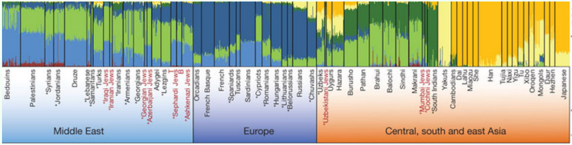
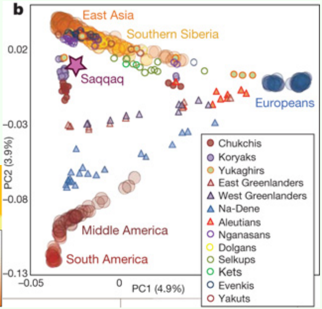
PCA , admixture 'proportions' and SFS for low depth NGS data

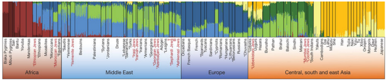
Anders Albrechtsen



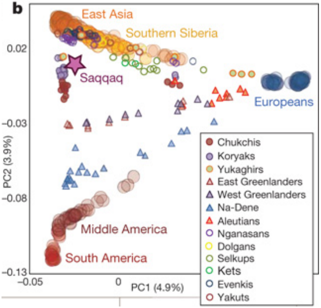


Analysis of low depth sequencing data

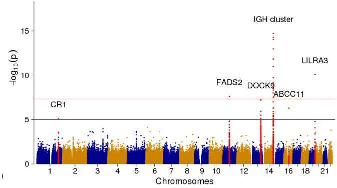
Admixture proportions



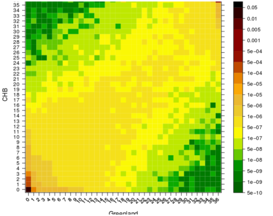
PCA



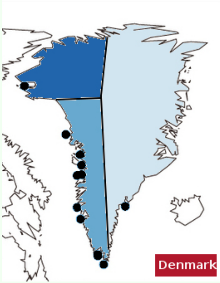
Individual allele frequencies (PCA)

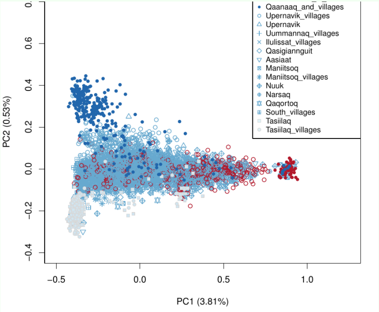


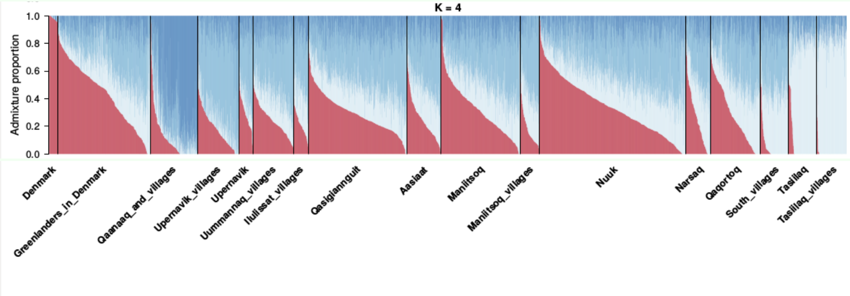
SFS and Fst



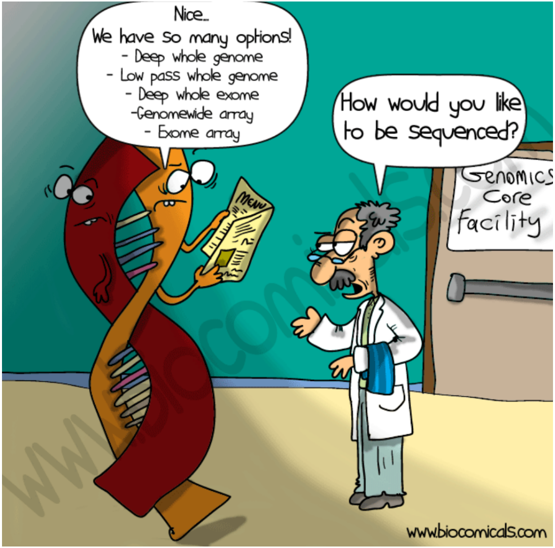
Admixture clustering /PCA - which is more informative?





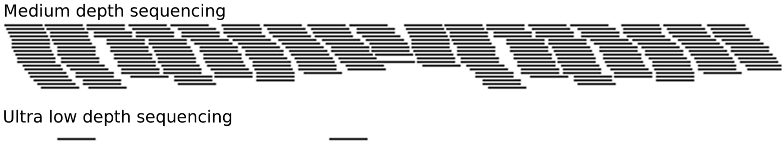


Sequencing types



What is low depth sequencing - my take on it

medium/high depth vs. ultra low depth



medium/low

- Depth lower than 10X
- Often a financial choice
- Ancient DNA

Ultra low sequencing

- Depth lower than 1X
- by product of capture data
-

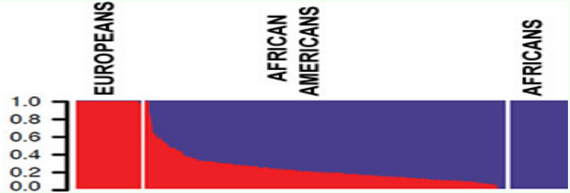
This morning

1. Admixture model
   - Intro to the model
   - likelihood based on called genotypes
2. NGSadmix
   - ML inference based on genotype likelihoods
3. Introduction to PCA
   - population structure and PCA
   - Problems with PCA analysis
   - NGS data
4. PCA for NGS - genotype likelihood approach
   - The expectation of the covariance
5. analysis based on individual allele frequencies
   - Admixture proportions vs. PCA
   - Inbreeding

Examples of known solutions and software

- Several methods:
  - Bayesian: e.g. *Structure* (Pritchard et al. 2000)
  - Maximum Likelihood: e.g. ADMIXTURE (Alexander et al. 2009)

- They all base their inference on called genotypes and infer
  1. Admixture proportions, Q



  2. Allele frequencies for all loci for all K populations, F

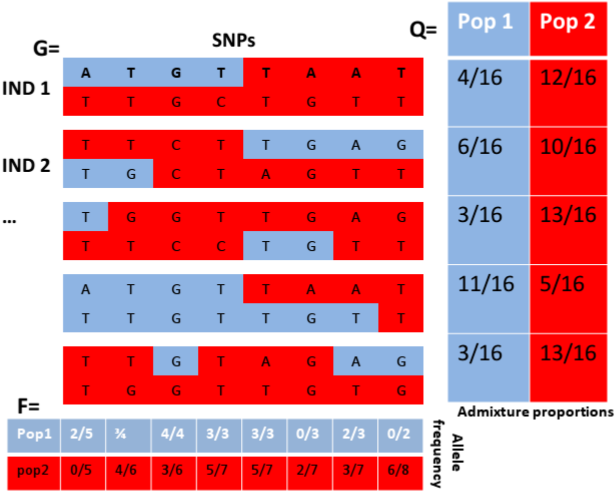
ML solution

- To find an ML solution we have to
  - Define a model/likelihood function $p(G|Q,F)$
  - Find an efficient way to find $\underset{(Q,F)}{\text{argmax}}\; p(G|Q,F)$
- The latter is usually solved using EM which I will no focus on
- I will spend time describing the model/likelihood function

G	the genotype data
F	the ancestral frequencies
Q	the admixture proportions

Visualized - if we know everything

known ancestry



G=

SNPs

IND 1	A	T	G	T	T	A	A	T
	T	T	G	C	T	G	T	T
IND 2	T	T	C	T	T	G	A	G
	T	G	C	T	A	G	T	T
...	T	G	G	T	T	G	A	G
	T	T	C	C	T	G	T	T
	A	T	G	T	T	A	A	T
	T	T	G	T	T	G	T	T
	T	T	G	T	A	G	A	G
	T	G	G	T	T	G	T	G

Q=

Pop 1	Pop 2
4/16	12/16
6/16	10/16
3/16	13/16
11/16	5/16
3/16	13/16

Admixture proportions

F=

Pop1	2/5	¾	4/4	3/3	3/3	0/3	2/3	0/2
pop2	0/5	4/6	3/6	5/7	5/7	2/7	3/7	6/8

Allele frequency

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- $Q^i = (q_1^i, q_2^i, ..., q_K^i)$ be i's genomewide admixture proportions
- G_{ij} be the genotype of i in j (measured in counts of allele A)
- $F^j = (f_1^j, f_2^j, ..., f_K^j)$ denote the allele frequencies of allele A

Then

- for one of i's alleles: $p(allele|Q^i, F^j) = q_1^i f_1^j + q_2^i f_2^j + ...q_K^i f_K^j = \pi^{ij}$
- π is also called the individual allele frequency
- all individual allele frequencies $\Pi = QF^T$
- Assuming HWE the probability of a observing genotype is:

$$p(G_{ij}|Q^i, F^j) = \begin{cases} (\pi^{ij})^2 & \text{if } G_{ij} = 2, \\ 2\pi^{ij}(1-\pi^{ij}) & \text{if } G_{ij} = 1, \\ (1-\pi^{ij})^2 & \text{if } G_{ij} = 0. \end{cases}$$

Likelihood function (N individuals, M diallelic loci)

- If we assume:
  - the individuals are unrelated and thus independent
  - loci are independent

  we can write the (composite) likelihood as

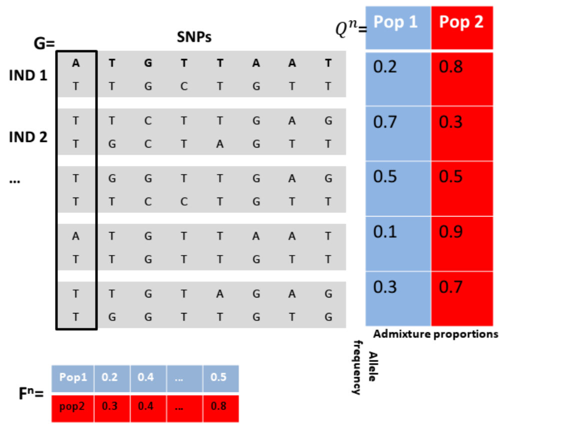
$$p(G|Q,F) = \prod_i^N \prod_j^M p(G_{ij}|Q^i, F^j)$$

- ML estimate (like ADMIXTURE): $(\hat{Q}, \hat{F}) = \underset{(Q,F)}{\operatorname{argmax}}\ p(G|Q,F)$.

Very large number of parameters

M × K + N × (K-1)

EM algorithm. A single site. New estimate of F



G=

SNPs

IND 1	A	T	G	T	T	A	A	T
	T	T	G	C	T	G	T	T
IND 2	T	T	C	T	T	G	A	G
	T	G	C	T	A	G	T	T
...	T	G	G	T	T	G	A	G
	T	T	C	C	T	G	T	T
	A	T	G	T	T	A	A	T
	T	T	G	T	T	G	T	T
	T	T	G	T	A	G	A	G
	T	G	G	T	T	G	T	G

Q^n=

Pop 1	Pop 2
0.2	0.8
0.7	0.3
0.5	0.5
0.1	0.9
0.3	0.7

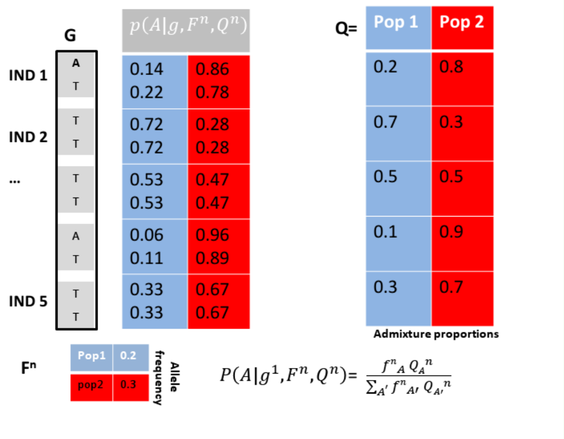
Admixture proportions

Allele frequency

Fⁿ=

Pop1	0.2	0.4	...	0.5
pop2	0.3	0.4	...	0.8

EM algorithm. A single site. New estimate of F



	G	$p(A\|g,F^n,Q^n)$	
IND 1	A	0.14	0.86
	T	0.22	0.78
IND 2	T	0.72	0.28
	T	0.72	0.28
...	T	0.53	0.47
	T	0.53	0.47
	A	0.06	0.96
	T	0.11	0.89
IND 5	T	0.33	0.67
	T	0.33	0.67

Q=

Pop 1	Pop 2
0.2	0.8
0.7	0.3
0.5	0.5
0.1	0.9
0.3	0.7

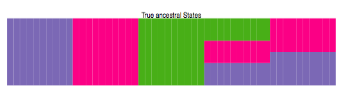
Admixture proportions

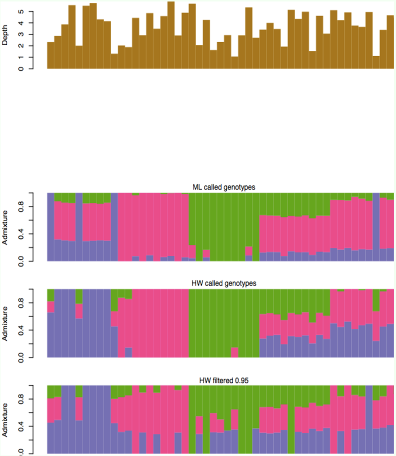
F^n

	Allele frequency
Pop1	0.2
pop2	0.3

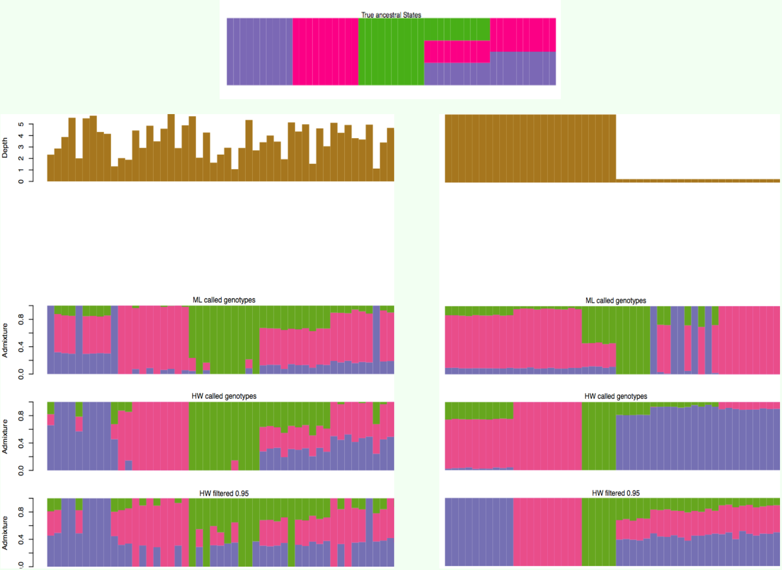
$$P(A|g^1,F^n,Q^n) = \frac{f^n{}_A\, Q_A{}^n}{\sum_{A'} f^n{}_{A'}\, Q_{A'}{}^n}$$

Some problems (NGS data, variable depth)





Some problems (NGS data, variable depth)

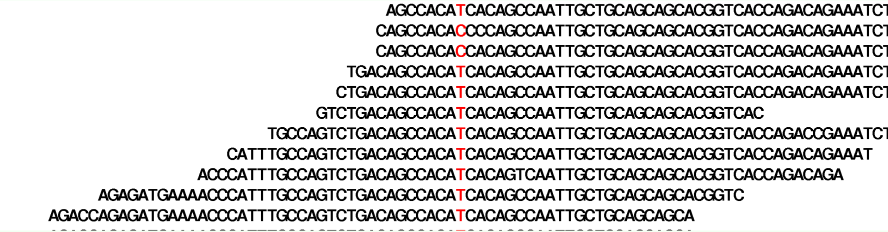


Why we have problems with genotype calling

This is not like Sanger sequencing

Sanger Both alleles are amplified and sequenced at the same time.

NGS Each allele is sequenced separately and the allele are sampled with replacement



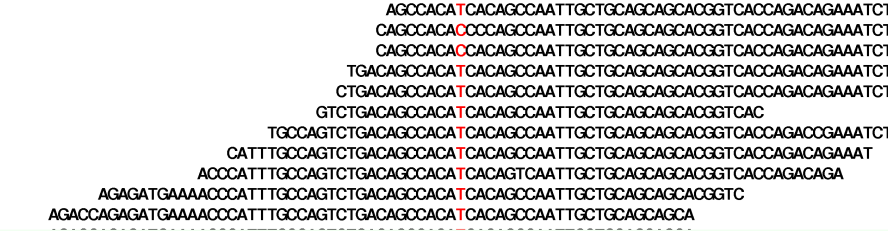
```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

why don’t we have genotypes?

Question?

Assuming an error rate of 1%

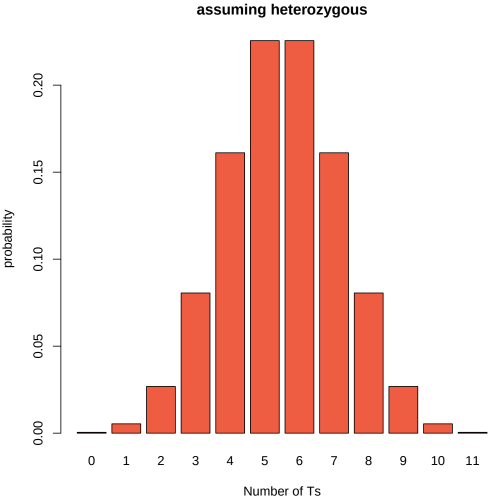
- Is the individual heterozygous C/T?



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

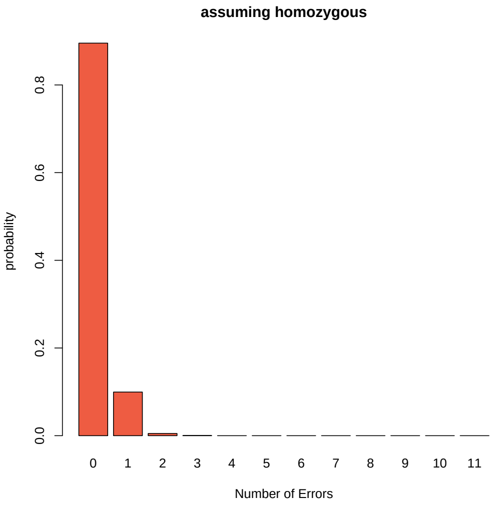
What do we expect

P(2 or less minor bases | heterozygous) = 0.065



What do we expect

P(2 or more errors | homozygous) = 0.00015

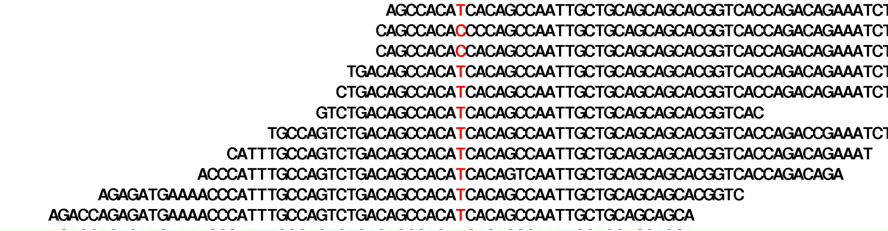


why don't we have genotypes?

Question?

Assuming an error rate of 1%

- Is the individual heterozygous C/T?
- P(2 or more errors | homozygous) = 0.00015
- P(2 or less minor bases | heterozygous) = 0.065



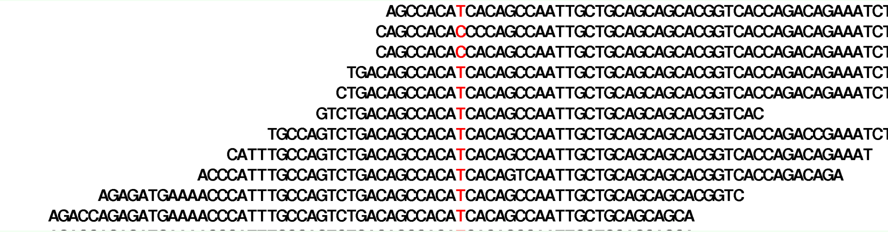
```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

why don't we have genotypes?

Question?

Assuming an error rate of 1%

- Is the individual heterozygous C/T?
- P(2 or more errors | homozygous) = 0.00015
- P(2 or less minor bases | heterozygous) = 0.065
- assuming on average there is about 1 heterozygous site per 1000 bases



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCT
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

Genotype likelihoods

Summarise the data in 10 genotype likelihoods

bases (b):
TCCTTTTTTTT
quality scores (Q):
GHSSBBTTTTG

⟼

	A	C	G	T
A	1	2	3	4
C		5	6	7
G			8	9
T				10

The genotype likelihood $P(X|G)$

$P(Data|G = \{A_1, A_2\}) = P(X|G = \{A_1, A_2\})$
where $A \in \{A, C, G, T\}$

Estimating genotype likelihoods

GATK (McKenna et al. 2010)

$$P(X|G) \propto \prod_{i=0}^{n} P(b_i|A_1, A_2) = \prod_{i=0}^{n} \left(\frac{1}{2} P(b_i|A_1) + \frac{1}{2} P(b_i|A_2) \right)$$

where $P(b|A) = \begin{cases} \frac{\epsilon}{3} & b \neq A \\ 1 - \epsilon & b = A \end{cases}$,

where $G = \{A_1, A_2\}$, b is the observed base and ϵ is the probability of error from the quality score.

Example of genotype likelihood calculations

b	Qasci	Qscore	ϵ	$p(b_i\|T)$	$p(b_i\|C)$	$p(b_i\|G/A)$
T	G	38	0.00016	1 - 0.00016	5.3e-05	5.3e-05
C	H	39	0.00013	4.2e-05	1 - 0.00013	4.2e-05
C	S	50	1e-05	3.3e-06	1 - 1e-05	3.3e-06
T	S	50	1e-05	1 - 1e-05	3.3e-06	3.3e-06
T	B	33	5e-04	1 - 5e-04	0.00017	0.00017
T	B	33	5e-04	1 - 5e-04	0.00017	0.00017
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	T	51	7.9e-06	1 - 7.9e-06	2.6e-06	2.6e-06
T	G	38	0.00016	1 - 0.00016	5.3e-05	5.3e-05

$$P(Data|G=TC) \propto \prod_{i=0}^{n} P(b_i|T,C) = \prod_{i=0}^{n} \left(\frac{1}{2} P(b_i|T) + \frac{1}{2} P(b_i|C) \right)$$

Genotype likelihoods with inferred major and minor alleles

The genotype likelihood

$$p(X \mid geno)$$

Summarise data for diallelic site

bases:
TTTCCTTTTTTTT
quality score:
BBGHSSBBTTTTG

$\rightarrowtail$

0	$p(X \mid geno = 0)$
1	$p(X \mid geno = 1)$
2	$p(X \mid geno = 2)$

Solution: NGSadmix

- Works on genotype likelihoods instead of called genotypes
- I.e. input is $p(X_{ij}|G_{ij})$ for all 3 possible values of G_{ij}, where X_{ij} is NGS data for individual i at locus j
- The previous likelihood is extended from

$$p(G|Q,F) = \prod_{i}^{N}\prod_{j}^{M} p(G_{ij}|Q^i,F^j)$$

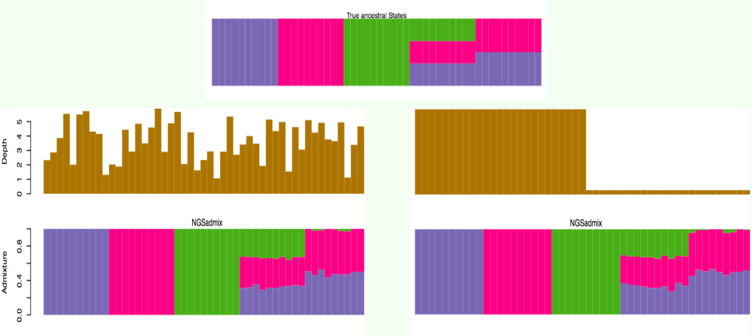
to

$$p(X|Q,F) = \prod_{i}^{N}\prod_{j}^{M} p(X_{ij}|Q^i,F^j) = \prod_{i}^{N}\prod_{j}^{M} \sum_{G_{ij}\in\{0,1,2\}} p(X_{ij}|G_{ij})p(G_{ij}|Q^i,F^j)$$

- Note that for known genotypes the two are equivalent
- A solution is found using an EM-algorithm

Solution: NGSadmix

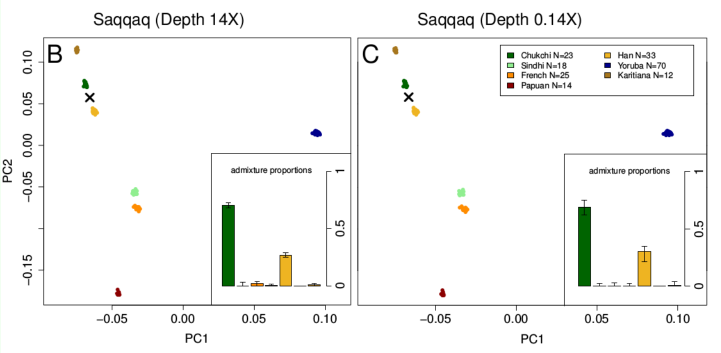
- Does well even for low depth and variable depth data:



Using reference data e.g. HGDP SNP chip

FastNGSadmix, Jorsboe *et al* 2016

- same model as NGSadmix, but uses a allele frequencies from reference panel
- similar to iAdmix (and ADMIXTURE projection) but takes reference size into account



Ancient Eskimo[a]

[a] Rasmussen *et. al.*, 2010

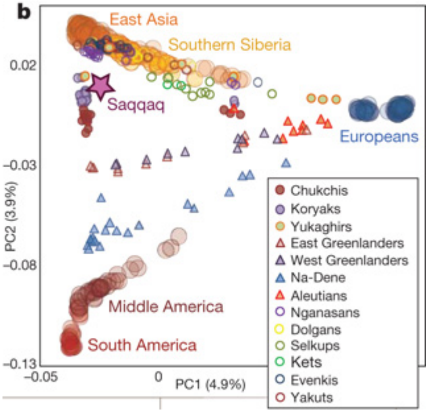


Figure: First principal components of selected populations.

singular value decomposition

SVD - singular value decomposition

$$G = UDV^T$$

- G does not have to be symmetric

PCA for a covariance matrix 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

Genotype data

5 individuals, genotypes

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

5 individuals, allele counts

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

IBS distances

5 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

Multidimensional scaling

Goal

Based on pairwise distances reduce the number of dimension by a transformation that preserves the pairwise distances as best as possible.

go from dimension N x N to N x S, where $S < N$

Principal component analysis for genetic data

SVD - singular value decomposition

$$\tilde{G} = UDV^T$$

PCA for a covariance matrix

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

- The first principal component/eigenvector accounts for as much of the variability in the data as possible
- Can be use to reduce the dimension of the data

Goal

Capture the population structure in a low dimensional space

Measure for pairwise differences

Identical by descent (IBS) matrix - used in MDS

Optimal way to represent pairwise distance in defined number dimensions

pros	fast
cons	Ignores allele frequency (bad weighting)
cons	Problems with some kinds of missingness

Covariance / correlation matrix - used in PCA

Optimal way to maximime the variance of the data

pros	better weighting scheme for each site
cons	Slower and cannot easily deal with missing data

Approximation of the genotype covariance

M number of sites

G genotypes

G^j genotypes for individual j

G^j_k genotypes for site k in individual j

f_k allele frequency for site k

variables (SNPs) should be identically distributed

- Same mean

  solution subtract the mean: $G^j_k - avg(G_k) = G^j_k - 2f_k$

- Same variance

  solution divide by standard deviation: $\frac{G^j_k}{\sqrt{var(G_k)}} = \frac{G^j_k}{\sqrt{2f_k(1-f_k)}}$

Approximation of the genotype covariance

M number of sites

G genotypes

G^j genotypes for individual j

G^j_k genotypes for site k in individual j

f_k allele frequency for site k

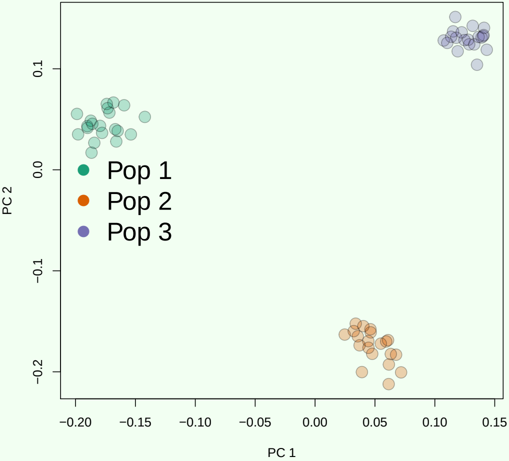
Known genotypes - covariance between individuals i and j[a]

[a] Patterson N, Price AL, Reich D, plos genet. 2006

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

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

The two first principal component



Early use of PCA in genetics

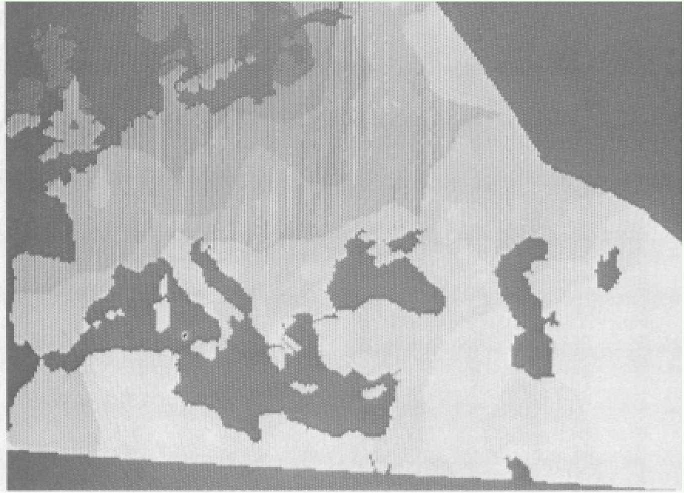


Fig. 1. The first principal component of gene frequencies from 38 independent alleles at the human loci: ABO, Rh, MNS, Le, Fy, Hp, PGM$_1$, HLA-A, and HLA-B. Shades indicate different intensities of the first principal component, which accounts for 27 percent of the tota variation and is represented with green shades in the photograph on the cover.

Shown

1 PC at 400 locations

Science 1978

Menozzi P, Piazza A, Cavalli-Sforza L.

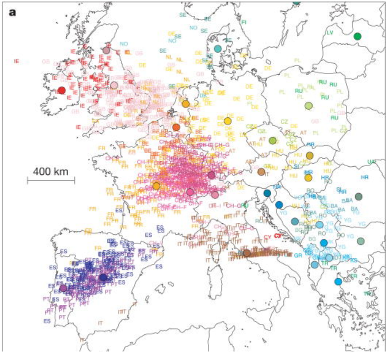
Data

38 loci

PCA mania

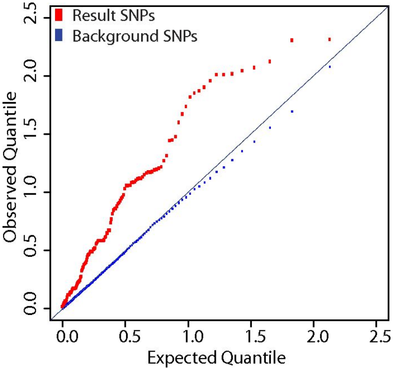
Genetic map from PCA[a]

[a] Novembre et. al, nat genet. 2008

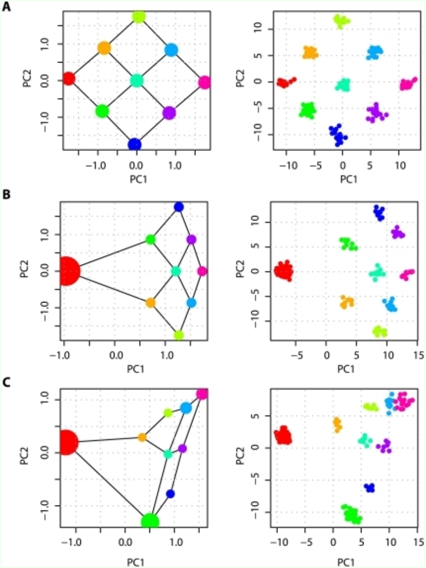


Eigenstrat[a]

[a] Price et. al, nat genet. 2008



Sample size/information bias



sample sizes will affect both the distance and the pattern [a]

[a] McVean G PLoS Genet. (2009)

Dealing with Missingness

Covariance matrix - Eigensoft [a]

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 ascertainment bias

[a] Patterson N, Price AL, Reich D, plos genet. 2006

IBS matrix

The site is skipped for the pair of individuals

- Missingness must be random

non random missingness

Major source

Using multiple source of data

- Two SNP chips with not all individuals typed using both.
- Using SNP chip for some and sequencing for others

other sources

- Differential missingness between individuals
- Sequencing at different depths

It is still kind of useful - and pretty

Ancient Eskimo[a]

[a] Rasmussen *et. al.*, Nature 2010

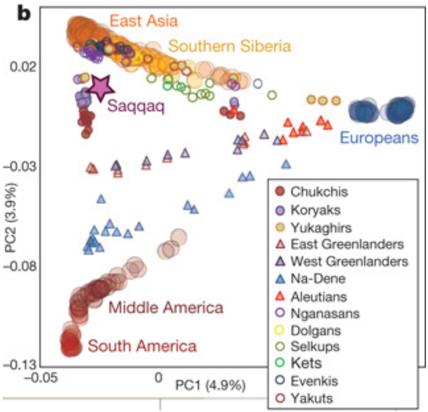


Figure: First principal components of selected populations.

PCA for NGS using genotype likelihoods

model with Known genotypes

$$cov(G^i, G^j) = \frac{1}{M}\sum_{m=1}^{M} \frac{(G^i_m - 2f_m)(G^j_m - 2f_m)}{2f_m(1-f_m)},$$

the model based on GL

$$cov(G^i, G^j) = \frac{1}{M}\sum_{m=1}^{M} \frac{\sum_{\{G^1,G^2\}}(G^1 - 2f_m)(G^2 - 2f_m)p(G^1, G^2|X^j_m, X^i_m)}{2f_m(1-f_m)},$$

PCA for NGS

the model based on GL[a]

[a] Skotte, *genet epi.* 2012

$$cov(G^i, G^j) = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{\{G^1, G^2\}} (G^1 - 2f_m)(G^2 - 2f_m) p(G^1|X_m^i) p(G^2|X_m^j)}{2f_m(1-f_m)},$$

were $p(G|X)$ is the posterior probability estimated using the allele frequency as a prior.
assumption: $p(G^1, G^2|X_k^j, X_k^i) = p(G^1|X_k^i, f_k)p(G^2|X_k^j, f_k)$,
with $p(G^1|X_k^i, f_k) \propto p(X_k^i|G_k^1)p(G_k^1|f_k)$

motivation is the same as eigensoft

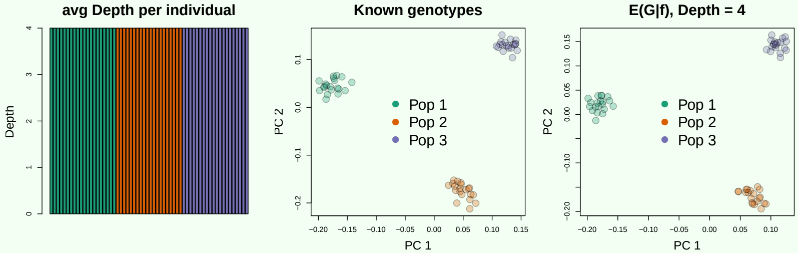
- $E[\tilde{G}_k^i] = 0$ for a random individual
- $E[cov(G^i, G^j)] = 0$ without relatedness or admixture.

PCA for NGS using genotyp7e likelihoods

the model based on GL[a]

[a] Skotte, *genet epi.* 2012

$$cov(G^i, G^j) = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{\{G^1,G^2\}} (G^1 - 2f_m)(G^2 - 2f_m) p(G^1|X_m^i) p(G^2|X_m^j)}{2f_m(1-f_m)},$$



PCA for NGS, ascertainment corrected

Model that work without inferring variable sites[a]

$$cov(G^i, G^j) = \frac{1}{M}\sum_{k=1}^{M} \frac{\sum_{\{G^1,G^2\}}(G^1-2f_k)(G^2-2f_k)p(G^1|X_k^i)p(G^2|X_k^j)p_{var}^k}{2f_k(1-f_k)\sum_{k'=1}^{M} p_{var}^{k'}},$$

were $p(G|X)$ is the posterior probability estimated using the allele frequency as a prior

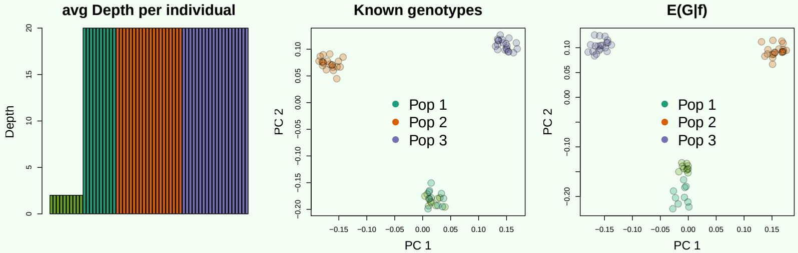
[a] Fumagalli, et al, Genetics, 2013

The assumption of independence can be problematic

the model based on GL[a]

[a] Skotte, *genet epi.* 2012

$$cov(G^i, G^j) = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{\{G^1,G^2\}} (G^1 - 2f_m)(G^2 - 2f_m) p(G^1|X_m^i) p(G^2|X_m^j)}{2f_m(1 - f_m)},$$



The problem under extreme depth differences

The assumption is valid under HWE[a] for unrelated individuals

[a] Fumagalli, et al, Genetics, 2013

$p(G^i, G^j | X_m^j, X_m^i) = p(G^i | X_m^i, f_m) p(G^j | X_m^j, f_m)$ assuming known allele frequency

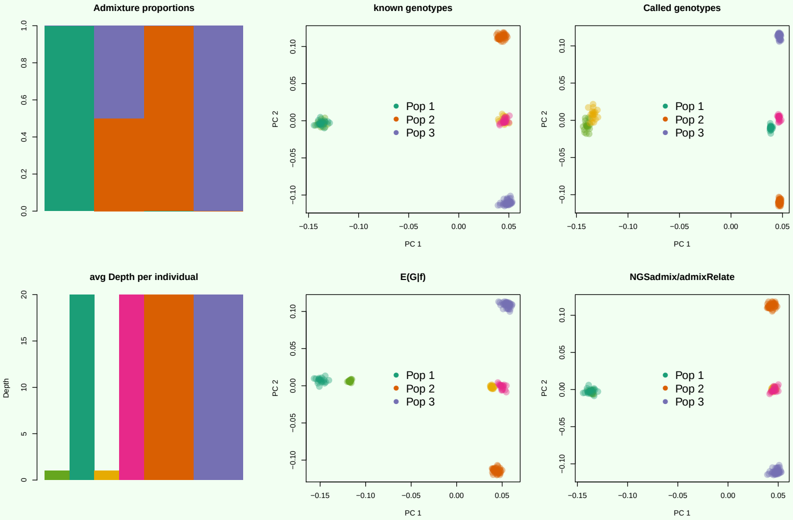
One solution - IBS/Cov matrix based on a sample of a single read

$d(g_m^i, g_m^j) = \begin{cases} 0 & if \quad g_m^i \neq g_m^j \\ 1 & if \quad g_m^i = g_m^j \end{cases}$ or $C = \frac{1}{M} \sum_{m=1}^{M} \frac{(g_m^i - f_m)(g_m^j - f_m)}{f_m(1-f_m)}$

GL solution - with better 'priors' based on NGSadmix model

$p(G^i, G^j | X_m^j, X_m^i) = p(G^i | X_m^i, \hat{F}, \hat{Q}_i) p(G^j | X_m^j, \hat{F}, \hat{Q}_j)$ same model as in NGSadmix

Admixture aware prior is not affected by depth bias



NGS framework for heterogenious samples

Admixture aware priors

Instead of a single allele frequency we will use a different prior for each individuals

Admixture proportions priors

individual allele frequency at site i: $\pi_i = q_i^1 f_i^1 + q_i^2 f_i^2 + ... + q_i^k f_i^k$

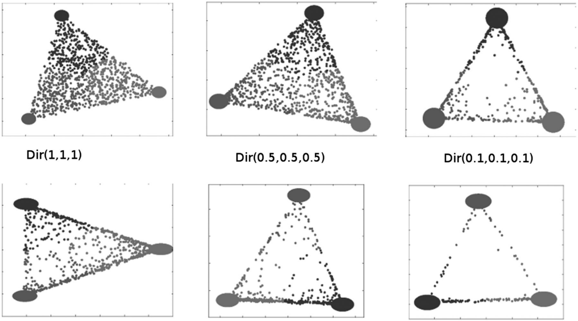
PCA based priors is also possible

individual allele frequency predicted from the PCA

Individual allele frequencies from PCA

Intuition by Popescu et al. 2014

There are some simplex or planes in the PCA that will represent admixture proportions



Dir(1,1,1)

Dir(0.5,0.5,0.5)

Dir(0.1,0.1,0.1)

Dir(0.2,0.2,0.5)

Dir(0.2,0.2,0.05)

Dir(0.05,0.05,0.01)

Individual allele frequencies from PCA

Many ways the principal components predict deviations from the joint allele frequency, Hao et. al (2015)

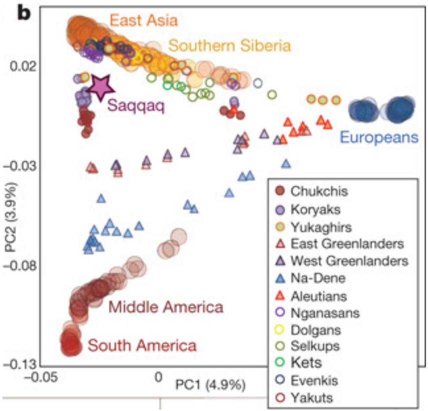


Figure: Rasmussen et al 2010

Individual allele frequencies from PCA

Concept used Conomos et al. 2016 (PC-relate)

The principal components predict deviations from the joint allele frequency

linear model: $\pi_i = \alpha + V\beta$

- π_i individual allele frequencies for site i
- α average allele frequency for site i
- V top principal components (coordinates)
- β allele frequency difference from the average allele frequency)

α and β estimated from the expected genotypes

$E[G|X,\pi]/2 = \alpha + V\beta$, where $E[G|X,\pi] = \sum_{g\in\{0,1,2\}} p(G=g|X,\pi)g$

remember that $p(G=g|X,\pi) = \frac{P(X|G)P(G|\pi)}{P(X|\pi)}$

individual allele frequencies from PCA

Hen and the Egg problem

- if we know the individual allele frequencies we can make the PCA
- if we know the PCA we can get the individual allele frequencies

One Solution

Iterative updating - PCangsd by jonas Meisner

individual allele frequencies from PCA

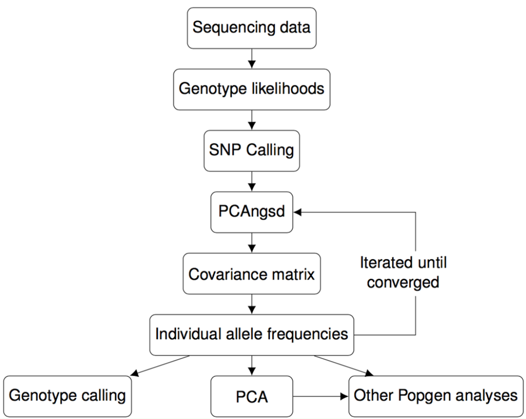


Figure: PCAngsd framework

1000 Genomes - true genotypes

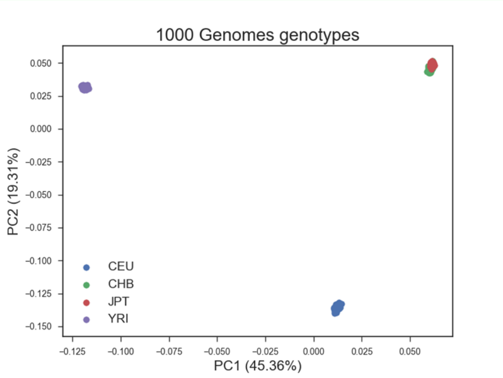


Figure: 1000 Genomes data

1000 Genomes - called genotypes from low depth

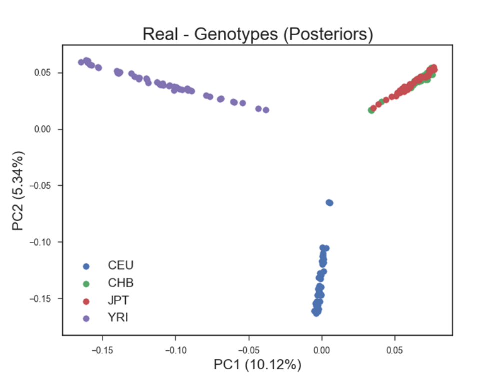


Figure: 1000 Genomes data

1000 Genomes - Genotype likelihood with frequency prior

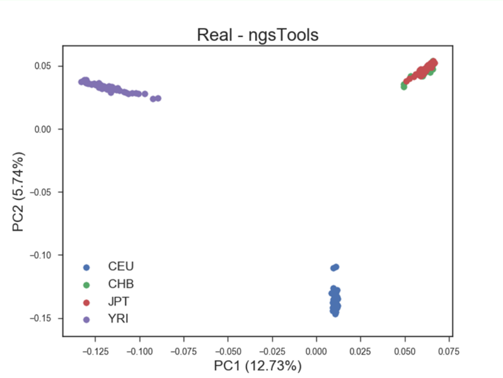


Figure: 1000 Genomes data

1000 Genomes - Genotype likelihood with individuals frequency prior

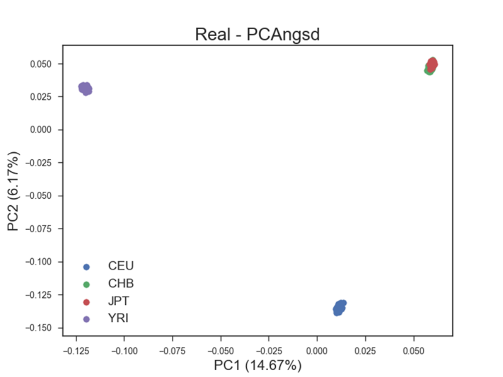


Figure: 1000 Genomes data

Admixture VS PCA

indirect goal of both ADMIXTURE and PCA

To predict the individual allele frequencies Π from lower dimensional matrices. $E(G) = 2\Pi$

ADMIXTURE

K=N-1 $G = 2QF^T$

K low $\Pi \approx QF^T$

ADMIXTURE → PCA

$$cov(\tilde{G}^i, \tilde{G}^j) = \frac{1}{M} \sum_{m=1}^{M} \frac{(\Pi^i_m - f_m)(\Pi^j_m - f_m)}{f_m(1-f_m)} = \frac{1}{M} \tilde{G}\tilde{G}^T$$

PCA

K=N-1 $\tilde{G} = UDV^T$

K low $\tilde{\Pi} \approx U_{[K]}DV^T_{[K]}$

PCA → ADMIXTURE

$argmin_{Q,F} ||\Pi - QF^T||^2_F$

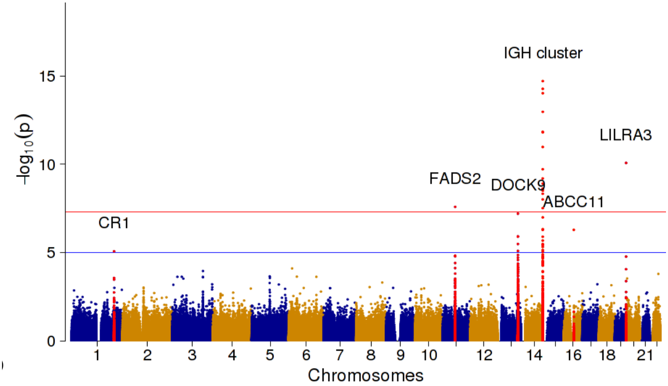
Solved with NMF with penalty

Selection scan from PCA for NGS data

FastPCA test statictic from Galinsky et al (2016)

$$\frac{M}{D_k^2}(2\Pi_m V_k)^2 \sim \chi^2$$

selection scan in >100k Han chinese with low depth sequencing $< 0.1X$



Inbreeding and admixture

joint allele frequencies f

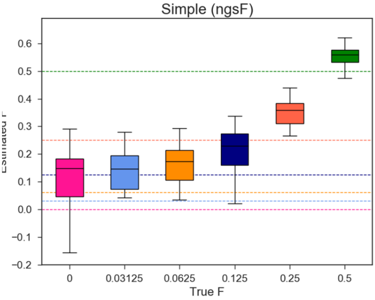


Figure: Simulated inbreeding from admixed 1000G individuals

Individual allele frequencies Π

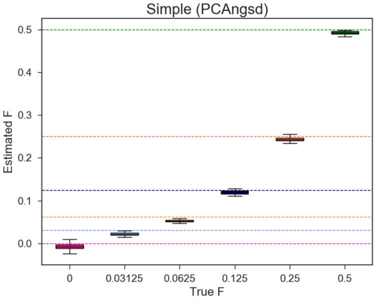


Figure: Simulated inbreeding from admixed 1000G individuals

The end

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 are useful when working with genotype likelihoods