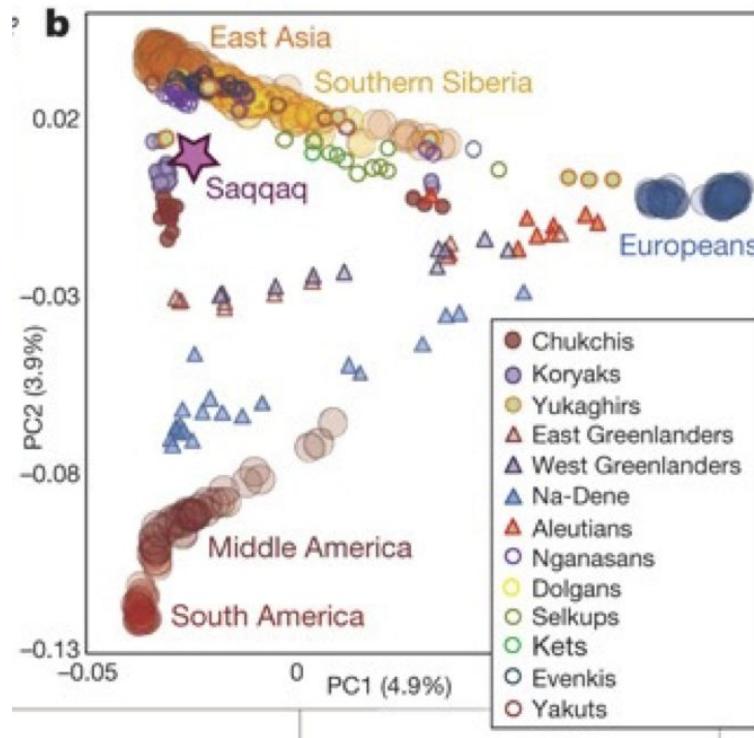


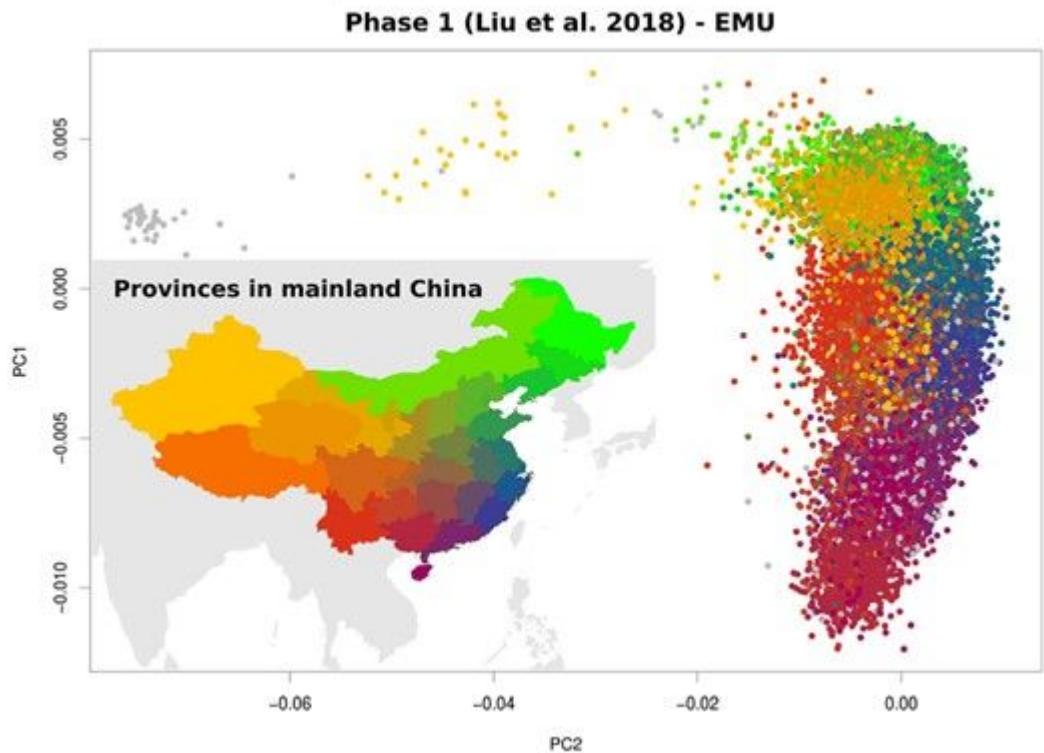
PCA and selection scan

Anders Albrechtsen

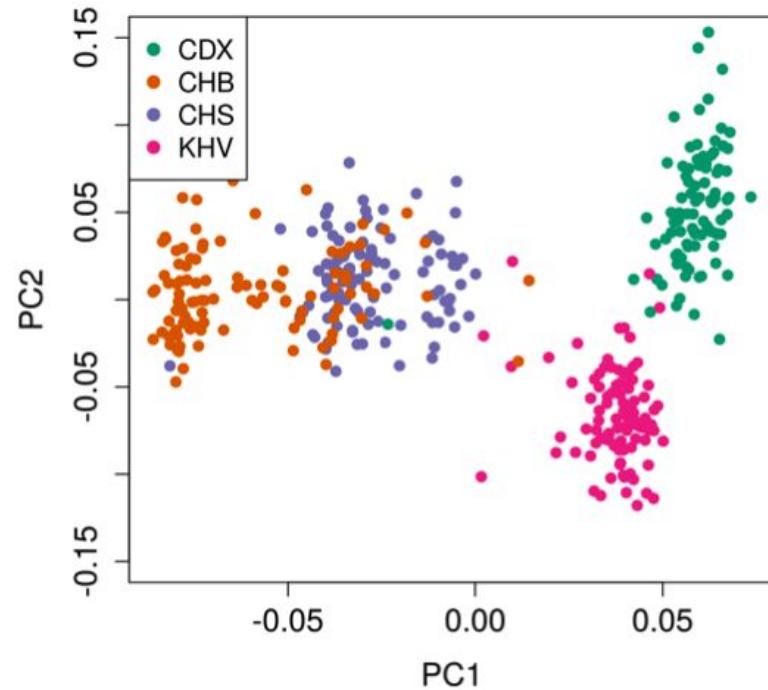
PCA for population structure



Continues structure / within population



	Ind1	Ind2	Ind3	Ind4	Ind5
SNPs	1	1	1	0	0
1	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 inds is a dot in pca plot

Genotype

left eigenvectors

singular values

right eigenvectors

G

$m \times n$

U

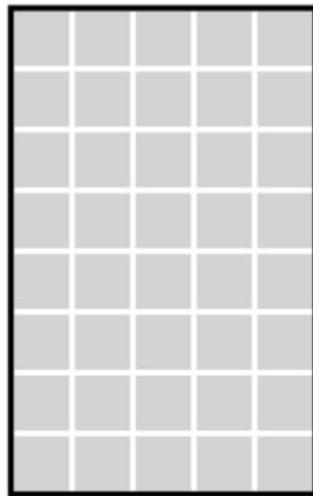
$m \times n$

Σ

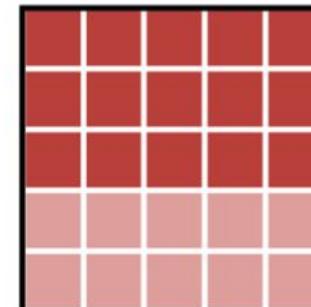
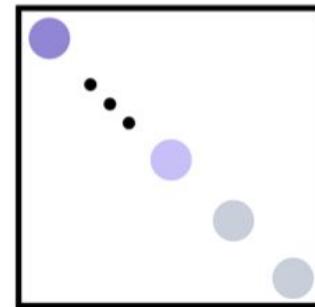
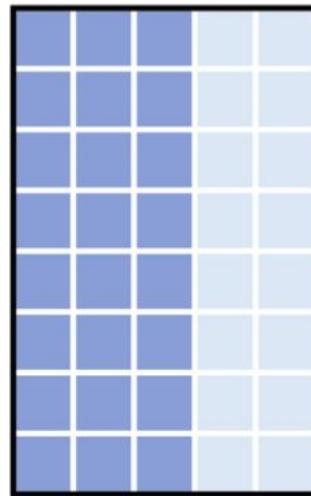
$n \times n$

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

Explain the approximation here

SVD - singular value decomposition

$$G = UDV^T$$

- G (genotype matrix) does not have to be symmetric

PCA for a covariance matrix (C) or pairwise distance

$$C = V\sqrt{D}V^T$$

- The first principal component/eigenvector accounts for as much of the variability in the data as possible
- C is symmetric
- Optimally the multidimensional data is identically distributed

PCA and distances

individuals

SNP	Ind1	Ind2	Ind3	Ind4	Ind5
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

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

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

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

Dealing with missingness

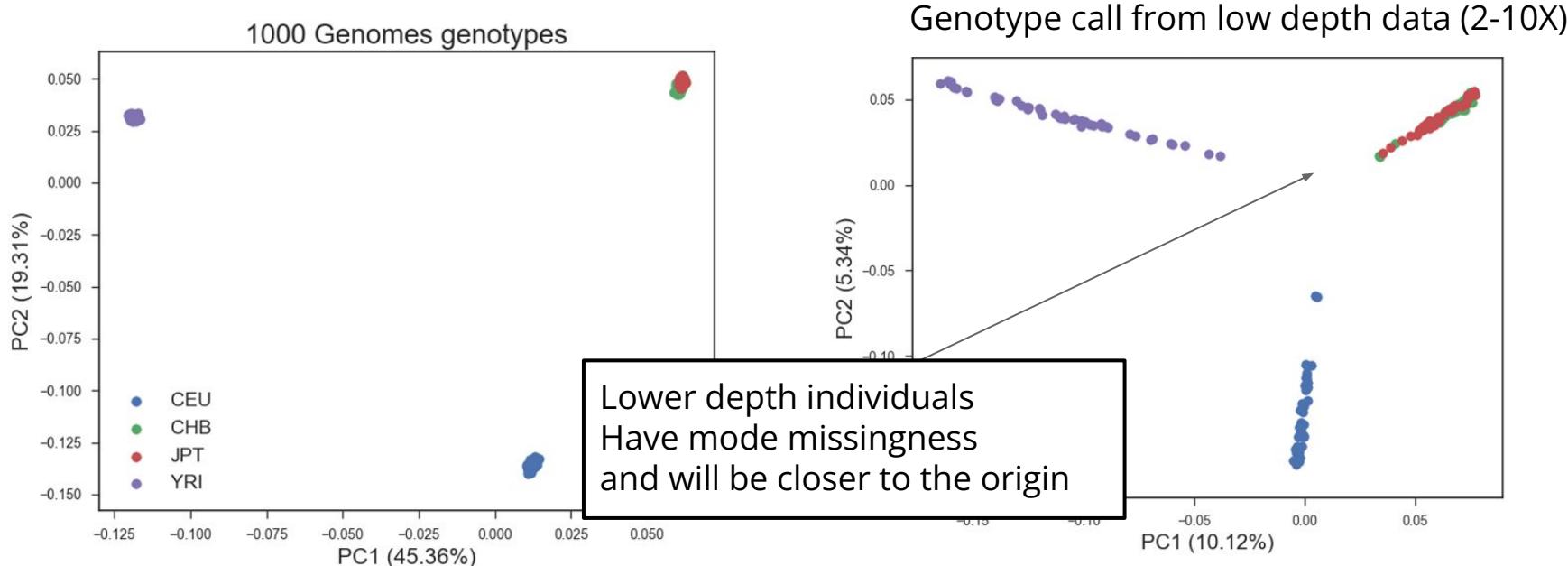
If a genotype is missing then \tilde{G}_k^i is set to zero

- $E[\tilde{G}_k^i] = 0$ 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

Issues with missingness/low depth data



D

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

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



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

E_0

0	1	0	1	0
0	0	0	1	0
0	0	0	0	0
0	0	1	0	1

E_1

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

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

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	0	0.9	0.1
0	0.9	1	0.3	1

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

 $\hat{\Pi}$

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

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

The Idea of PCangsd/EMU

EMU - PCA

D is the data matrix with missing information.

E is the imputed data matrix which can be decomposed as

The missing data is imputed from the first K PCs

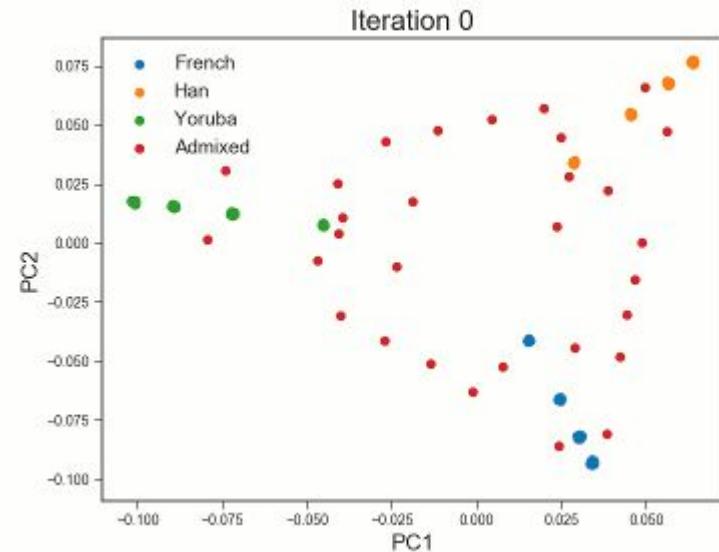
$$\mathbf{E} = \mathbf{WSU}^T$$

From the PCs that minimize the nonmissing part

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

Where C is a matrix with 0 for missing and 1 otherwise

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



Low depth sequencing

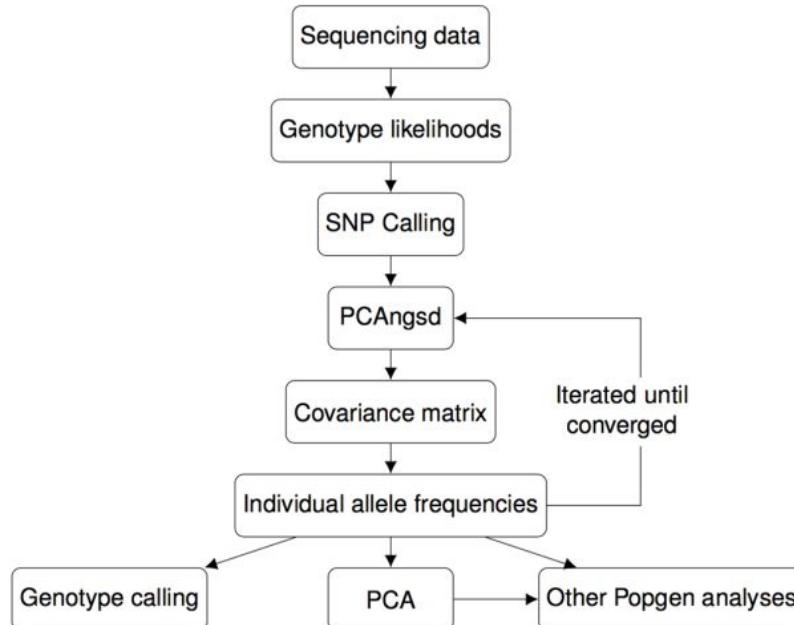
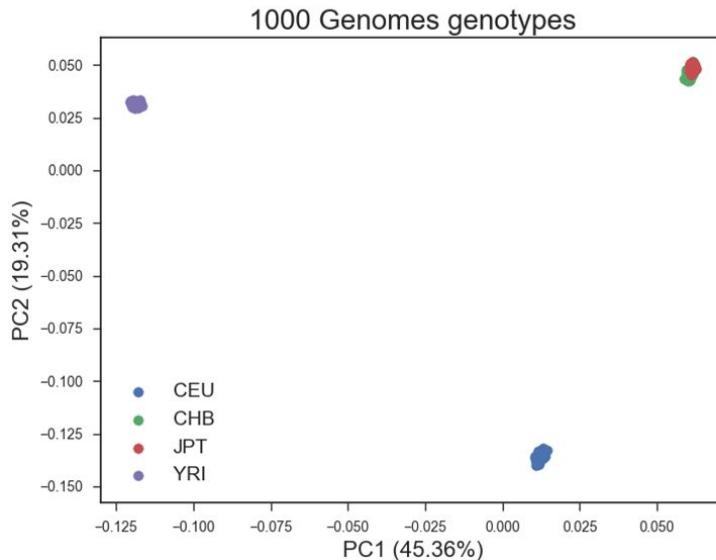
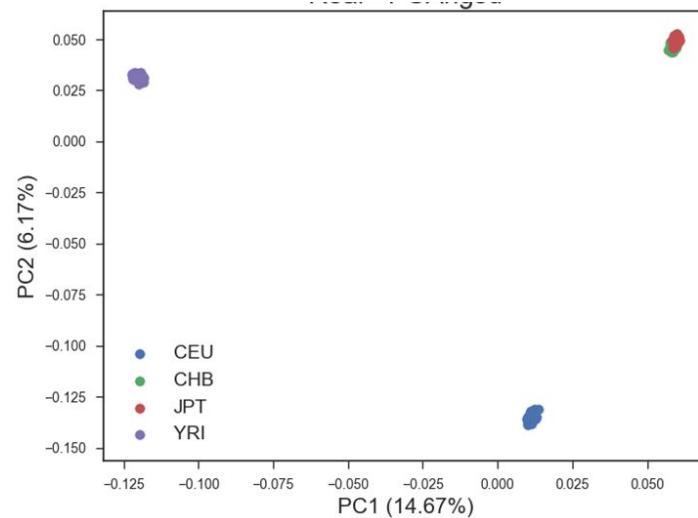


Figure: PCAngsd framework

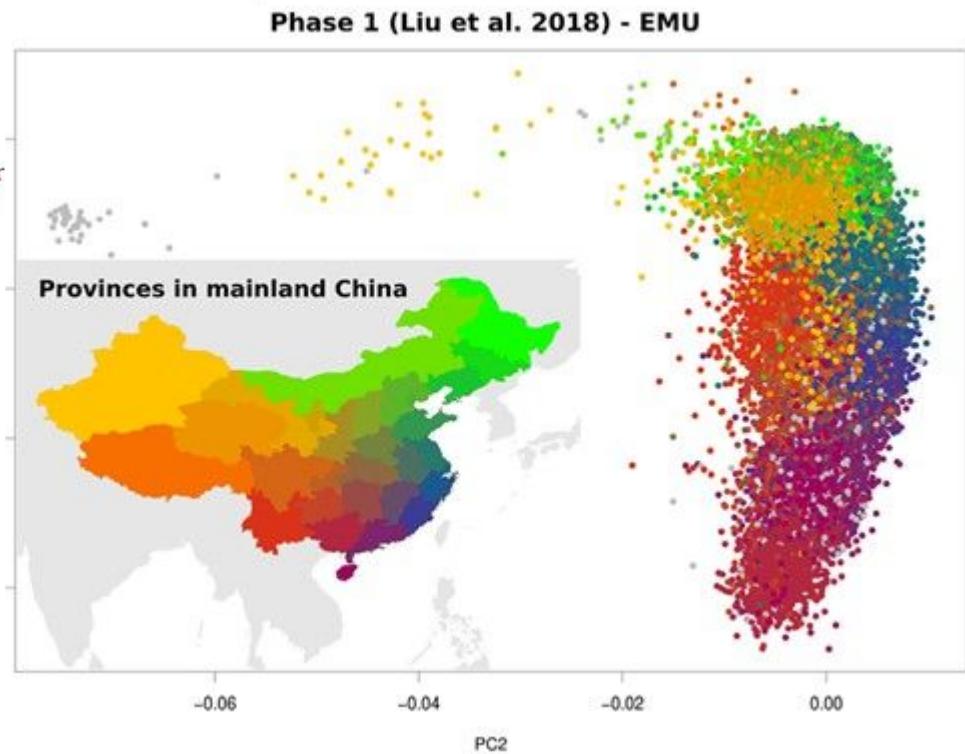
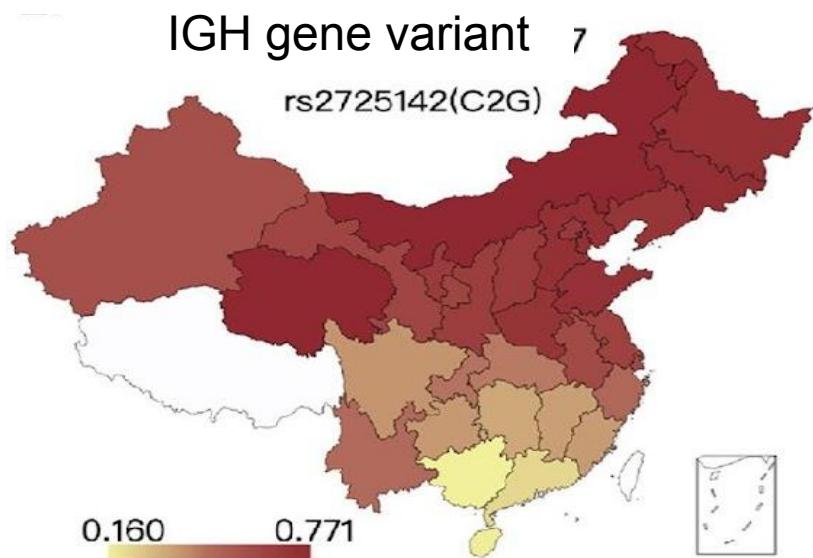
Issues with missingness



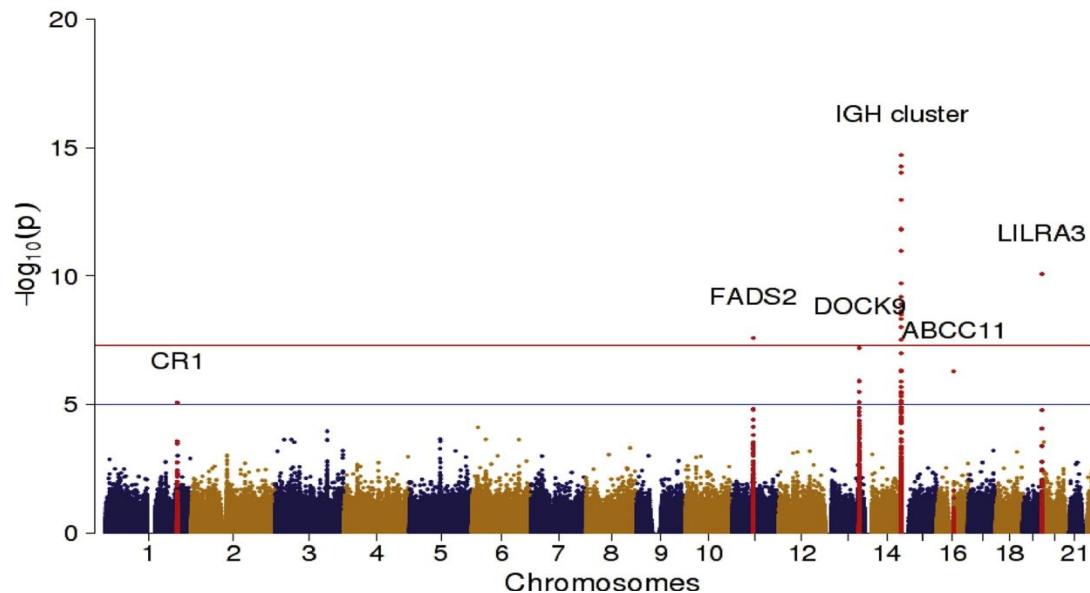
PCAngsd from low depth data (2-10X)



Selection - genotype correlated with PCs

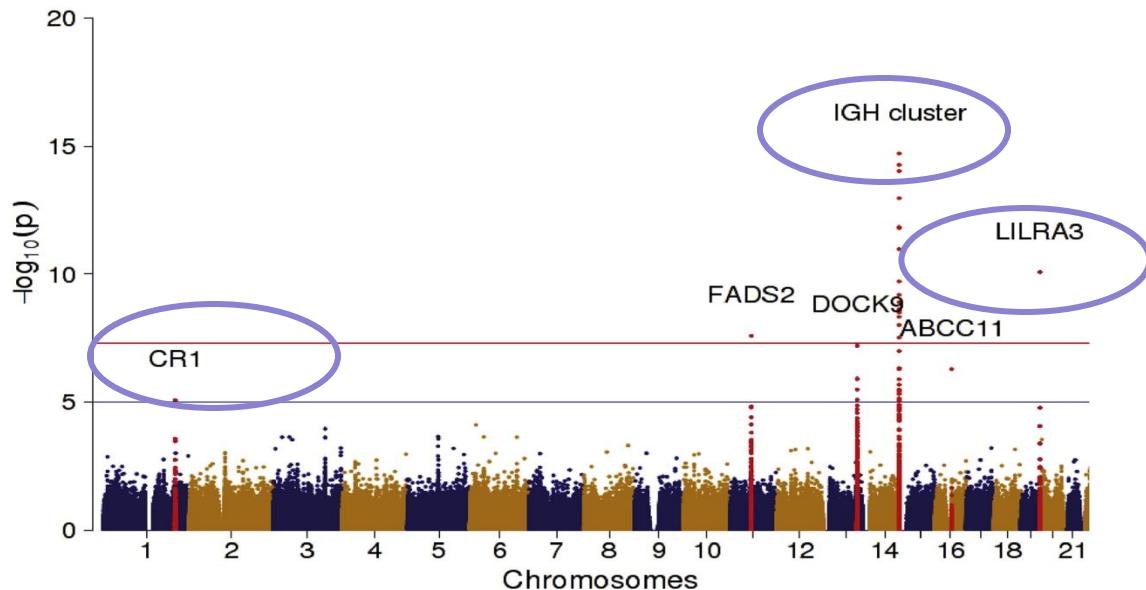


SELECTION SCAN - GWAS on first PC



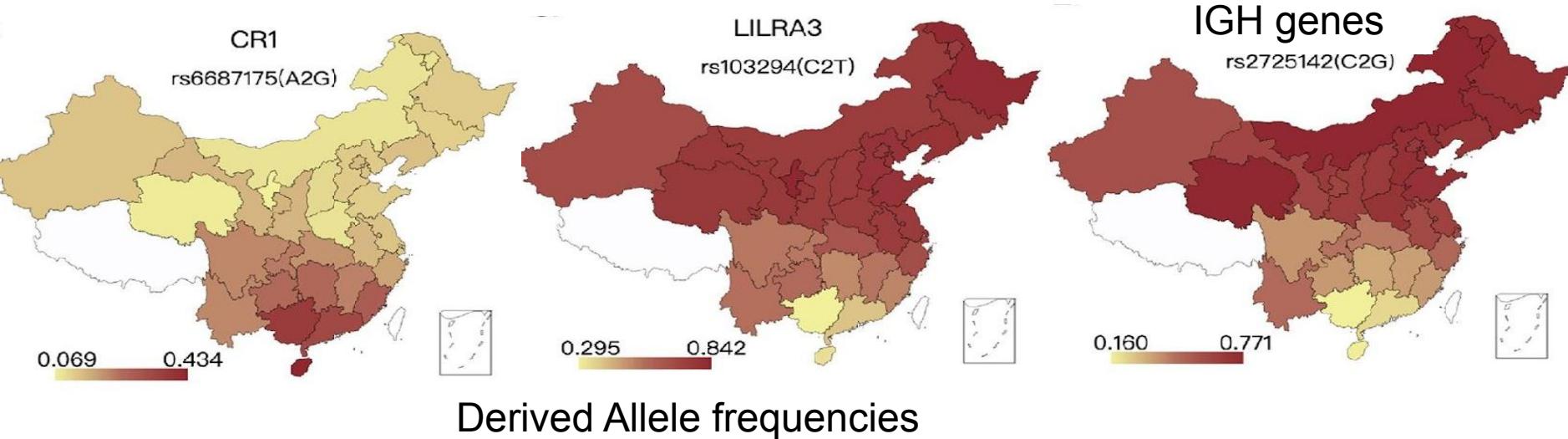
PC1

SELECTION SCAN

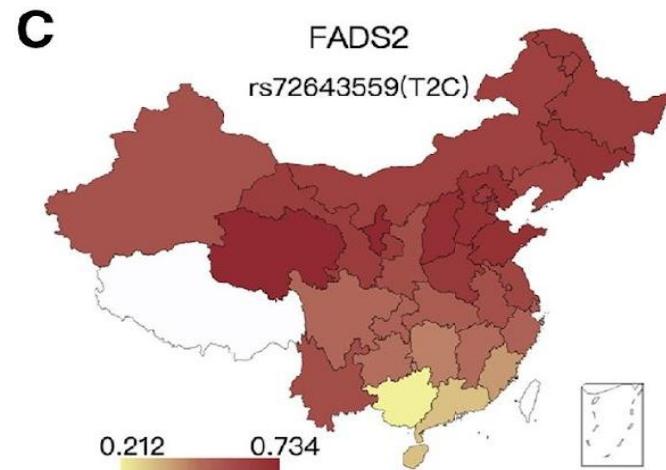
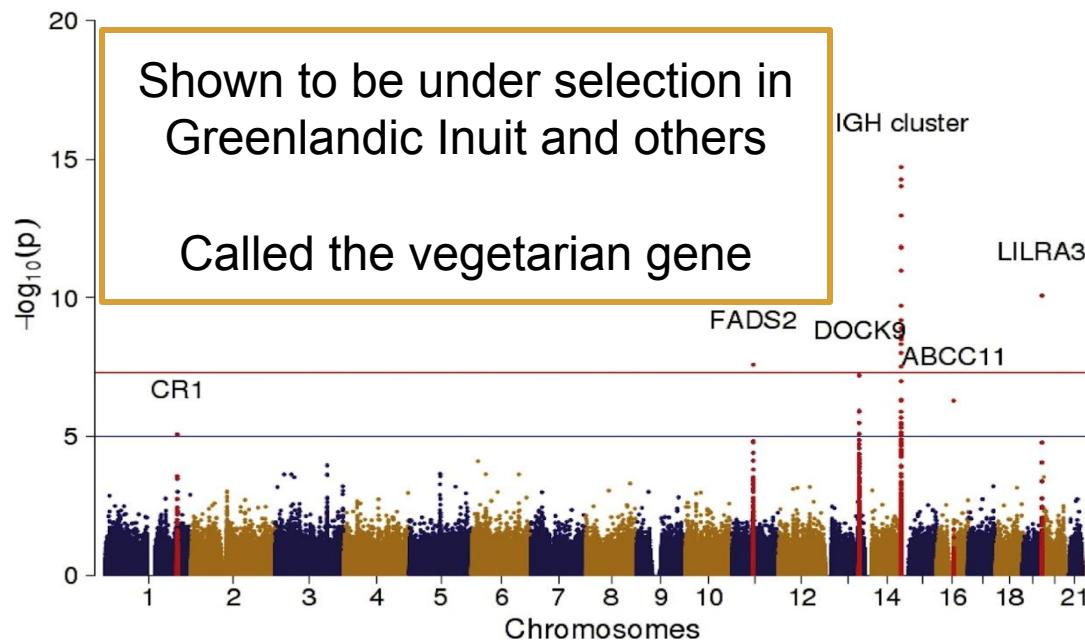


Immune genes
previously
Shown to be under
selection
In East Asians

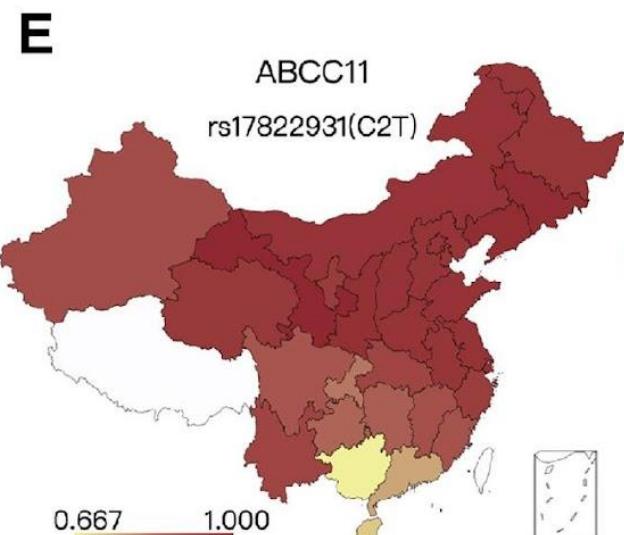
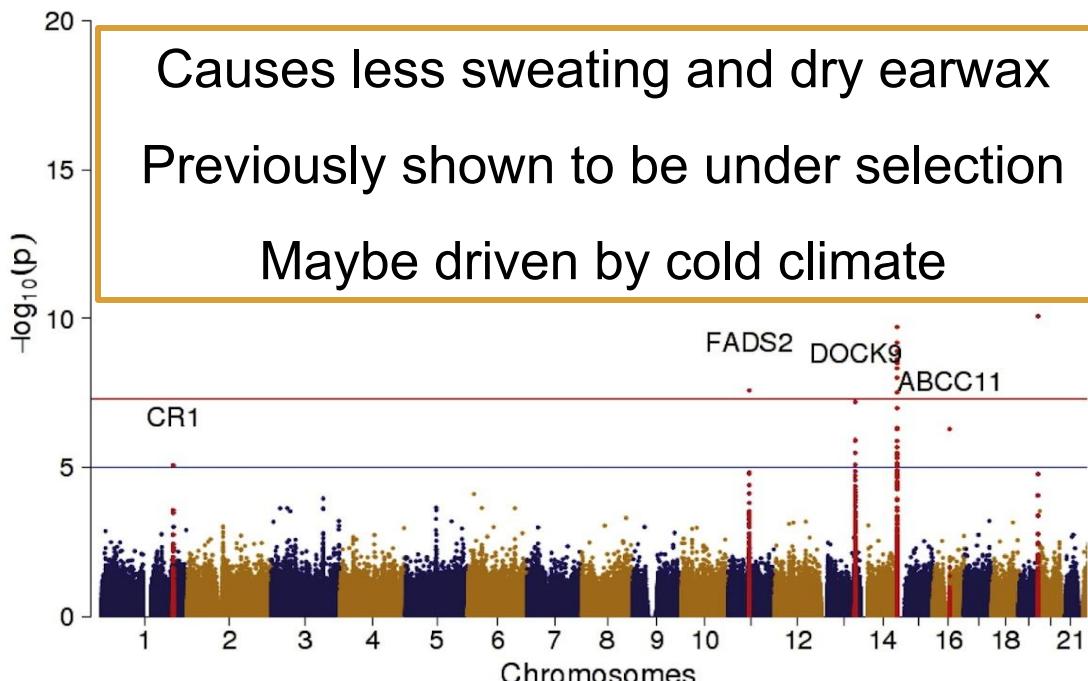
IMMUNE RESPONSE GENES



FATTY ACID METABOLISM (FAD2)



SWEAT AND EARWAX(ABCC11)



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

go to

<https://github.com/aalbrechtsen/embo2022>