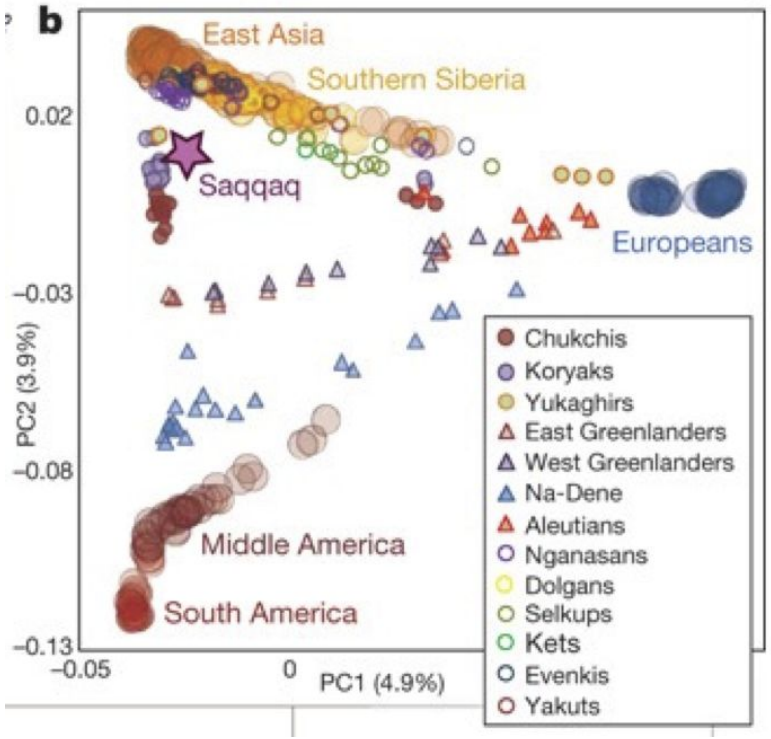
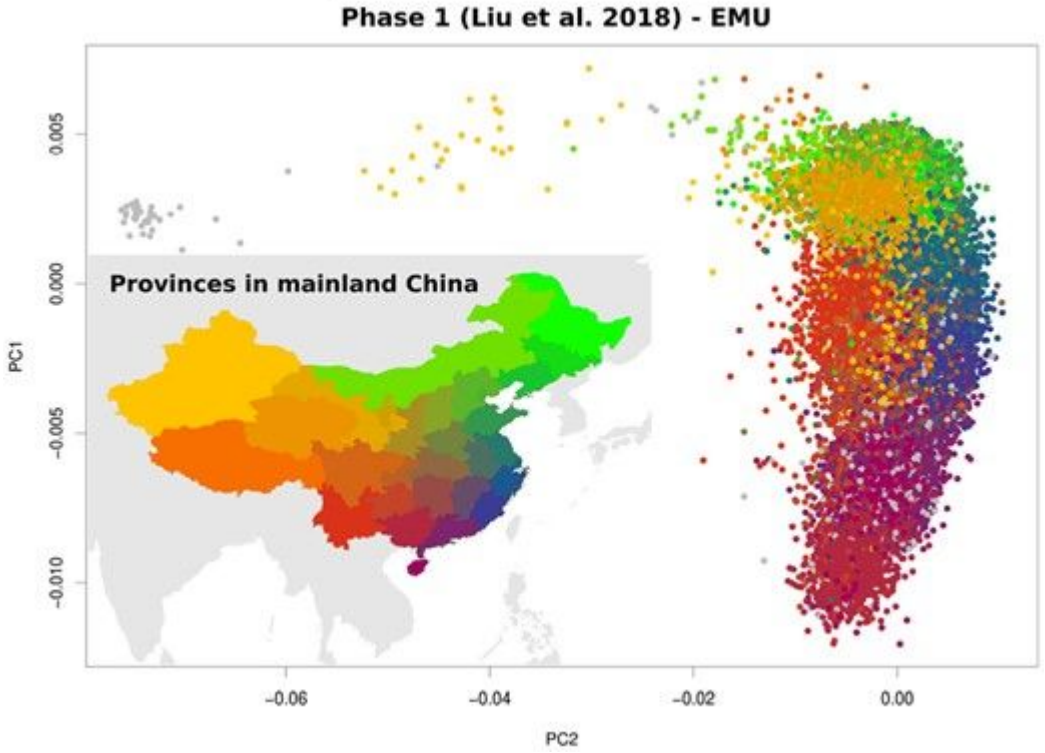
PCA and selection scan

Anders Albrechtsen

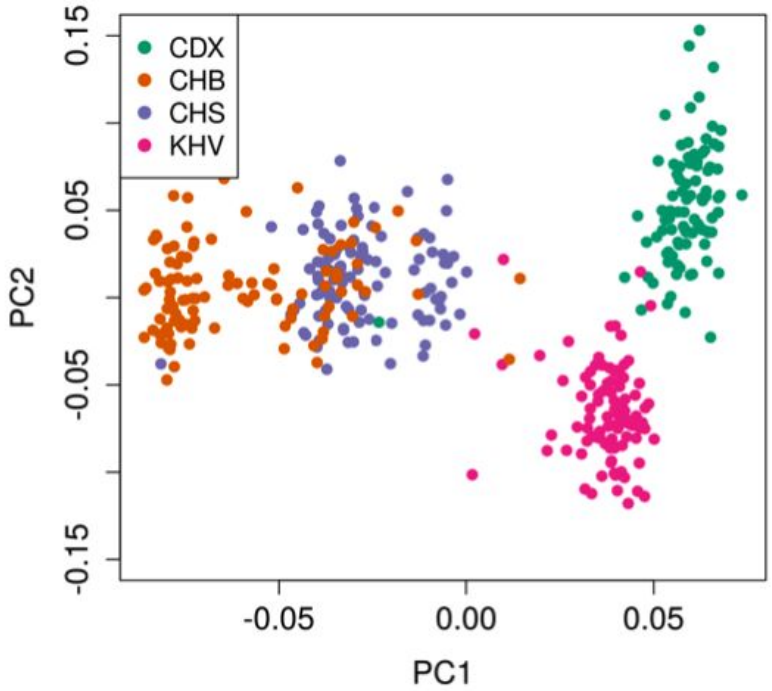
PCA for population structure



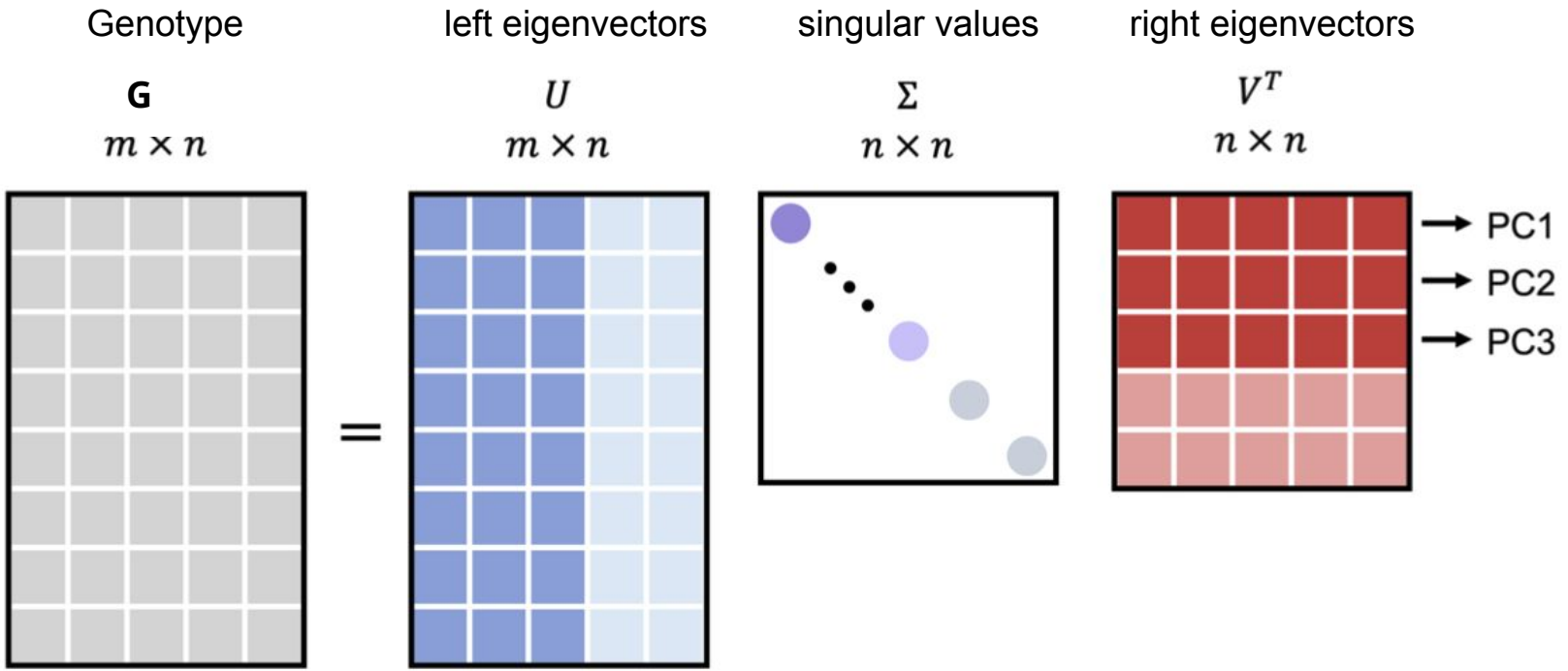
Continues structure / within popuation



	Ind1	Ind2	Ind3	Ind4	Ind5
SNPs	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



Each individuals is a dot in PCA plot



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

PCA and distances

individuals

SNP1	1	1	1	0	0
SNP2	0	1	2	1	2
SNP3	2	1	1	0	1
SNP4	0	0	1	2	2
SNP5	2	1	1	0	0
SNP6	0	0	1	1	1
SNP7	2	2	1	1	0

Total Distance

	Ind1	Ind2	Ind3	Ind4	Ind5
Ind1	0	3	7	10	11
Ind2	3	0	4	7	8
Ind3	7	4	0	5	4
Ind4	10	7	5	0	3
Ind5	11	8	4	3	0

1 dimensional projection

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

Genotype covariance matrix

M number of sites

G genotypes

G^j genotypes for individual j

G_k^j genotypes for site k in individual j

f_k allele frequency for site k

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

$$\tilde{G}_k^i = \frac{G_k^i - 2f_k}{\sqrt{2f_k(1-f_k)}}, \quad var(G_k) = 2f_k(1-f_k)$$

After normalization all SNPs have the same mean and variance

Dealing with missingness (most software)

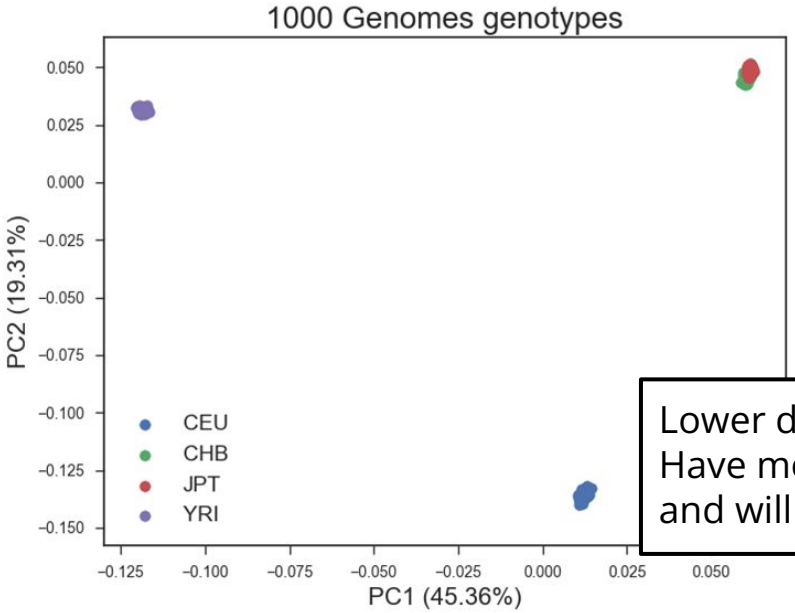
If a genotype is missing then $\tilde{G}_k^i$ is set to zero

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

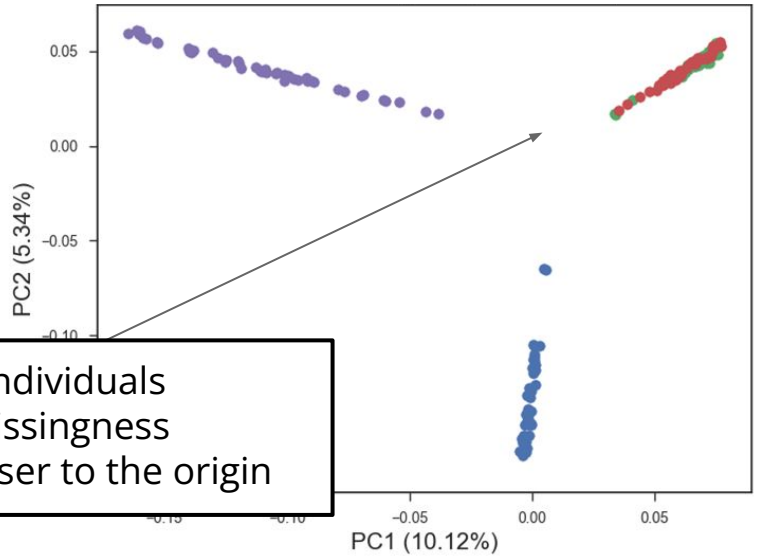
or a site is discarded

- Not possible for large samples
- Will likely cause bias

Issues with missingness/low depth data



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



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

D

0	1	na	1	na
na	0	na	1	0
na	na	0	na	na
0	na	1	na	1

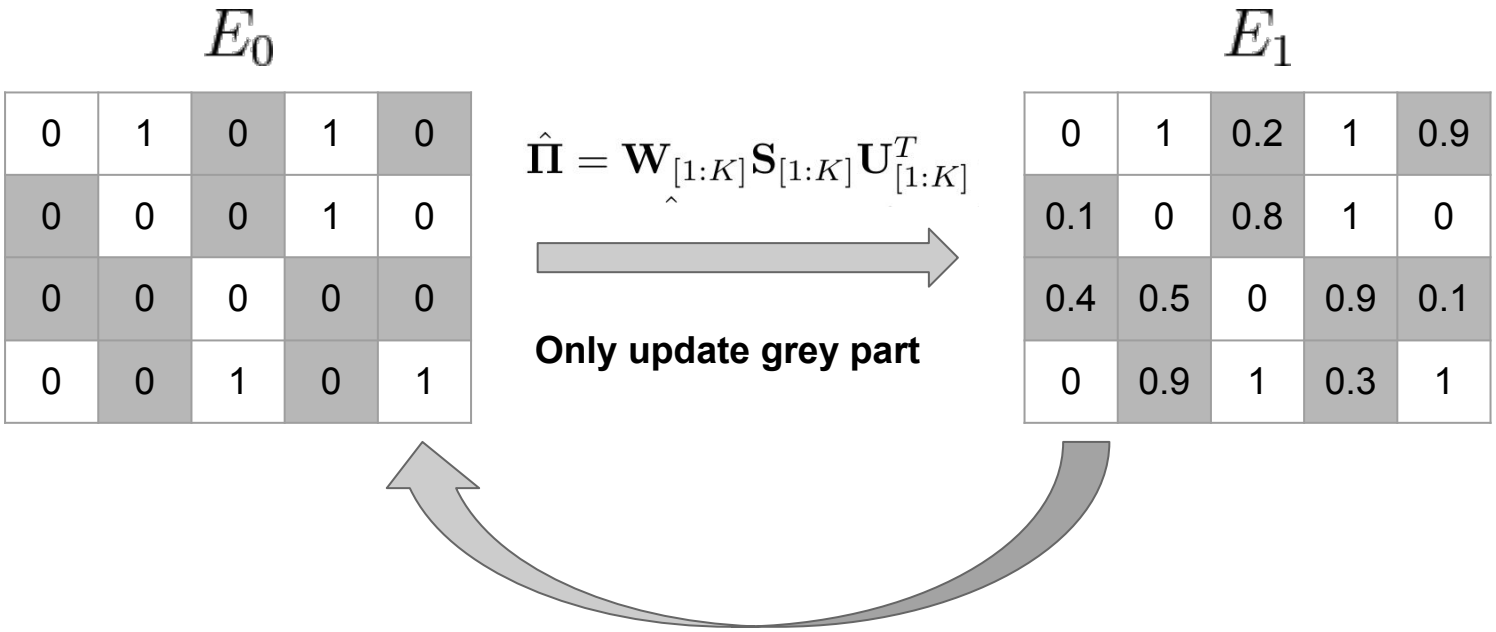
PCA to predict data

$$\hat{\mathbf{\Pi}} = \mathbf{W}_{[1:K]}\mathbf{S}_{[1:K]}\mathbf{U}^T_{[1:K]}$$

Only update grey part

$\hat{E}$

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



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

$\hat{E}$

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

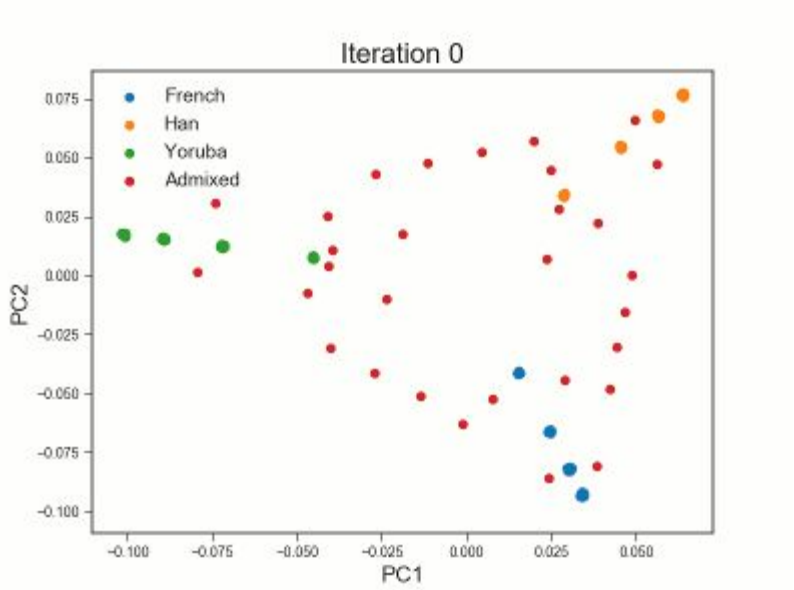
$$\hat{\mathbf{\Pi}} = \mathbf{W}_{[1:K]}\mathbf{S}_{[1:K]}\mathbf{U}^T_{[1:K]}$$

$\hat{\Pi}$

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

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

The Idea of PCangsd/EMU



Low depth sequencing

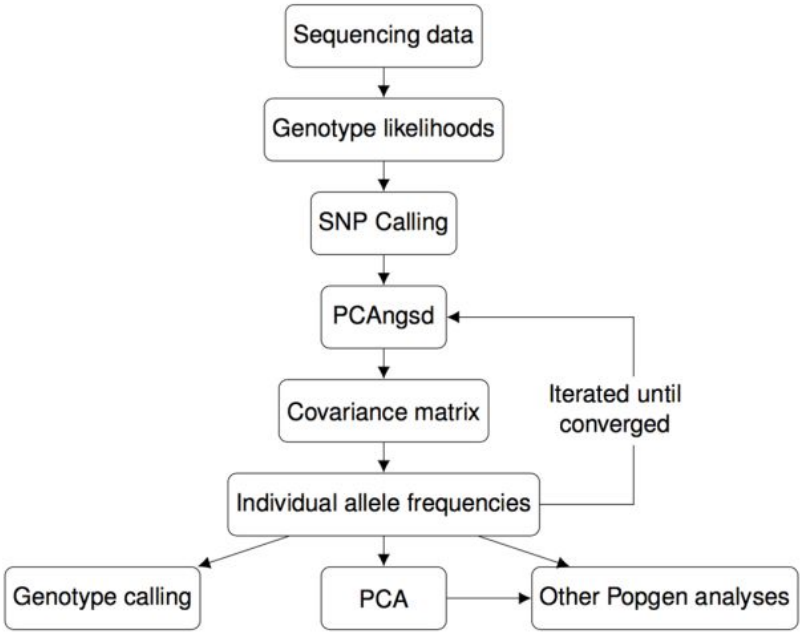
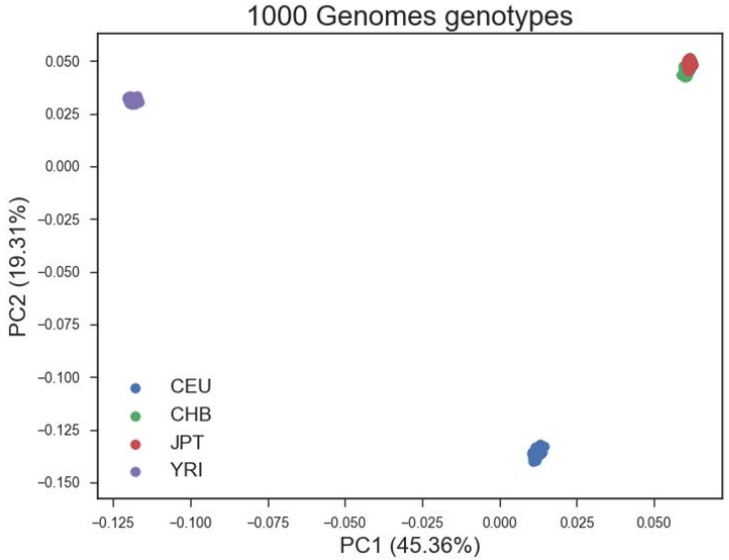
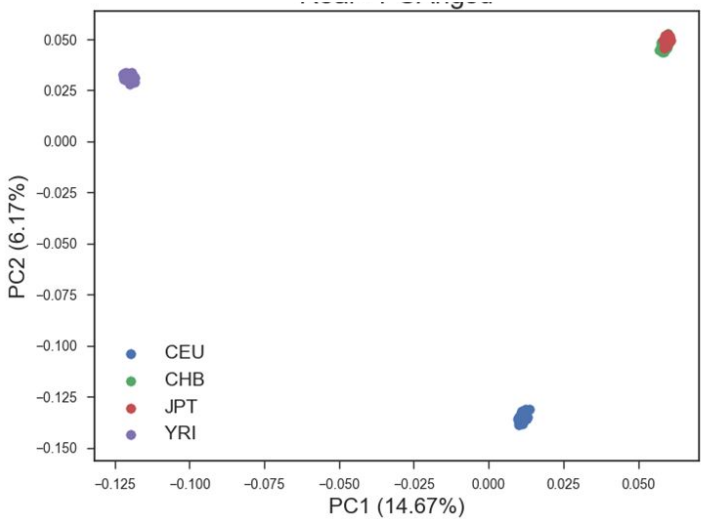


Figure: PCAngsd framework

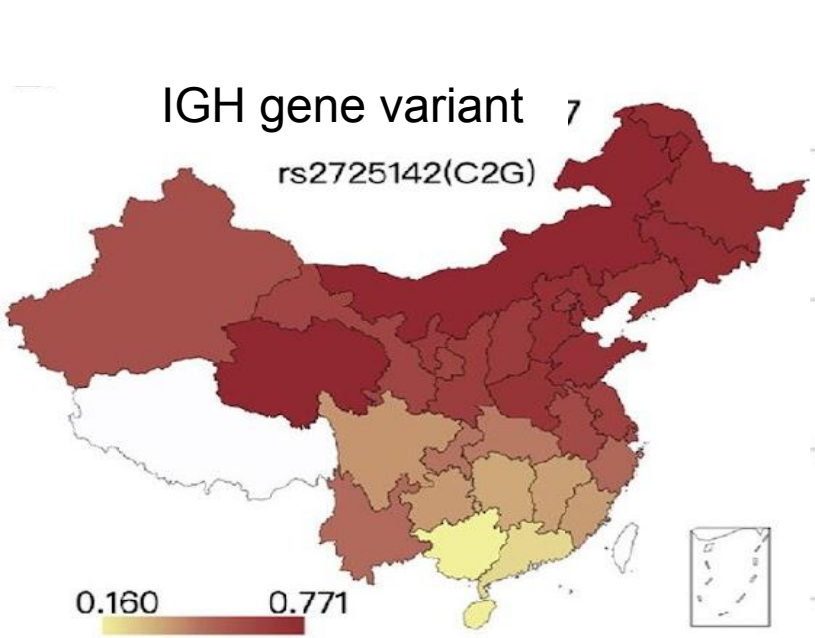
Issues with missingness

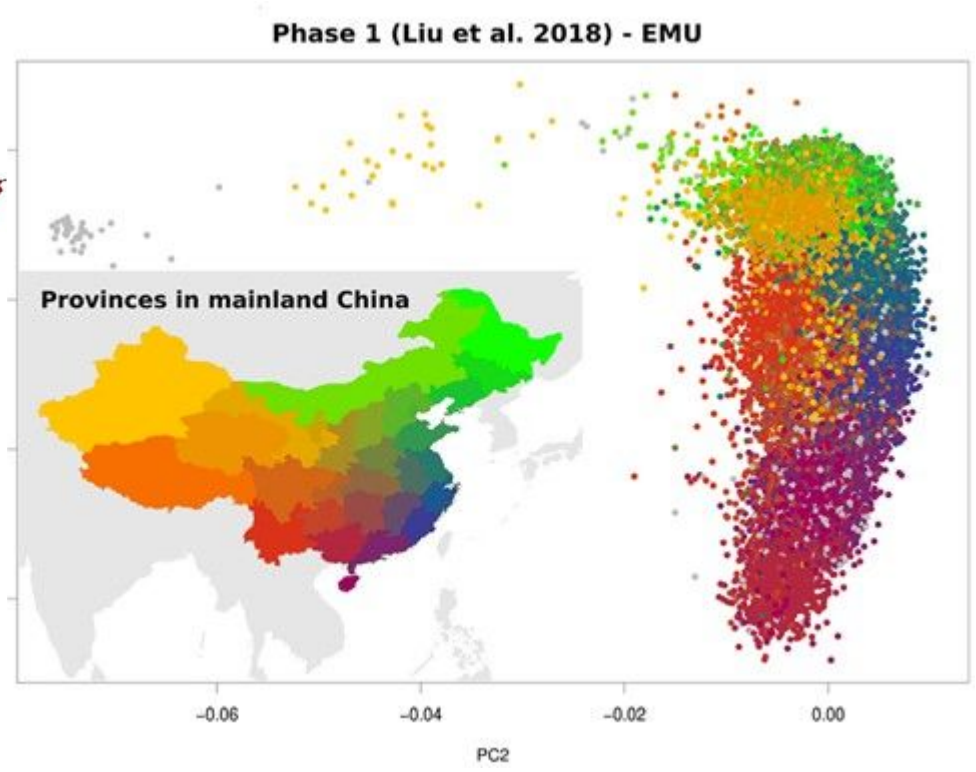


PCAngsd from low depth data (2-10X)

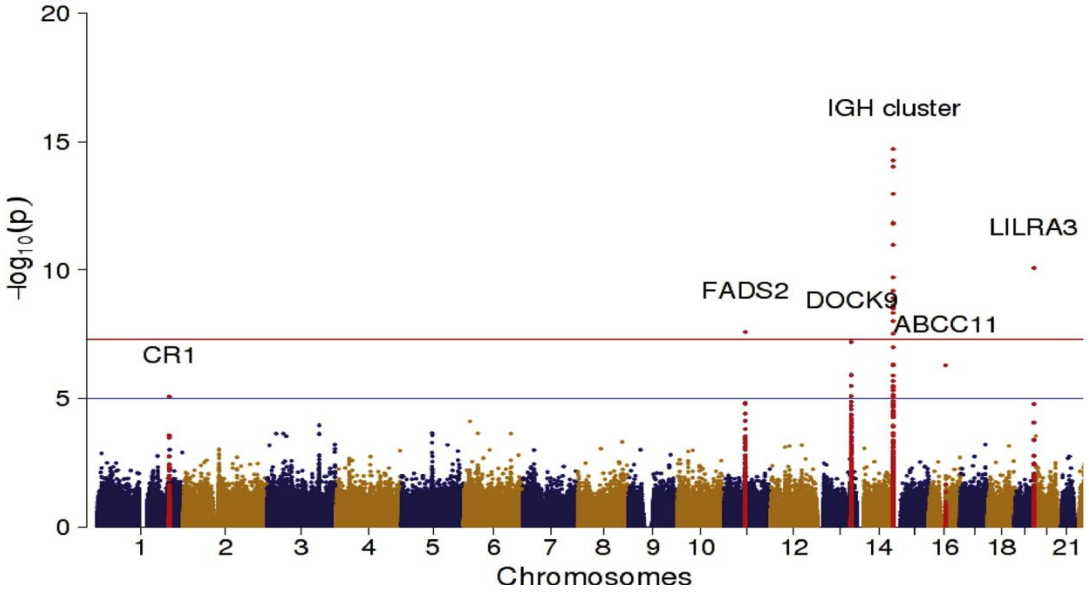


Selection - genotype correlated with PCs



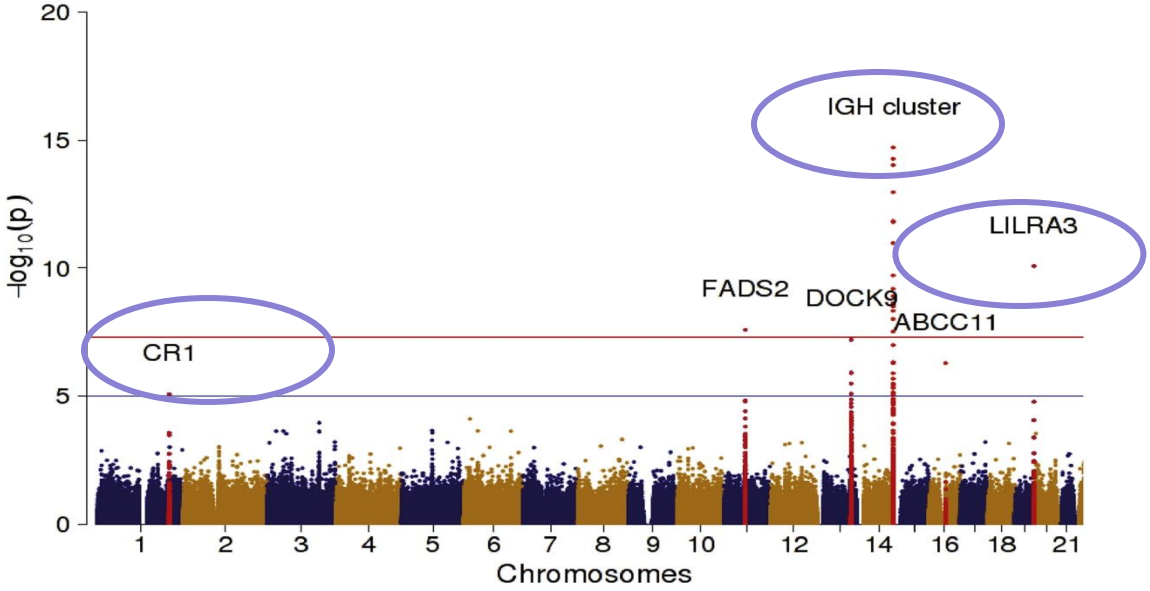


SELECTION SCAN - GWAS on first PC



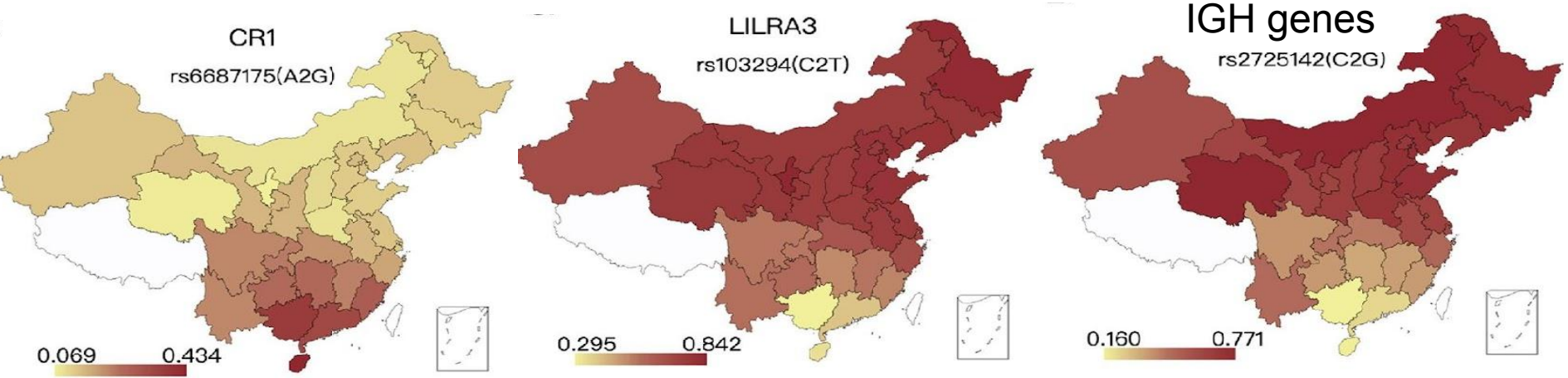
PC1

SELECTION SCAN



Immune genes previously
Shown to be under selection
In East Asians

IMMUNE RESPONSE GENES

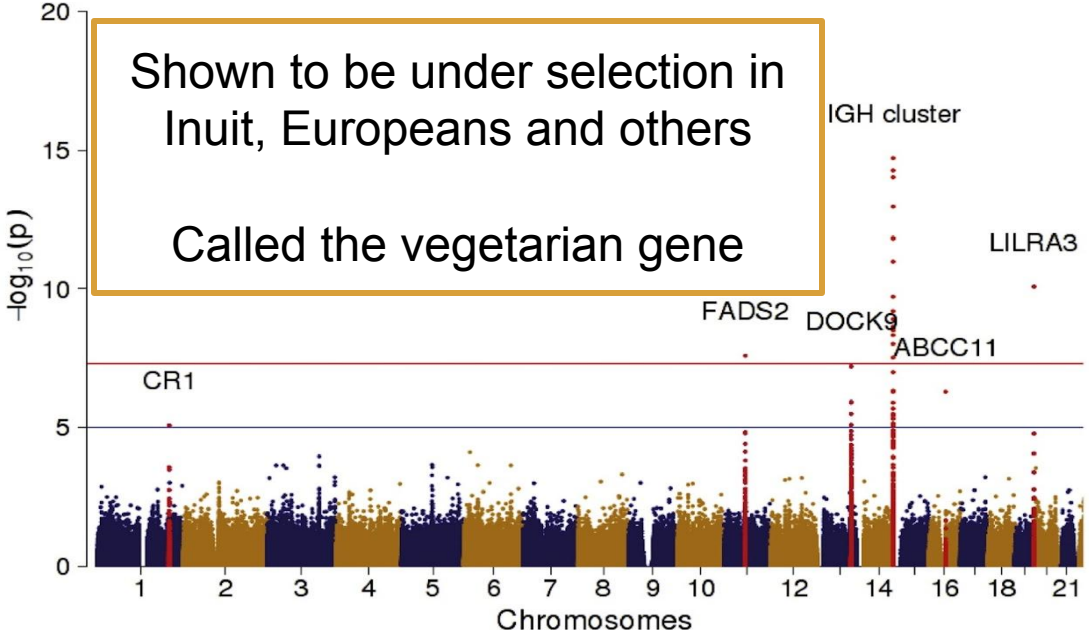


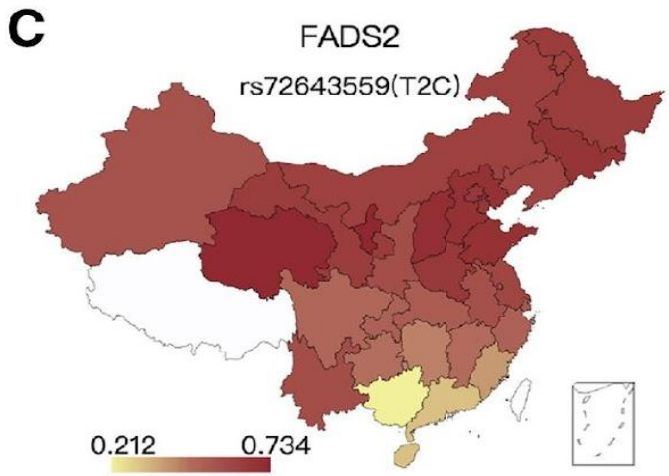
Derived Allele frequencies

FATTY ACID METABOLISM (FAD2)

Shown to be under selection in Inuit, Europeans and others

Called the vegetarian gene



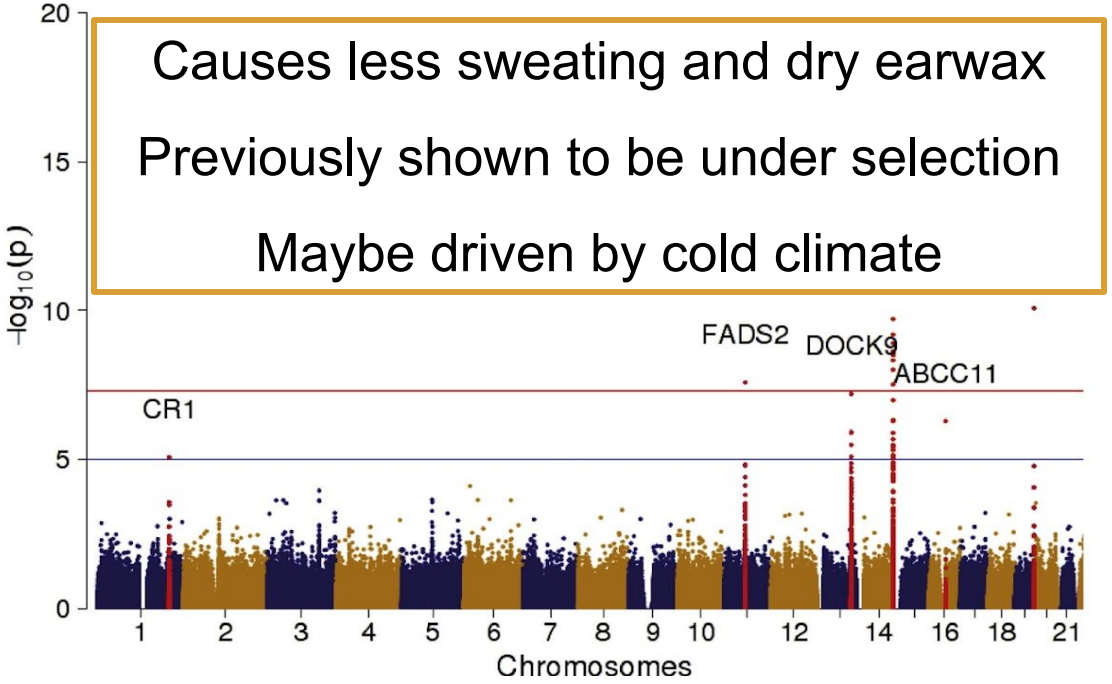


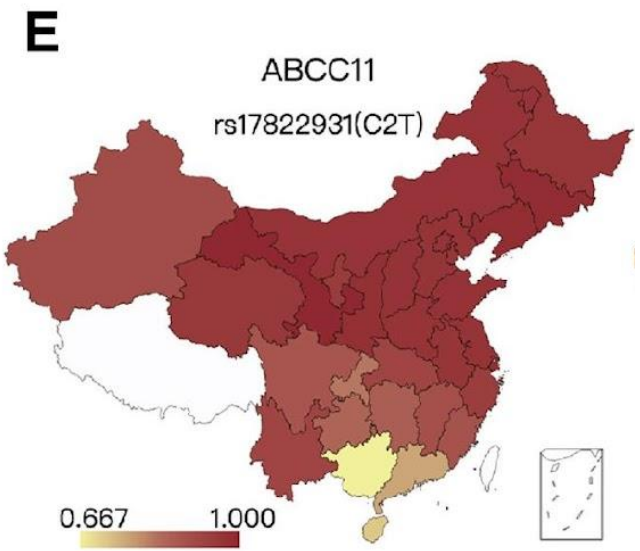
SWEAT AND EARWAX(ABCC11)

Causes less sweating and dry earwax

Previously shown to be under selection

Maybe driven by cold climate





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

Exercises - Selection scan in Europeans

go to

https://github.com/aalbrechtsen/embo2022

Group 1. Selection scan using Tajimas Pi

Group 2. Selection scan using EHH

Group 3. Selection scan using EP-EHH

Group 4. Selectin scan using PCA

Data sets

Selection scan using Tajimas Pi, EHH and EP-EHH

- Data is phased high quality haplotypes
  - VCF file with haplotypes

Selection scan using PCA

- Data is crappy low depth sequencing data
  - Beagle file with genotype likelihoods (no genotype calls)

PCAngsd and selScan

- High depth Sequencing data
- VCF file with genotype
- Phasing
- VCF with haplotypes
- SelScan
- (tajimas, EHH, XP-EHH

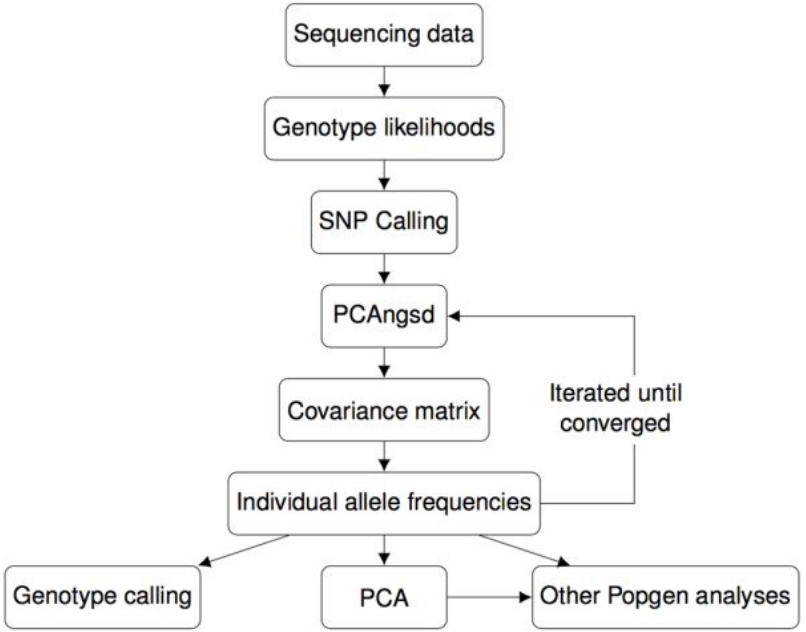


Figure: PCAngsd framework