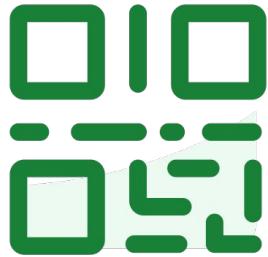


# Population structure II

## PCA

Anders Albrechtsen

**Do not edit**  
How to change the design

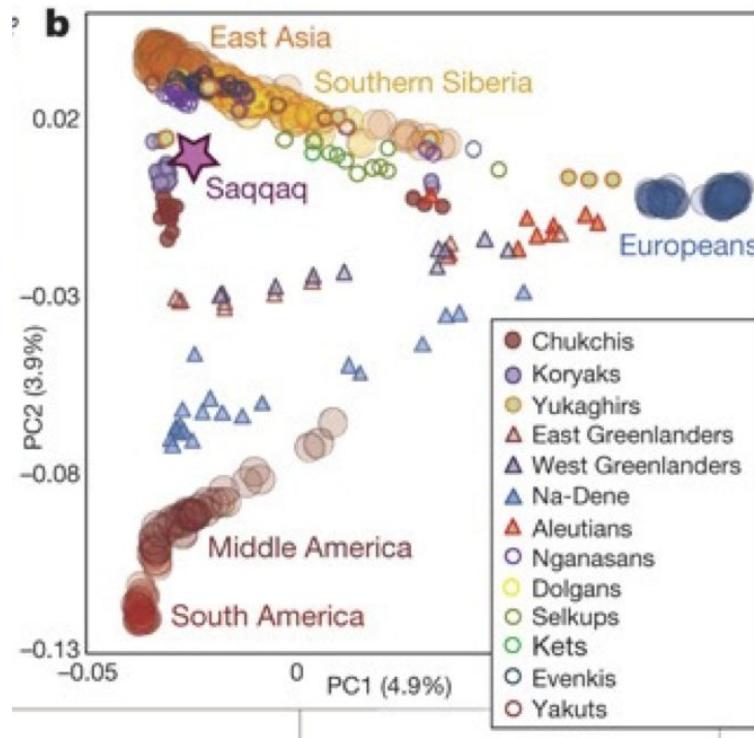


**Join at [slido.com](https://www.slido.com)  
#3351809**

- ⓘ Presenting with animations, GIFs or speaker notes? Enable our [Chrome extension](#)

**slido**

# PCA for population structure



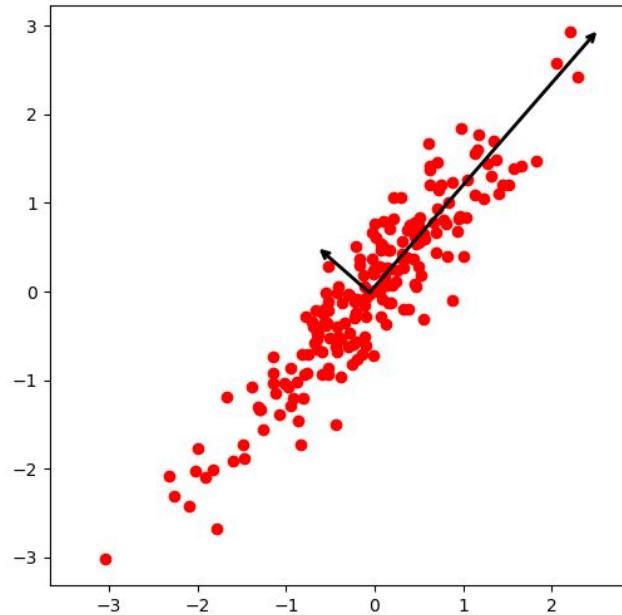


**what can we infer from the plot  
about the Saqqaq?**

- Who Saqqaq close to?
- Who is admixed?

# Principal component analysis (PCA)

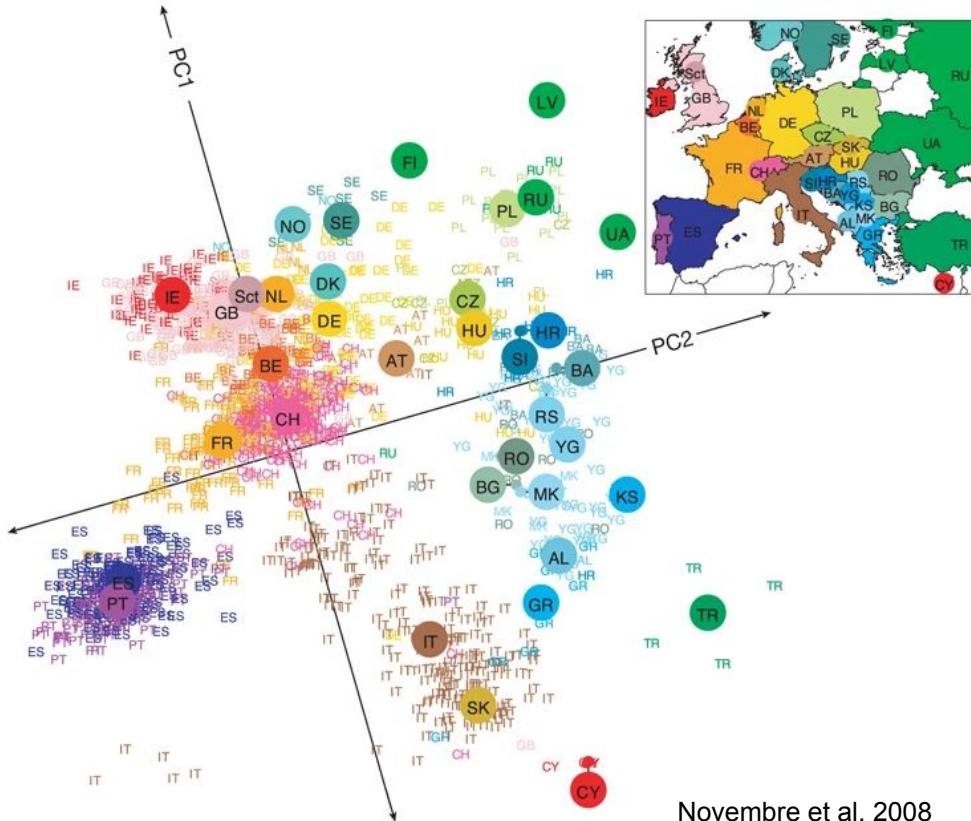
- Dimensionality reduction
- Axis of variation
- Principal components
- Models more ***Continuous\**** population structure than ADMIXTURE



**\*not really true**

# Principal component analysis (PCA)

- Genetic data
- $m > 1$  million
- Captures genetic structure



Novembre et al. 2008

# Today you will learn

- The underlying “model” of PCA and MDS
  - What these two methods are trying to achieve
- The relationship between admixture proportions and PCA
- How PCA predict genotypes
- Issues with missingness
  - For call genotypes and for low depth sequencing
- How to deal with missingness
- How PCA can be used for selection scan ( teaser )

# Genotype data

	Ind1	Ind2	Ind3	Ind4	Ind5
SNP1	AG	AG	AG	AA	AA
SNP2	TT	TA	AA	AT	AA
SNP3	AA	AC	AC	CC	AC
SNP4	GG	GG	GC	CC	CC
SNP5	TT	TC	TC	CC	CC
SNP6	AA	AA	AC	AC	AC
SNP7	TT	TT	TC	TC	CC



SNPs

Ind	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	0	0	
0	1	2	1	2	
2	1	1	0	1	
0	0	1	2	2	
2	1	1	0	0	
0	0	1	1	1	
2	2	1	1	0	

Genotype

left eigenvectors

singular values

right eigenvectors

**G**

$m \times n$

**U**

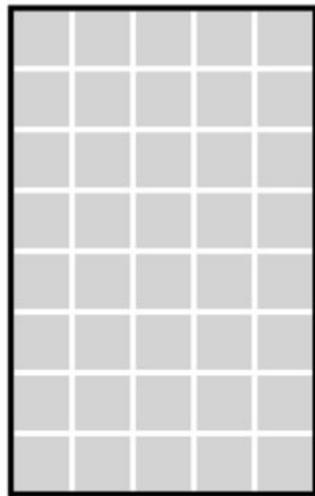
$m \times n$

**$\Sigma$**

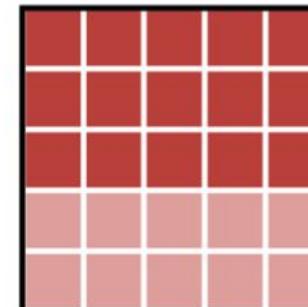
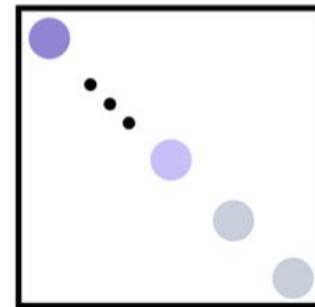
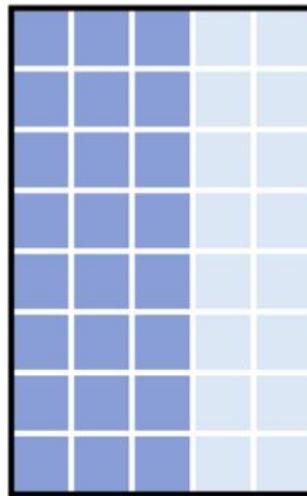
$n \times n$

**$V^T$**

$n \times n$



=

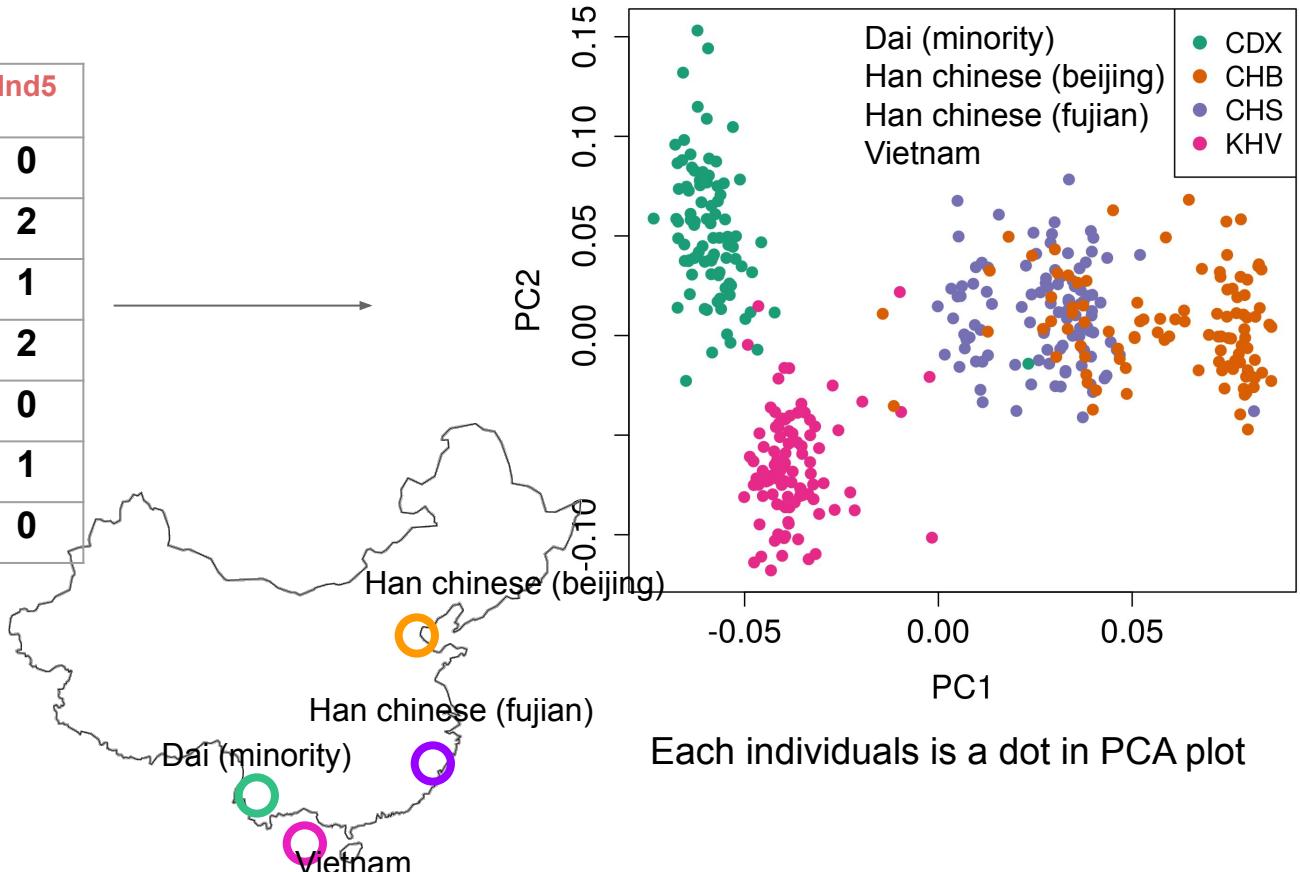


→ PC1  
→ PC2  
→ PC3

**G** is a genotype matrix,  $n$  is the number of samples,  $m$  is the number of SNPs

# PCA of 1000G Asian Populations

SNPs	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	1	0	0
0	1	1	2	1	2
2	1	1	1	0	1
0	0	1	1	2	2
2	1	1	1	0	0
0	0	1	1	1	1
2	2	1	1	1	0



# Multi-dimensional scaling (MDS)

Goal: Project the data into a low dimensional space that preserves distances

- Choose a distance
- Choose a dimension (K)

# Multi-dimensional scaling

SNPs

	Ind <sub>1</sub>	Ind <sub>2</sub>	Ind <sub>3</sub>	Ind <sub>4</sub>	Ind <sub>5</sub>
Ind <sub>1</sub>	1	1	1	0	0
Ind <sub>2</sub>	0	1	2	1	2
Ind <sub>3</sub>	2	1	1	0	1
Ind <sub>4</sub>	0	0	1	2	2
Ind <sub>5</sub>	2	1	1	0	0
Ind <sub>1</sub>	0	0	1	1	1
Ind <sub>2</sub>	2	2	1	1	0

Pairwise distance

Manhattan distance

	Ind <sub>1</sub>	Ind <sub>2</sub>	Ind <sub>3</sub>	Ind <sub>4</sub>	Ind <sub>5</sub>
Ind <sub>1</sub>	0	3	7	10	11
Ind <sub>2</sub>	3	0	4	7	8
Ind <sub>3</sub>	7	4	0	5	4
Ind <sub>4</sub>	10	7	5	0	6
Ind <sub>5</sub>	11	8	4	3	0

Project into 1 dimension

	Ind1	Ind2	Ind3	Ind4	Ind5
Dim 1	6.1	3.08	-0.62	-3.7	-4.85



# What is the distance

# Multi-dimensional scaling

Manhattan distance

	Ind <sub>1</sub>	Ind <sub>2</sub>	Ind <sub>3</sub>	Ind <sub>4</sub>	Ind <sub>5</sub>
Ind <sub>1</sub>	0	3	7	10	11
Ind <sub>2</sub>	3	0	4	7	8
Ind <sub>3</sub>	7	4	0	5	4
Ind <sub>4</sub>	10	7	5	0	6
Ind <sub>5</sub>	11	8	4	3	0

Project into 1 dimension

	Ind1	Ind2	Ind3	Ind4	Ind5
Dim <sub>1</sub>	6.1	3.08	-0.62	-3.7	-4.85

# Two ways to do PCA

$$\tilde{G} \underset{\text{M x N}}{=} \underset{\text{M x N}}{U} \underset{\text{N x N}}{S} \underset{\text{N x N}}{V^T}$$

The diagram illustrates the decomposition of a genotype matrix  $\tilde{G}$  into three components:  $U$ ,  $S$ , and  $V^T$ . The matrix  $\tilde{G}$  is  $M \times N$ . The matrix  $U$  is  $M \times N$  and has a blue-to-white gradient. The matrix  $S$  is  $N \times N$  and contains five points forming a diagonal line. The matrix  $V^T$  is  $N \times N$  and has a red-to-white gradient. To the right of  $V^T$ , three arrows point to labels: PC1, PC2, and PC3.

Directly on the  
genotypes

$$C \underset{\text{N x N}}{=} \underset{\text{N x N}}{V} \underset{\text{N x N}}{S^2} \underset{\text{N x N}}{V^T}$$

The diagram illustrates the decomposition of a covariance matrix  $C$  into three components:  $V$ ,  $S^2$ , and  $V^T$ . The matrix  $C$  is  $N \times N$  and has a grey grid pattern. The matrix  $V$  is  $N \times N$  and has a blue-to-white gradient. The matrix  $S^2$  is  $N \times N$  and contains five points forming a diagonal line. The matrix  $V^T$  is  $N \times N$  and has a red-to-white gradient. To the right of  $V^T$ , three arrows point to labels: PC1, PC2, and PC3.

On the covariance  
matrix

# Principal component analysis

Goal: Project the data into a low dimensional space that explains the largest amount of variance

- Choose a dimension (K)

# Principal component analysis

SNPs

Ind	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	1	0	0
0	1	2	1	1	2
2	1	1	0	0	1
0	0	1	2	2	2
2	1	1	0	0	0
0	0	1	1	1	1
2	2	1	1	1	0

Calculate Covariance



Covariance matrix

	Ind1	Ind2	Ind3	Ind4	Ind5
Ind1	12.6	5.6	-2.0	-7.0	-9.1
Ind2	5.6	4.7	-0.8	-3.7	-5.8
Ind3	-2.0	-0.8	2.3	-0.8	1.3
Ind4	-7.0	-3.7	-0.8	6.7	4.7
Ind5	-9.1	-5.8	1.3	4.7	8.9

Project into 1 dimension

	Ind1	Ind2	Ind3	Ind4	Ind5
Dim 1	0.65	0.36	-0.08	-0.4	-0.53



# Genotype covariance matrix

$M$  number of sites

$G$  genotype

$G_i$  genotype for individual i

$G_{ij}$  genotype for individual i site j

$f_j$  frequency for site j

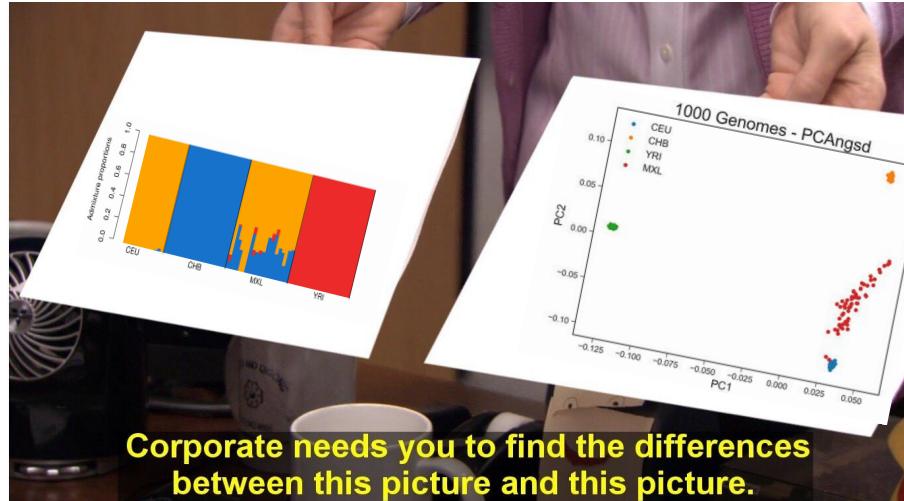
$$\tilde{G}_{ij} = \frac{G_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}$$

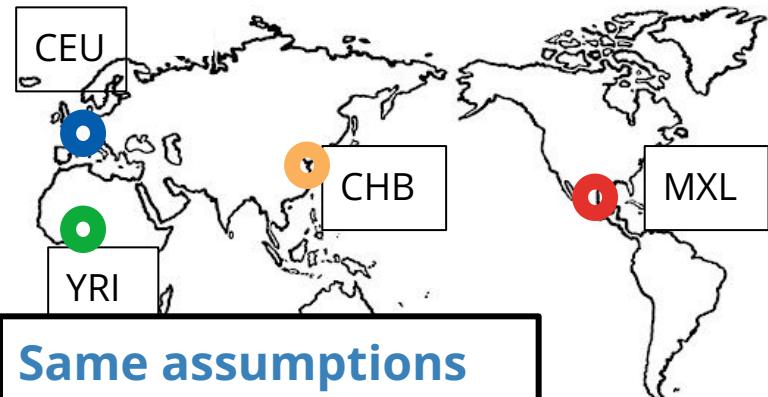
$$var(G_{ij}) = 2f_j(1 - f_j)$$

After normalization  
all SNPs have the  
same mean and  
variance

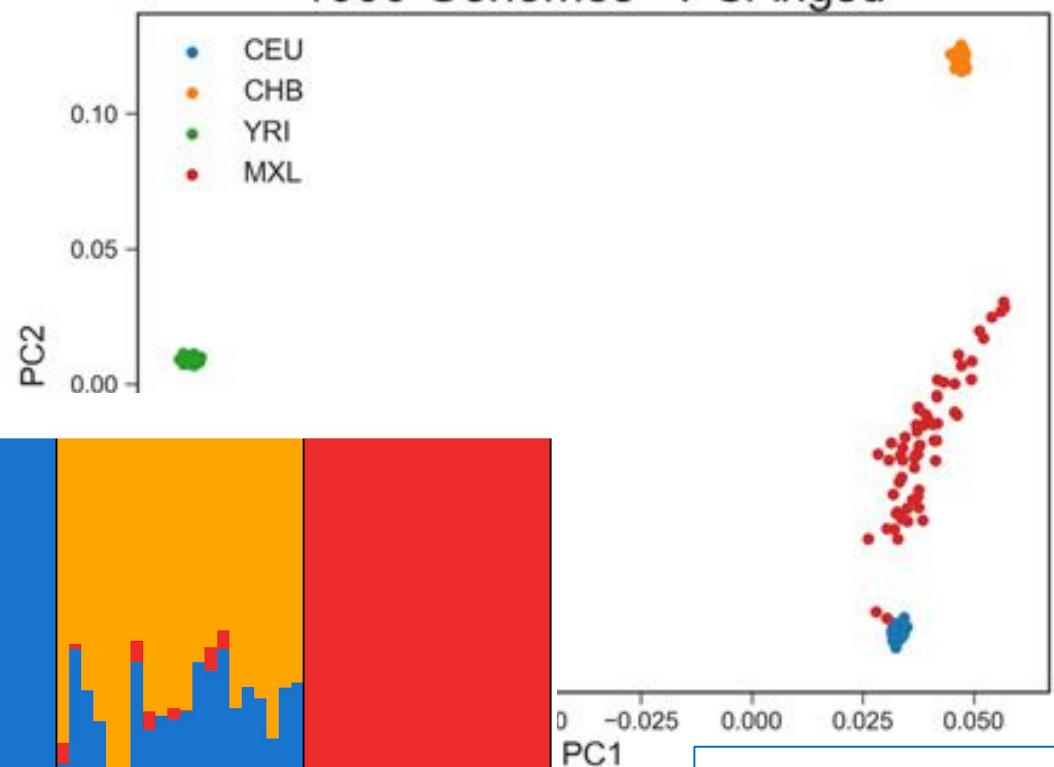
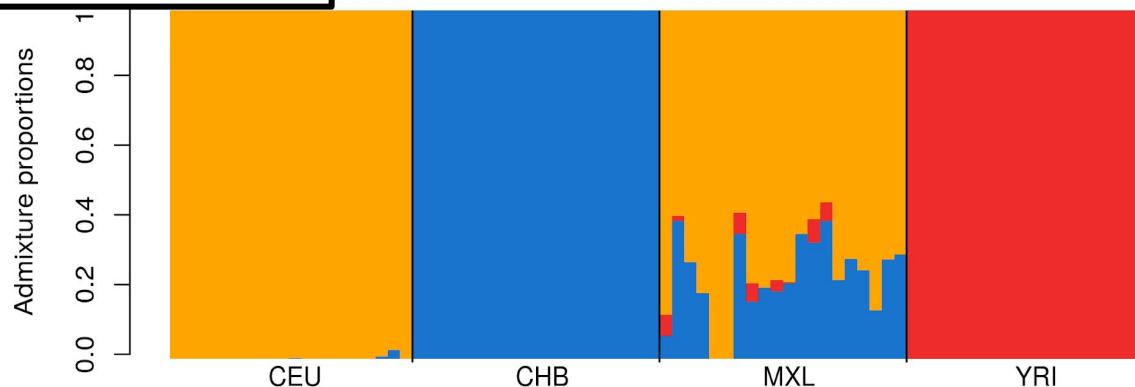
$$cov(\tilde{G}_i, \tilde{G}_l) = \frac{1}{M} \sum_{j=1}^M \frac{(G_{ij} - 2f_j)(G_{lj} - 2f_j)}{2f_j(1-f_j)} = \frac{1}{M} \tilde{G} \tilde{G}^T$$

# Connection between Admixture analysis and PCA





Same assumptions  
=  
same issues



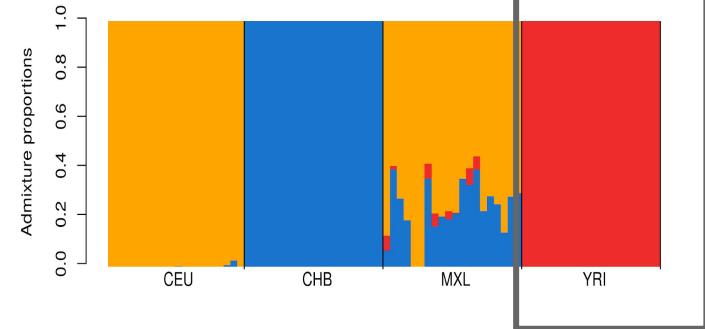
EvalAdmix

# Individual allele frequencies

- Population allele frequency:  $\mathbb{E}[g] = 2p$
- Individual allele frequency:  $\mathbb{E}[g_i] = 2\pi_i = 2p$

$$p(g) = \begin{cases} p^2 & g = 0 \\ 2p(1 - p) & g = 1 \\ (1 - p)^2 & g = 2 \end{cases}$$

**No admixture**  
 $p$ = frequency  
in YRI



# Individual allele frequencies

- Individual allele frequencies

$$\mathbb{E}[g_i] = 2$$

$$\pi_{ij} = \sum_{k=1}^K f_{jk} q_{ik}$$

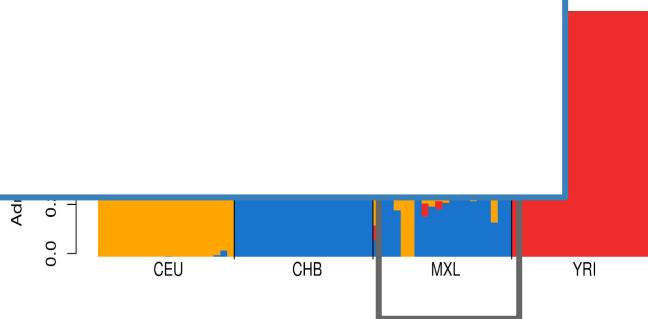
$K$  = #populations

$i$  = individual i

$$p(g_i) =$$

$j$  = site j

$\Pi = QF$



# Individual allele frequencies from PCA

$$\begin{matrix} \mathbf{G} \\ M \times N \end{matrix} = \begin{matrix} \mathbf{U} \\ M \times N \end{matrix} \begin{matrix} \mathbf{S} \\ N \times N \end{matrix} \begin{matrix} \mathbf{V}^T \\ N \times N \end{matrix}$$

→ PC1  
→ PC2  
→ PC3

Directly on the genotypes

$$\begin{matrix} 2\Pi \\ M \times N \end{matrix} = \begin{matrix} \mathbf{U}_{1:k} \\ M \times K \end{matrix} \begin{matrix} \mathbf{S}_{1:k} \\ K \times K \end{matrix} \begin{matrix} \mathbf{V}_{1:k}^T \\ K \times N \end{matrix}$$

→ PC1  
→ PC2  
→ PC3

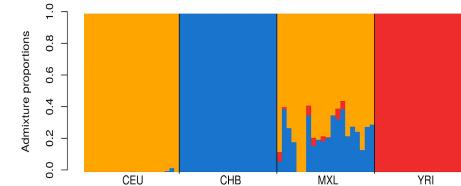
Directly on the genotypes

# Individual allele frequencies

- Allele frequency:
- Low-rank approximation

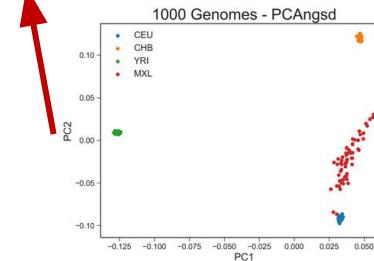
- Admixture
- PCA
- Truncated SVD

$$\mathbb{E}[g] = 2p$$



$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{Q}\mathbf{F}$$

$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$



# Admixture and PCA from $\Pi$

$\Pi$  is the matrix of individual frequencies

ADMXTURE  $\rightarrow$  PCA

$$\begin{aligned} & cov(\tilde{G}_i, \tilde{G}_l) \\ & \approx \frac{1}{M} \sum_{j=1}^M \frac{(\Pi_{ij} - f_j)(\Pi_{lj} - f_j)}{f_j(1-f_j)} \\ & \approx \frac{1}{M} \tilde{G} \tilde{G}^T \end{aligned}$$

PCA  $\rightarrow$  ADMIXTURE

$$argmin_{Q,F} \|\Pi - QF\|$$

Solved with NMF

# Issues with PCA: missingness

Most genotype data sets has missingness and all low depth sequencing data has missingness everywhere

Medium depth sequencing



Ultra low depth sequencing



# Dealing with missingness (most software)

If a genotype is missing then  $\tilde{G}_k^i$  is set to zero

- $E[\tilde{G}_k^i] = 0$  for a random individual
- $E[\text{cov}(G^i, G^j)] = 0$  i.e. relatedness or population structure.

or a site is discarded

- Not possible for large samples
- Will likely cause bias

## Projection:

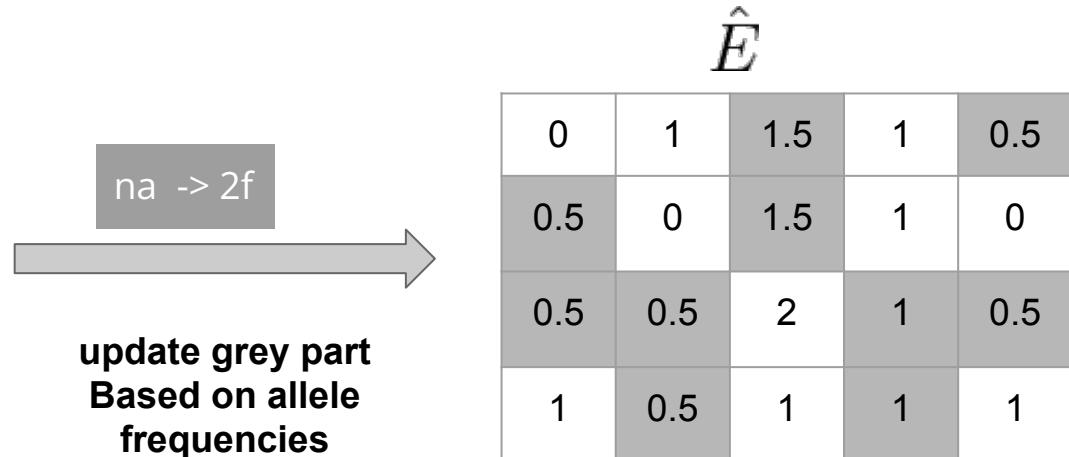
You can make a PCA with your good data and then project individuals with lots of missingness on top

# Mean imputation

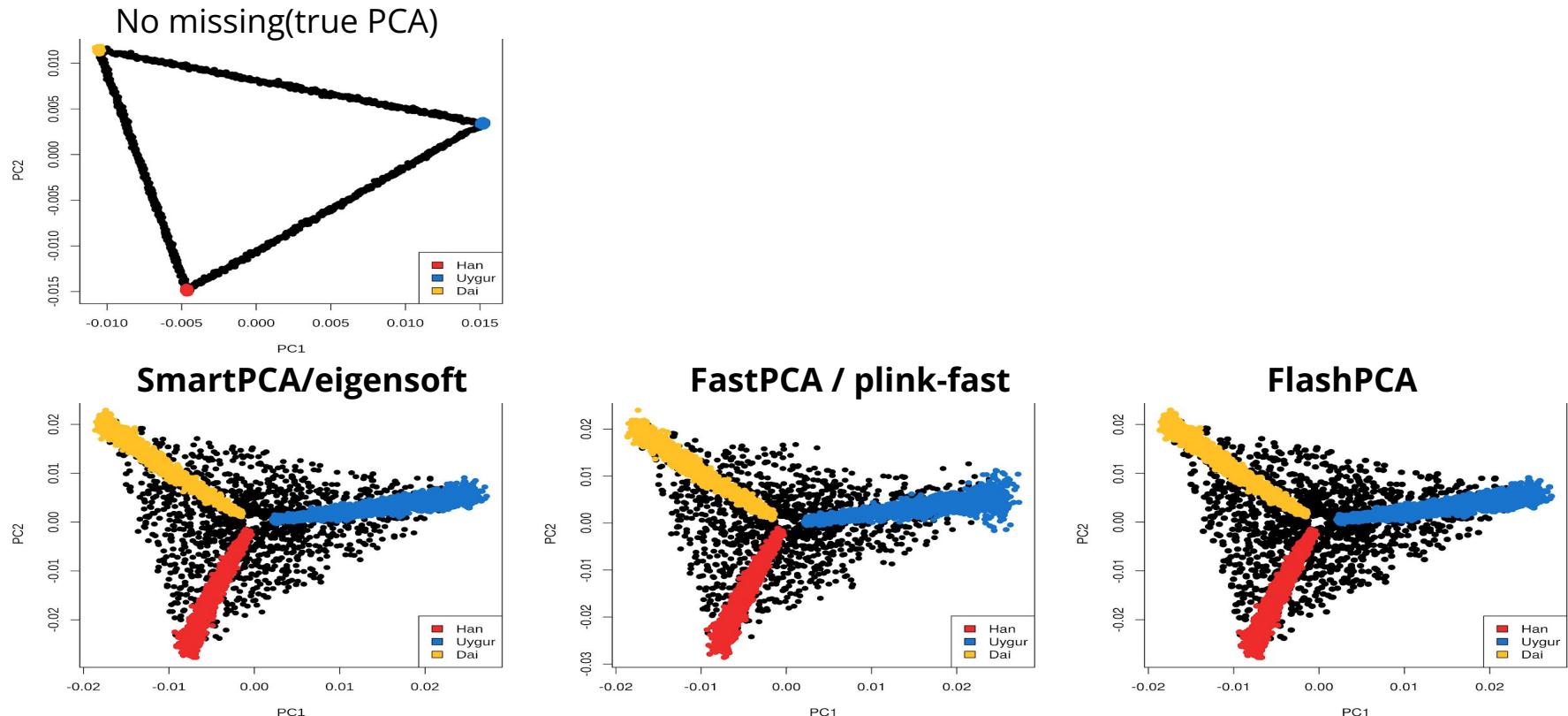
How almost all software deal with missingness

SmartPCA / eigensoft, plink2 -fast, fastPCA, flashPCA ect.  
(exceptions are: plink2, proPCA, emu, pcangsd)

	<b>G</b>				
Ind1	0	1	na	1	na
Ind2	na	0	na	1	0
Ind3	na	na	2	na	na
Ind4	1	na	1	na	1
	0.25	0.25	0.75	0.5	0.25
	Allele frequency (f)				



# Simulated 90-99% missingness





# which individuals are mapped close to the origin?

# How to solve it - EMU

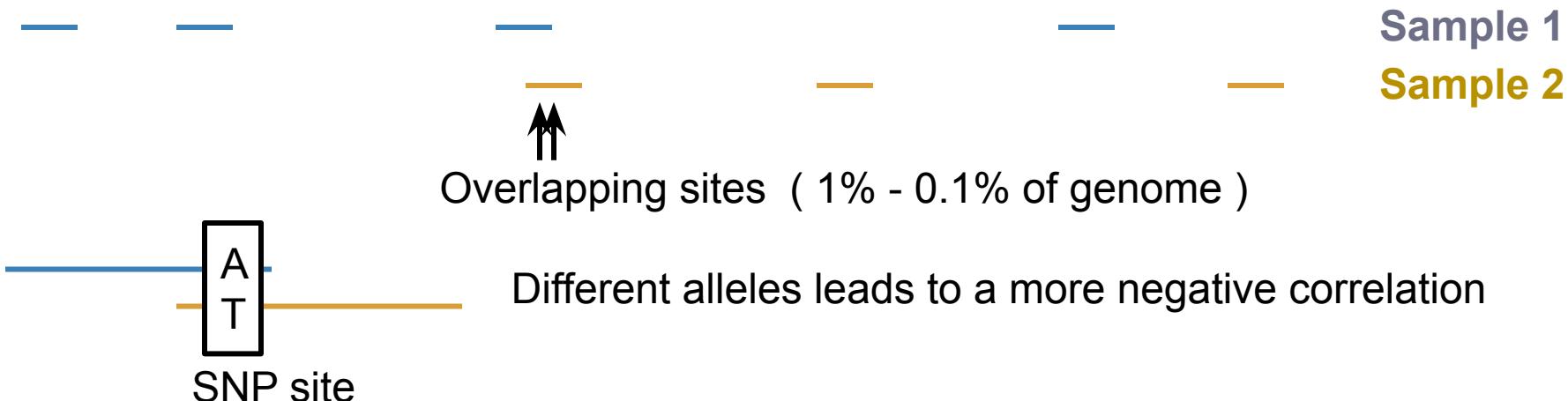
If we have the PCA we can calculate the individual allele frequencies

$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$

If we have the individual allele frequencies we can calculate the PCA

# FIRST APPROACH ( e.g. plink)

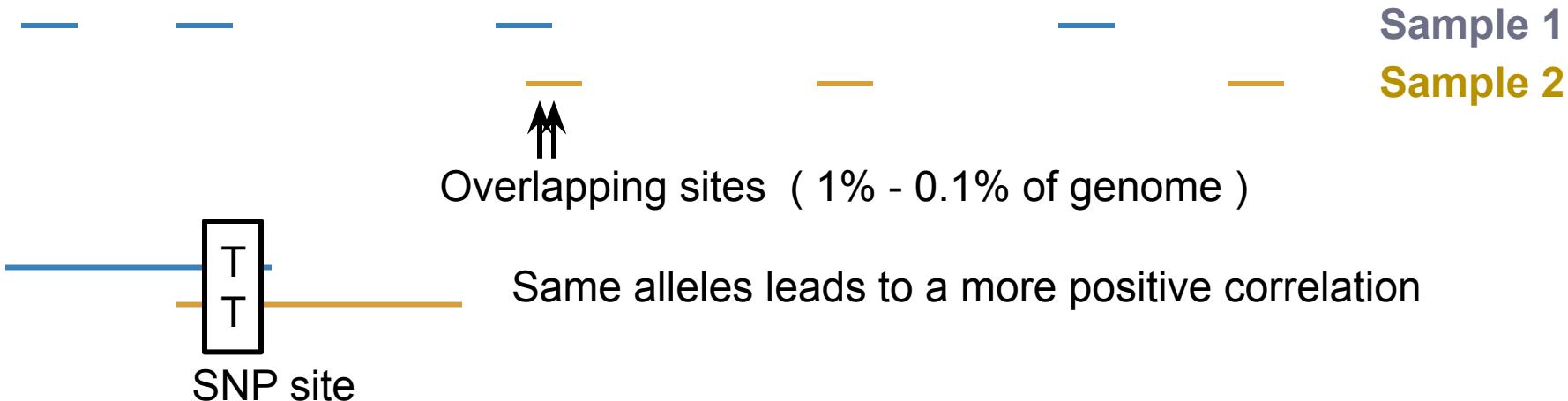
- 1) Estimate allele frequencies for each SNP
- 2) For each pair of individual estimate the covariance for overlapping sites



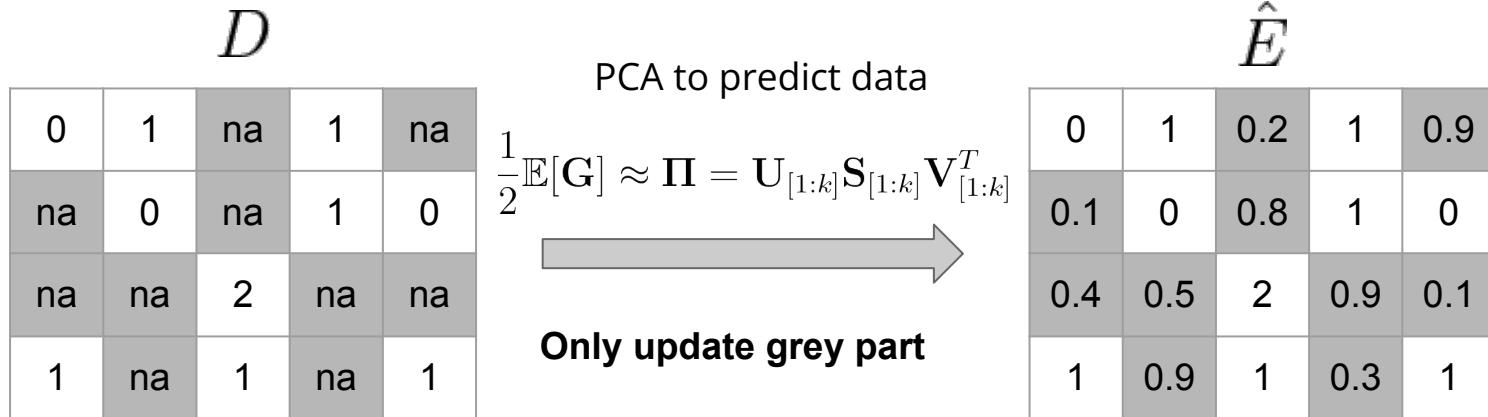
# FIRST APPROACH

$$\text{cov}(G^i, G^j) = \frac{1}{M} \sum_{k=1}^M \frac{(G_k^i - 2f_k)(G_k^j - 2f_k)}{2f_k(1 - f_k)}$$

- 1) Estimate allele frequencies for each SNP
- 2) For each pair of individual estimate the covariance for overlapping sites



# Second approach - EMU



$E_0$ 

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

 $E_1$ 

0	1	0.2	1	0.9
0.1	0	0.8	1	0
0.4	0.5	2	0.9	0.1
1	0.9	1	0.3	1

$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$

Only update grey part

Repeat until convergence :  $\sqrt{\text{mean}(U_K^{n+1} - U_K^n)^2} < 5e^{-7}$

$\hat{E}$ 

0	1	0.2	1	0.9
0.1	0	0.8	1	0
0.4	0.5	2	0.9	0.1
1	0.9	1	0.3	1

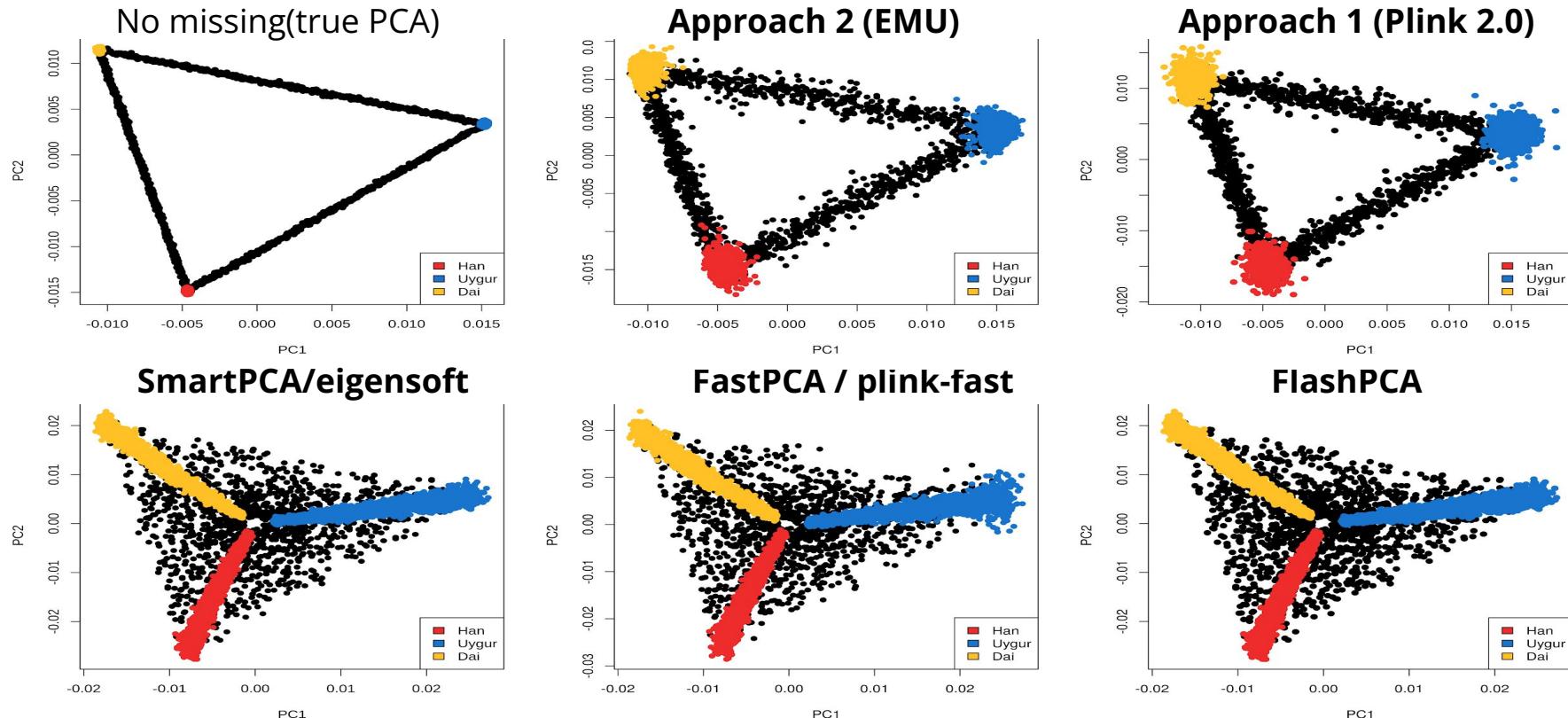
$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$

 $\hat{\boldsymbol{\Pi}}$ 

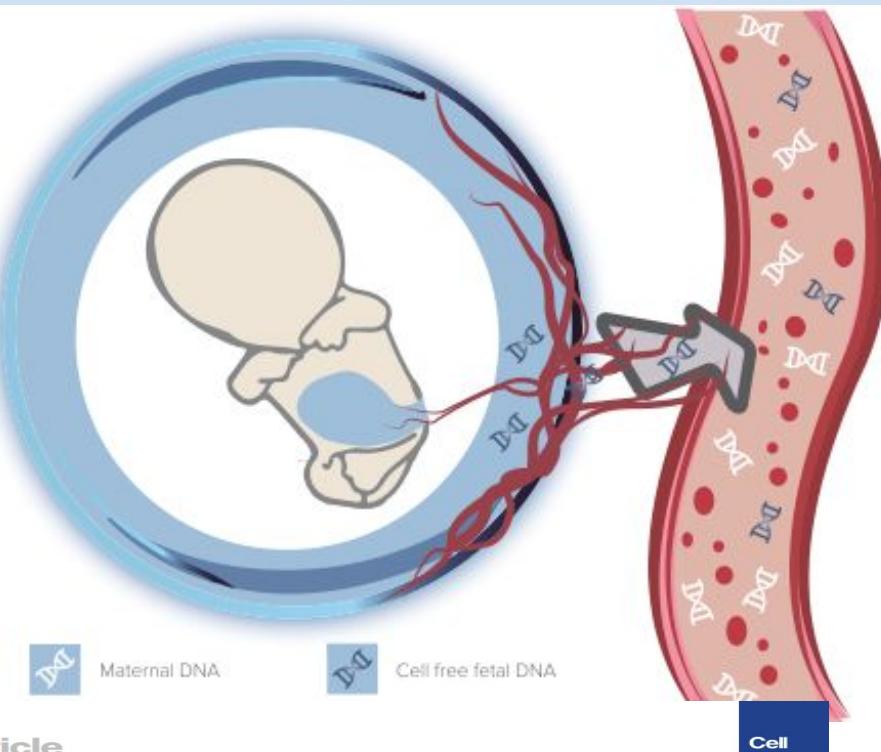
0.1	0.9	0.2	0.9	0.4
0.3	0.1	0.9	1.0	0.1
0.1	0.2	0.1	0.8	0.6
0.1	0.1	0.9	0.8	1.0

$$\min_{\boldsymbol{\Pi}} \left\| \mathbf{C} \odot (\mathbf{D} - \boldsymbol{\Pi}) \right\|_F^2$$

# Simulated 90-99% missingness



# THE NON-INVASIVE FETAL TRISOMY TEST



Blood contains some fetus DNA

Does the fetus have large chromosomal anomalies?

E.g. Trisomy21 (Downs syndrome)

## Article

### Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Siyang Liu,<sup>1,25</sup> Shujia Huang,<sup>1,3,25</sup> Fang Chen,<sup>1,23,24,25</sup> Lijian Zhao,<sup>1,25</sup> Yuying Yuan,<sup>1,25</sup> Stephen Starko F

# THE PILOT PROJECT

## Participants

N=141,431

Recruited in 2012-2013

Written informed  
consent

## Self-reported ethnicity\*

~66,000 Han

~2000 minorities

\*Consent based

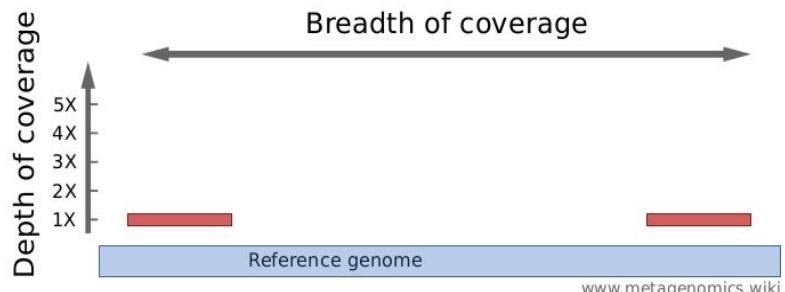
## Sequencing

Illumina Hiseq 2000

49bp reads (~20,000)

35bp reads (~120,000)

## Ultra-low depth (~0.06X)



Ethnic approval of our study: NO. BGI-IRB 17088

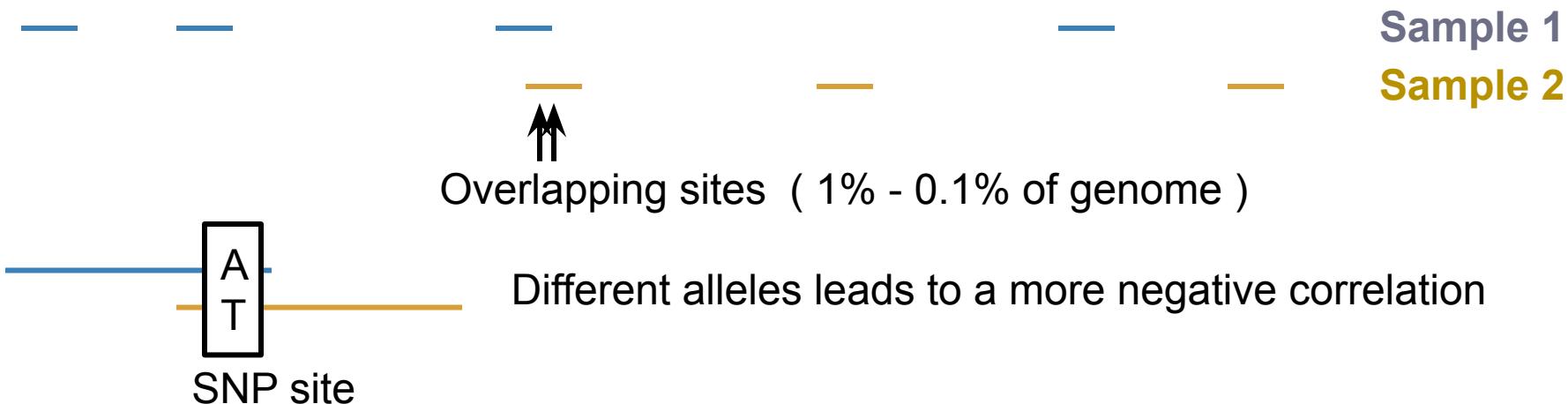
# Population structure in China?

Can we perform PCA on ultra low depth?

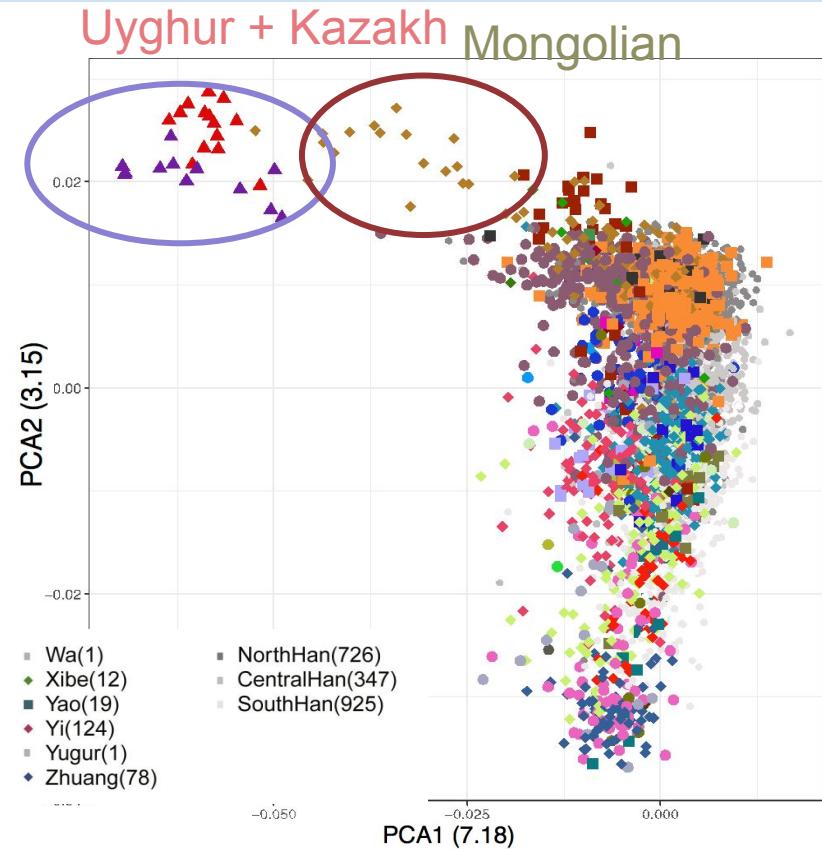
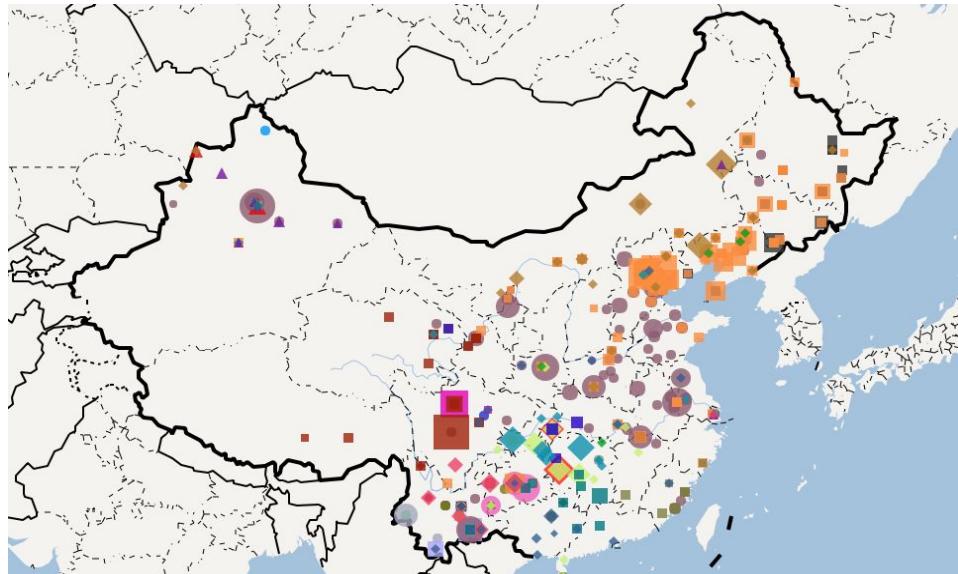


# FIRST APPROACH

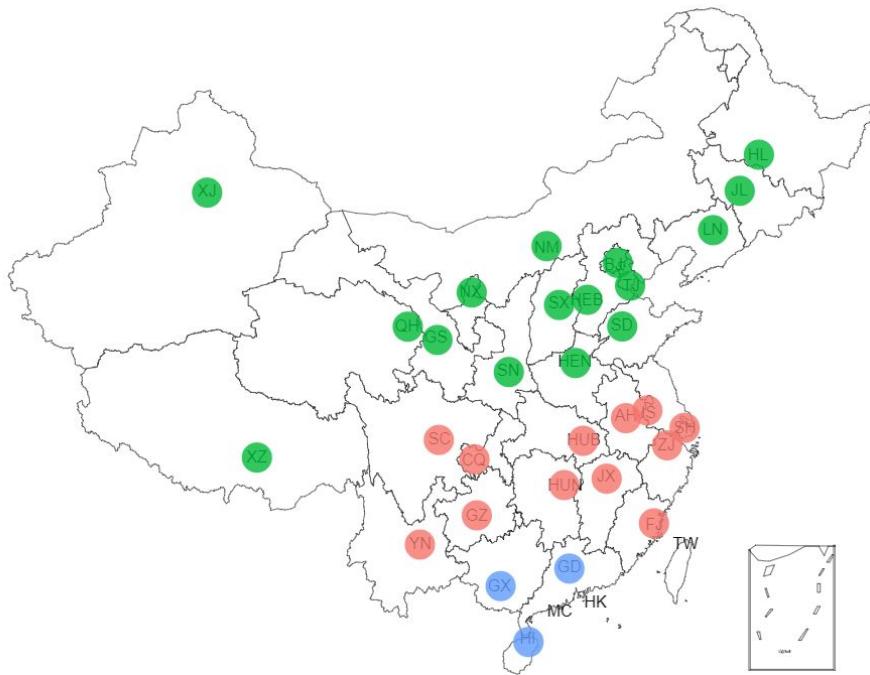
- 1) Estimate allele frequencies for each SNP
- 2) For each pair of individual estimate the covariance for overlapping sites



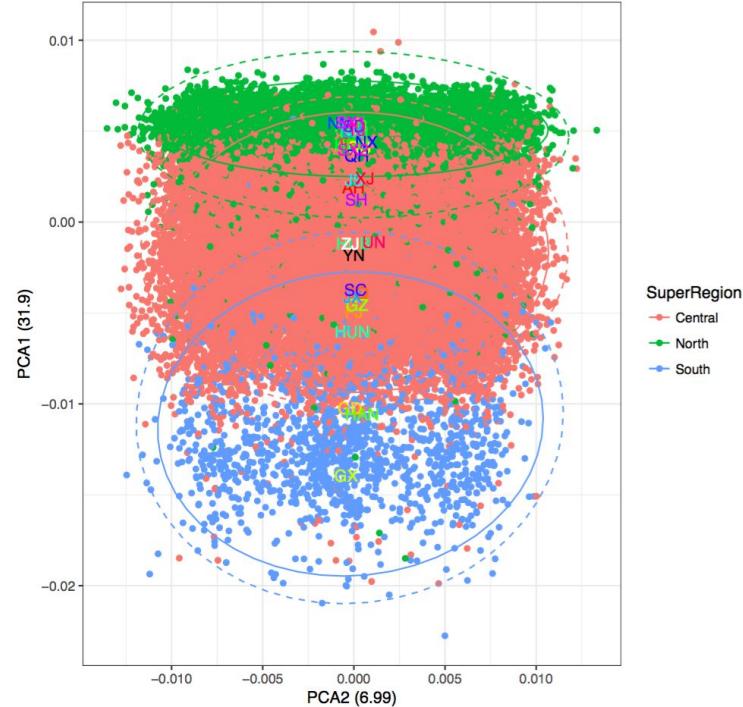
# PCA with Han and minorities



# HAN showed little differentiation



PCA from covariance ( as in plink2)



# EMU algorithm

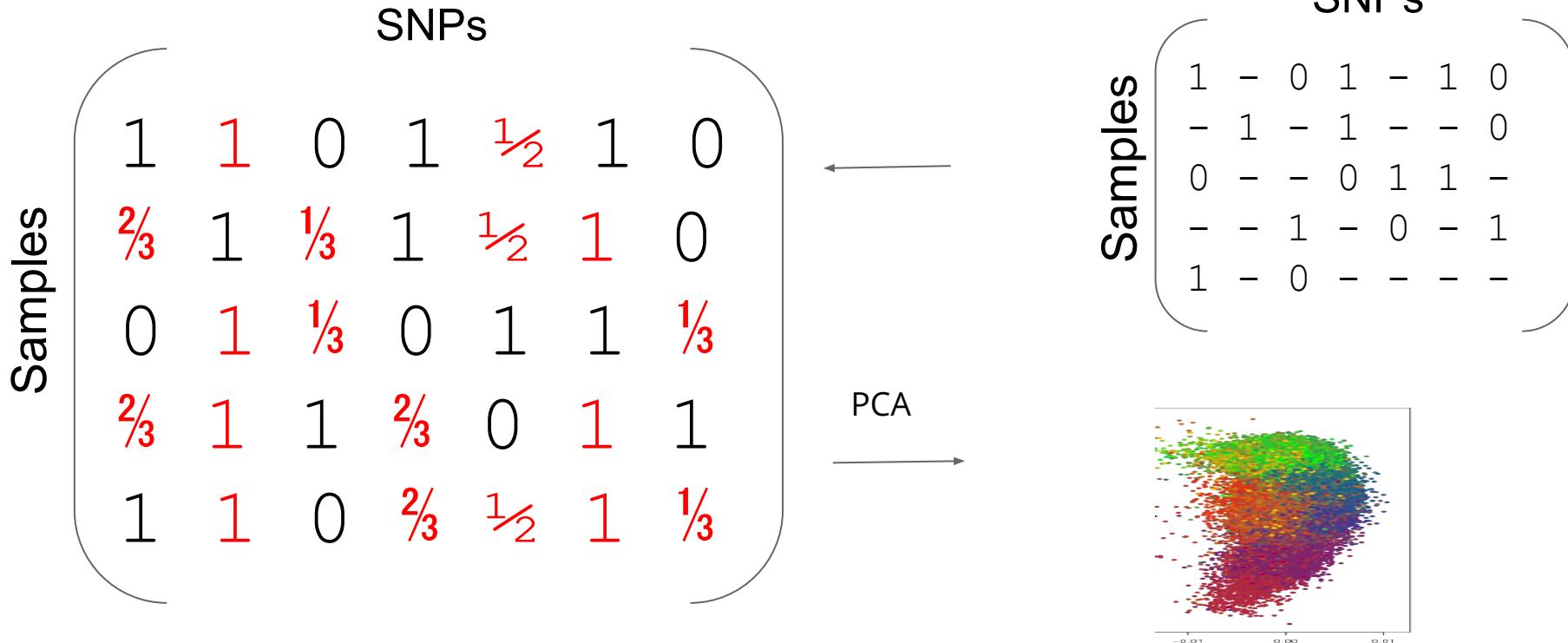
0 : Major allele

1 : Minor allele

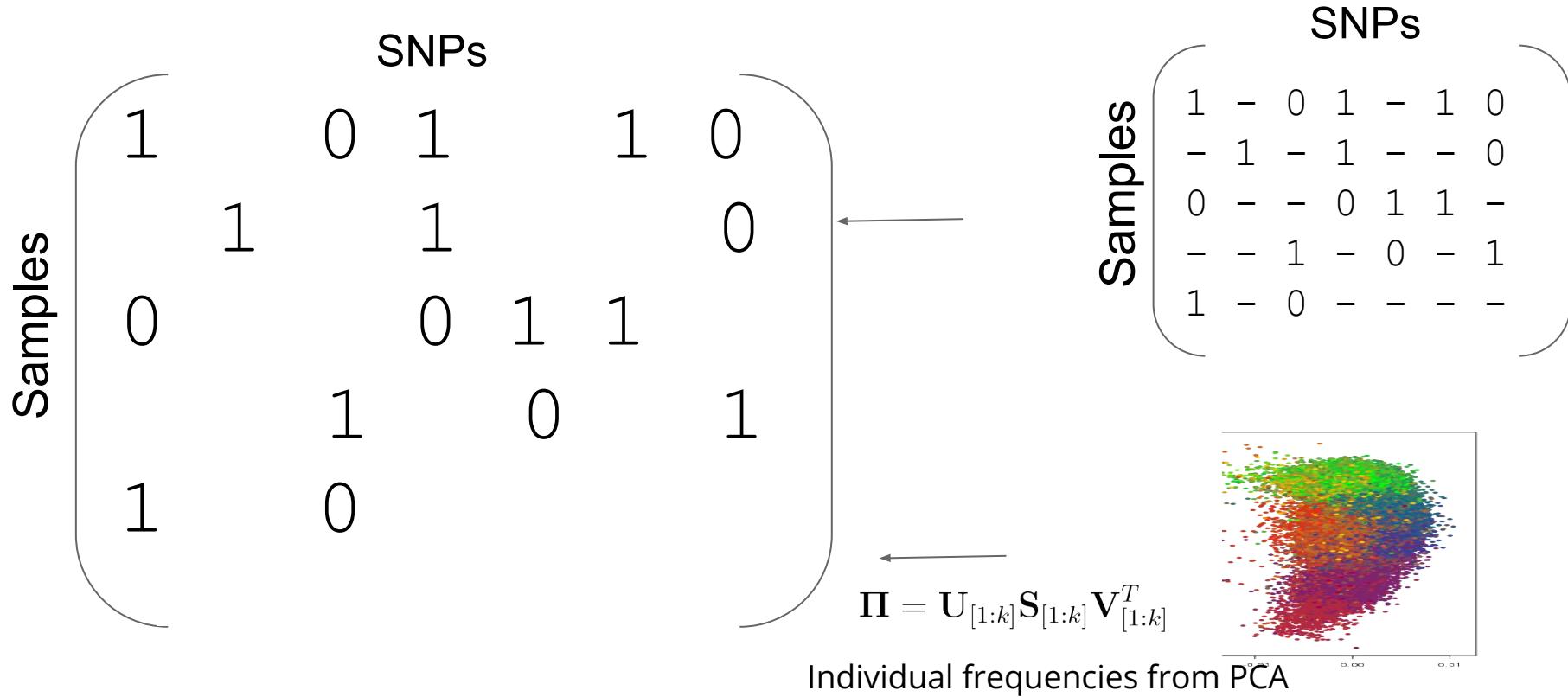
- : Missing (90-99% missing)

	SNPs							
Samples	1	-	0	1	-	1	0	
	-	1	-	1	-	-	0	
	0	-	-	0	1	1	-	
	-	-	1	-	0	-	1	
	1	-	0	-	-	-	-	

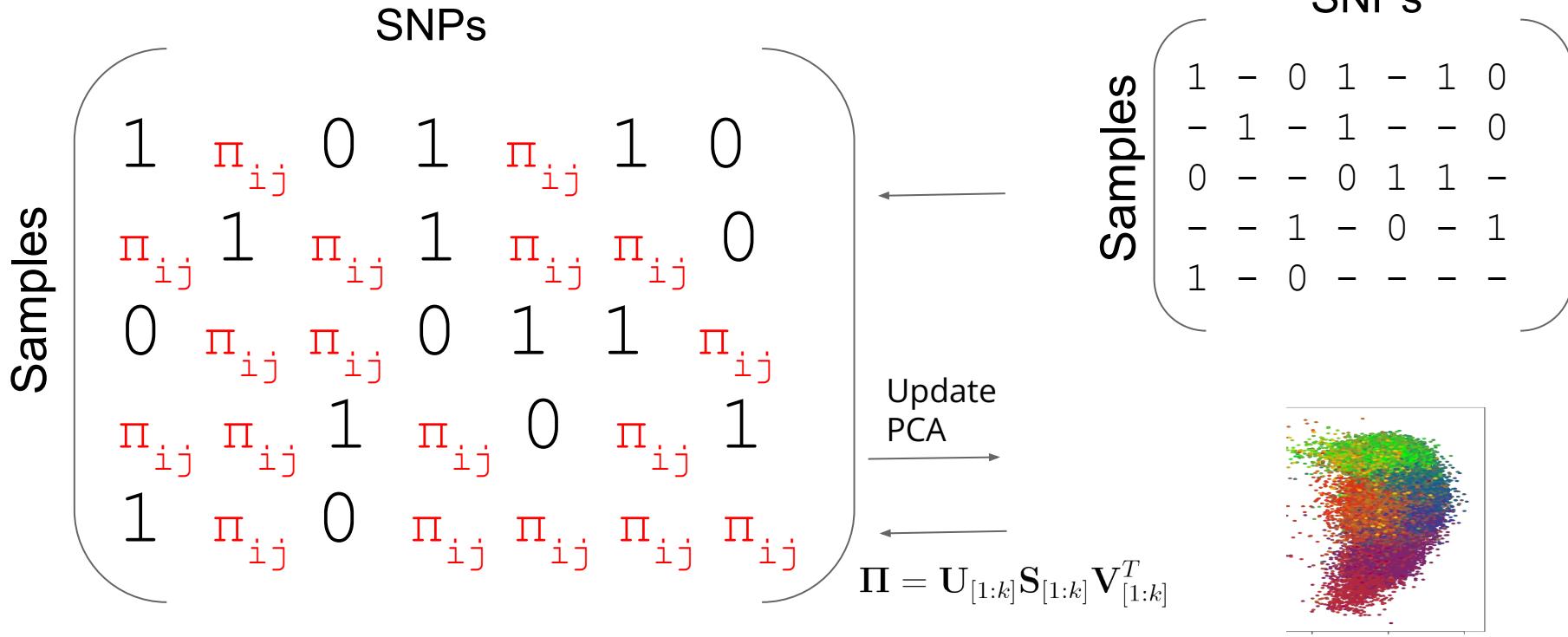
# EMU - step 1: mean imputation



# EMU - step (n): PCA imputation

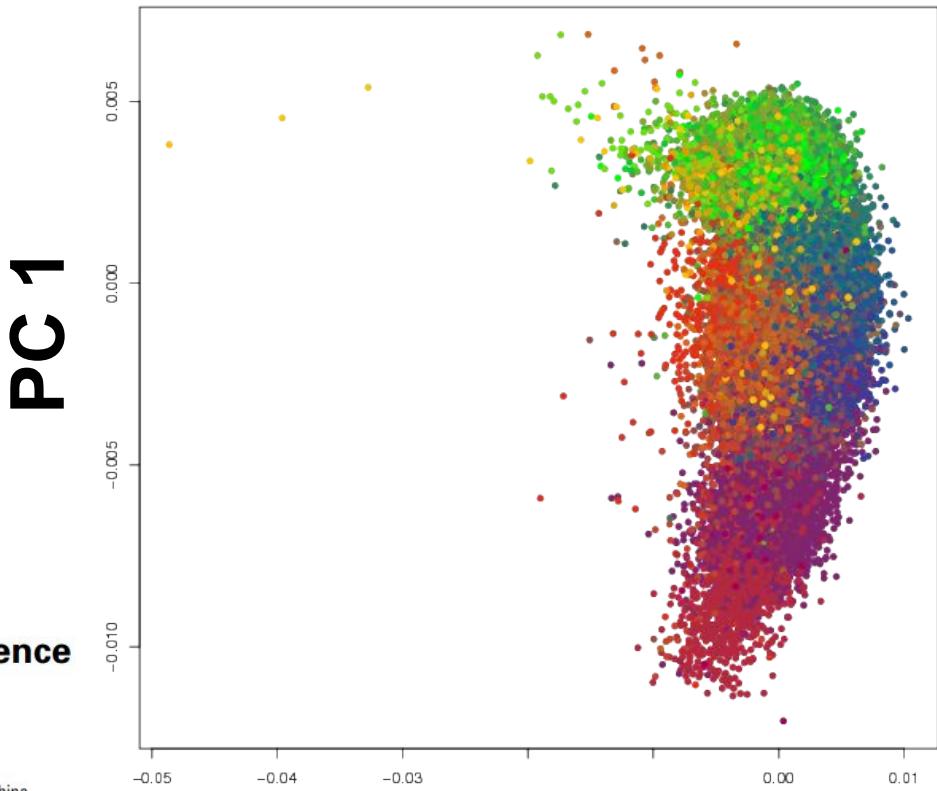


# EMU - step (n): PCA imputation



# PCA shows Longitudinal differentiation

## Mainland China



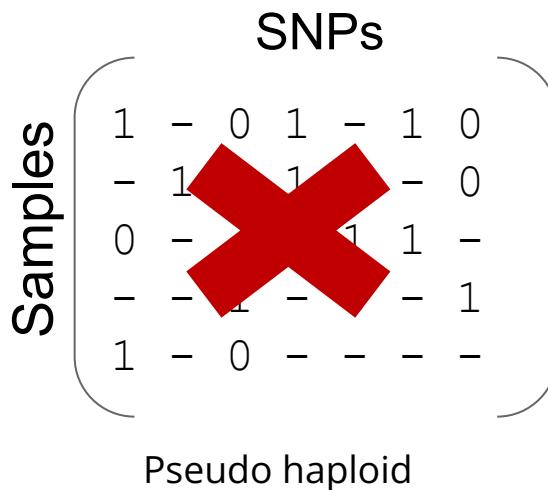
**Large-scale inference of population structure in presence of missingness using PCA**

Jonas Meisner <sup>1,\*</sup>, Siyang Liu<sup>2</sup>, Mingxi Huang<sup>2</sup> and Anders Albrechtsen<sup>1</sup>

<sup>1</sup>Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark and <sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China

# PCA for low/medium depth sequencing

Low depth sequencing has partial missing information for all genotypes.



marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1
chr1_787290	3	1	0.888891		0.111109		0.0	
chr1_832873	1	0	0.941178		0.058822		0.0	
chr1_855316	1	3	0.984616		0.015384		0.0	
chr1_872843	1	2	0.888891		0.111108		0.0	
chr1_893503	2	0	0.984616		0.015384		0.0	
chr1_914838	0	3	2.5e-05	0.999975		0.0	0.666446	
chr1_931513	1	3	0.0	0.999684		0.000315		
chr1_954724	2	0	0.984616		0.015384		0.0	
chr1_975014	1	3	0.0	1.0	0.0		0.800001	
chr1_993036	2	0	0.0	0.833765		0.166235		
chr1_1017048	0	2	0.0	0.001949		0.998051		
chr1_1047561	1	3	0.0	0.999747		0.000253		
chr1_1065910	1	3	0.969698		0.030302		0.0	
chr1_1089921	1	3	0.969698		0.030302		0.0	
chr1_1110693	1	0	0.0	0.995057		0.004943		
chr1_1125786	3	1	0.0	0.058822		0.941178		

Genotype likelihoods

# PCA for low depth sequencing

Low depth sequencing has partial missing information for all genotypes.

Similar to mean imputation we can use the overall allele frequency to fill in the missing information

$$P(G = g | X, \pi) = \frac{P(X | G = g)P(G = g | \pi)}{\sum_{g'=0}^2 P(X | G = g')P(G = g' | \pi)}$$

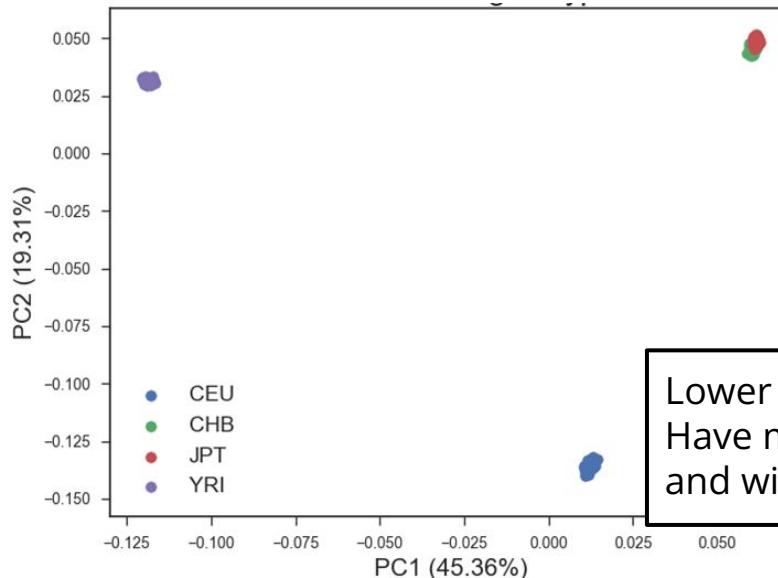
**Genotype likelihood**      **Individual Allele frequency**

**Dosage based on  
Allele frequency and  
Genotype likelihoods**

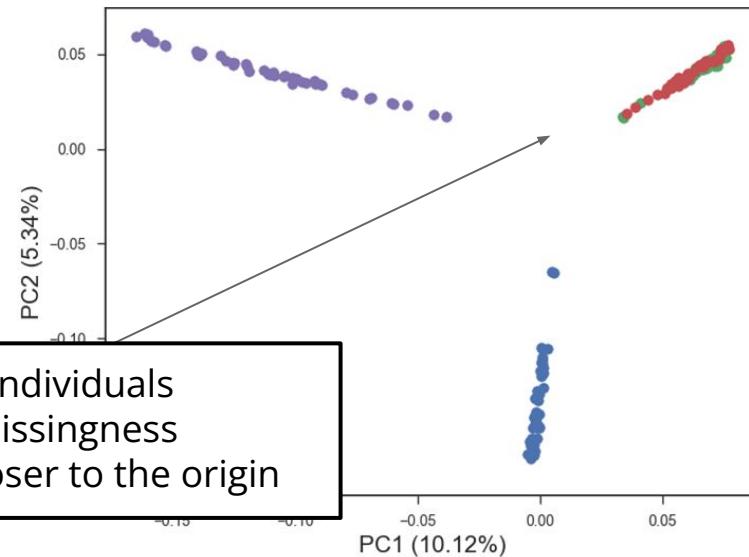
$$\rightarrow \mathbb{E}[G | X, \pi] = \sum_{g=0}^2 P(G = g | X, \pi)$$

# Issues with missingness/low depth data

PCA from high depth genotypes



Genotype call from low depth data (2-10X)



Lower depth individuals  
Have mode missingness  
and will be closer to the origin

# Low depth sequencing

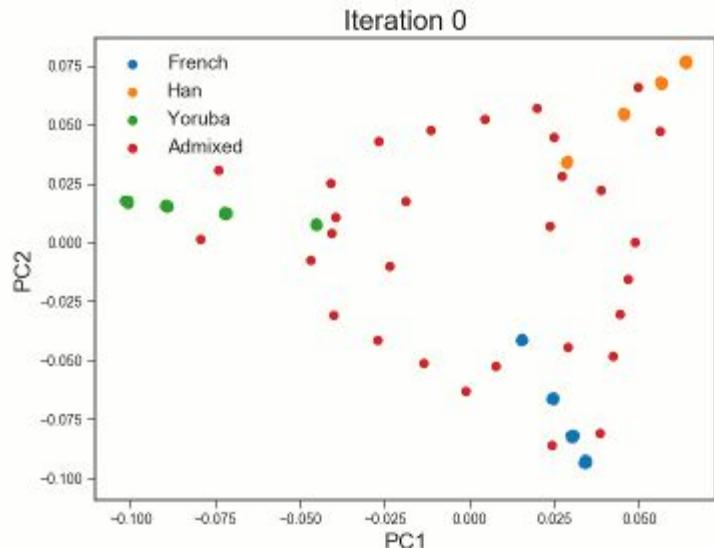
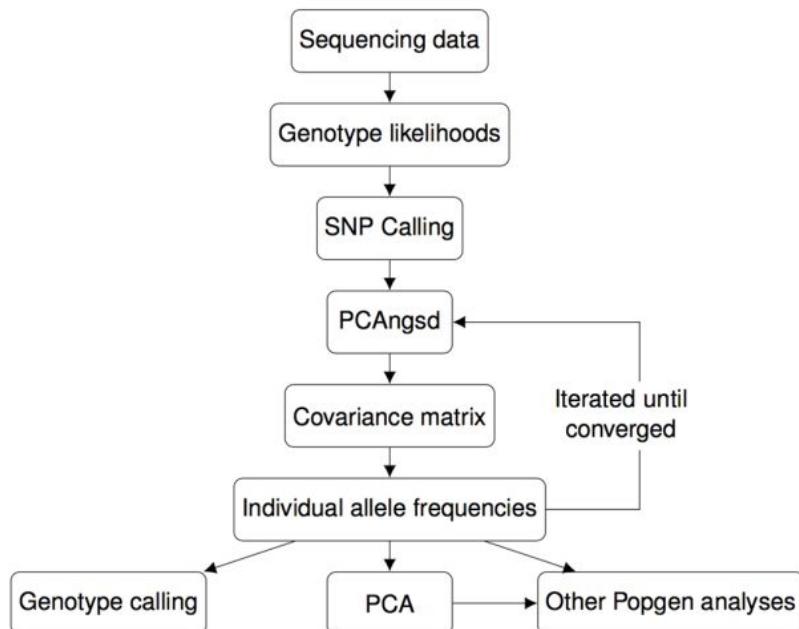
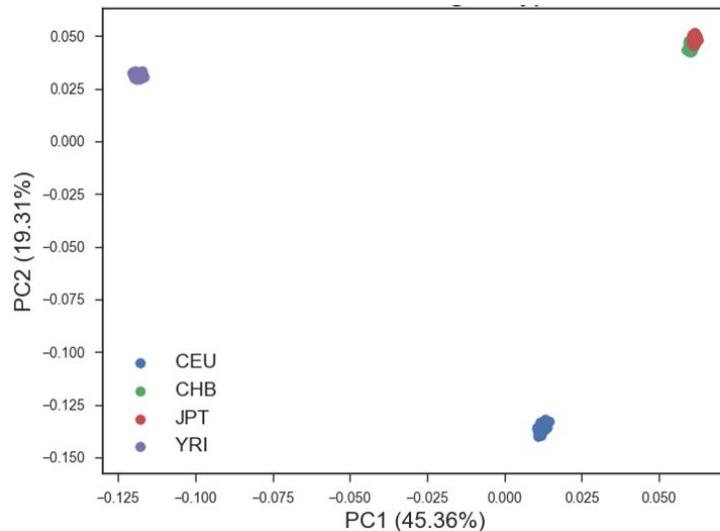


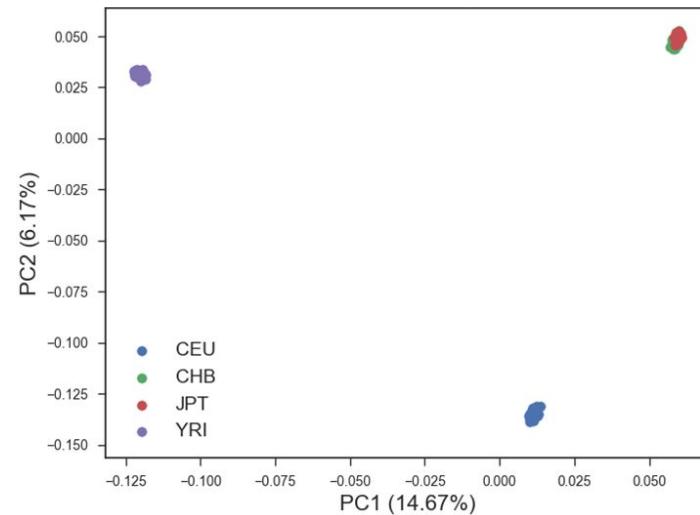
Figure: PCAngsd framework

# Issues with missingness

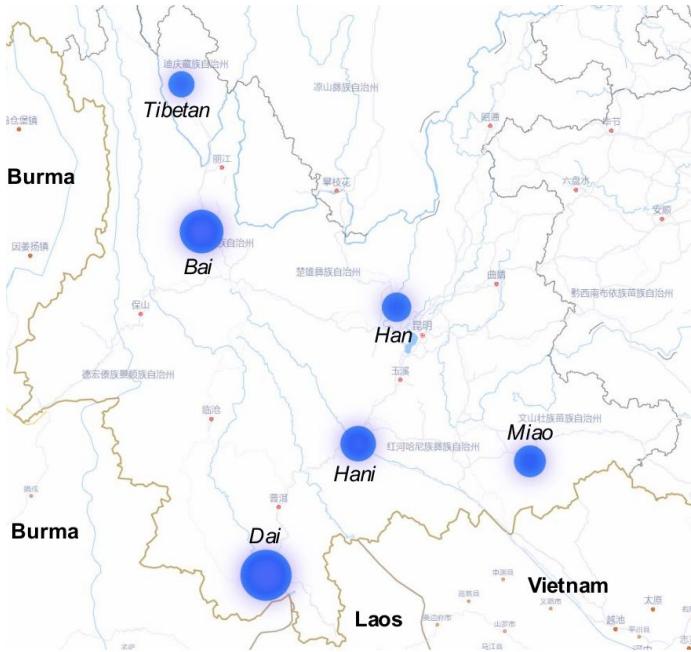
PCA from high depth genotypes



PCAngsd from low depth data (2-10X)



# Host genome from gut microbiome



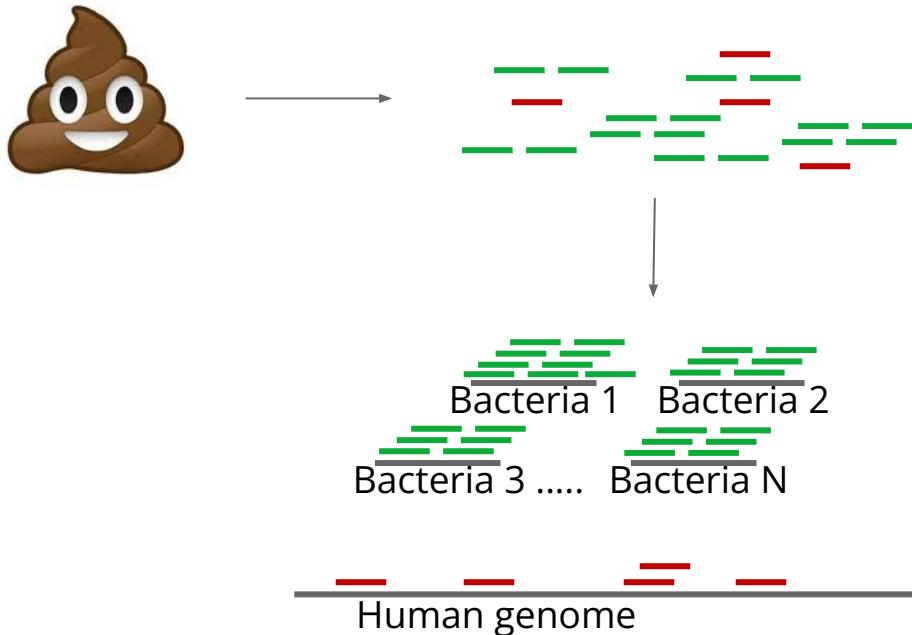
## 6 Minorities (n=671)

- Bai (n=120)
- Dai (n=107)
- Han (n=43)
- Hani (n=98)
- Miao (n=85)
- Tibetan (n=84)
- Unknown (n=134)

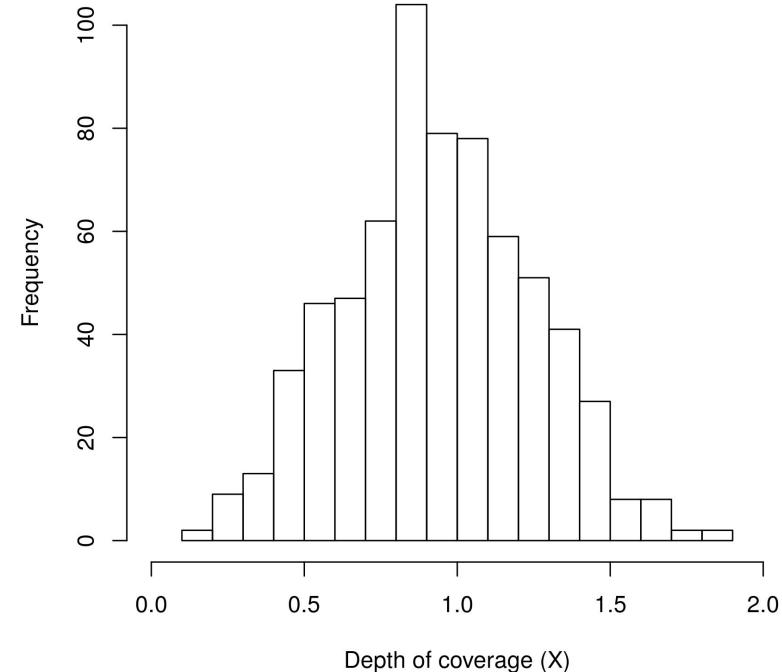


Project lead by Mo Han and Karsten Kristiansen ()

# PCA on host metagenomics

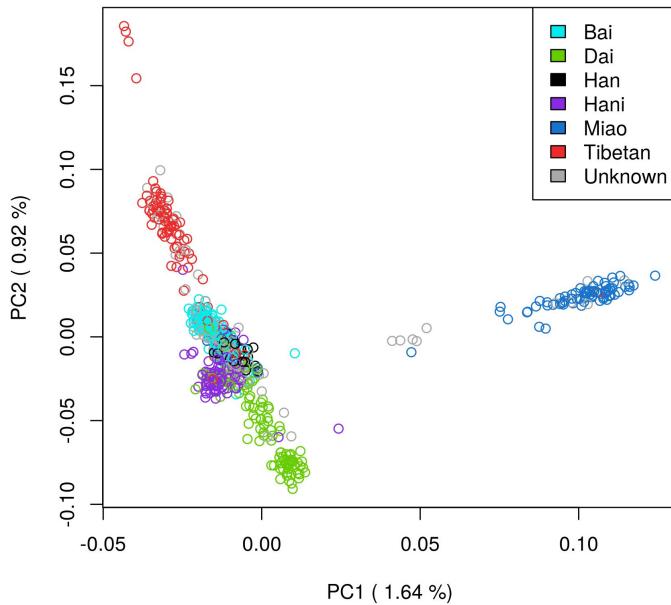


Sequencing depth of host mapping

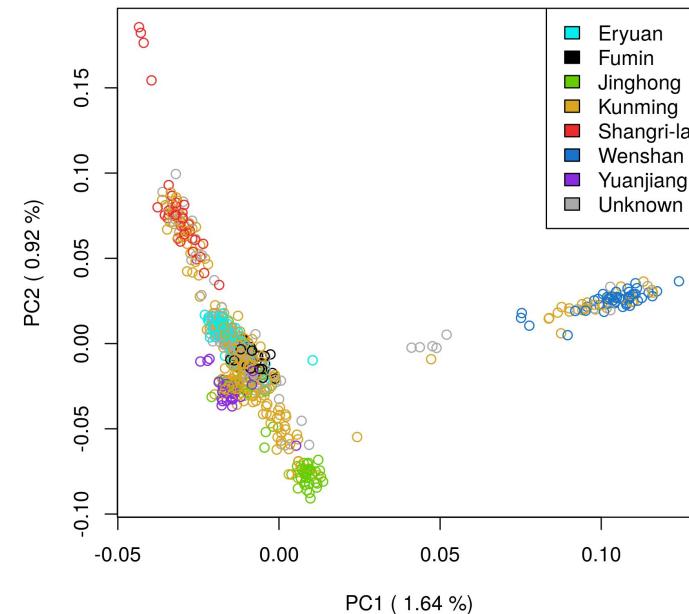


# PCAngsd on host genetics

Minority labels



Location in labels



# PCA and selection

# Continues structure / within population

FastPCA selection

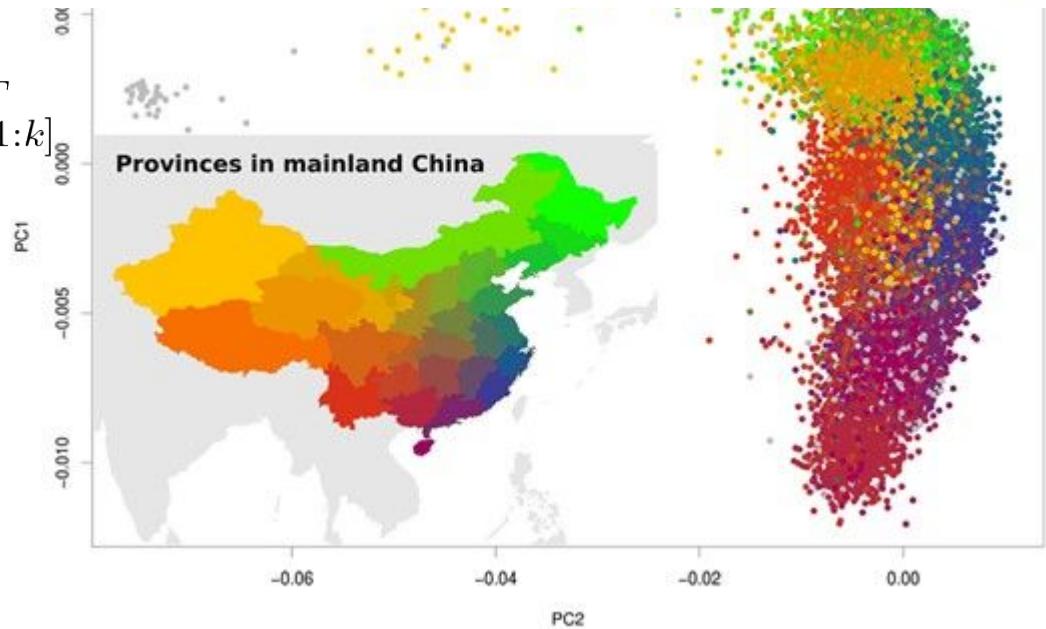
$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \boldsymbol{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$

$$M/S(2\boldsymbol{\Pi}V_k) \sim \chi^2$$

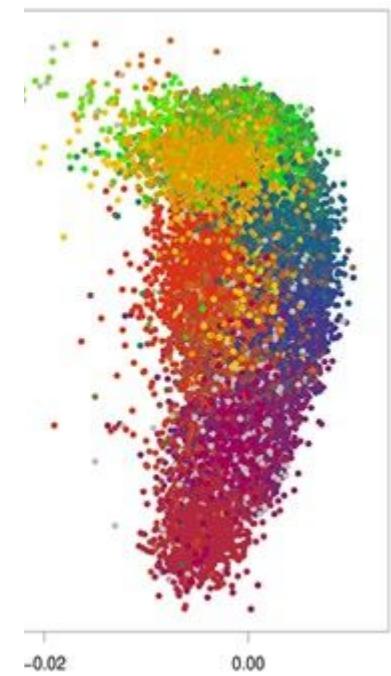
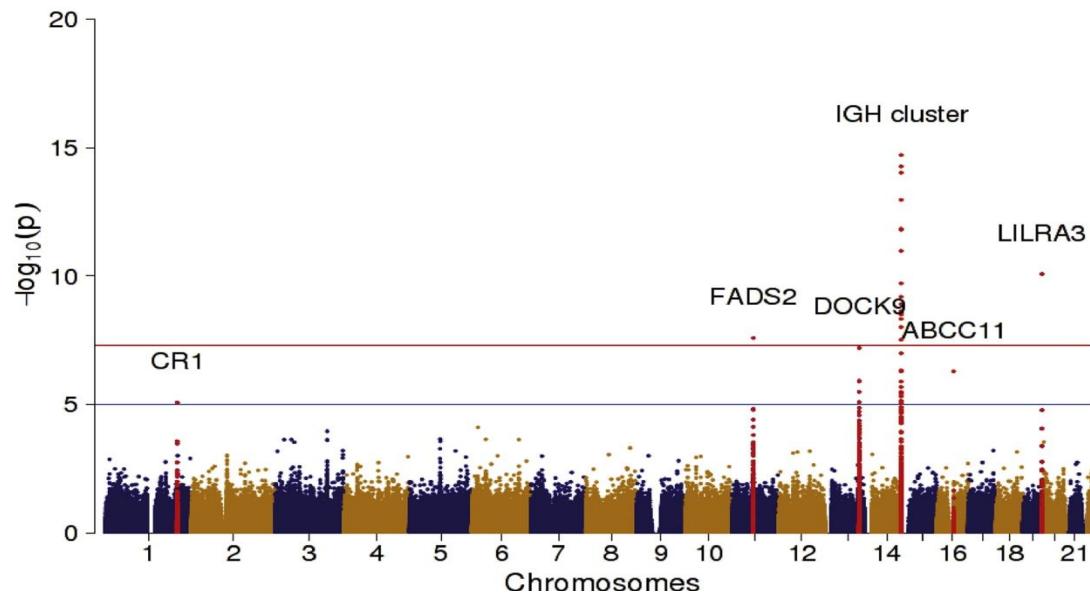
Galinsky et al (2016)

Article

Cell

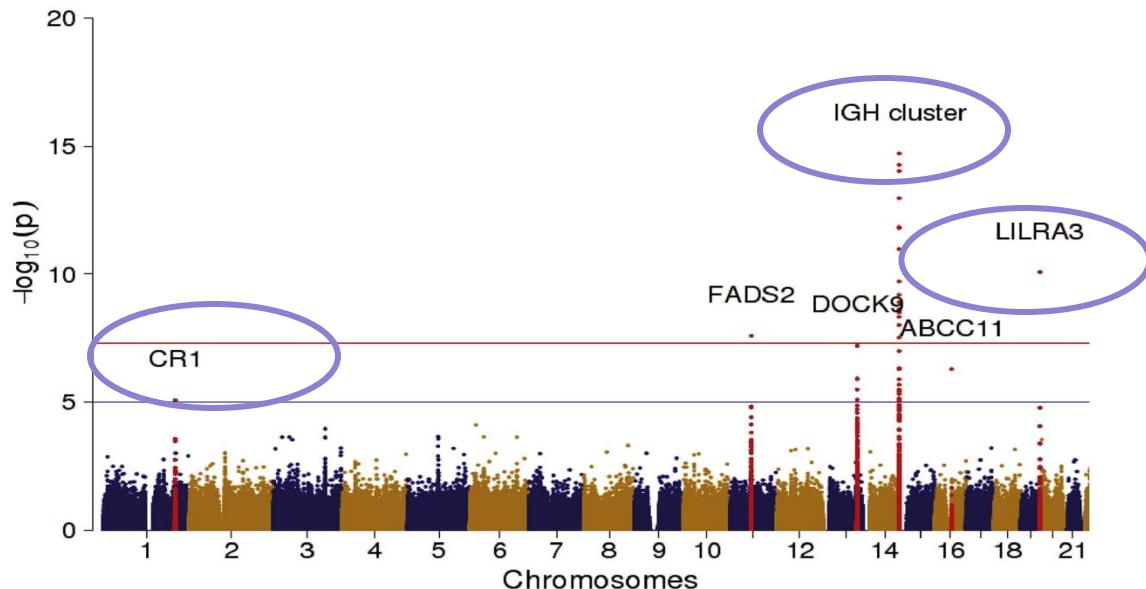


# SELECTION SCAN - GWAS on first PC



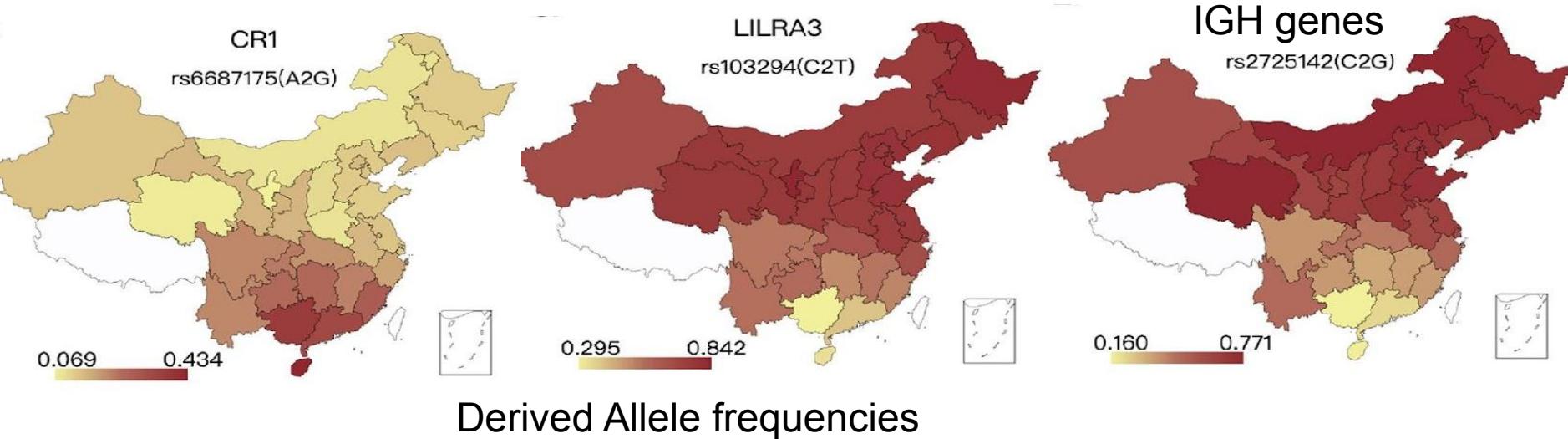
PC1

# SELECTION SCAN

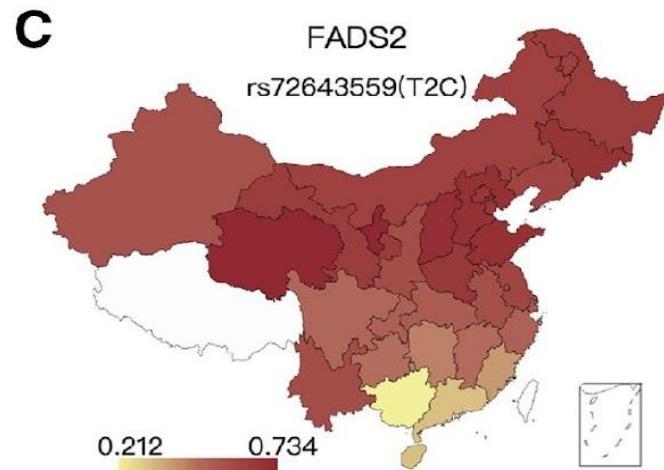
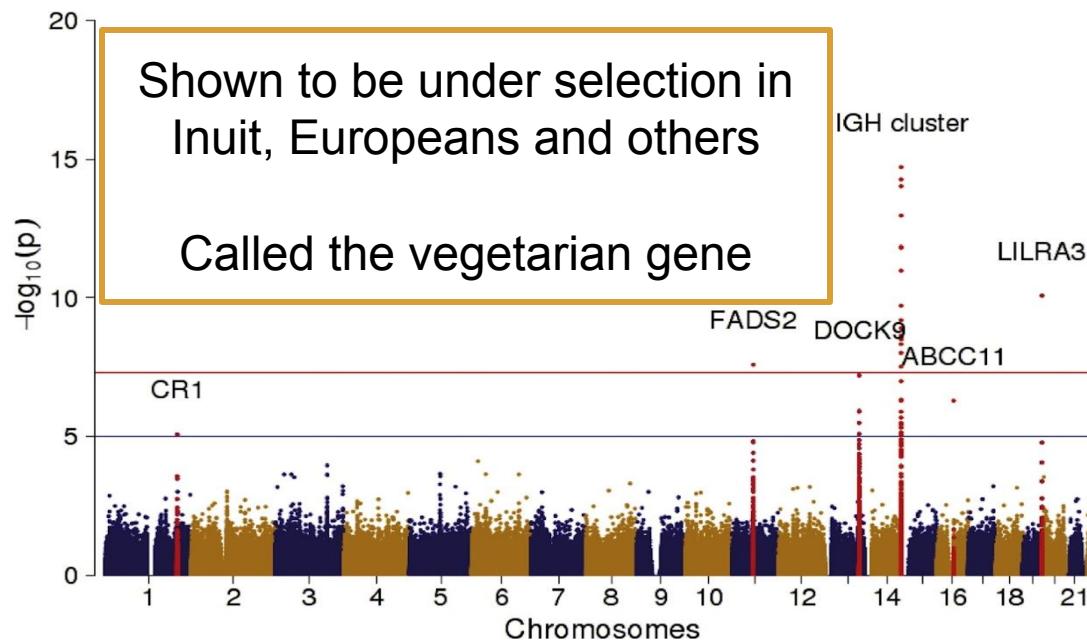


Immune genes  
previously  
Shown to be under  
selection  
In East Asians

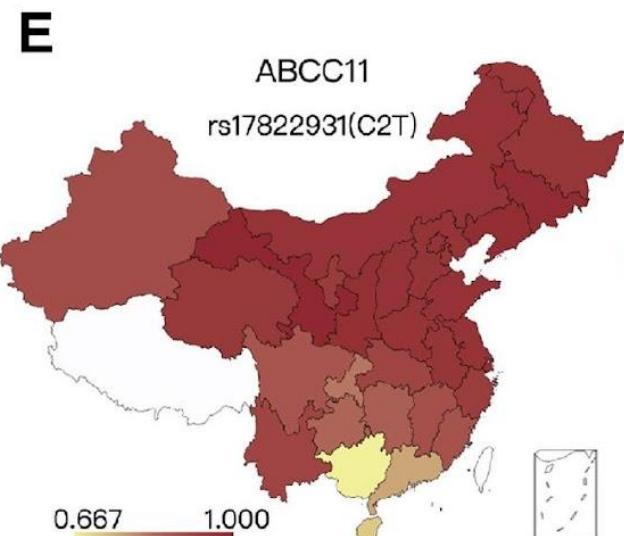
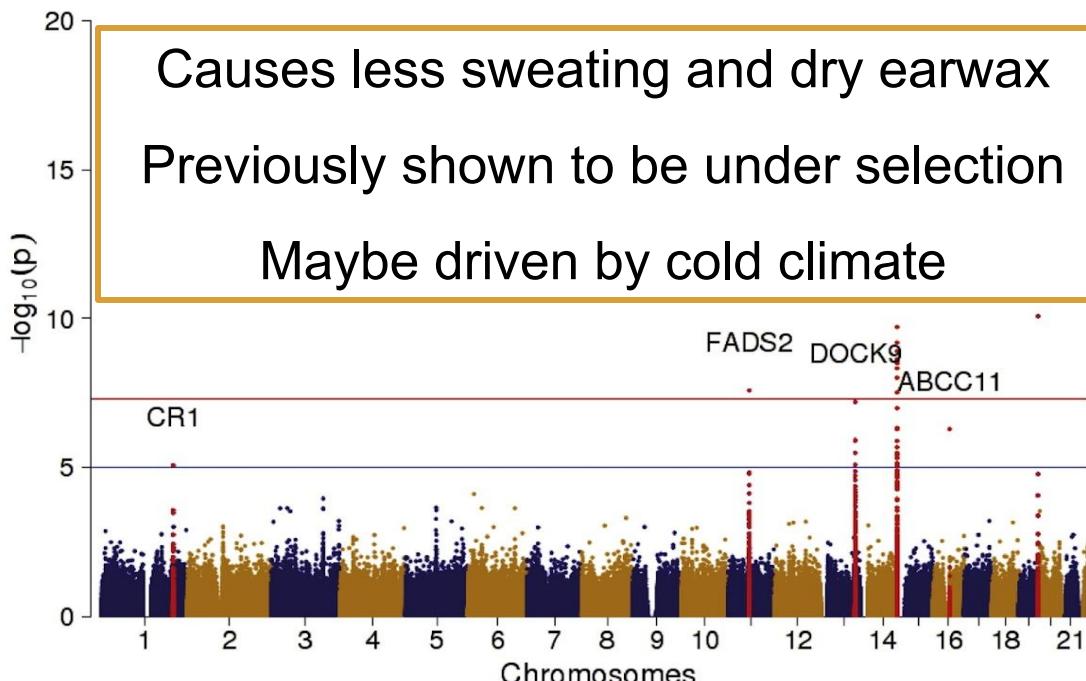
# IMMUNE RESPONSE GENES



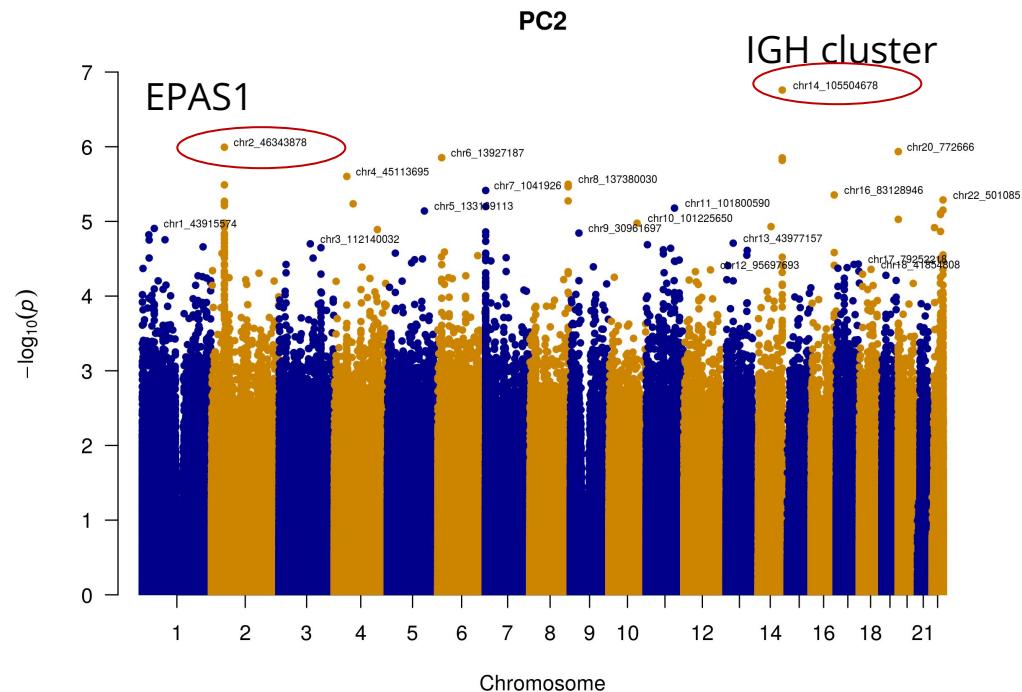
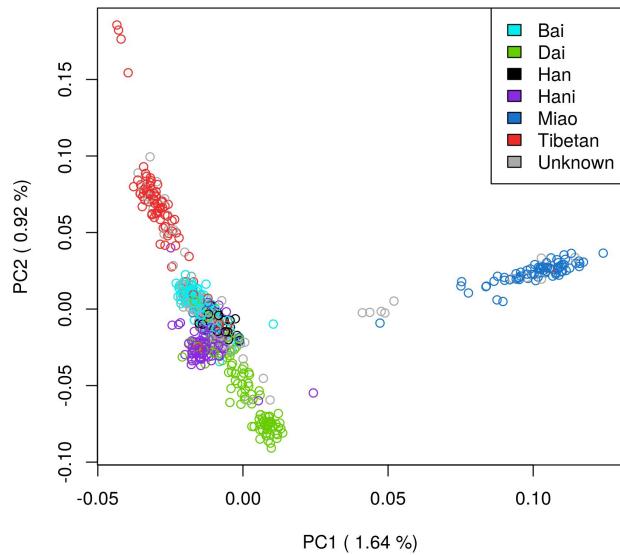
# FATTY ACID METABOLISM (FAD2)



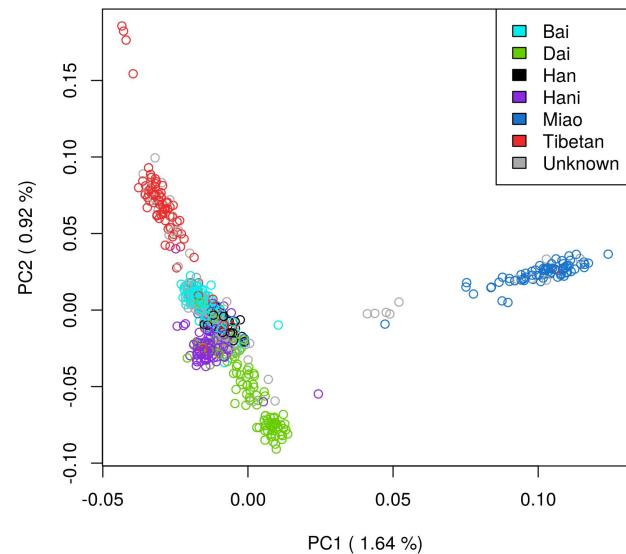
# SWEAT AND EARWAX(ABCC11)



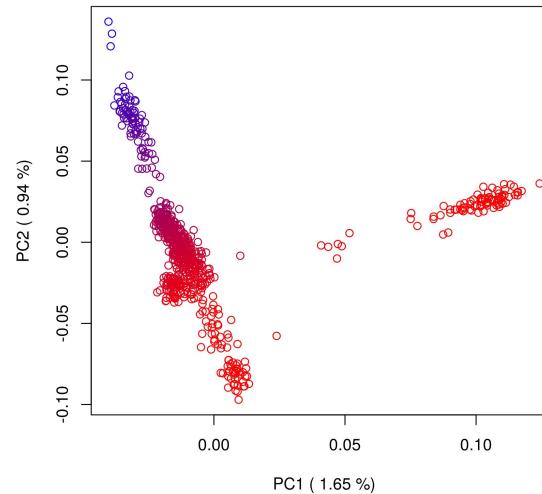
# Selection scan on PC 2



# PCA and expected allele frequencies



EPAS1 top variant



Expected allele frequency

0%

$$\Pi = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$

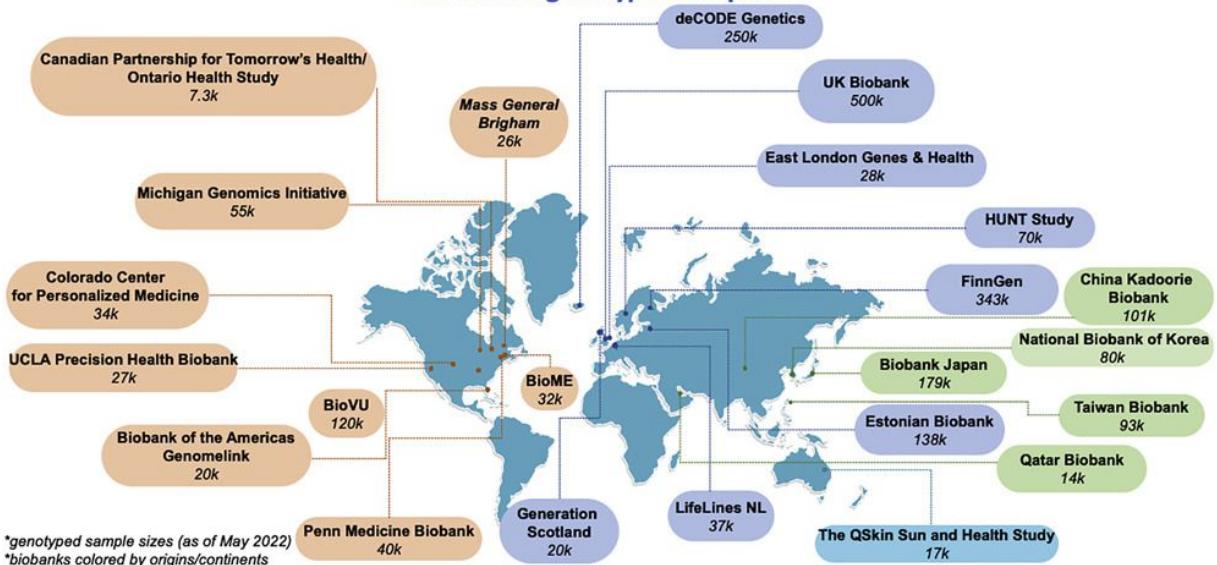
100%

# Time for exercises

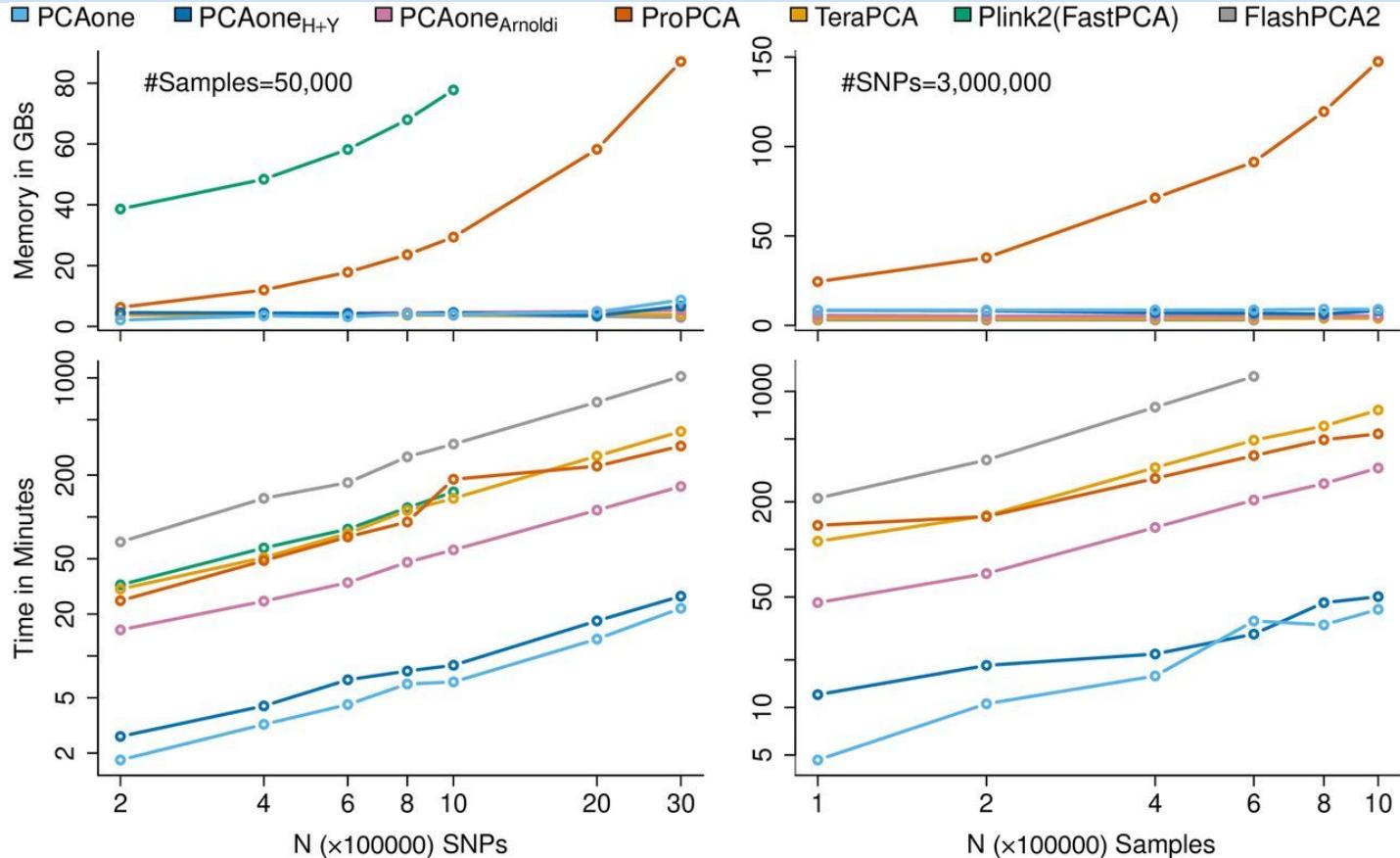
Run the pca notebook

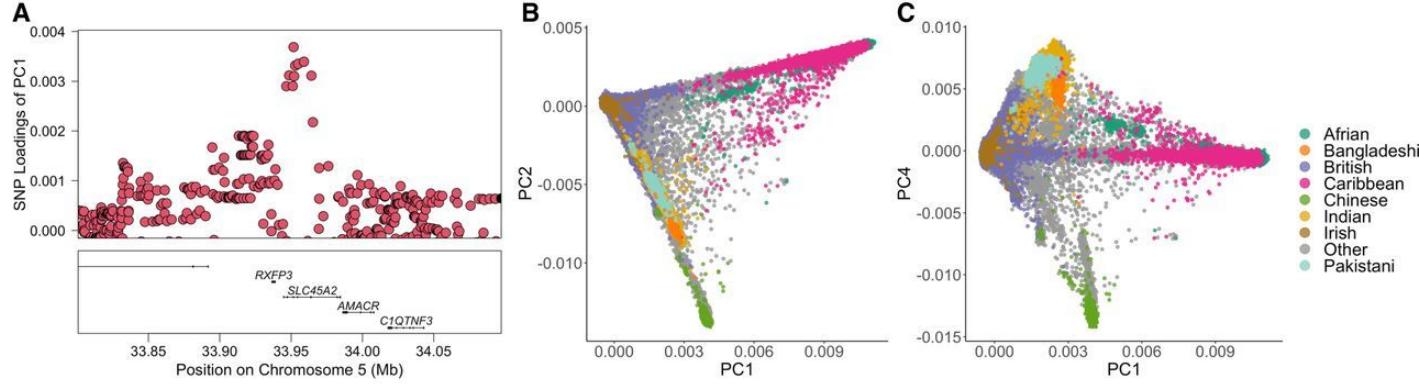
# When you have large data set

23 biobanks with different origins and ancestries have joined GBMI  
→ 2.2 million genotyped samples



# PCAone is fast, accurate and low RAM

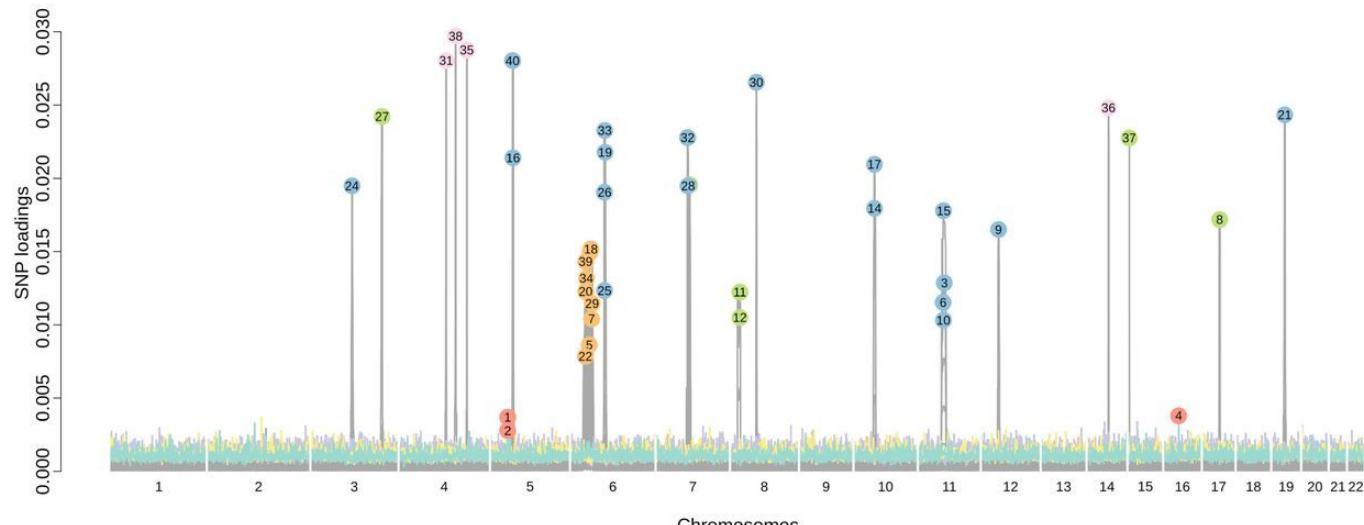




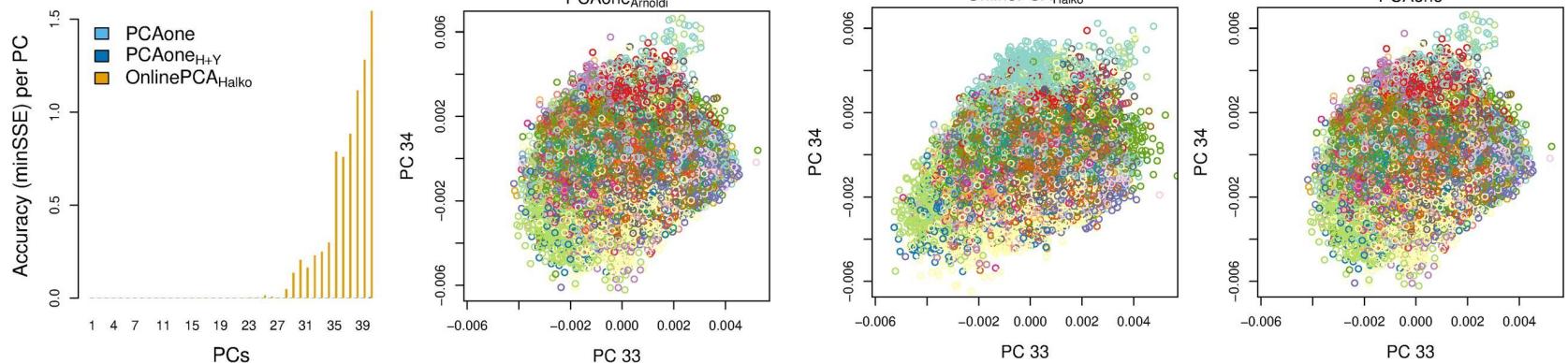
40 PCs

0.5M individuals

6 Million common SNPs

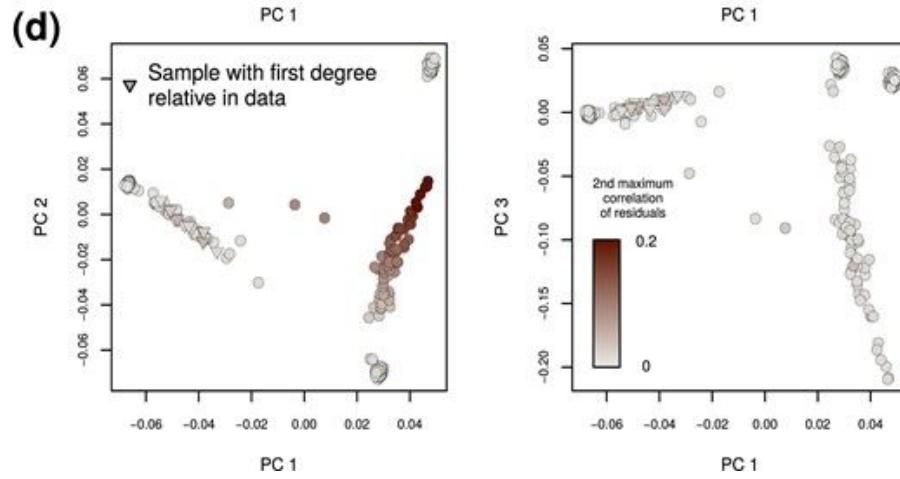
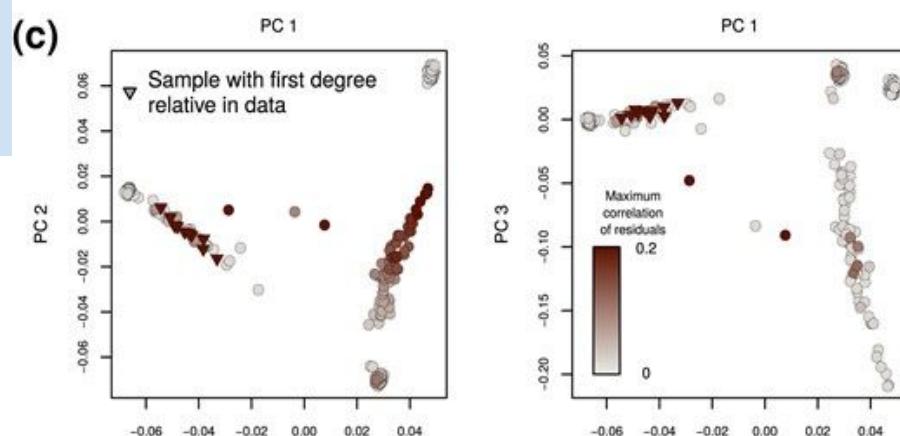
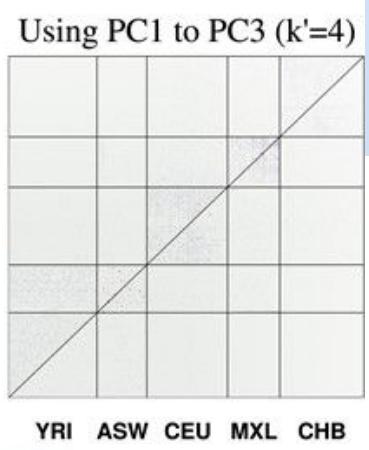
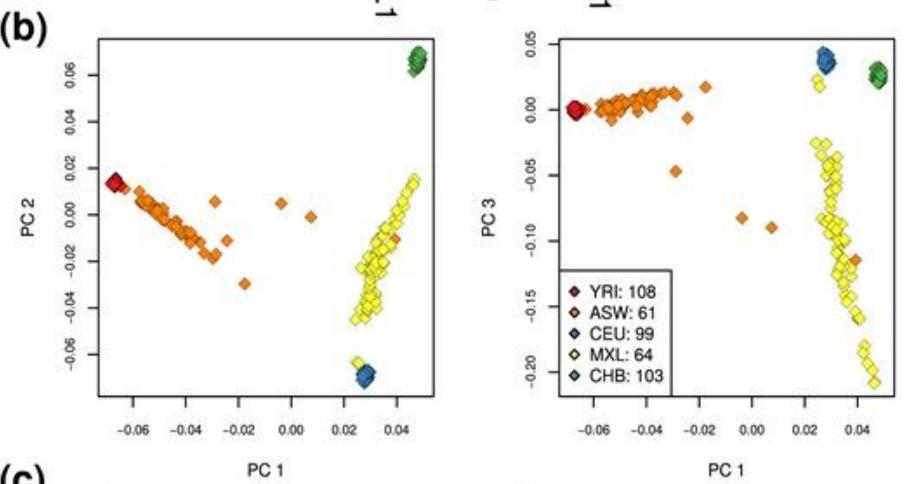
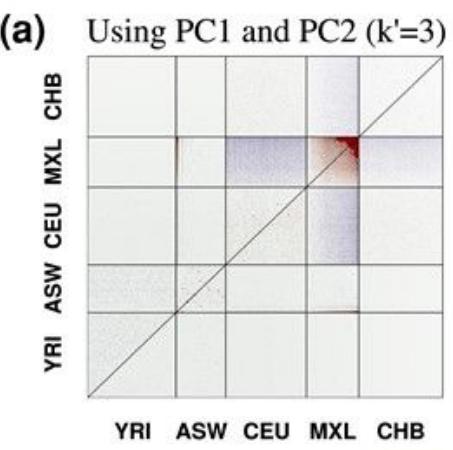


# Single Cell RNAseq



**Table 2.** Performance on scRNA-seq data with **1,306,127 cells and 23,771 genes** for estimating the top K = 40 PCs

Program	Wall Time(m)	IO Time(m)	Epochs	RAM(GB)	MEV	minSSE
PCAone	49	31	10	8.97	0.999957	0.00417
PCAone <sub>H+Y</sub>	42	30	8	6.82	0.999925	0.00732
OnlinePCA.jl <sub>Halko</sub>	461 + 100 <sup>a</sup>	300	8	1.71	0.954103	7.10293
PCAone <sub>Arnoldi</sub>	484	469	103	5.73	-	-



# Conclusion

- Calling genotypes can cause major bias for PCA and Admixture analysis
- Using genotype likelihoods instead can solve the problems
- Admixture analysis and PCA are related and can both be used to estimate individual allele frequencies
- individual allele frequencies can be used for selection scans

# Time for exercises

Run the pca notebook