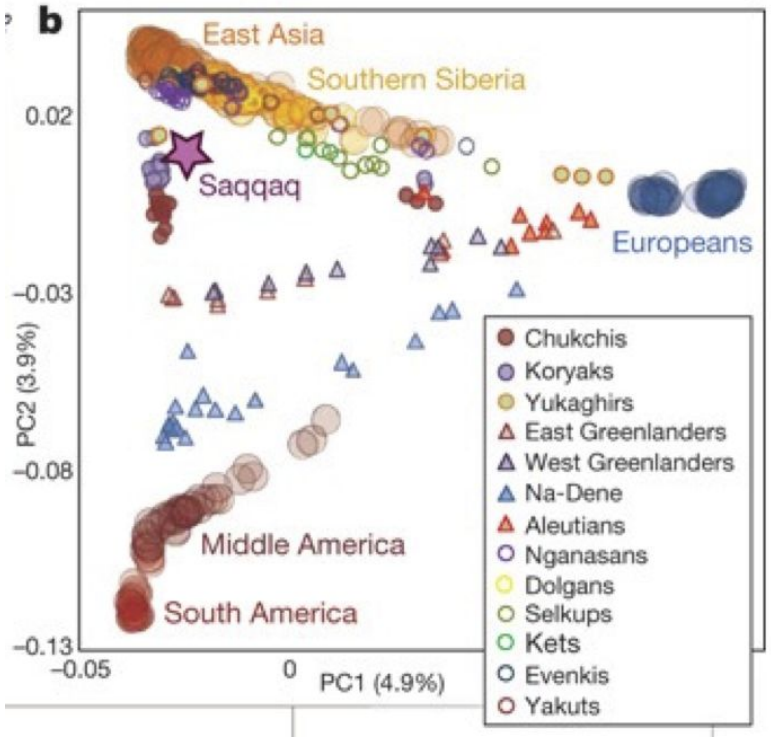
Population structure II
PCA

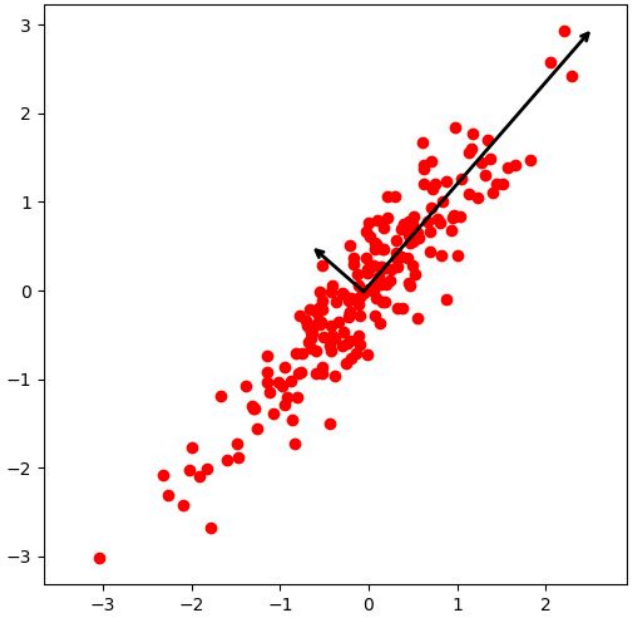
Anders Albrechtsen

PCA for population structure



Principal component analysis (PCA)

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

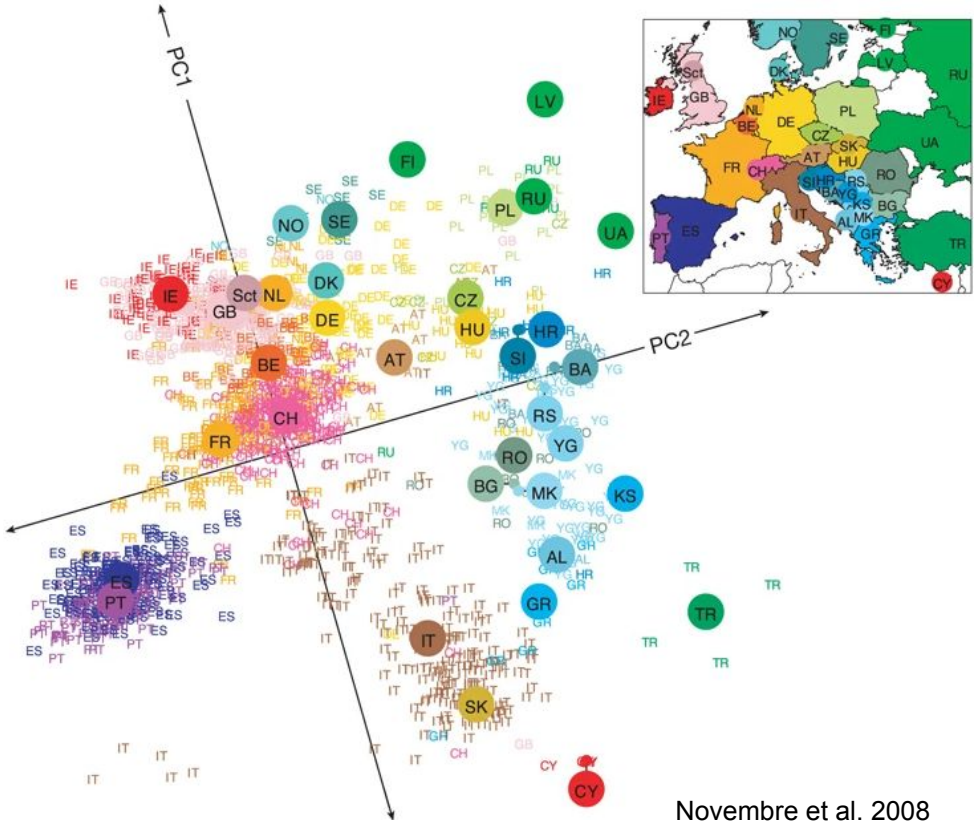


***not really true**

Principal component analysis (PCA)

Genetic data

- $m > 1$ million
- Captures genetic structure



Novembre et al. 2008

Today you will learn

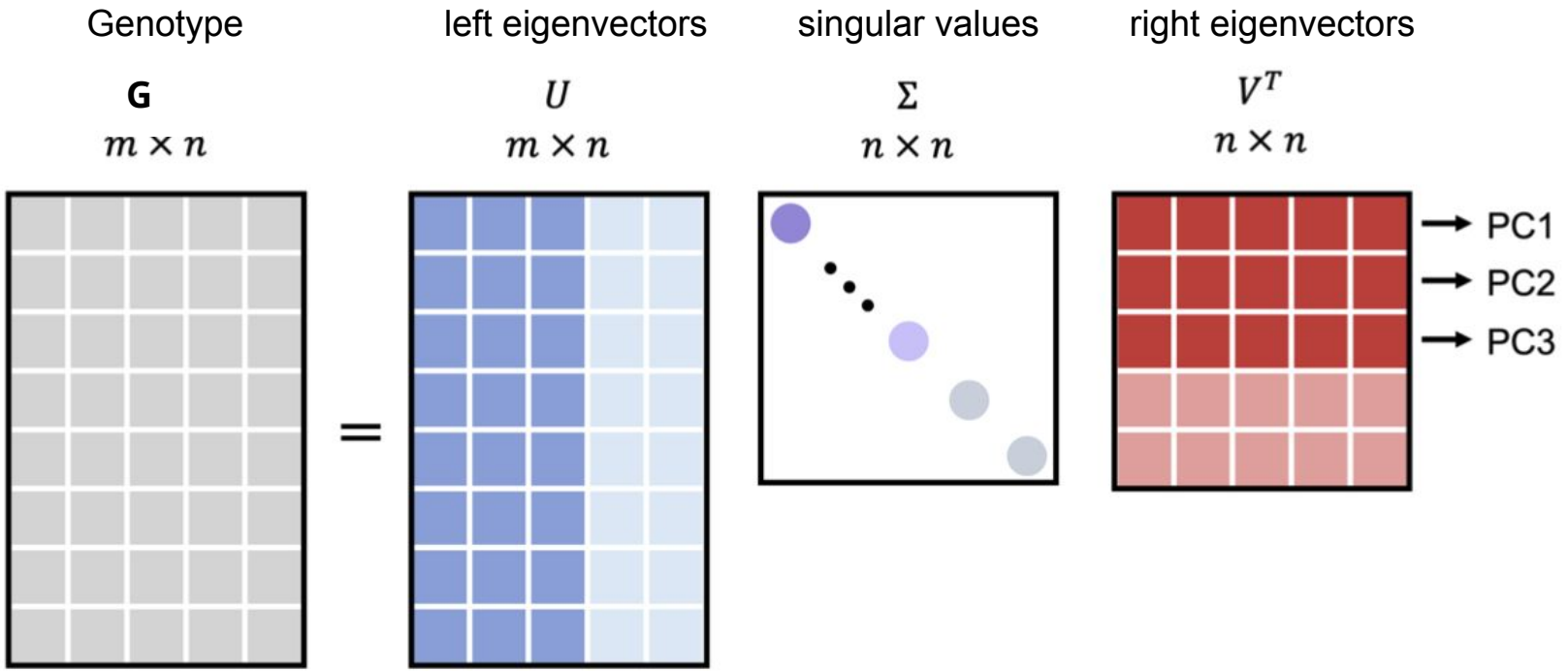
- The underlying “model” of PCA and MDS
  - What these two methods are trying to achieve
- The relationship between admixture proportions and PCA
- How PCA predict genotypes
- Issues with missingness
  - For call genotypes and for low depth sequencing
- How to deal with missingness
- How PCA can be used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

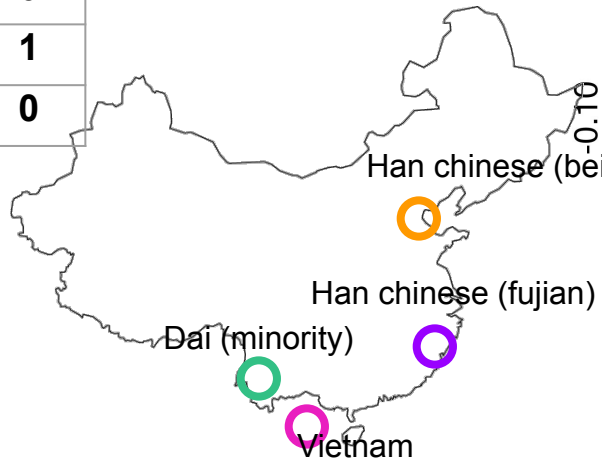
SNPs

Ind_1	Ind_2	Ind_3	Ind_4	Ind_5
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

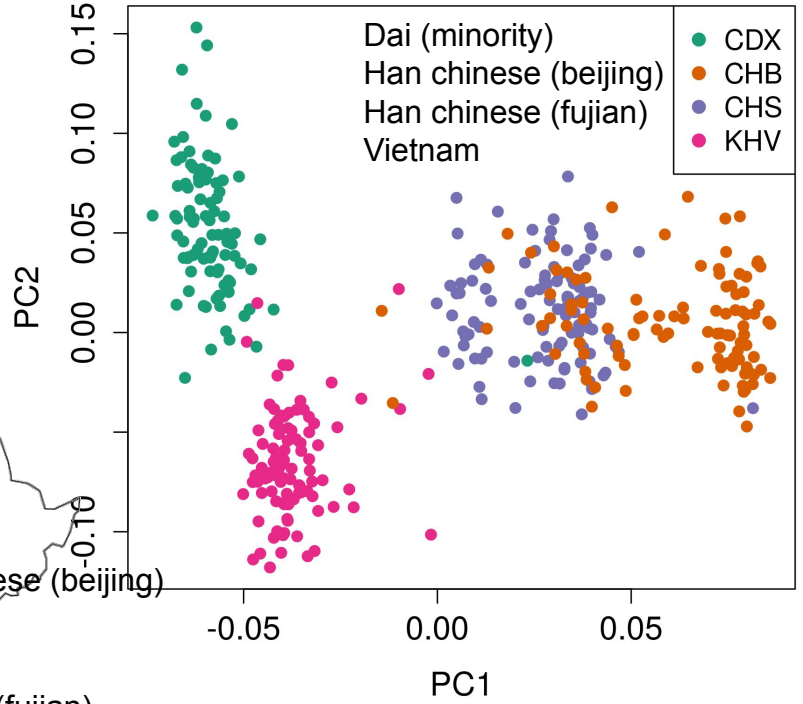


G is a genotype matrix, n is the number of samples, m is the number of SNPs

SNPs	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



PCA of 1000G Asian Populations



Each individuals is a dot in PCA plot

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_1	Ind_2	Ind_3	Ind_4	Ind_5
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

Pairwise distance

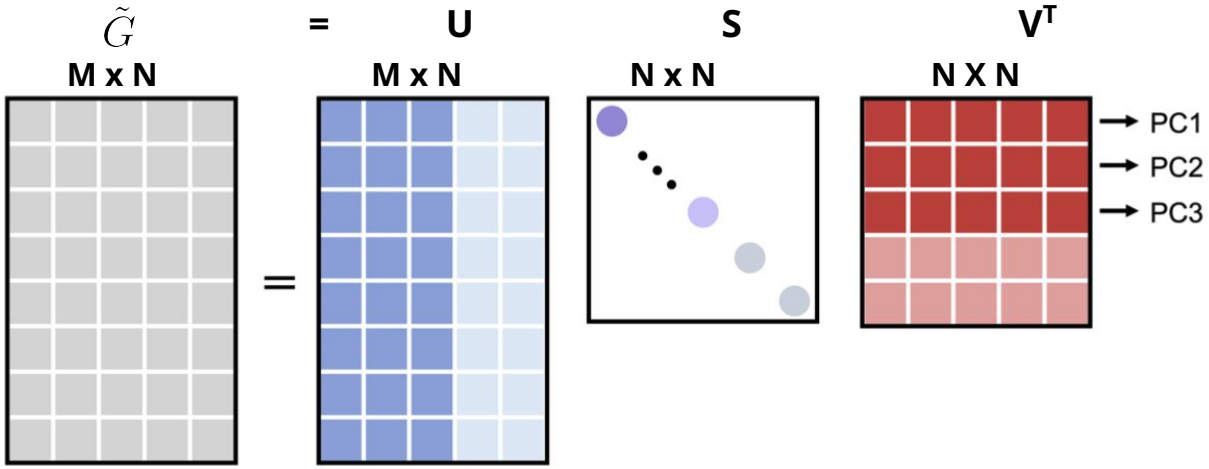
Manhattan distance

	Ind_1	Ind_2	Ind_3	Ind_4	Ind_5
Ind_1	0	3	7	10	11
Ind_2	3	0	4	7	8
Ind_3	7	4	0	5	4
Ind_4	10	7	5	0	6
Ind_5	11	8	4	3	0

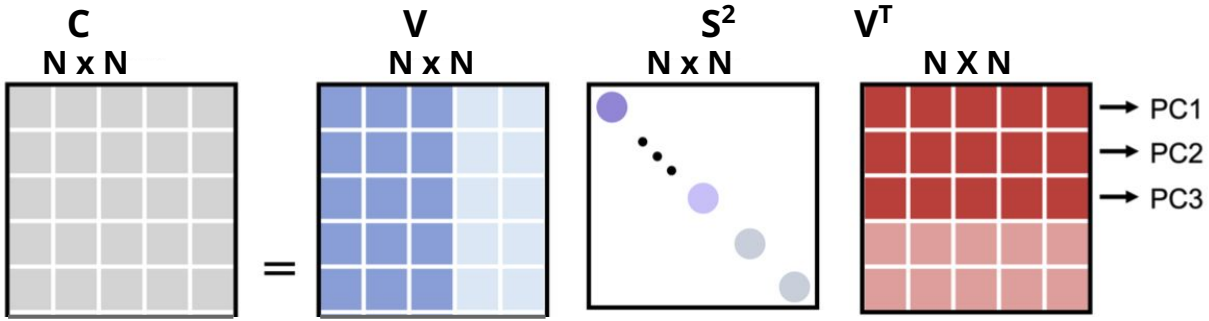
Project into 1 dimension

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

Two ways to do PCA



Directly on the genotypes



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

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

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

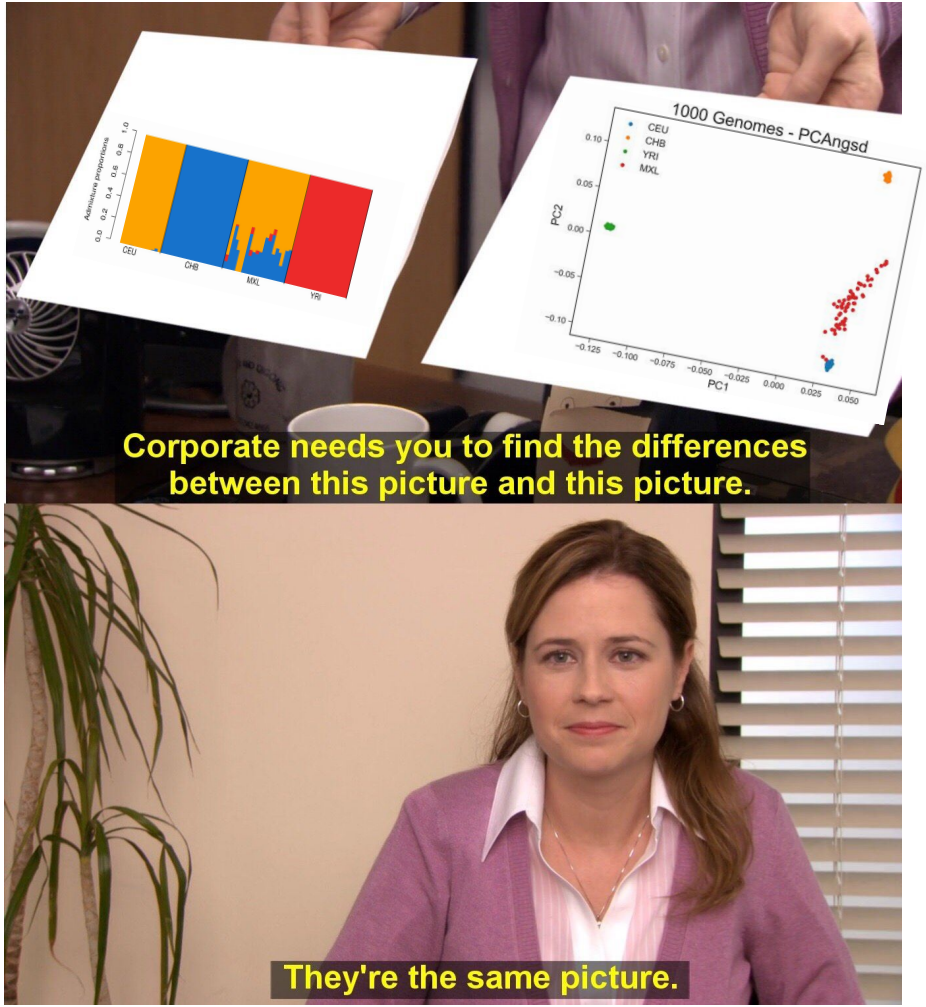
$$\tilde{G}_{ij} = \frac{G_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}$$

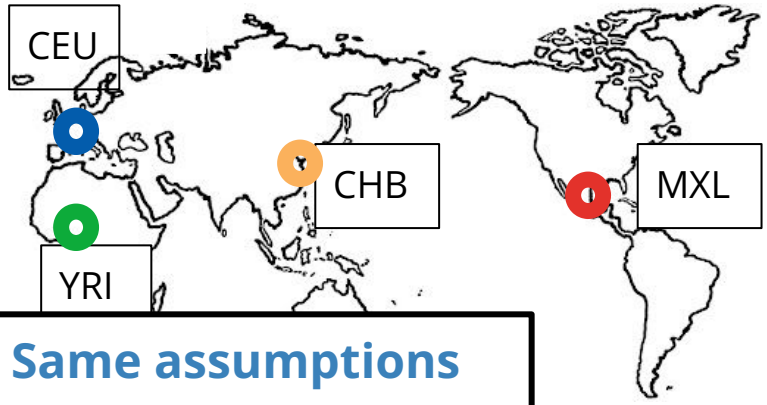
$$var(G_{ij}) = 2f_j(1 - f_j)$$

After normalization all SNPs have the same mean and variance

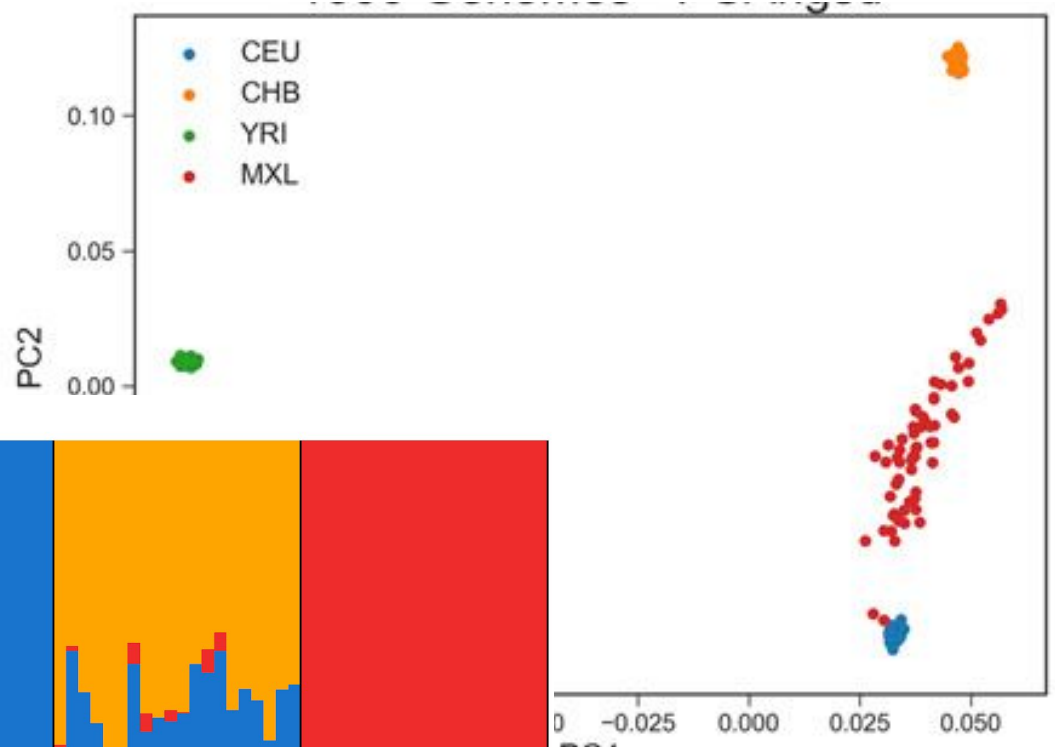
$$cov(\tilde{G}_i, \tilde{G}_l) = \frac{1}{M} \sum_{j=1}^{M} \frac{(G_{ij} - 2f_j)(G_{lj} - 2f_j)}{2f_j(1-f_j)} = \frac{1}{M} \tilde{G}\tilde{G}^T$$

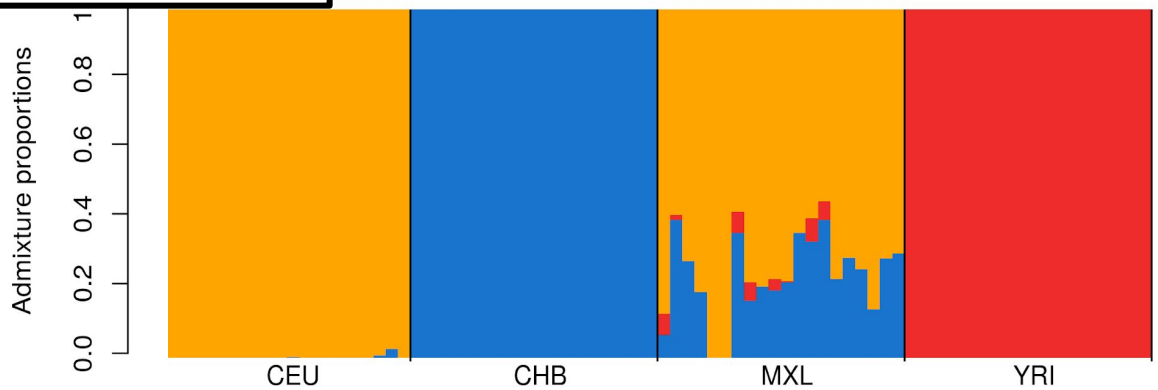
Connection between Admixture analysis and PCA





Same assumptions
=
same issues





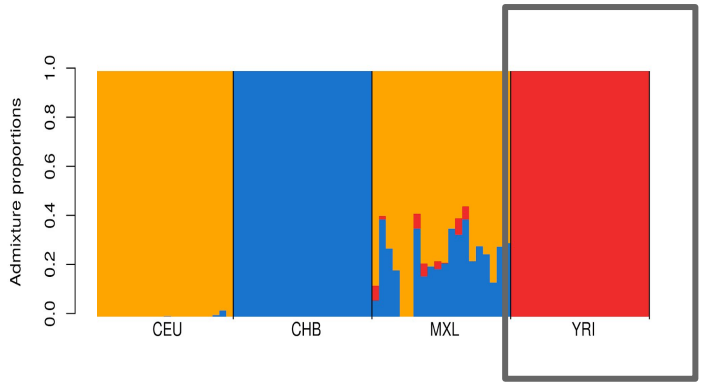
EvalAdmix

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

- Individu

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

$$p(g_i) =$$

$$IB)$$

Individual allele frequencies

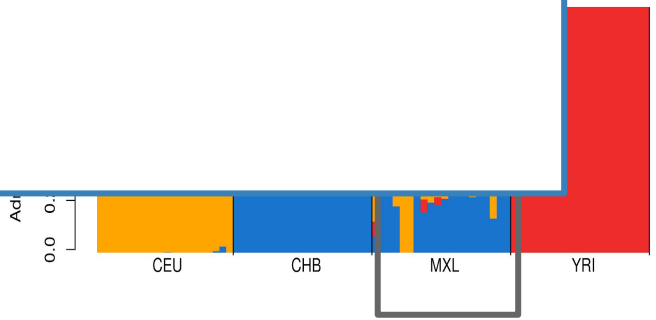
$$\pi_{ij} = \sum_{k=1}^{K} f_{jk} q_{ik}$$

$K = \#\text{populations}$

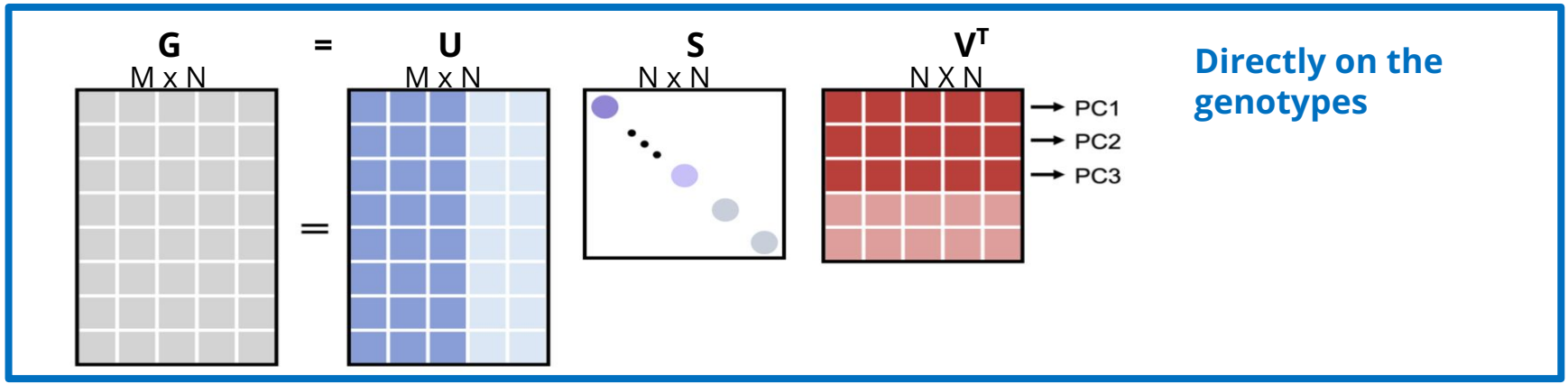
$i = \text{individual i}$

$j = \text{site j}$

$\Pi = QF$

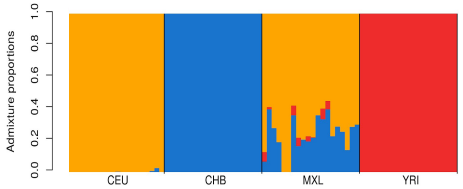


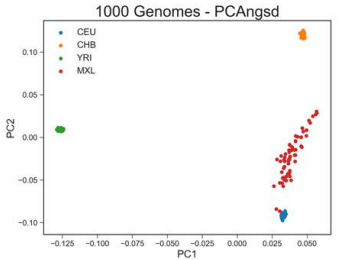
Individual allele frequencies from PCA



Individual allele frequencies

- Allele frequency: $\mathbb{E}[g] = 2p$
- Low-rank approximation
  - Admixture $\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{QF}$
  - PCA
    Truncated
    SVD $\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{U}_{[1:k]}\mathbf{S}_{[1:k]}\mathbf{V}^T_{[1:k]}$





Admixture and PCA from Π

Π is the matrix of individual frequencies

ADMIXTURE → PCA

$$cov(\tilde{G}_i, \tilde{G}_l)$$

$$\approx \frac{1}{M} \sum_{j=1}^{M} \frac{(\Pi_{ij} - f_j)(\Pi_{lj} - f_j)}{f_j(1-f_j)}$$

$$\approx \frac{1}{M} \tilde{G}\tilde{G}^T$$

PCA → 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

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

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)

G

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)

na -> 2f

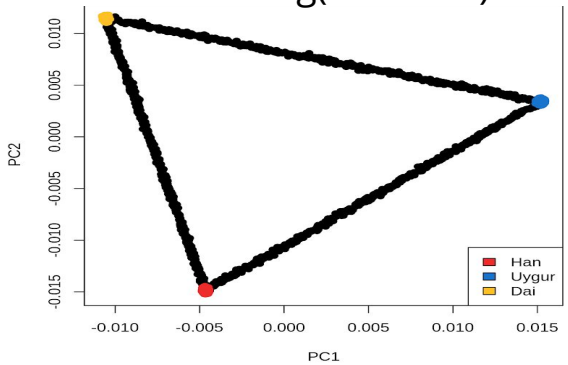
update grey part Based on allele frequencies

$\hat{E}$

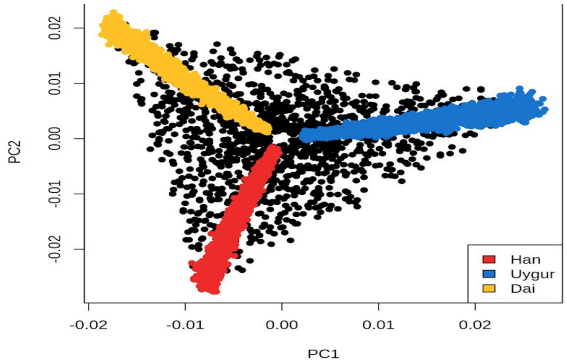
0	1	1.5	1	0.5
0.5	0	1.5	1	0
0.5	0.5	2	1	0.5
1	0.5	1	1	1

Simulated 90-99% missingness

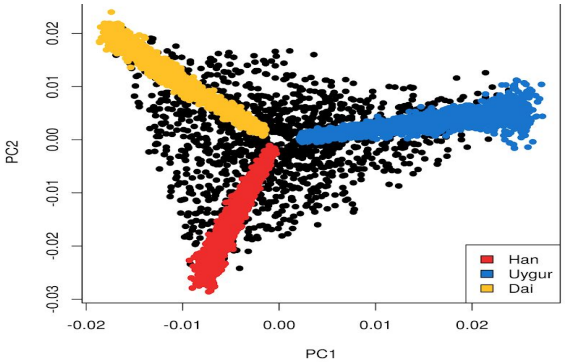
No missing(true PCA)



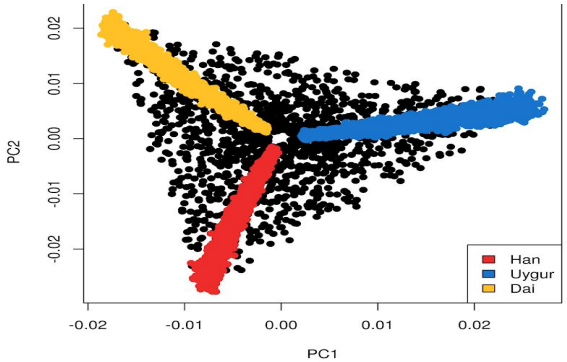
SmartPCA/eigensoft



FastPCA / plink-fast



FlashPCA



How to solve it - EMU

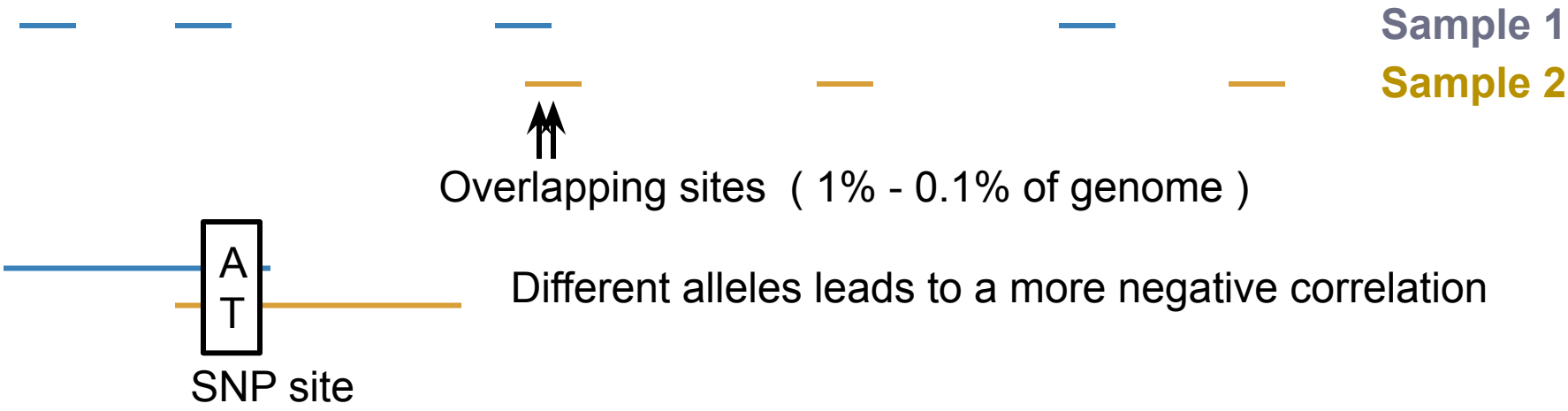
If we have the PCA we can calculate the individual allele frequencies

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

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

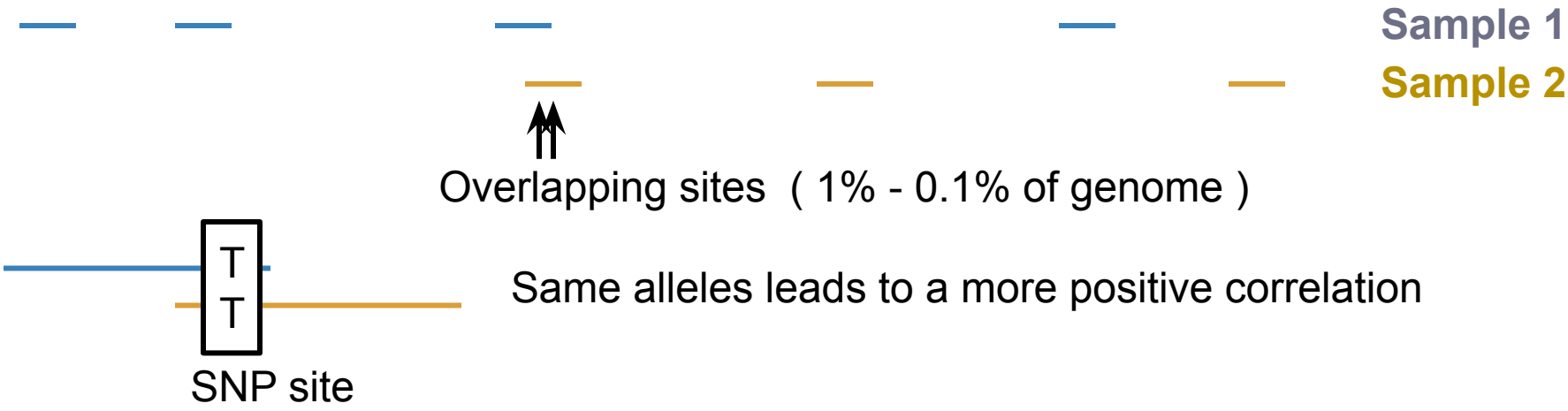


FIRST APPROACH

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

1) Estimate allele frequencies for each SNP

2) For each pair of individual estimate the covariance for overlapping sites



Second approach - EMU

D

0	1	na	1	na
na	0	na	1	0
na	na	2	na	na
1	na	1	na	1

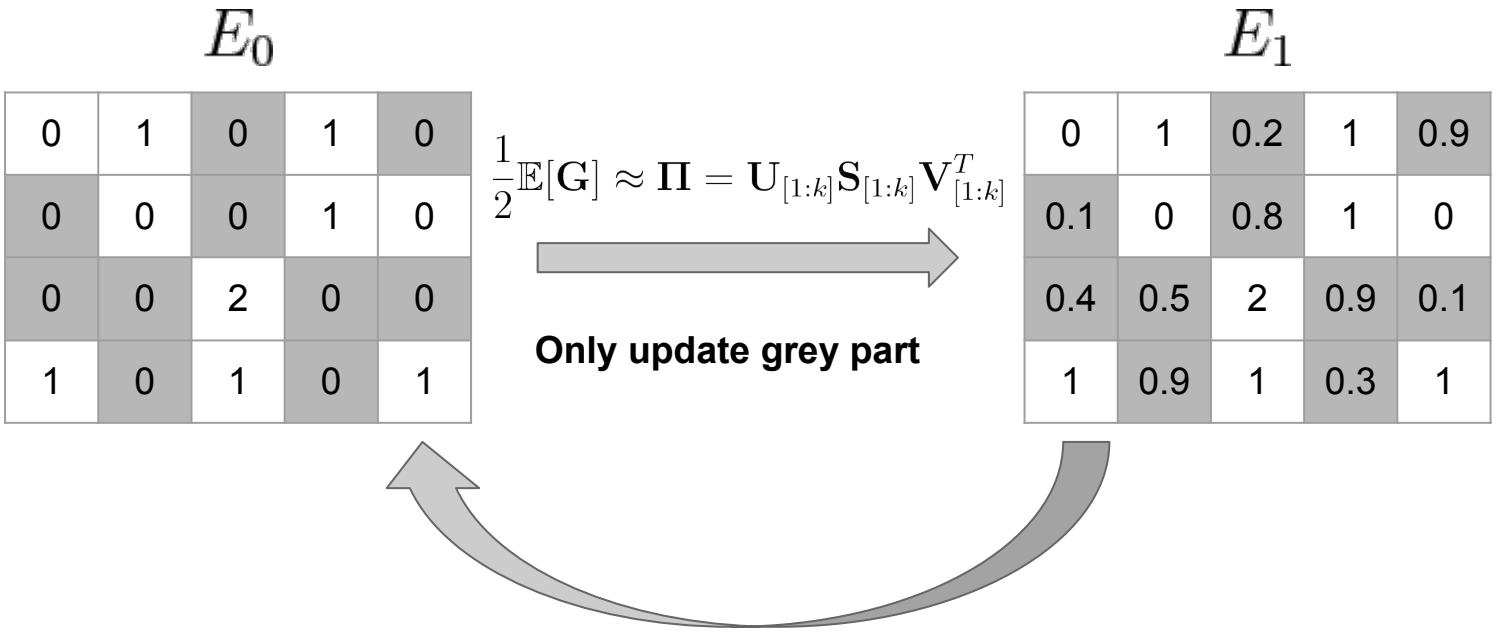
PCA to predict data

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

Only update grey part

$\hat{E}$

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

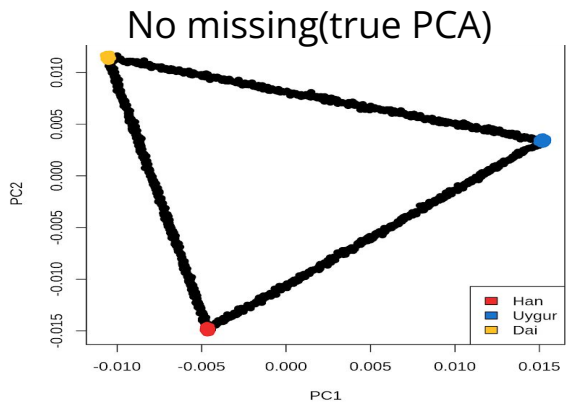
$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{U}_{[1:k]}\mathbf{S}_{[1:k]}\mathbf{V}^T_{[1:k]}$$

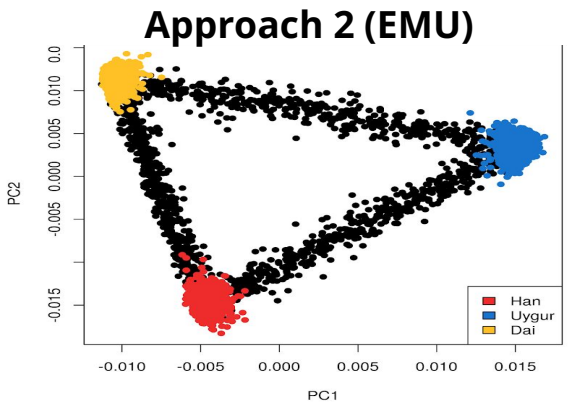
$\hat{\Pi}$

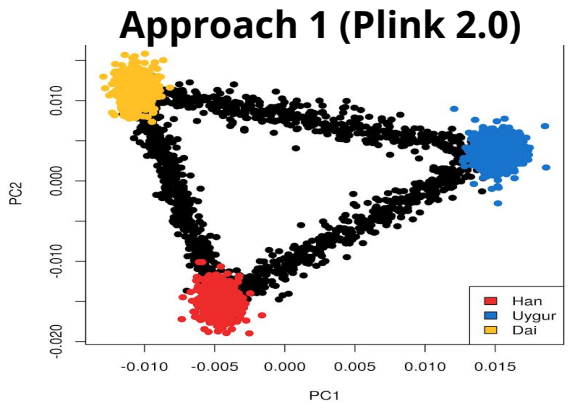
0.1	0.9	0.2	0.9	0.4
0.3	0.1	0.9	1.0	0.1
0.1	0.2	0.1	0.8	0.6
0.1	0.1	0.9	0.8	1.0

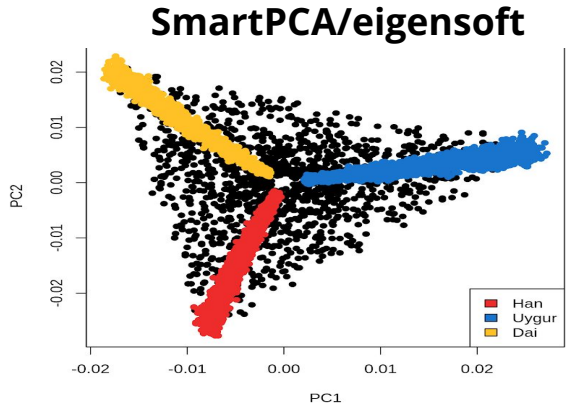
$$\min_{\mathbf{\Pi}} \left\| \mathbf{C} \odot (\mathbf{D} - \mathbf{\Pi}) \right\|_F^2$$

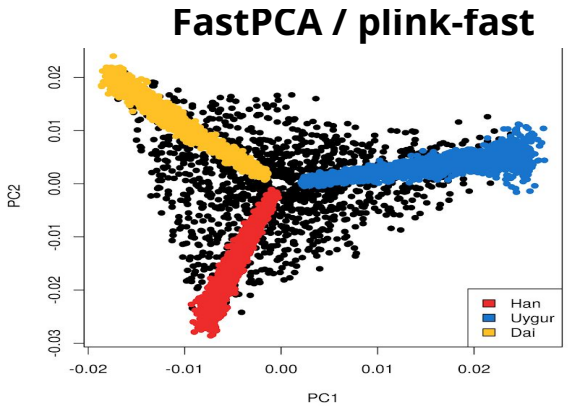
Simulated 90-99% missingness

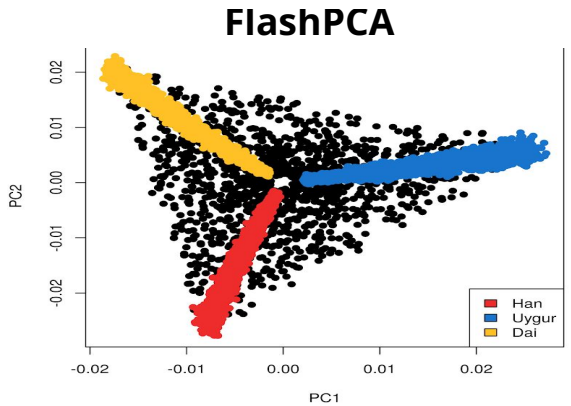




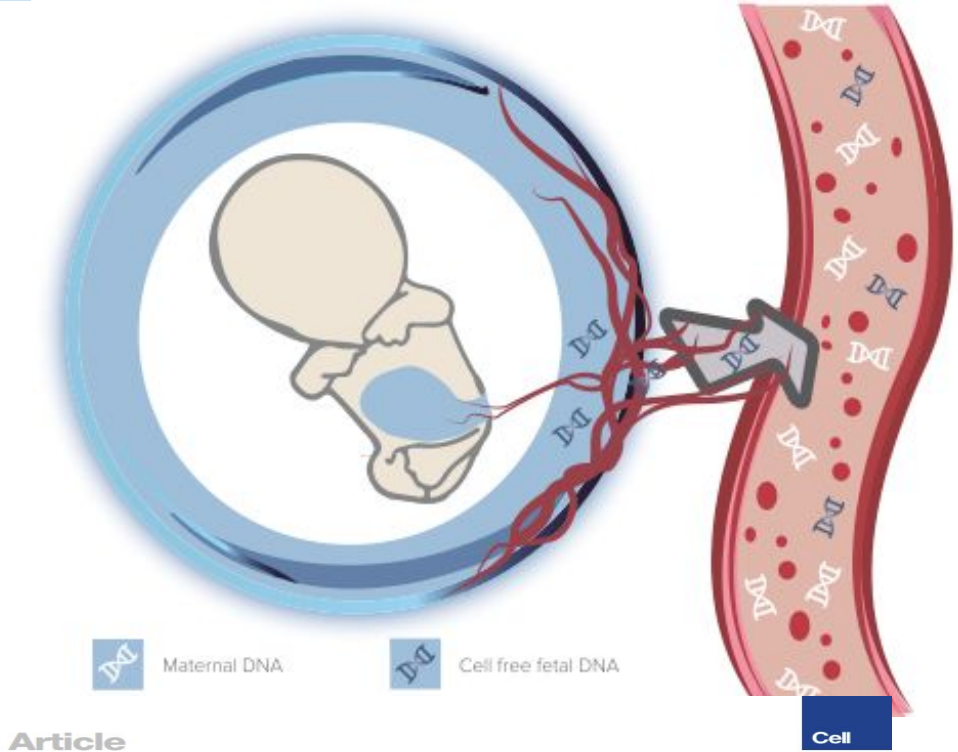








THE NON-INVASIVE FETAL TRISOMY TEST



Article

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

Siyang Liu,[1,25] Shujia Huang,[1,3,25] Fang Chen,[1,23,24,25] Lijian Zhao,[1,25] Yuying Yuan,[1,25] Stephen Starko F

Cell

Blood contains some fetus DNA

Does the fetus have large chromosomal anomalies?

E.g. Trisomy21 (Downs syndrome)

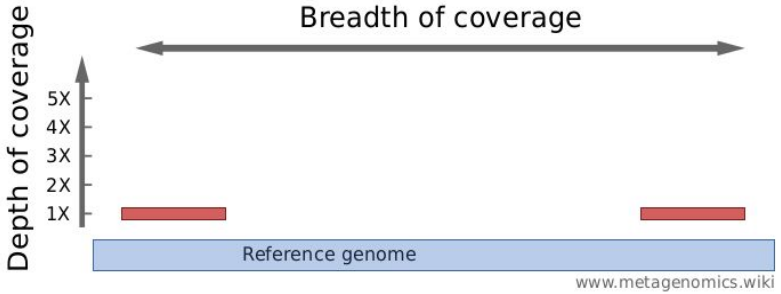
THE PILOT PROJECT

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

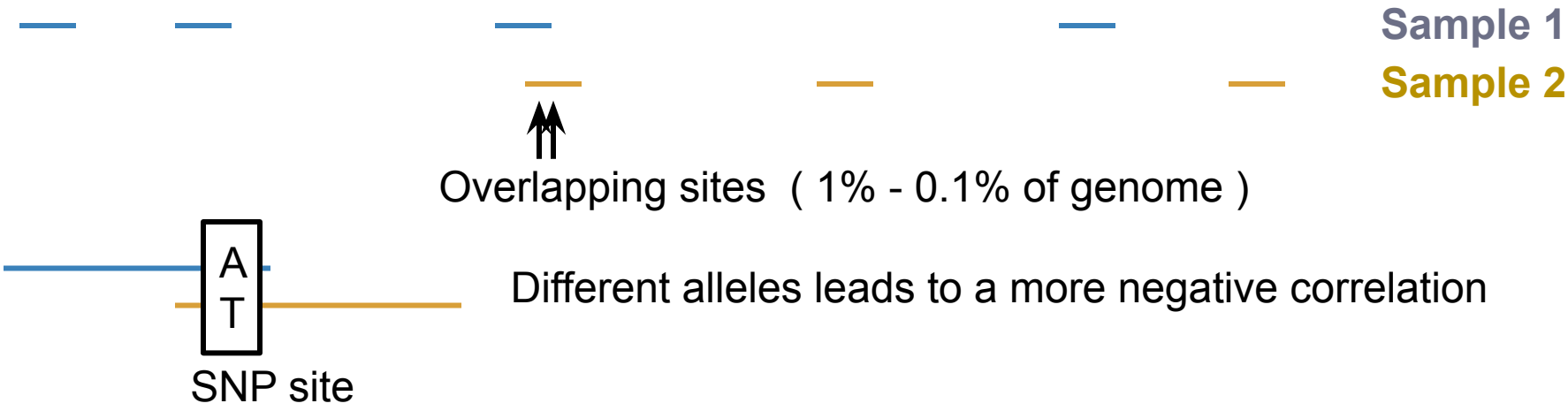
Population structure in China?

Can we perform PCA on ultra low depth?

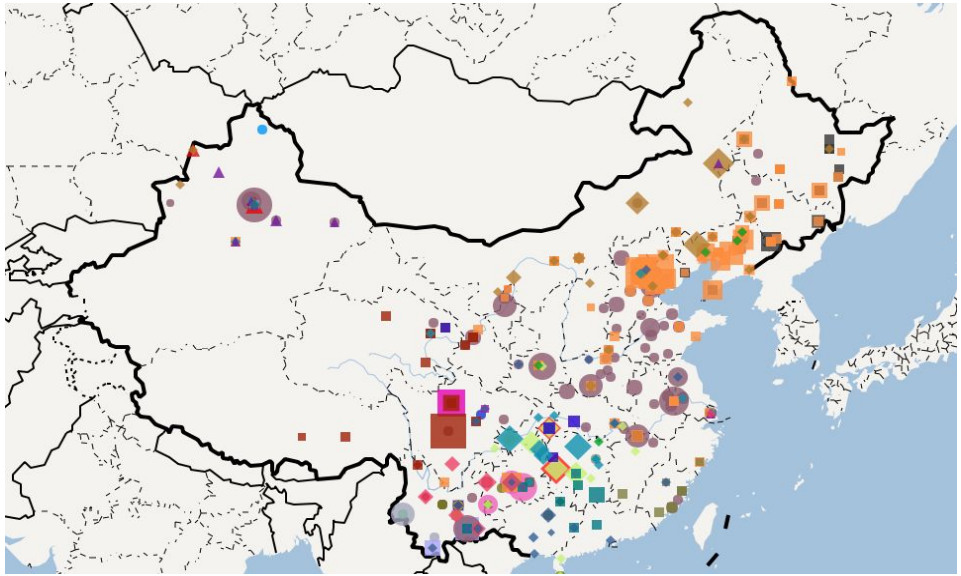
FIRST APPROACH

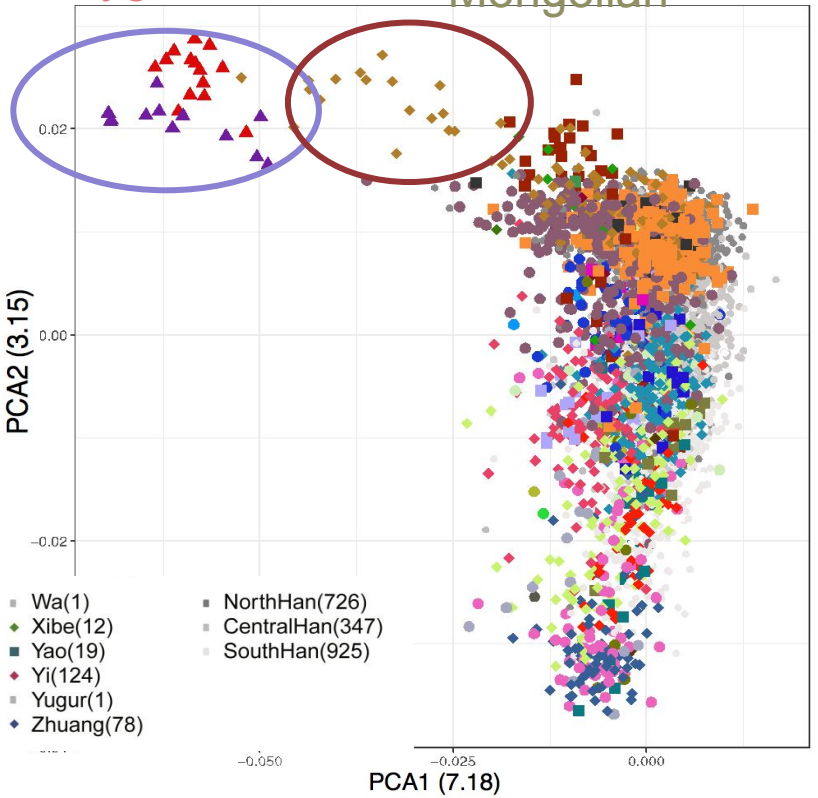
1) Estimate allele frequencies for each SNP

2) For each pair of individual estimate the covariance for overlapping sites



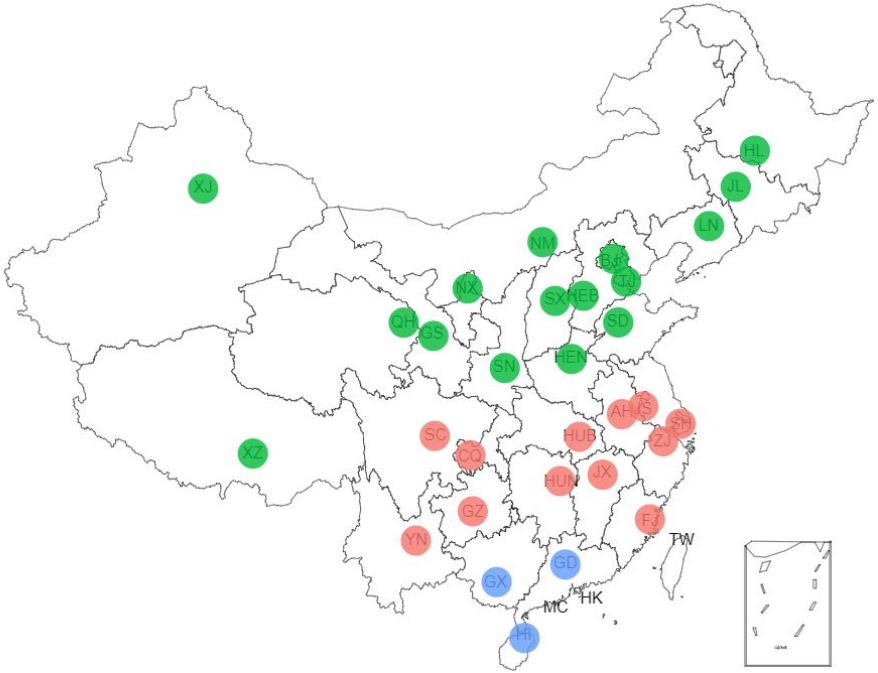
PCA with Han and minorities



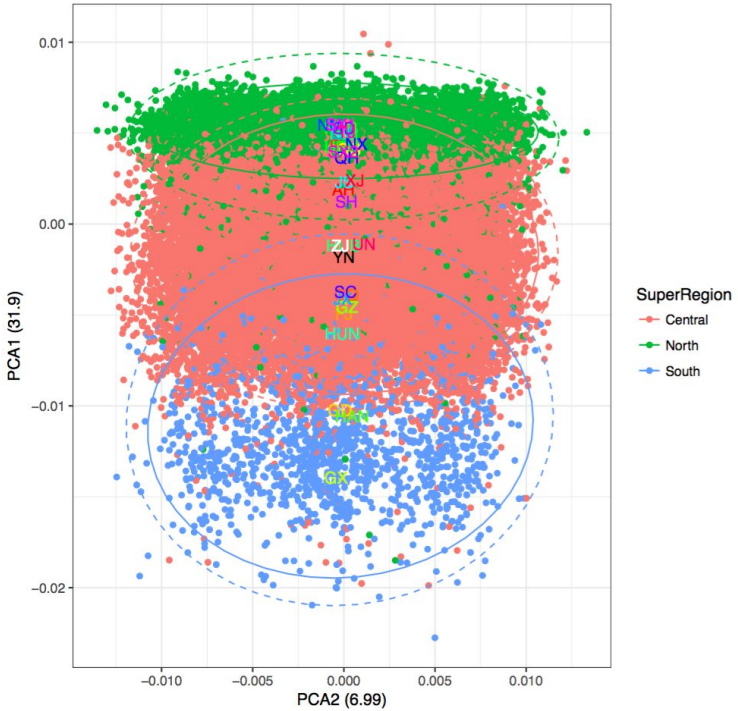


Achang(1) Bai(59) Blang(1) Bouyei(73) Dai(23) Dong(55) Evenki(1) Gaoshan(1) Gelao(27) Hani(18) Hui(389) Jingpo(1) Kazakh(14) Korean(38) Lahu(5) Li(10) Manchu(320) Miao(146) Mongol(134) Mulao(2) Nakhi(10) Qiang(8) Russian(2) Salar(1) She(17) Sui(2) Tibetan(42) Tu(16) Tujia(229) Uyghur(12) Wa(1) Xibe(12) Yao(19) Yi(124) Yugur(1) Zhuang(78) NorthHan(726) CentralHan(347) SouthHan(925)

HAN showed little differentiation



PCA from covariance (as in plink2)

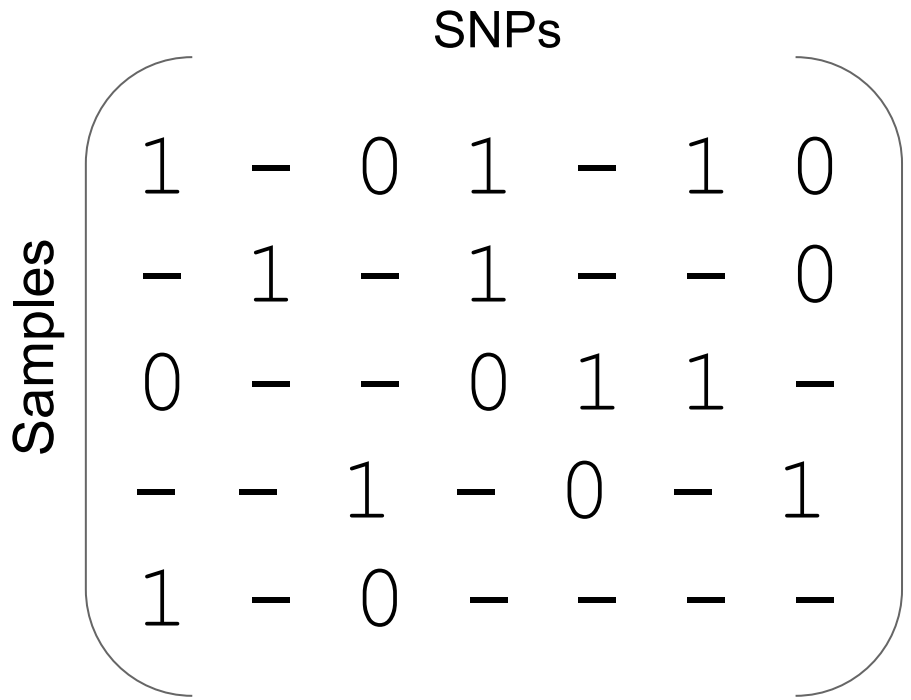


EMU algorithm

0 : Major allele

1 : Minor allele

\- : Missing (90-99% missing)



EMU - step 1: mean imputation

SNPs

Samples

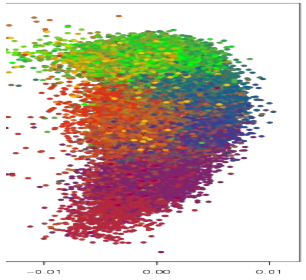
1	1	0	1	½	1	0
⅔	1	⅓	1	½	1	0
0	1	⅓	0	1	1	⅓
⅔	1	1	⅔	0	1	1
1	1	0	⅔	½	1	⅓

SNPs

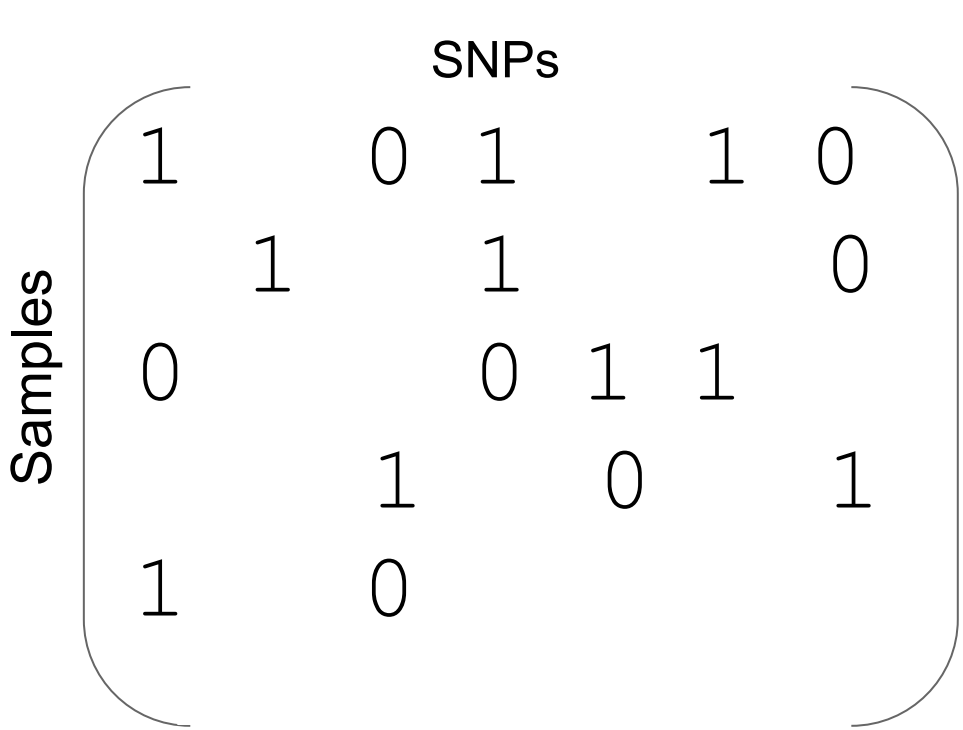
Samples

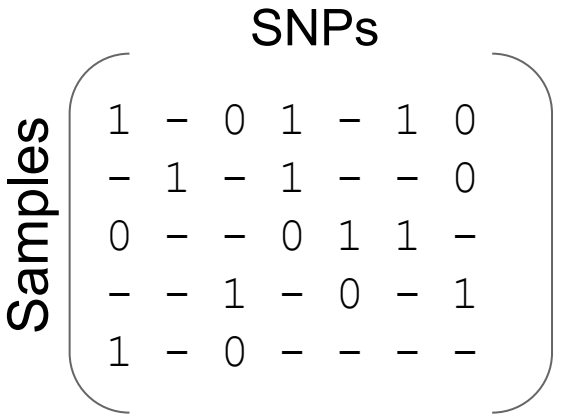
1	–	0	1	–	1	0
–	1	–	1	–	–	0
0	–	–	0	1	1	–
–	–	1	–	0	–	1
1	–	0	–	–	–	–

PCA

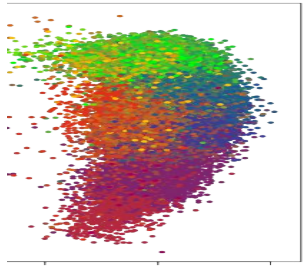


EMU - step (n): PCA imputation





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



Individual frequencies from PCA

EMU - step (n): PCA imputation

SNPs

Samples

$$\begin{pmatrix} 1 & \pi_{ij} & 0 & 1 & \pi_{ij} & 1 & 0 \\ \pi_{ij} & 1 & \pi_{ij} & 1 & \pi_{ij} & \pi_{ij} & 0 \\ 0 & \pi_{ij} & \pi_{ij} & 0 & 1 & 1 & \pi_{ij} \\ \pi_{ij} & \pi_{ij} & 1 & \pi_{ij} & 0 & \pi_{ij} & 1 \\ 1 & \pi_{ij} & 0 & \pi_{ij} & \pi_{ij} & \pi_{ij} & \pi_{ij} \end{pmatrix}$$

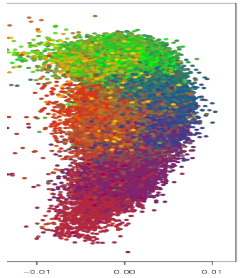
Update PCA

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

SNPs

Samples

$$\begin{pmatrix} 1 & - & 0 & 1 & - & 1 & 0 \\ - & 1 & - & 1 & - & - & 0 \\ 0 & - & - & 0 & 1 & 1 & - \\ - & - & 1 & - & 0 & - & 1 \\ 1 & - & 0 & - & - & - & - \end{pmatrix}$$

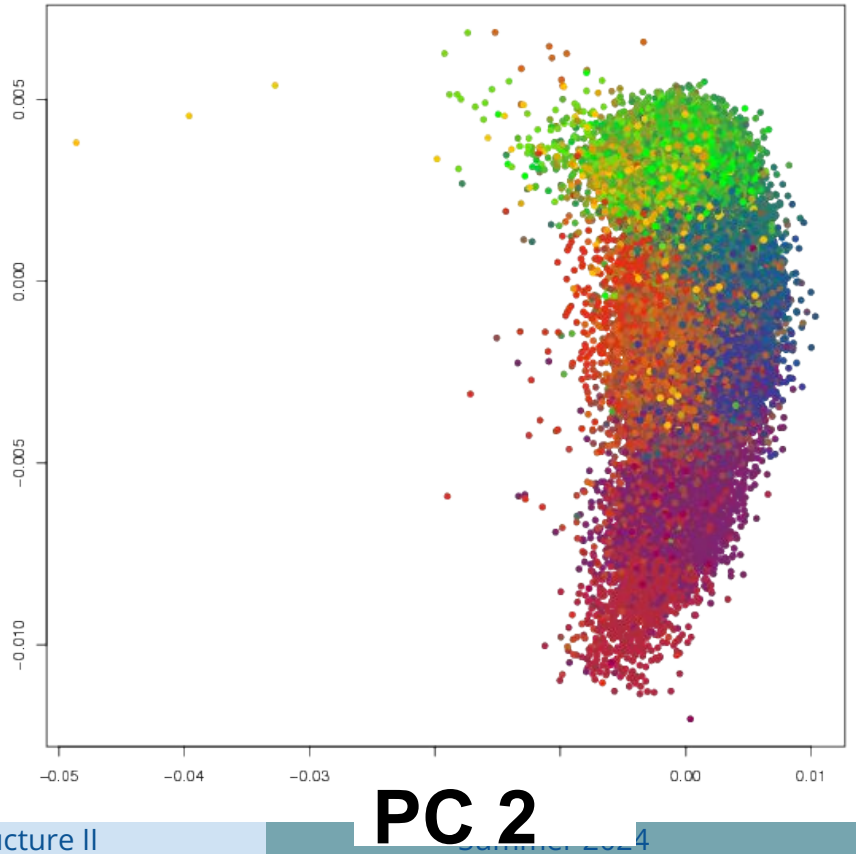


PCA shows Longitudinal differentiation

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

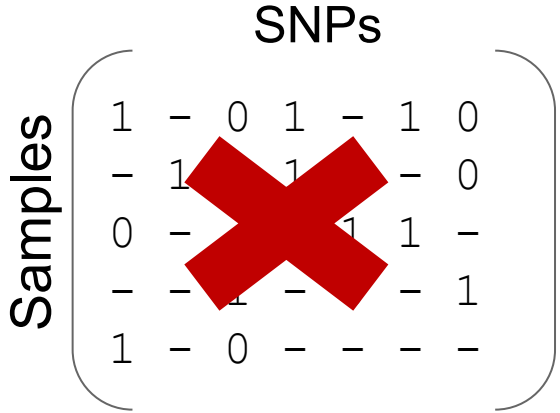
Jonas Meisner [1,*], Siyang Liu[2], Mingxi Huang[2] and Anders Albrechtsen[1]

[1]Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark and [2]BGI-Shenzhen, Shenzhen 518083, China



PCA for low/medium depth sequencing

Low depth sequencing has partial missing information for all genotypes.



Pseudo haploid

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

Genotype likelihood

Individual Allele frequency

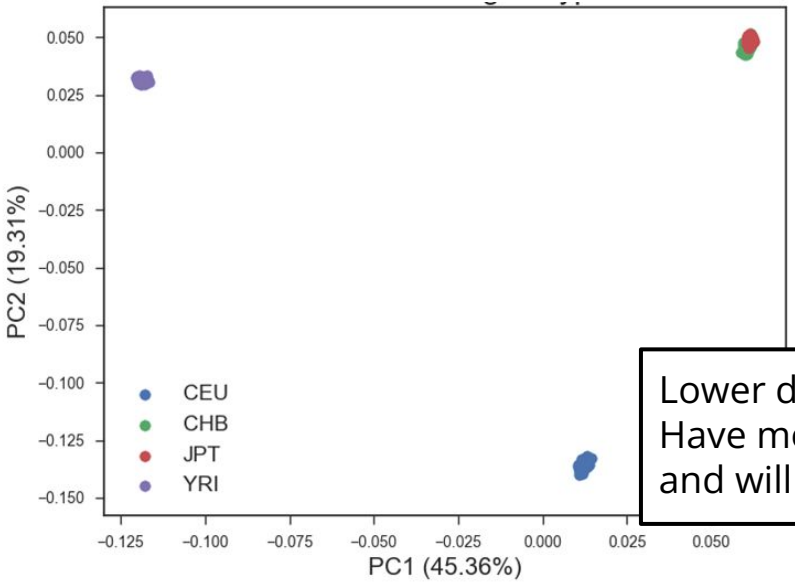
$$P(G = g \mid X, \pi) = \frac{P(X \mid G = g)P(G = g \mid \pi)}{\sum_{g'=0}^{2} P(X \mid G = g')P(G = g' \mid \pi)}$$

Dosage based on Allele frequency and Genotype likelihoods

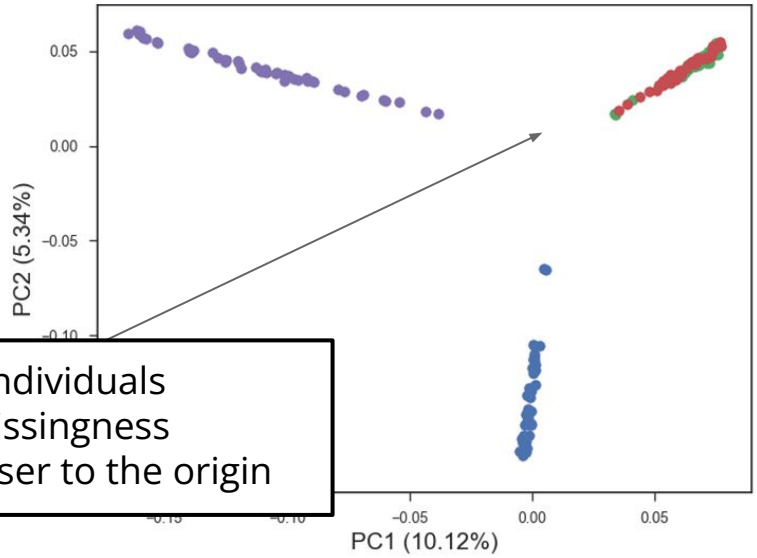
$$\longrightarrow \mathbb{E}[G \mid X, \pi] = \sum_{g=0}^{2} P(G = g \mid X, \pi)$$

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

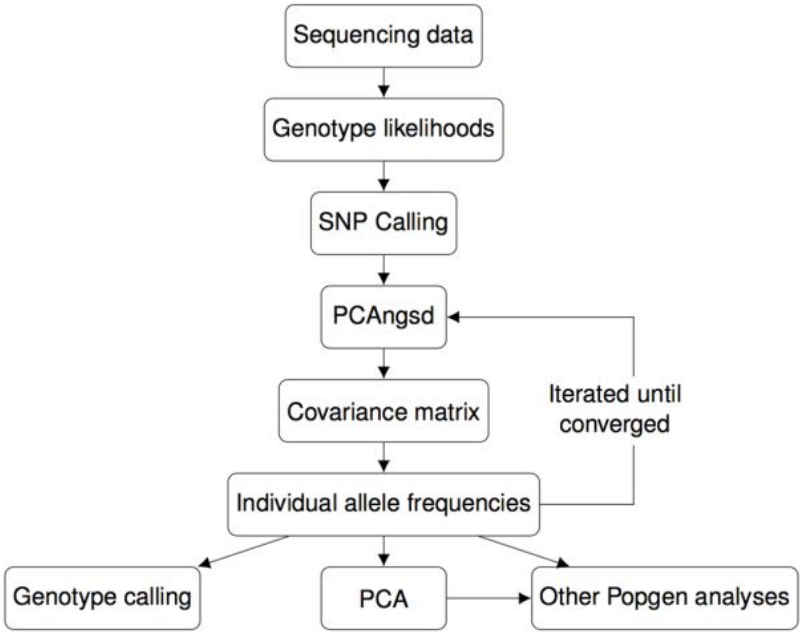
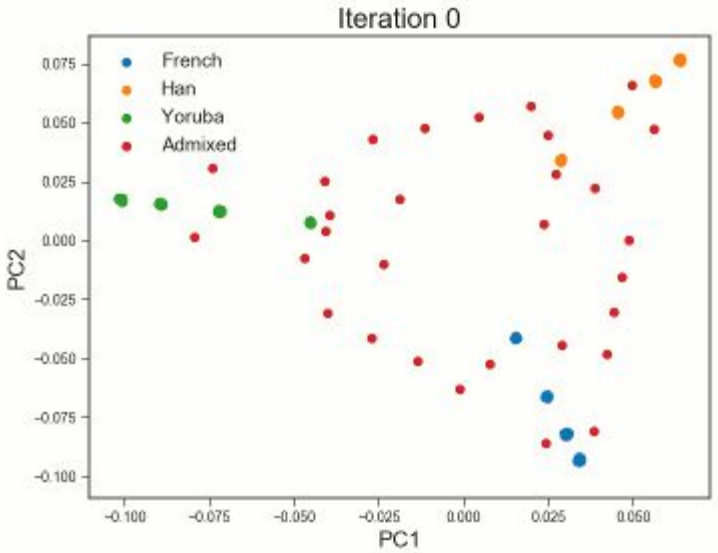
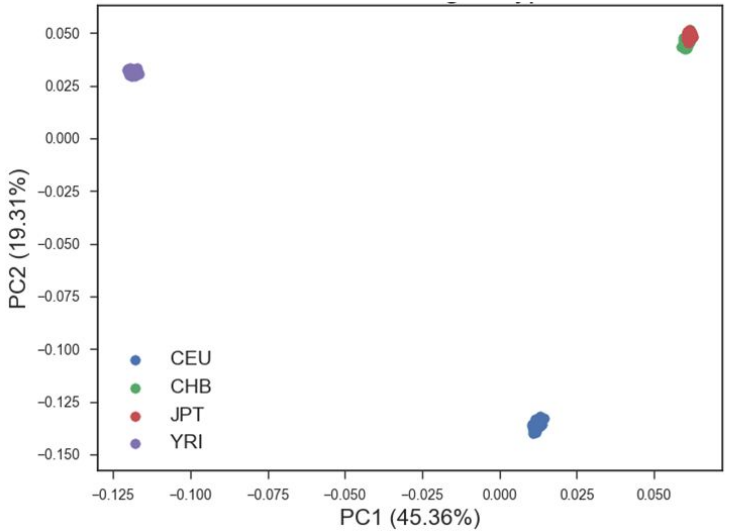


Figure: PCAngsd framework

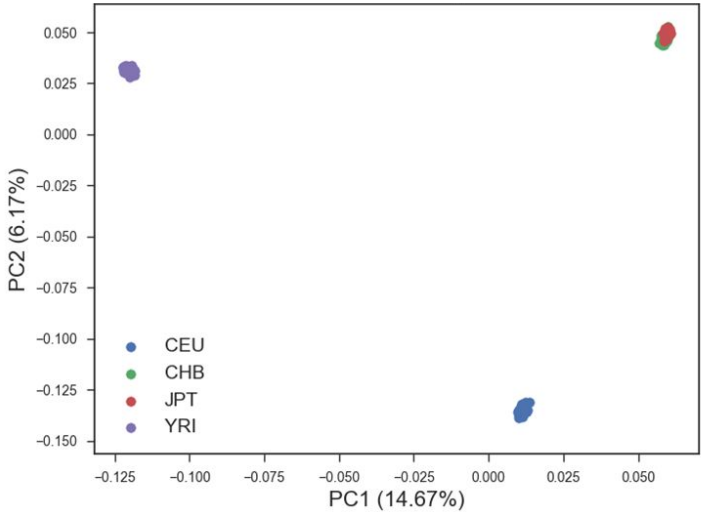


Issues with missingness

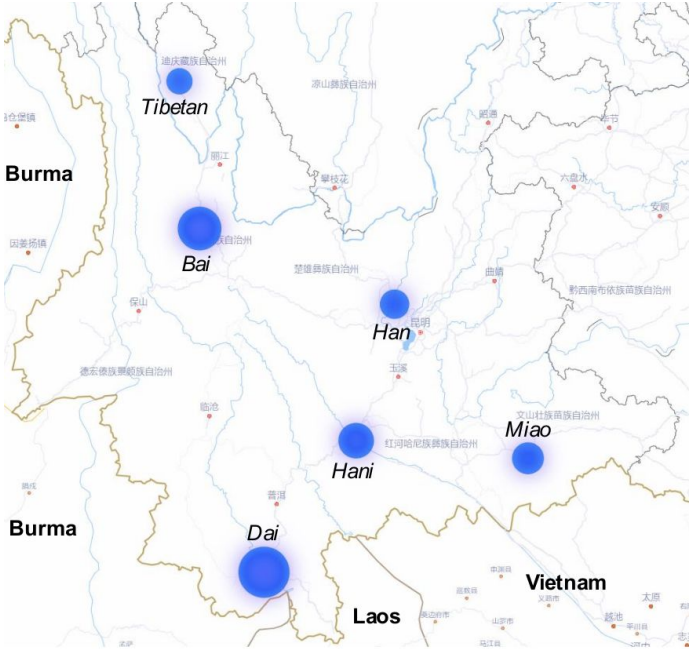
PCA from high depth genotypes



PCAngsd from low depth data (2-10X)



Host genome from gut microbiome

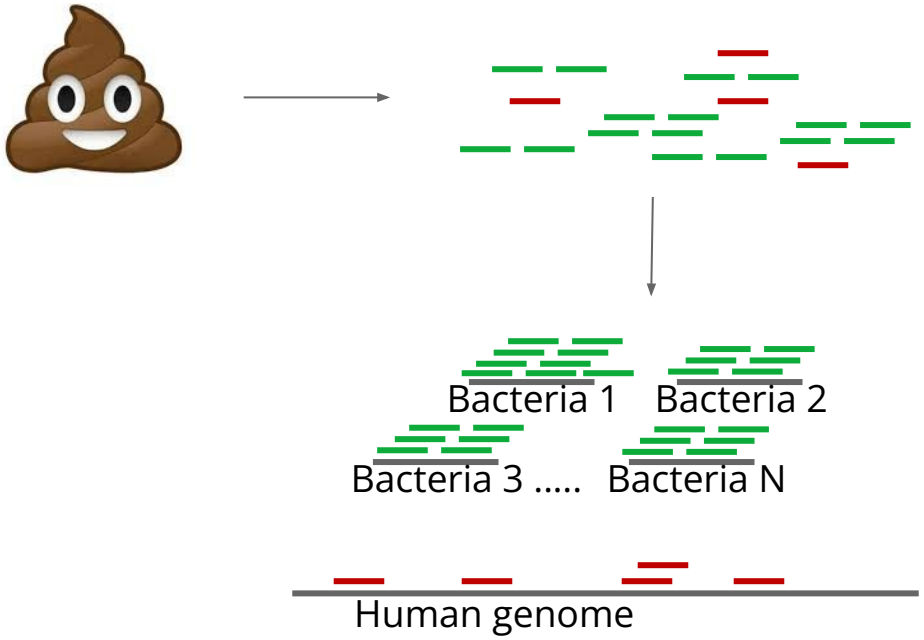


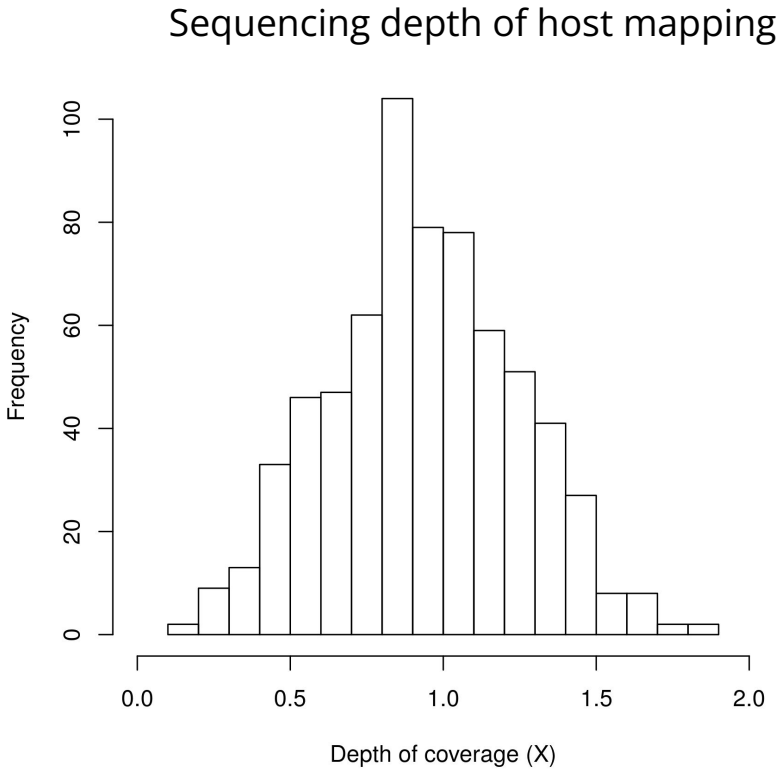
6 Minorities (**n=671**)

- 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 ()

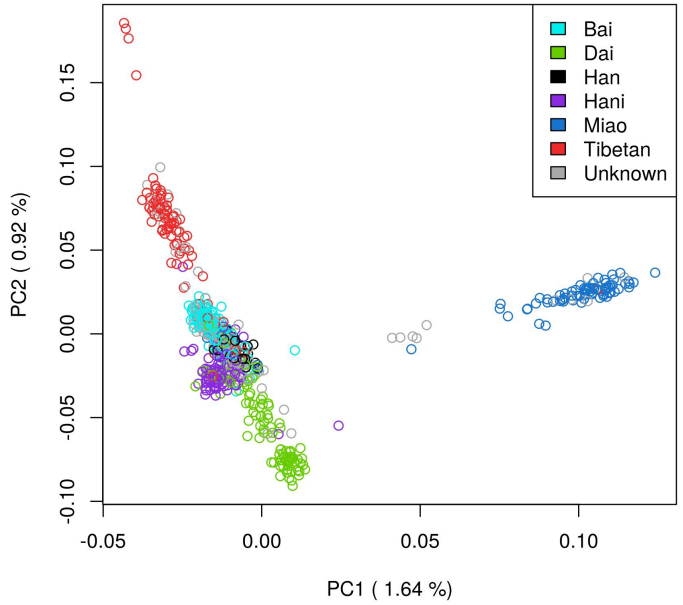
PCA on host metagenomics



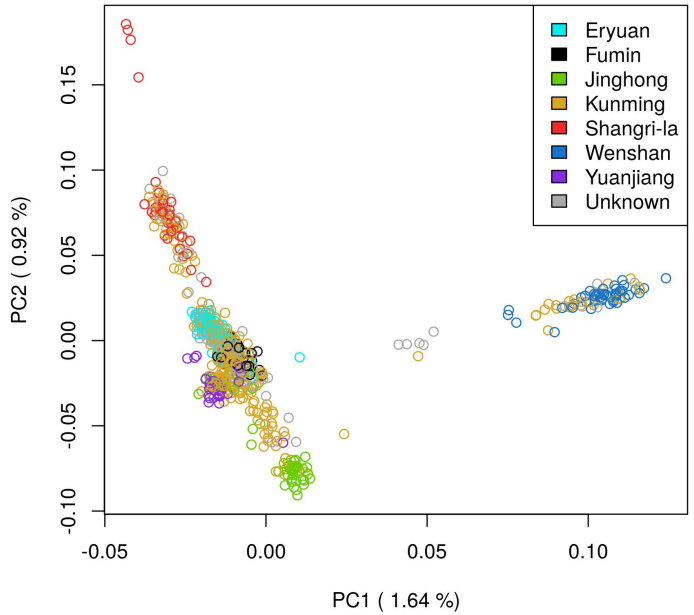


PCANGSD on host genetics

Minority labels



Location in labels



PCA and selection

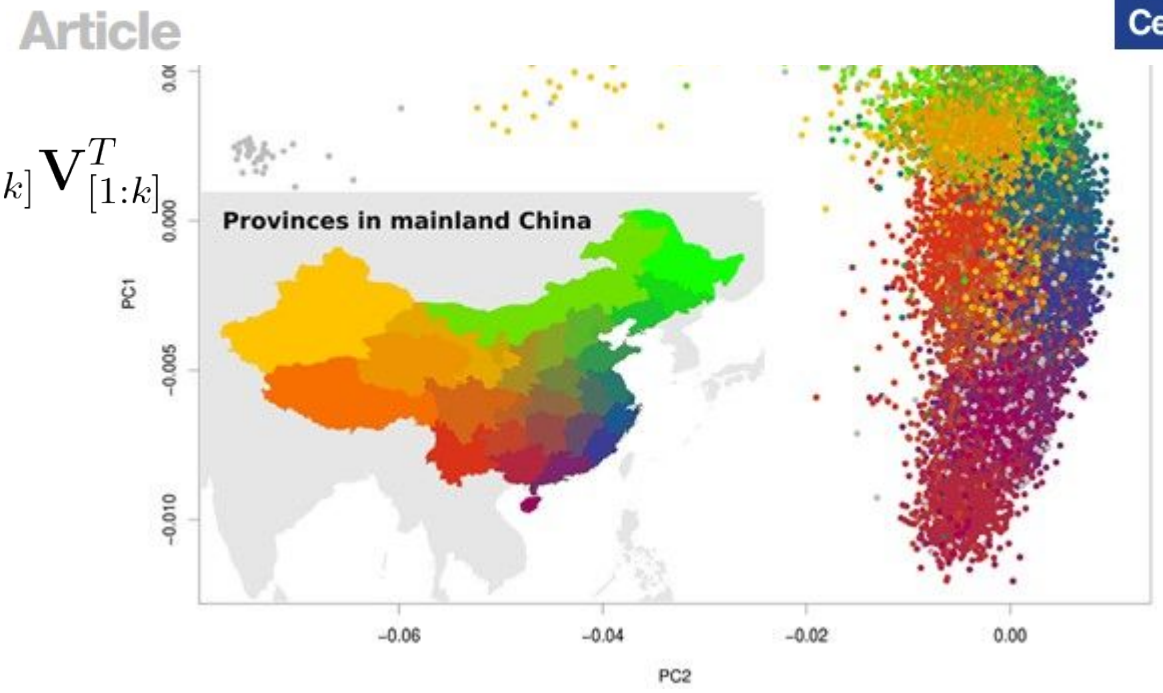
Continues structure / within population

FastPCA selection

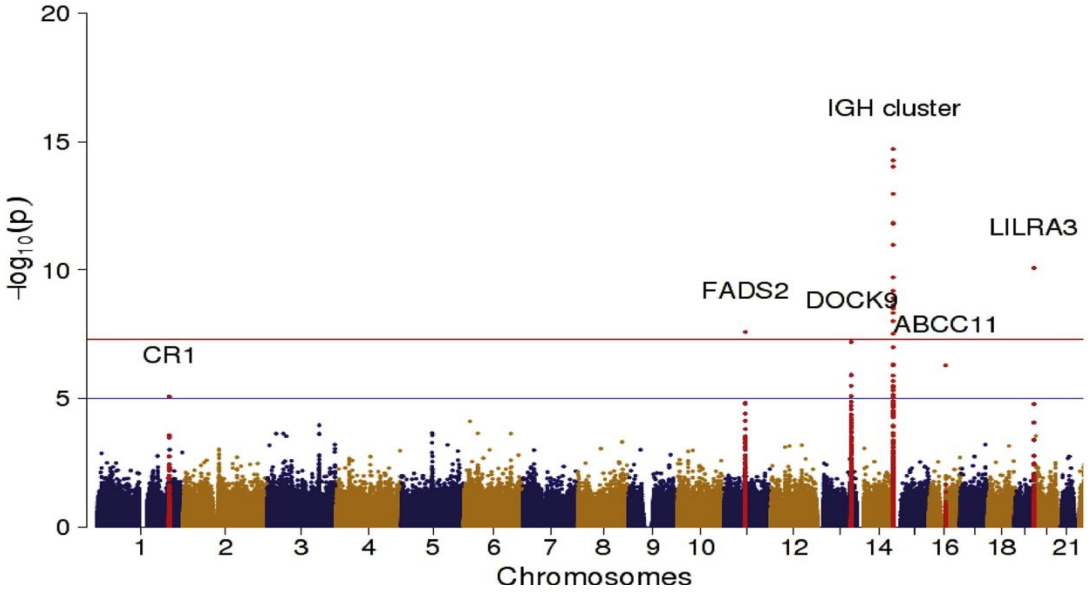
$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{U}_{[1:k]}\mathbf{S}_{[1:k]}\mathbf{V}^T_{[1:k]}$$

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

Galinsky et al (2016)

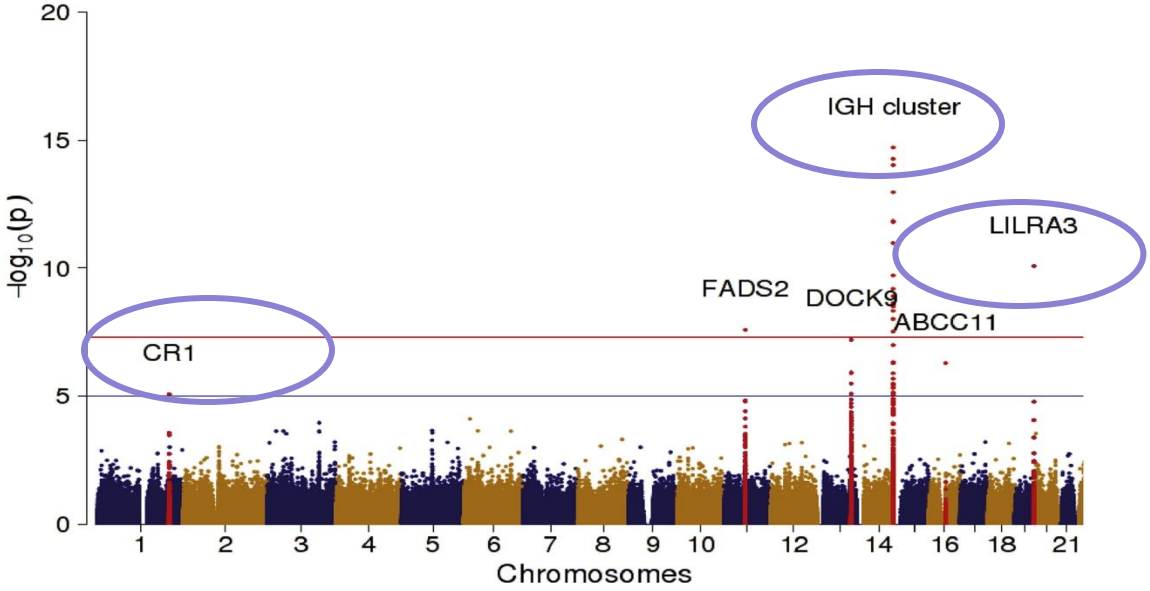


SELECTION SCAN - GWAS on first PC



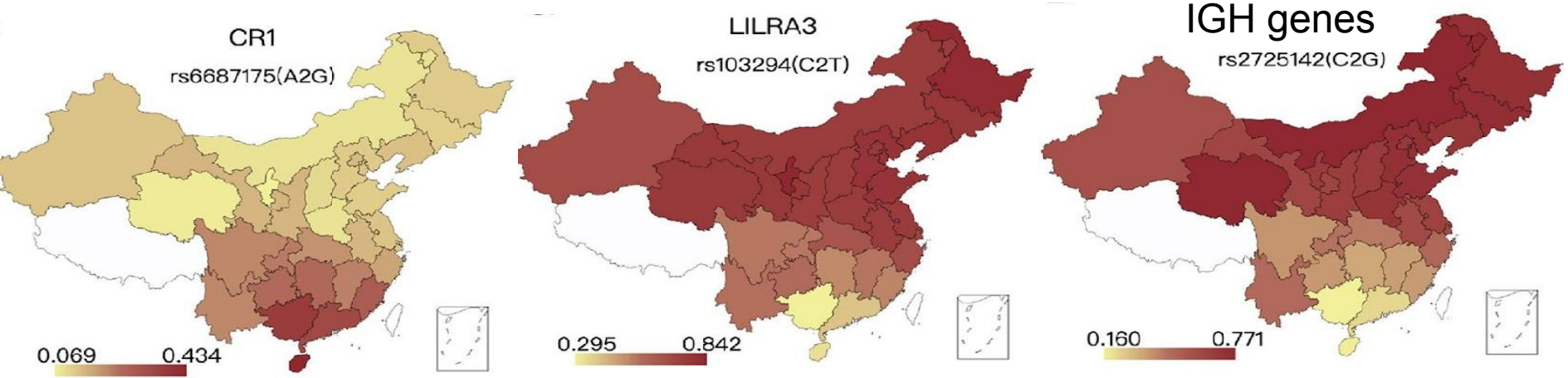
PC1

SELECTION SCAN



Immune genes previously
Shown to be under selection
In East Asians

IMMUNE RESPONSE GENES

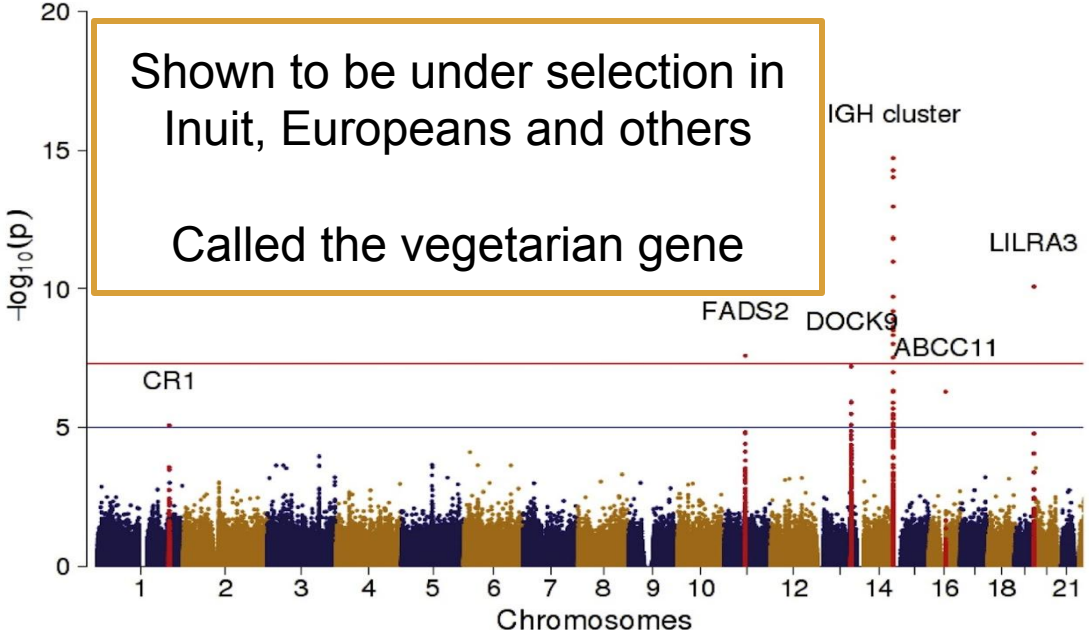


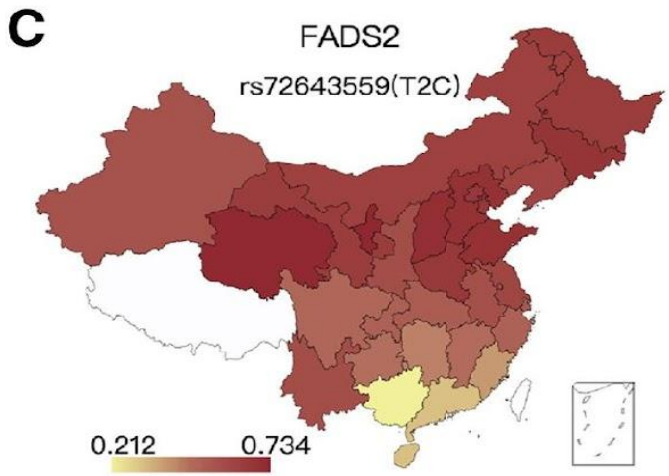
Derived Allele frequencies

FATTY ACID METABOLISM (FAD2)

Shown to be under selection in Inuit, Europeans and others

Called the vegetarian gene



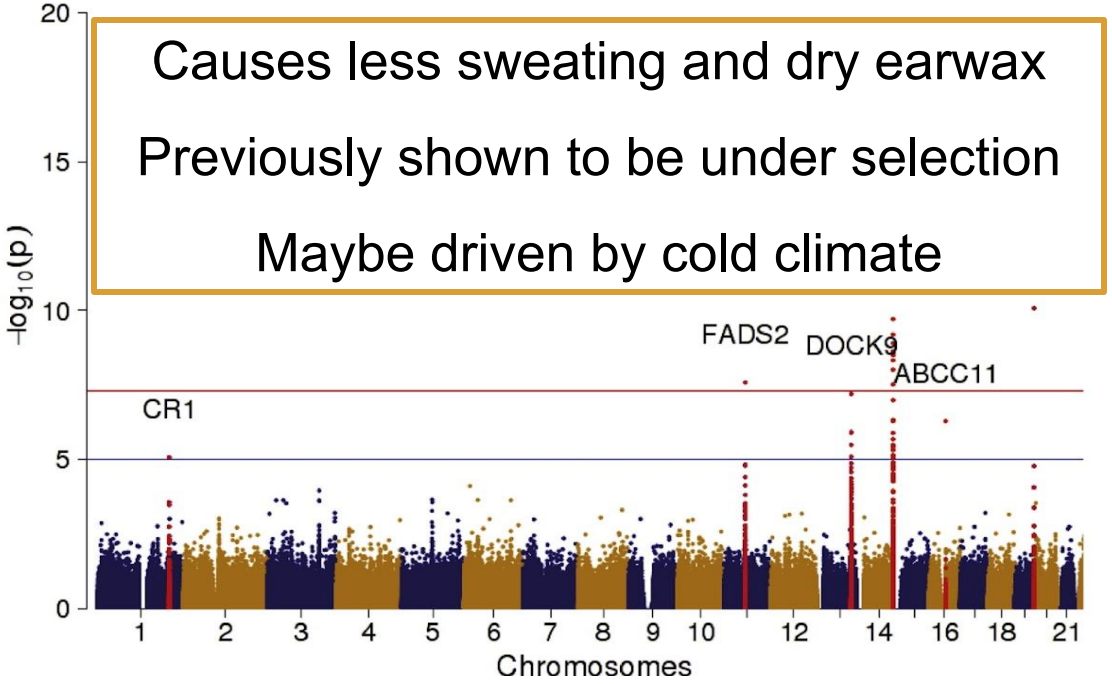


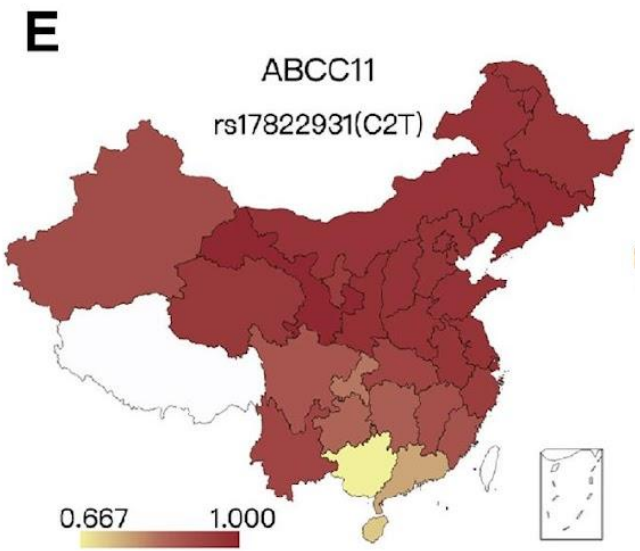
SWEAT AND EARWAX(ABCC11)

Causes less sweating and dry earwax

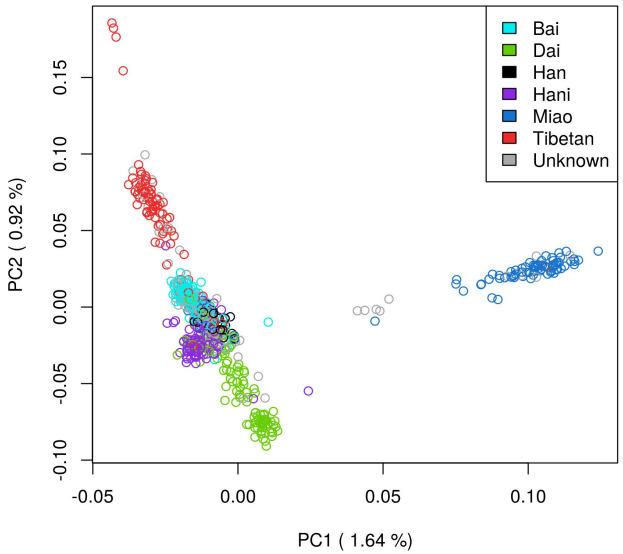
Previously shown to be under selection

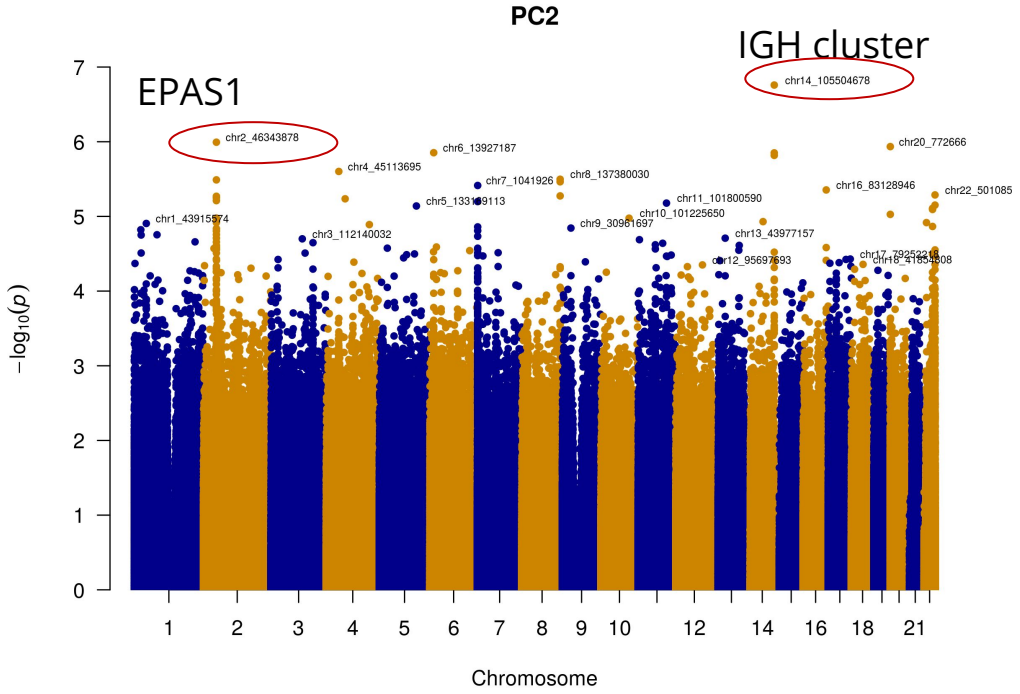
Maybe driven by cold climate



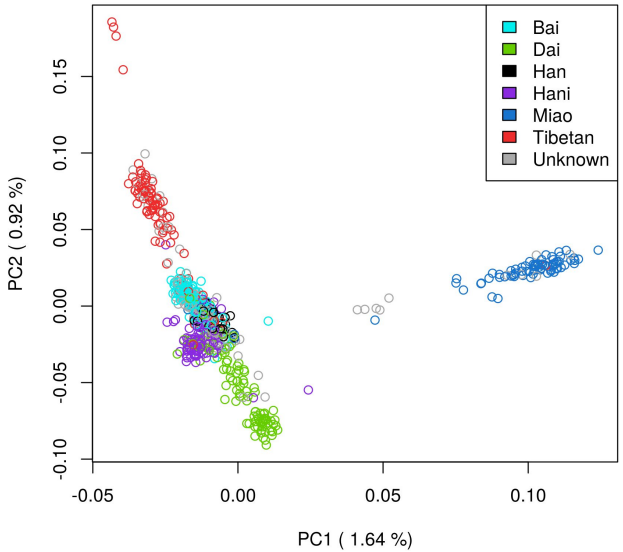


Selection scan on PC 2

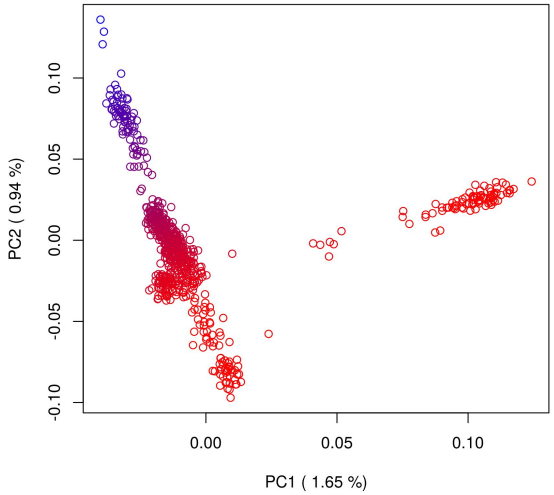




PCA and expected allele frequencies



EPAS1 top variant



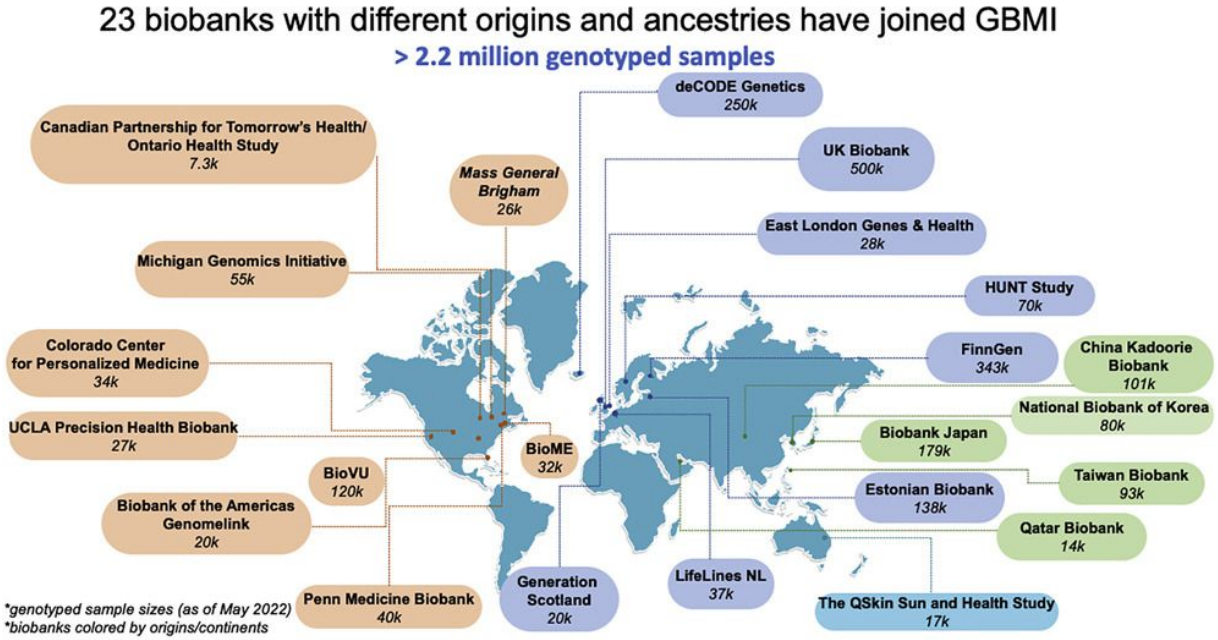
Expected allele frequency

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

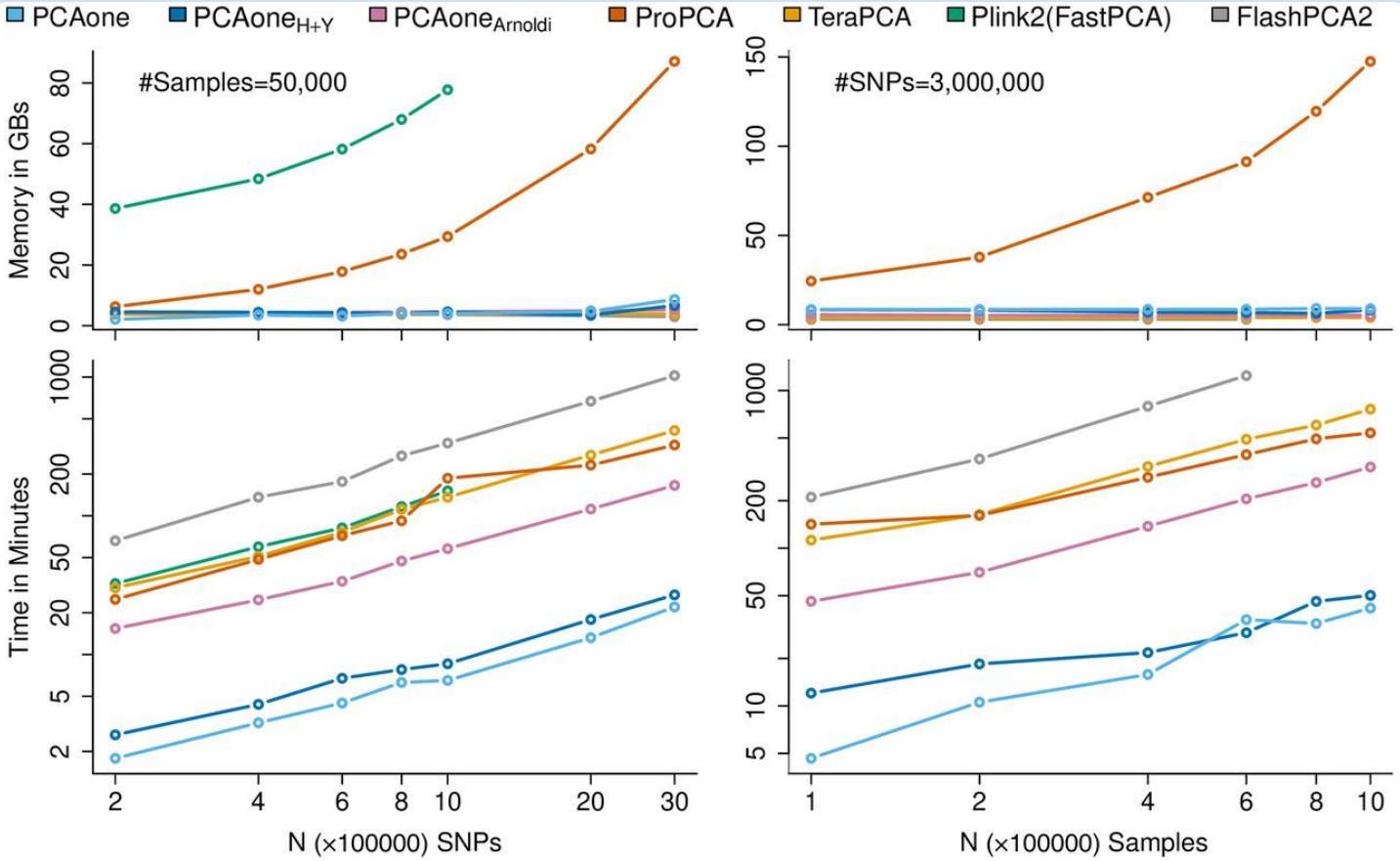
Time for exercises

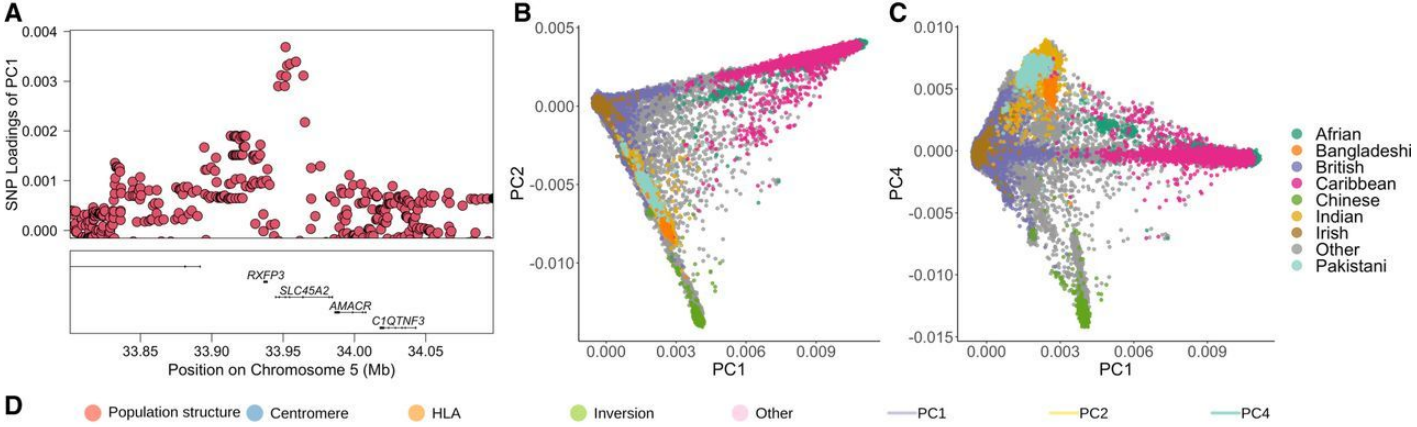
Run the pca notebook

When you have large data set

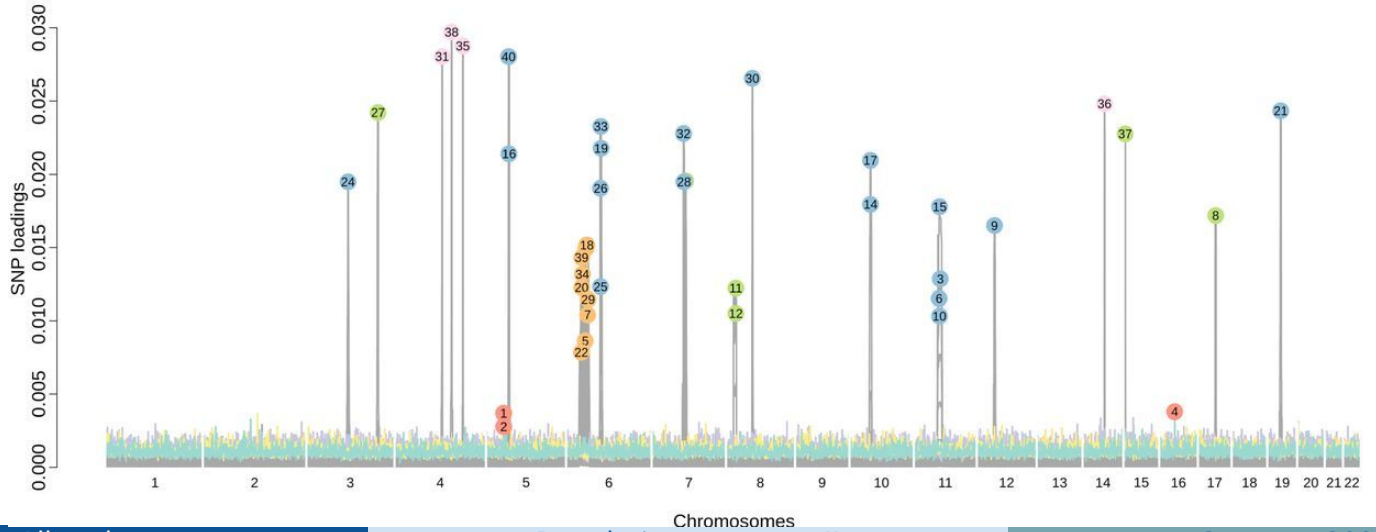


PCAone is fast, accurate and low RAM





40 PCs 0.5M individuals 6Million common SNPs



Single Cell RNAseq

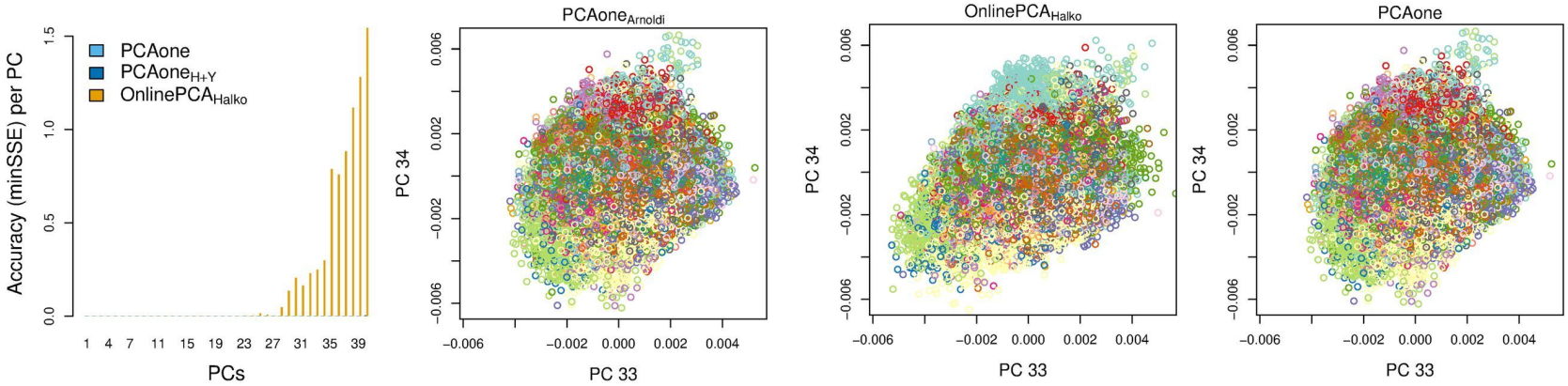


Table 2. Performance on scRNA-seq data with 1,306,127 cells and 23,771 genes for estimating the top K = 40 PCs

Program	Wall Time(m)	IO Time(m)	Epochs	RAM(GB)	MEV	minSSE
PCAone	49	31	10	8.97	0.999957	0.00417
PCAone$_{H+Y}$	42	30	8	6.82	0.999925	0.00732
OnlinePCA.jl$_{Halko}$	461 + 100[a]	300	8	1.71	0.954103	7.10293
PCAone$_{Arnoldi}$	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