{is}, X_{js}, \hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)}. \tag{3}$$

ngsTools splits up the joint posterior probability, $P(G_{is}, G_{js} \mid X_{is}, X_{js}, \hat{p}_s)$, into $P(G_{is} \mid X_{is}, \hat{p}_s) P(G_{js} \mid X_{js}, \hat{p}_s)$ for $i \neq j$ by assuming conditional independence between individuals given the estimated population allele frequencies.

$$c_{ij} = \frac{1}{m} \sum_{s=1}^{m} \frac{\sum_{g_i=0}^{2} \sum_{g_j=0}^{2} (g_i - 2\hat{p}_s)(g_j - 2\hat{p}_s) P(G_{is} = g_i, G_{js} = g_j \mid X_{is}, X_{js}, \hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)}. \tag{3}$$

ngsTools splits up the joint posterior probability, $P(G_{is}, G_{js} \mid X_{is}, X_{js}, \hat{p}_s)$, into $P(G_{is} \mid X_{is}, \hat{p}_s) P(G_{js} \mid X_{js}, \hat{p}_s)$ for $i \neq j$ by assuming conditional independence between individuals given the estimated population allele frequencies.

The problem with this approach is that the assumption of conditional independence between individuals given the population allele frequency is only valid when there is no population structure. Here we propose a novel approach of estimating the covariance matrix using iteratively estimated individual allele frequencies to update the prior information of the posterior genotype probability. Thereby we condition on the individual allele frequencies as in the clustering-based approaches such as Pritchard *et al.* (2000); Tang *et al.* (2005); Alexander *et al.* (2009); Skotte *et al.* (2013).

Meisner and Albechtsen, Genetics, 2018

Introduction to PCAngsd

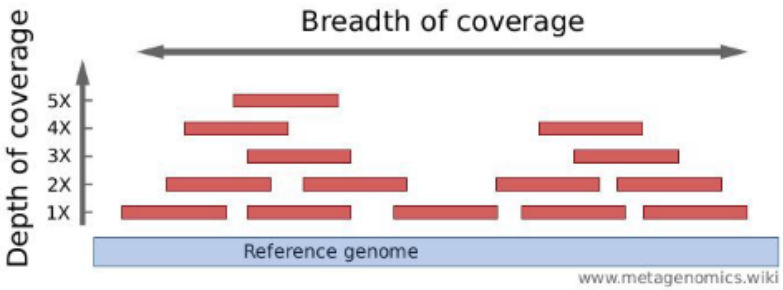
- Principal Component Analysis of Next-Generation Sequencing Data (PCAngsd)
- Multithreaded Python framework
- Infers population structure using PCA
- Structured populations
- Low and medium sequencing depth
- Published in Meisner and Albechtsen, Genetics, 2018

- DNA data
  - *n* **diploid** samples
  - *m* SNPs

- Diallelic ($G = \{0, 1, 2\}$)
- Genotype matrix

$$\#\text{ of SNPs}\left\{\begin{array}{cccccccccc} 0 & 1 & 2 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 2 & 0 & 0 & \cdots & 1 & 0 & 0 & 2 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 2 & 0 & \cdots & 2 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 2 & \cdots & 0 & 2 & 0 & 0 & \cdots & 0 \end{array}\right.$$

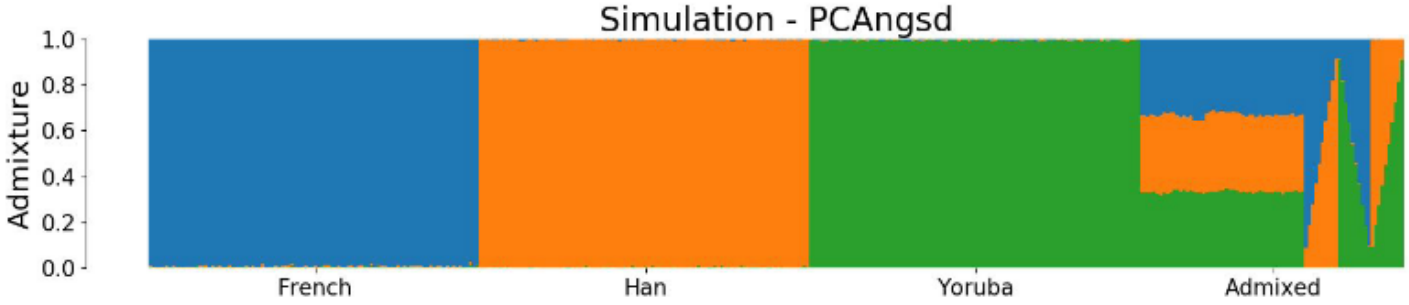
- Sequencing data
  - Low (<5X) and medium (<15X) sequencing depth
  - Genotype uncertainty



- Genotype likelihoods
  - Retain information of sequencing process
  - Take uncertainty into account

$$P(X_{is} \mid G = g)$$

- Population allele frequencies
  - Average in discrete populations



- PCA
  - Dimension reduction method
  - Construct axes of genetic variation
- Individual allele frequencies
  - Infer underlying sampling parameter

$$g_{is} \sim \text{Binomial}(2, \pi_{is})$$

- Posterior genotype probability

$$P(G = g \mid X_{is}, p_s) = \frac{P(X_{is} \mid G = g) P(G = g \mid p_s)}{\sum_{g'=0}^{2} P(X_{is} \mid G = g') P(G = g' \mid p_s)}$$

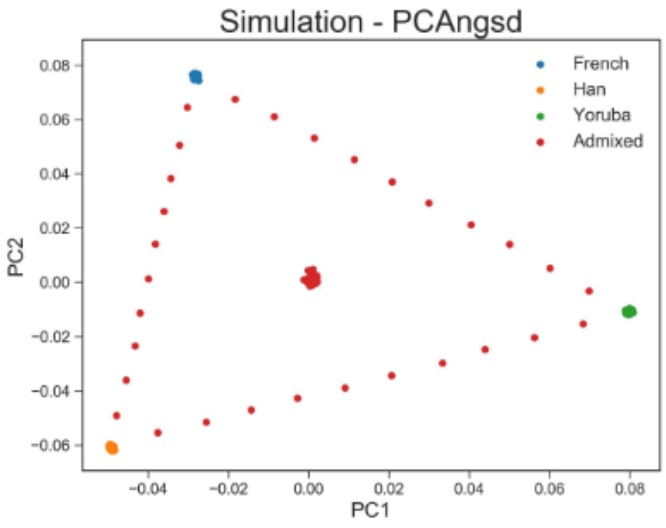
- Low-rank SVD reconstruction

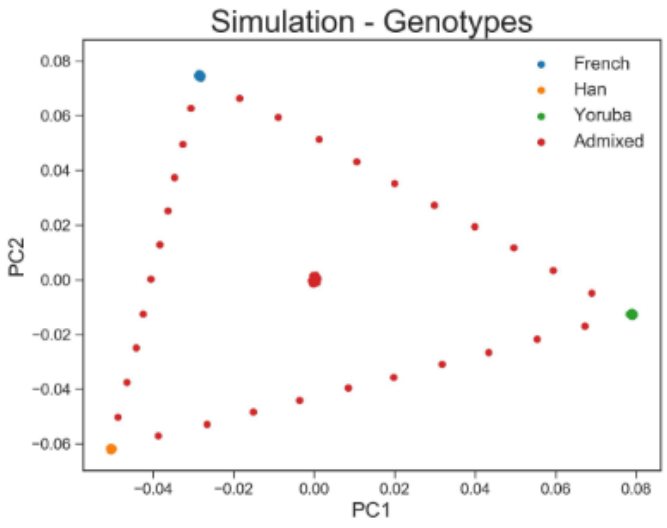
$$\mathbb{E}[\mathbf{G} \mid \mathbf{X}] \approx \mathbf{U}_{1:D} \boldsymbol{\Delta}_{1:D} \mathbf{V}_{1:D}$$

$$\pi_{is} = \frac{1}{2} \mathbf{U}_{[i,1:D]} \, \boldsymbol{\Delta}_{1:D} \, \mathbf{V}_{[s,1:D]}$$

- Update prior information and iterate!

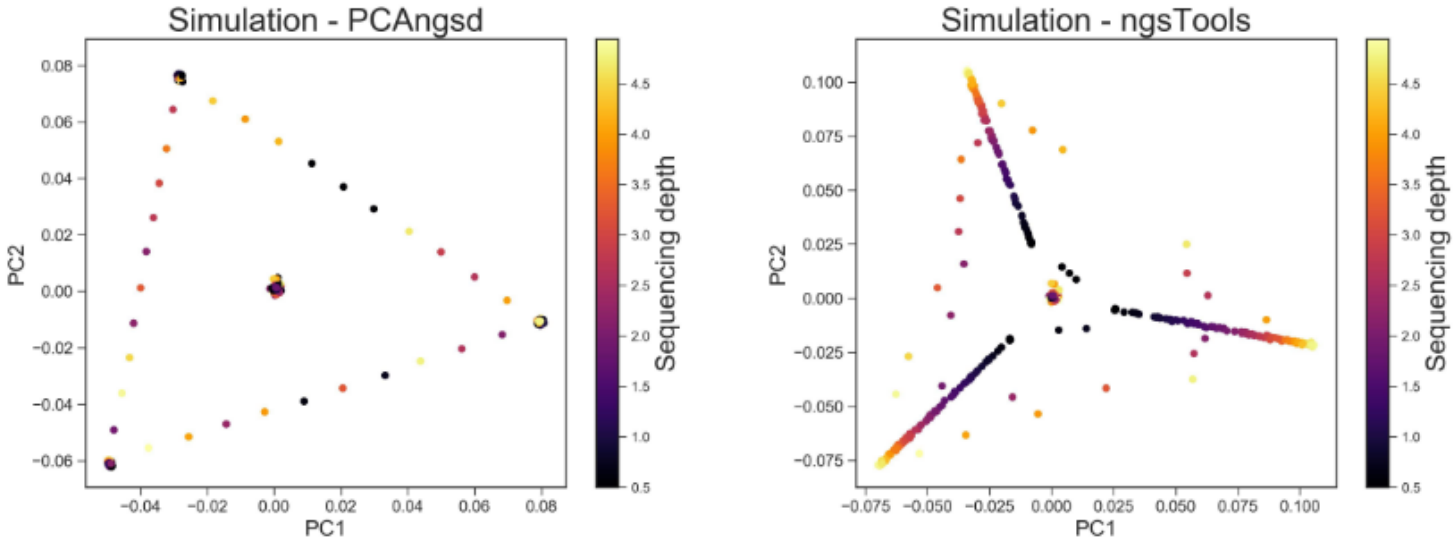
Inferring population structure (**0.5 - 5X**)





Jonas Meisner

Existing methods are biased by sequencing depth!



Jonas Meisner

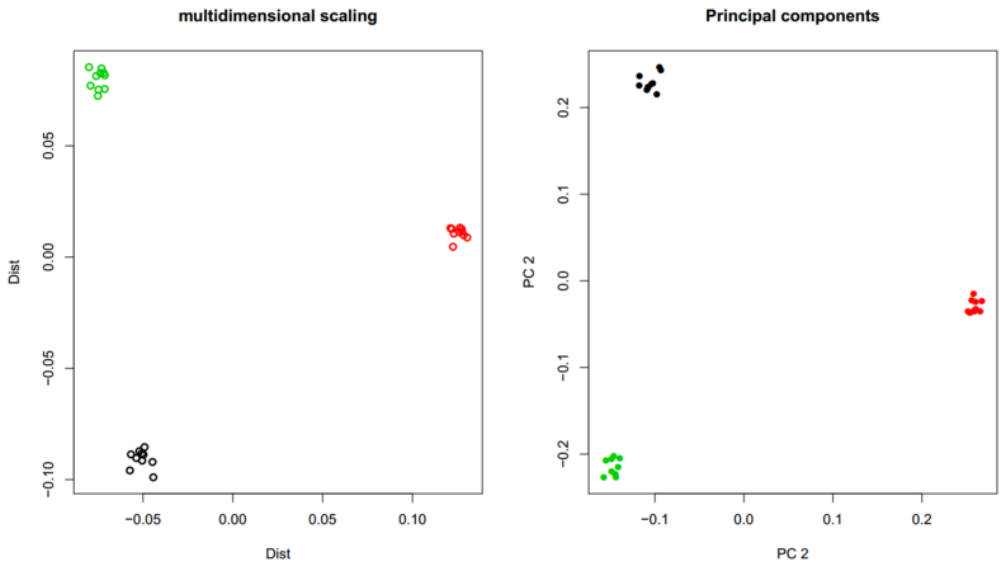
PCAngsd is the current state-of-the-art for PCA from low-depth data

Imperial College London

PCA/MDS

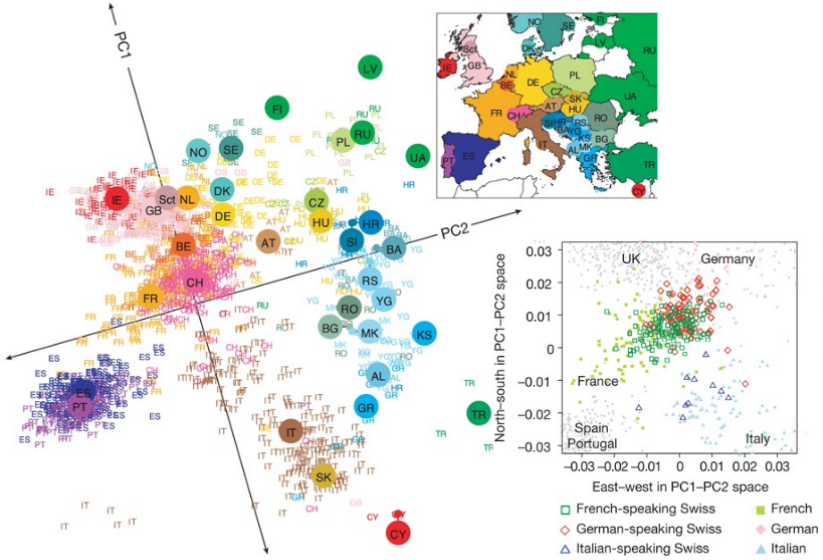
with random sampling of read for IBS (for MDS) or covariance (for PCA) matrix.

It works even with very low depth data (¡1X), but it requires low error rate and known polymorphic sites.



Imperial College London

Beyond PCA?



Novembre et al. 2008

PCA represents genetic drift as allele frequency differences between populations.

There is a good theory behind it but possibly still ongoing issues on interpretability and quantification of admixture/gene flow.

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

AA: observed genotype

Pr(AA | pop=1) = ?

Pr(AA | pop=2) = ?

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

AA: observed genotype

Pr(AA | pop=1) = ?

Pr(AA | pop=2) = ?

Assuming HWE

P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

$Pr(AA \mid pop=1) = f_{A.1}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.90$

$Pr(AA \mid pop=2) = f_{A.2}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.10$

Assuming HWE
P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

Pr(AG | pop=1) = ?

Pr(AG | pop=2) = ?

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

$Pr(AA \mid pop=1) = f_{A.1}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.90$

$Pr(AA \mid pop=2) = f_{A.2}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.10$

$Pr(AG \mid pop=1) = 2*f_{A.1}*(1-f_{A.1}) / ... = 0.60$

$Pr(AG \mid pop=2) = ... = 0.40$

Assuming HWE
P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

Bayesian (STRUCTURE) or Maximum Likelihood (ADMIXTURE) approaches.

Genomes of n samples, p mutations stored in $n \times p$ matrix **X**

x_{ij} contains the number of non-reference mutations for individual i at position ("SNP") j, so that $x_{ij} \in \{0,1,2\}$

Assume there are $k = 1 \ldots c$ ancestral populations. The frequency μ_{jk} of mutation j in population k is unknown

Goal: infer the fraction of ancestry α_{ik} for individual i and all values of k, subject to $\sum_{k=1}^{K} \alpha_{ik} = 1$

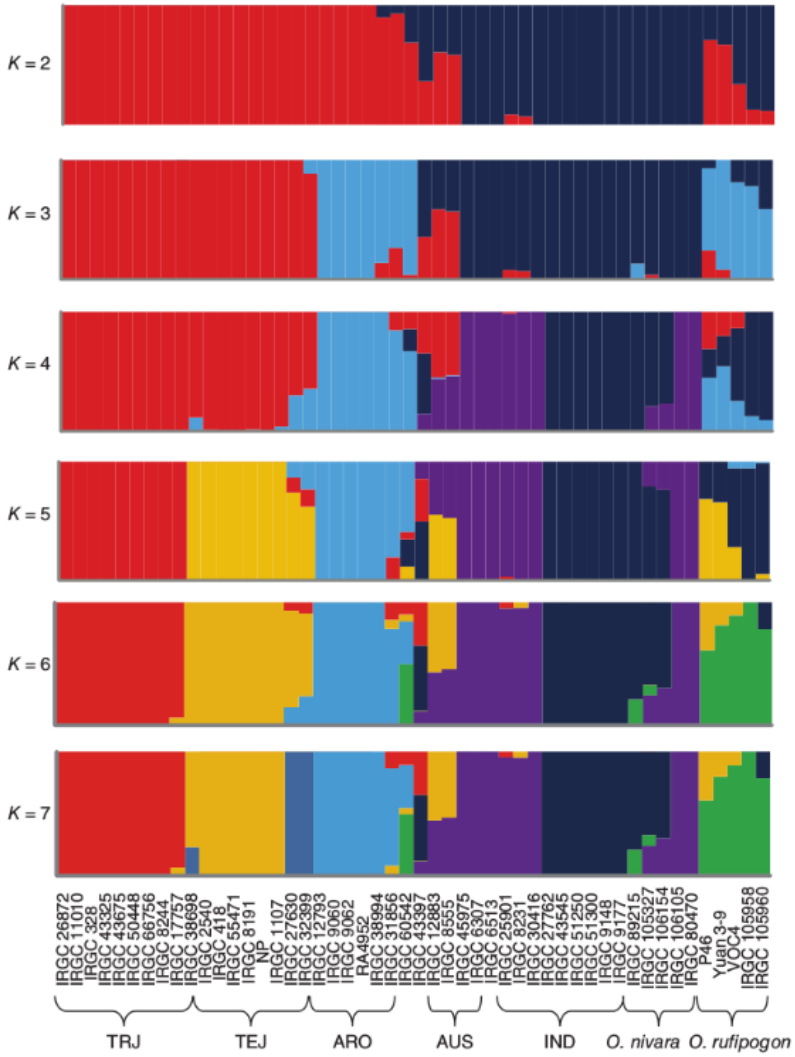
We need to find α_{ik} and μ_{jk} that maximize P(DATA | α_{ik}, μ_{jk} $\forall$ i, j, k), which is proportional to

$$\prod_{i=1}^{n} \prod_{j=1}^{p} \left[\sum_{k=1}^{c} \alpha_{ik}\, \mu_{jk}\right]^{x_{ij}} \left[\sum_{k=1}^{c} \alpha_{ik}\, (1 - \mu_{jk})\right]^{2-x_{ij}}$$

STRUCTURE: Pritchard et al. Genetics, 2000
LDA: Blei et al. JMLR 2003, LDA notes

Adaptation of example by A. Price

Admixture analyses



Xu et al. Nature Biotech. 2012

Imperial College London

NGSadmix

When the genotypes are observed, assuming that sites are independent, the likelihood is written as

$$p(G|Q,F) = \prod_{j=1}^{M}\prod_{i=1}^{N} p(G_{ij}|Q,F) = \prod_{j=1}^{M}\prod_{i=1}^{N} p\left(G_{ij}|h^{ij}\right). \tag{5}$$

If the sites are not independent, then this is a composite likelihood that will still have consistent estimates. This likelihood corresponds to the likelihood used in Tang *et al.* (2005) and Alexander *et al.* (2009) and will be used when dealing with called genotypes.

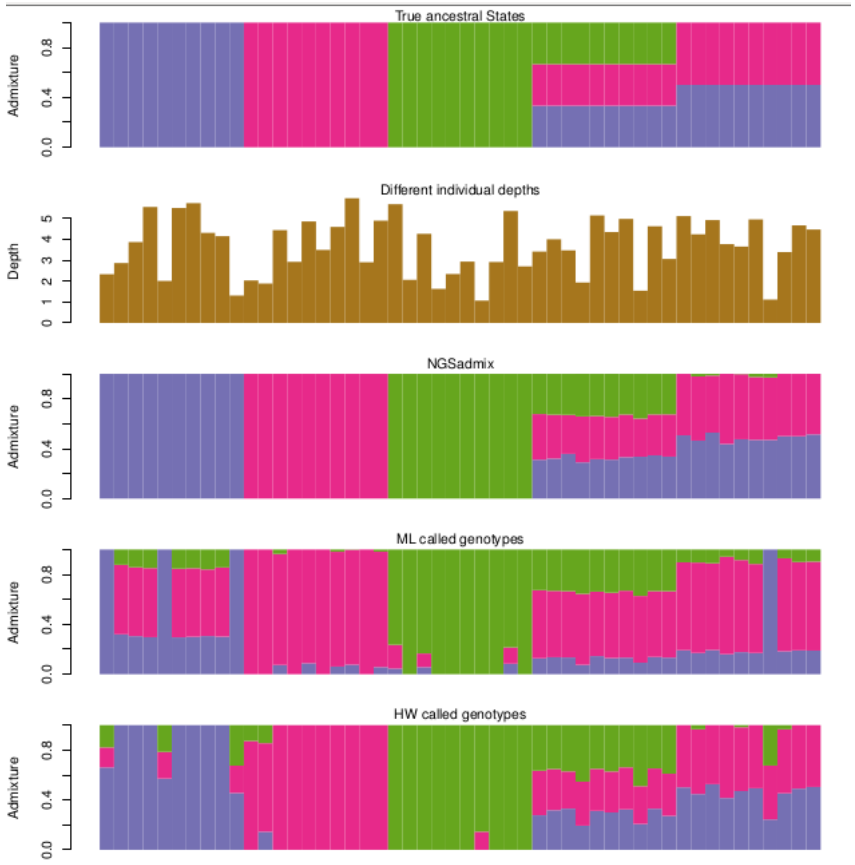
When using NGS data, the genotypes are not observed and we instead work with genotype likelihoods. The above likelihood is extended by summing over all possible genotypes:

$$\begin{aligned} p(X|Q,F) &= \prod_{j=1}^{M}\prod_{i=1}^{N} p(X_{ij}|Q,F) = \prod_{j=1}^{M}\prod_{i=1}^{N} p\left(X_{ij}|h^{ij}\right) \\ &= \prod_{j=1}^{M}\prod_{i=1}^{N} \sum_{G_{ij}\in\{0,1,2\}} p(X_{ij}|G_{ij})\, p\left(G_{ij}|h^{ij}\right). \end{aligned} \tag{6}$$

Skotte et al. Genetics 2013

Imperial College London

NGSadmix



in practice, PCAngsd extends its framework to admixture analysis

Intended Learning Outcomes

At the end of this session you are be able to

- understand the theory underlying distance and covariance matrices
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring population structure and admixture from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses