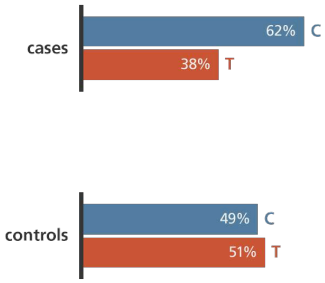
Population Structure in GWAS

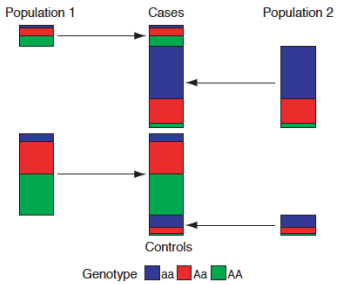
H3ABioNet 2018 Genotyping Chip Data Analysis and GWAS lecture series

Ananyo.Choudhury@wits.ac.za

Section I
Population structure

- Association testing
- Stratification in Association
- Basic idea of confounders
- Genetic associations
- Population structure
- How population structure affects GWAS?



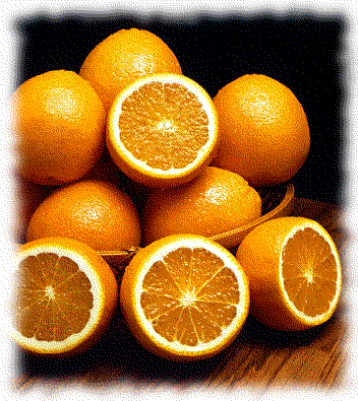


Association testing

Fertilizer increases higher vitamin C content in fruits?

Site A
Fertilizer
Applied

Site B
No
Fertilizer



Measure vitamin C content

Compare - estimate statistical significance of difference

Measure vitamin C content

Fertilizer application is associated to Higher Vitamin C production or not

Site A

Site B

Site A

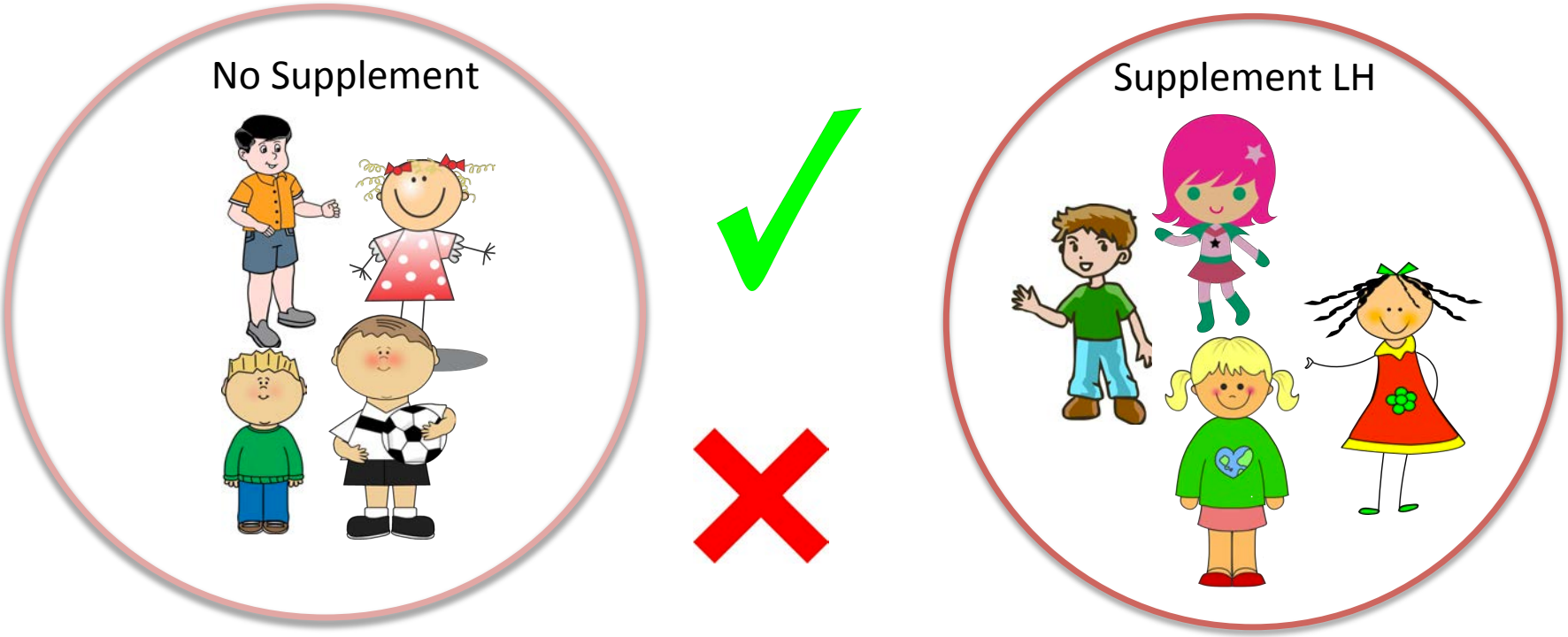
Site B

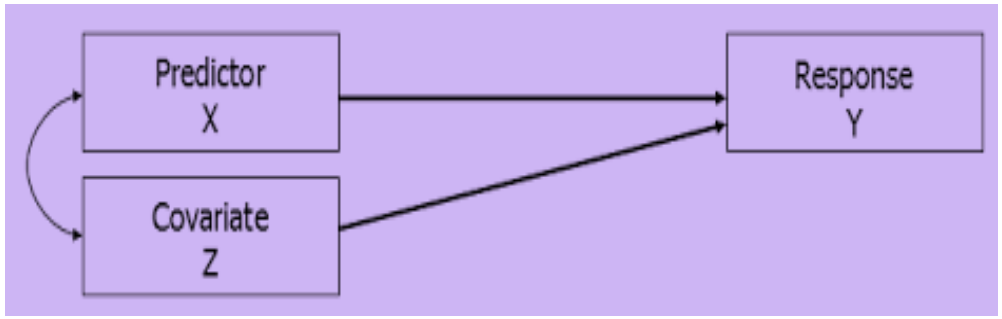
Significant statistic difference in Vitamin C content

Not due to fertilizer application, but due to stratification

Stratification - where association is observed due to systemic differences in the groups rather than differences in the outcome variable

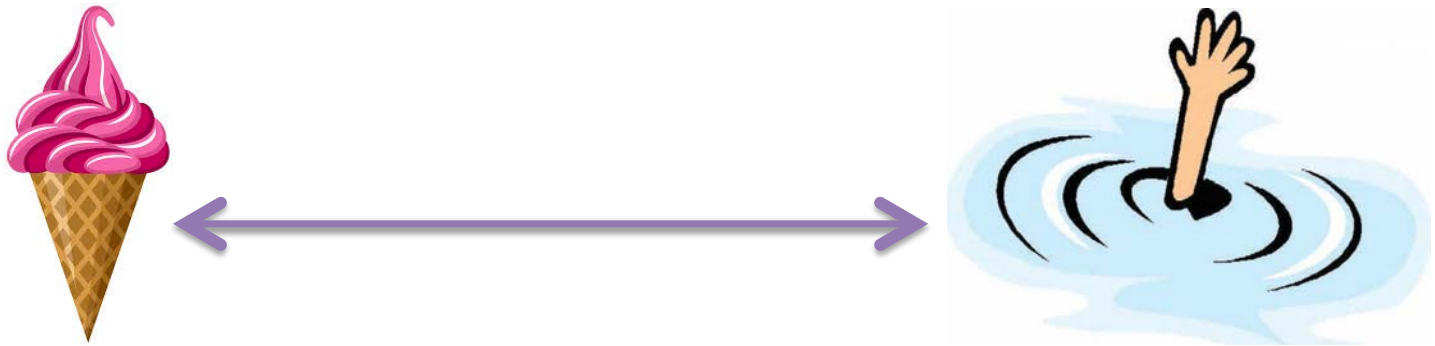
Kids taking diet supplement LH grow longer hair ?

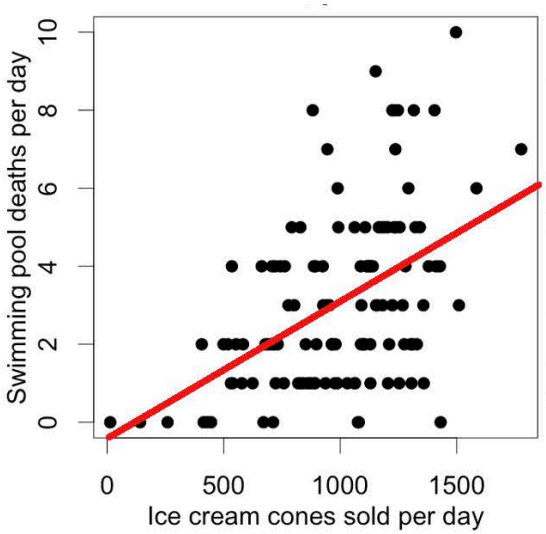




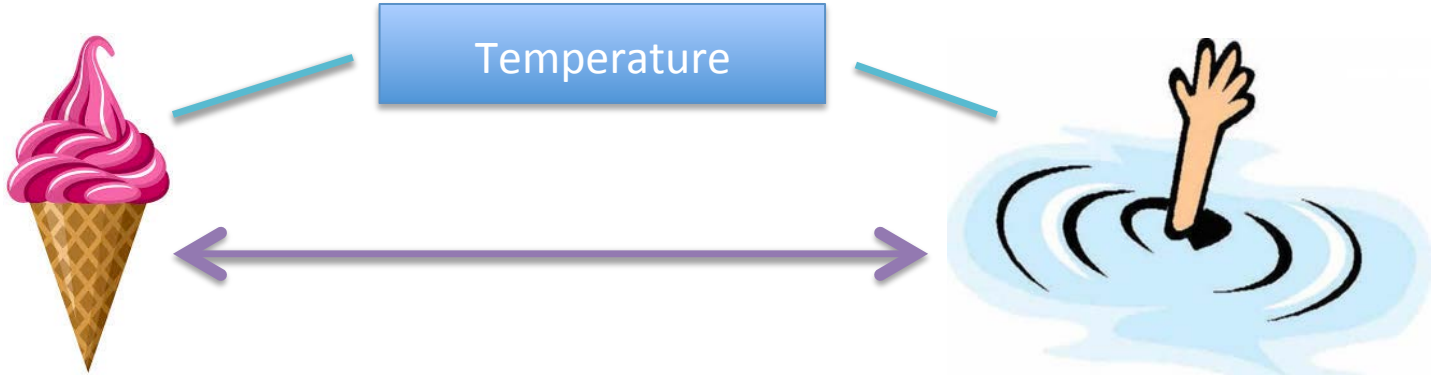
Confounding : Factors other than the major variable that could effect the outcome.

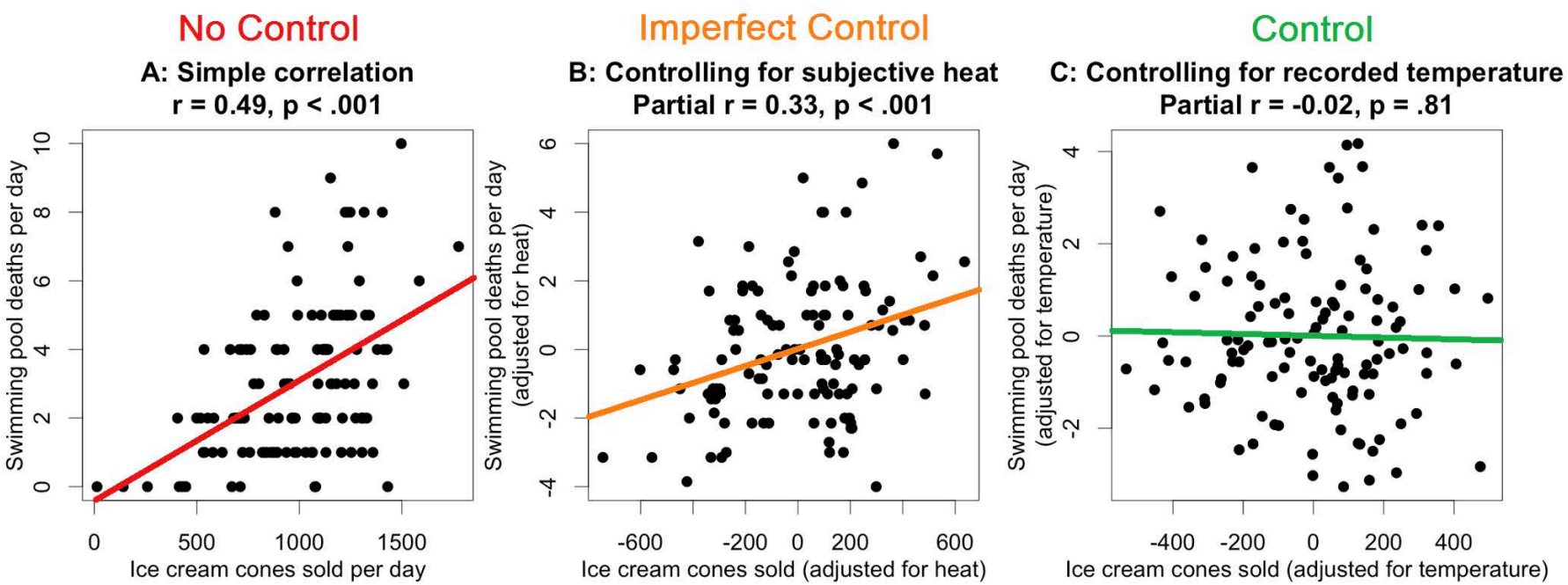
Example of controlling structure with covariates



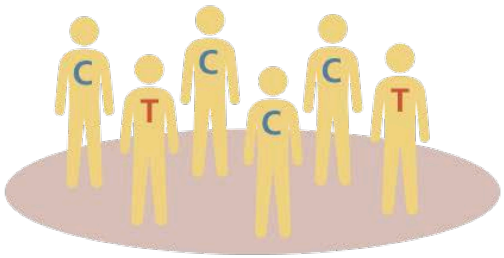


Example of controlling structure with covariates

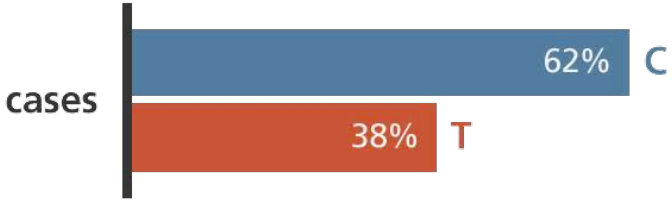


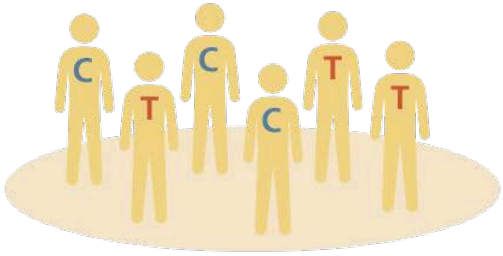


Genetic Association Studies



cases (n=1,000)
people with heart disease





controls (n=1,000)
people without heart disease

Results from a **case-control** genome-wide association study investigating genetic variants associated with heart disease.

Image credit: Genome Research Limited

Continuous trait

A linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

Data:

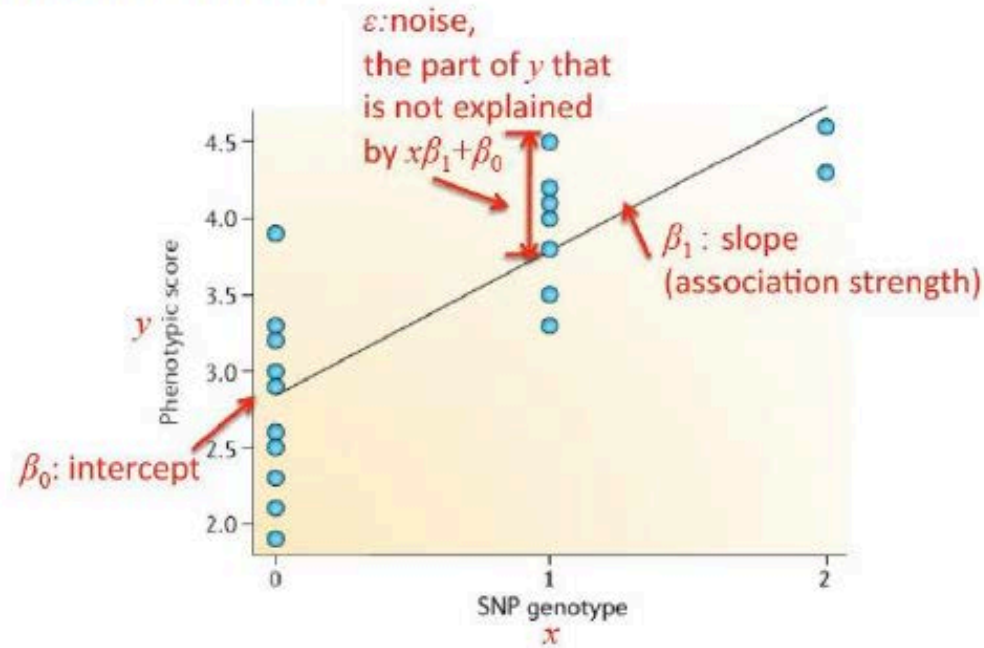
- y: a continuous trait
- x: SNP genotype at a given locus

Parameters:

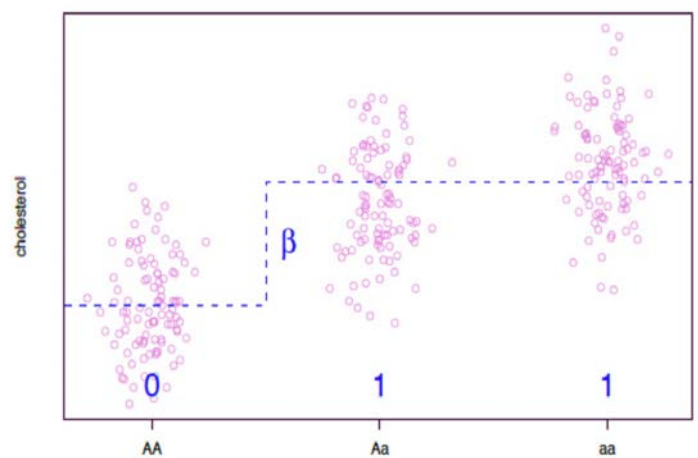
- β_1: regression coefficient, represents the strength of association between x and y
- β_o: intercept term (is 0 or ignored)
- ε: noise or the part of y that is not explained by x (e.g., environmental effect)

Assumptions:

- The individuals in the study are not related
- The phenotype y has a normal distribution



$$y = \beta_0 + \beta \times (G \neq AA)$$



GWAS for the chopstick gene

High chopstick skills

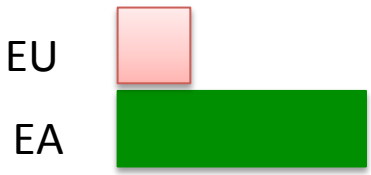
Low chopstick skills

SUSHI

Successful Use of Selected Hand Instrument !!

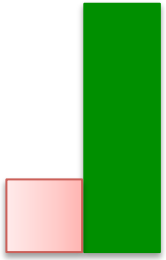
GWAS for the chopstick gene

High chopstick skills



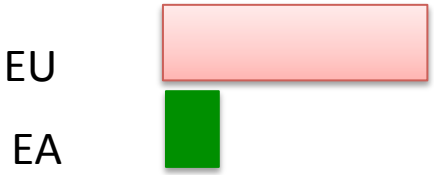
Ancestral composition HCS

SUSHI



Allele frequency of variant - HLA-A in European (EU) and East Asians (EA)

Low chopstick skills



Ancestral composition LCS

Successful Use of Selected Hand Instrument !!

Molecular Psychiatry (2000) 5, 11–13

NEWS & VIEWS

Beware the chopsticks gene

NATURE REVIEWS | GENETICS

The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] and Magnus Nordborg[3,4]

VOLUME 14 | JANUARY 2013

Population stratification/structure is the presence of multiple subpopulations (e.g., individuals with different ethnic background) in a study.

Subpopulations in addition to differing in allele frequencies might also differ in disease rate, trait variable, cultural practices, diet etc.

If both allele frequencies and trait variables differ between subpopulations, PS can lead to false positive associations and/or mask true associations.

Associations masked

Vol. 11, 505–512, June 2002 Cancer Epidemiology, Biomarkers & Prevention 505

Point/Counterpoint

Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations?[1]

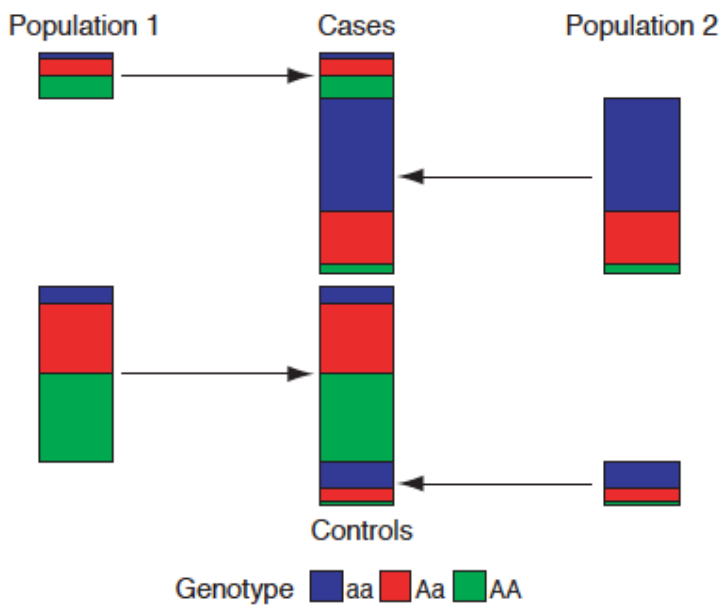
VOLUME 36 | NUMBER 5 | MAY 2004 NATURE GENETICS

The effects of human population structure on large genetic association studies

Jonathan Marchini[1], Lon R Cardon[2], Michael S Phillips[3] & Peter Donnelly[1]

Spurious associations

Confounding by Population Structure



		Phenotype			
Population	**Allele**	**Case**	**Control**	**Total**	**Association**
1	A	**270**	30	**300**	no
	B	630	70	700	
	Total	**900**	100	1000	
2	A	**80**	720	**800**	no
	B	20	180	200	
	Total	**100**	900	1000	
Pooled	A	**350**	750	**1100**	Yes
	B	650	250	900	$P < .0001$
	Total	**1000**	1000	2000	

- There could be a number of other confounders such as age, sex and BMI which could affect GWAS inference

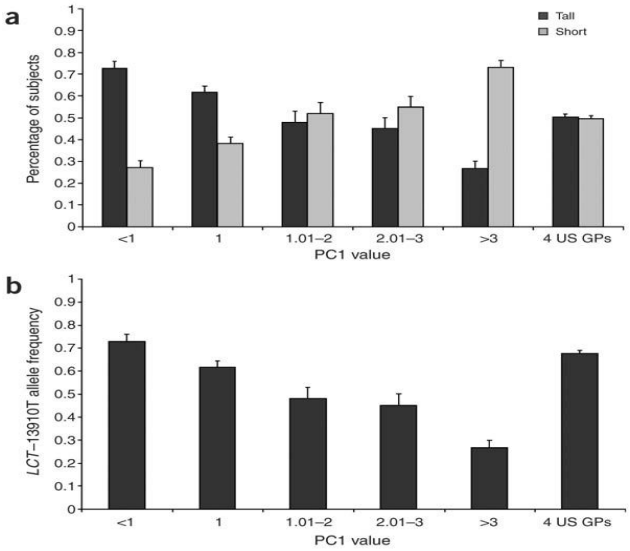
Effect of PS on GWAS

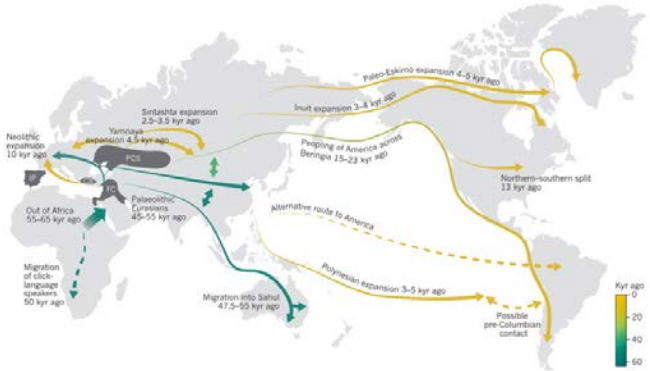
- Campbell et al. 2005, conducted a GWAS for height based on a set of European American individuals.

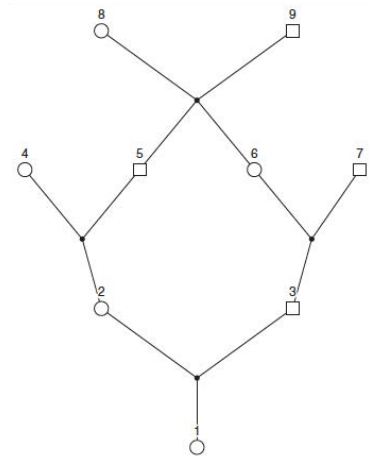
- They found a SNP in the gene LCT strongly to be associated associated with height (P < 10−6)".

- They also noted this SNP to show high allele frequency variation among European populations.

- Grouping individuals on the basis of European ancestry greatly reduced the apparent association that was due to population stratification.







Section II
Types of Population Structure and their origin

- Types of Structure
- Human history for the last few hundred millennia
- Cryptic Relatedness

Types of PS

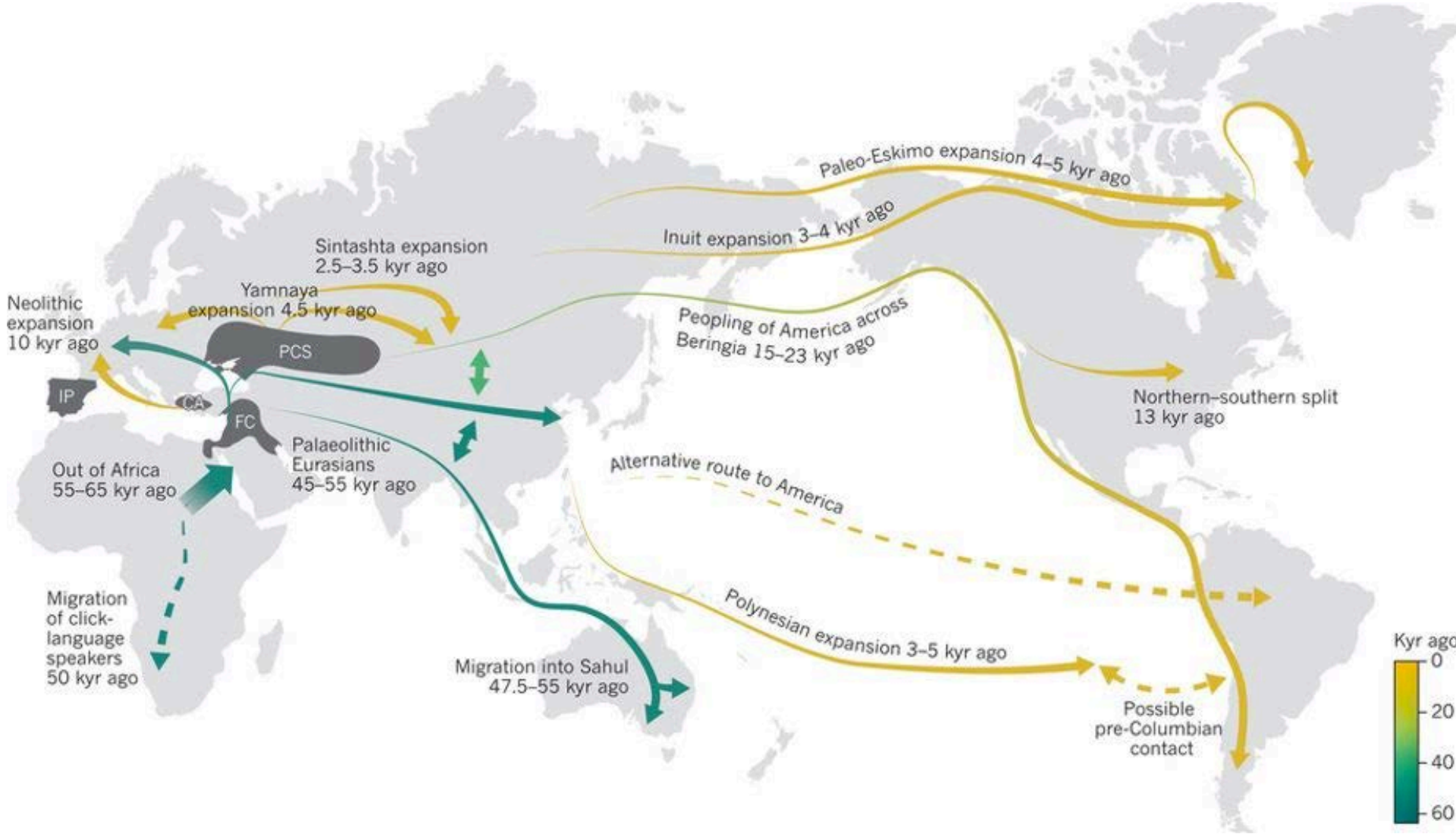
- Structure due to inclusion of usually geographically distinct populations
  - NW Europeans and SE Europeans

- Structure due to variation in ancestral contributions (differential admixture) from genetically distinct populations.
  - Gila river community

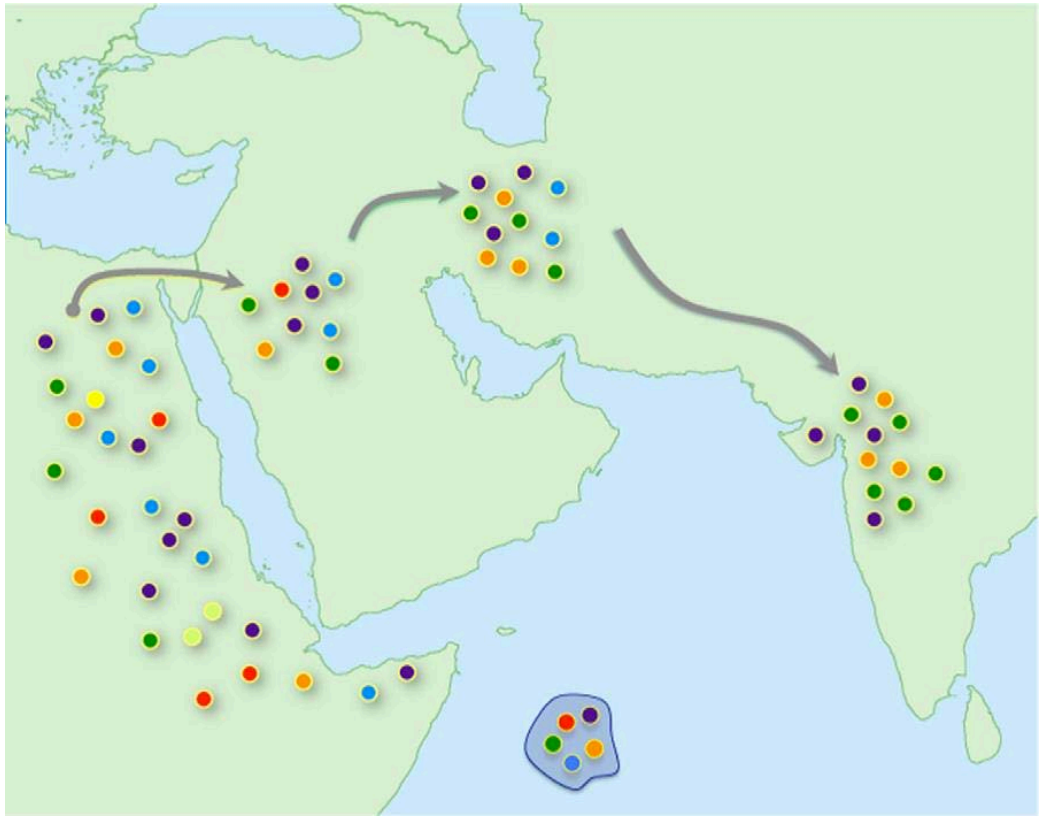
Indian Heritage	$Gm^{3;5;13;14}$ %	% Diabetes age adjusted
0	69%	18.5%
4	45%	28.6%
8	.01%	39.2%

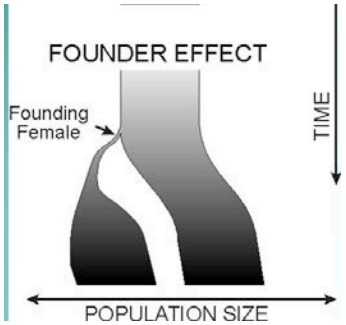
Adapted from Knowler, 1988

Peopling of the world

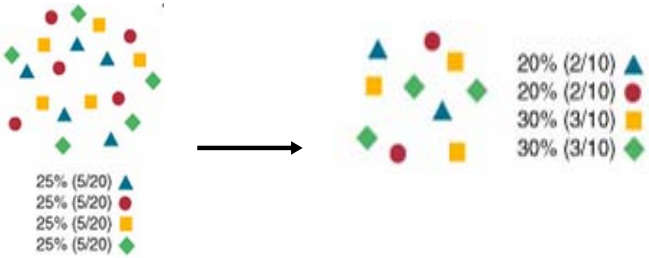


Major human migrations across the world inferred through analyses of genomic data. (Nielsen et al. 2017)

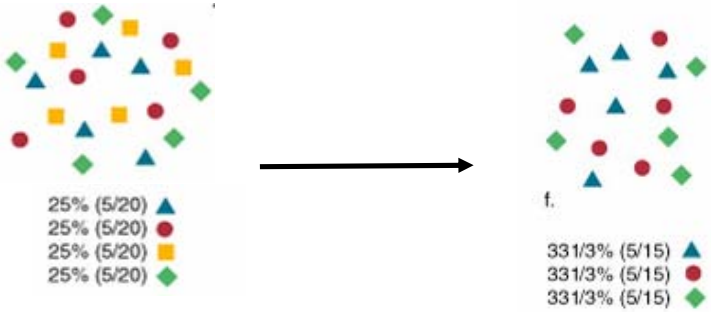




Genetic Drift



Natural Selection



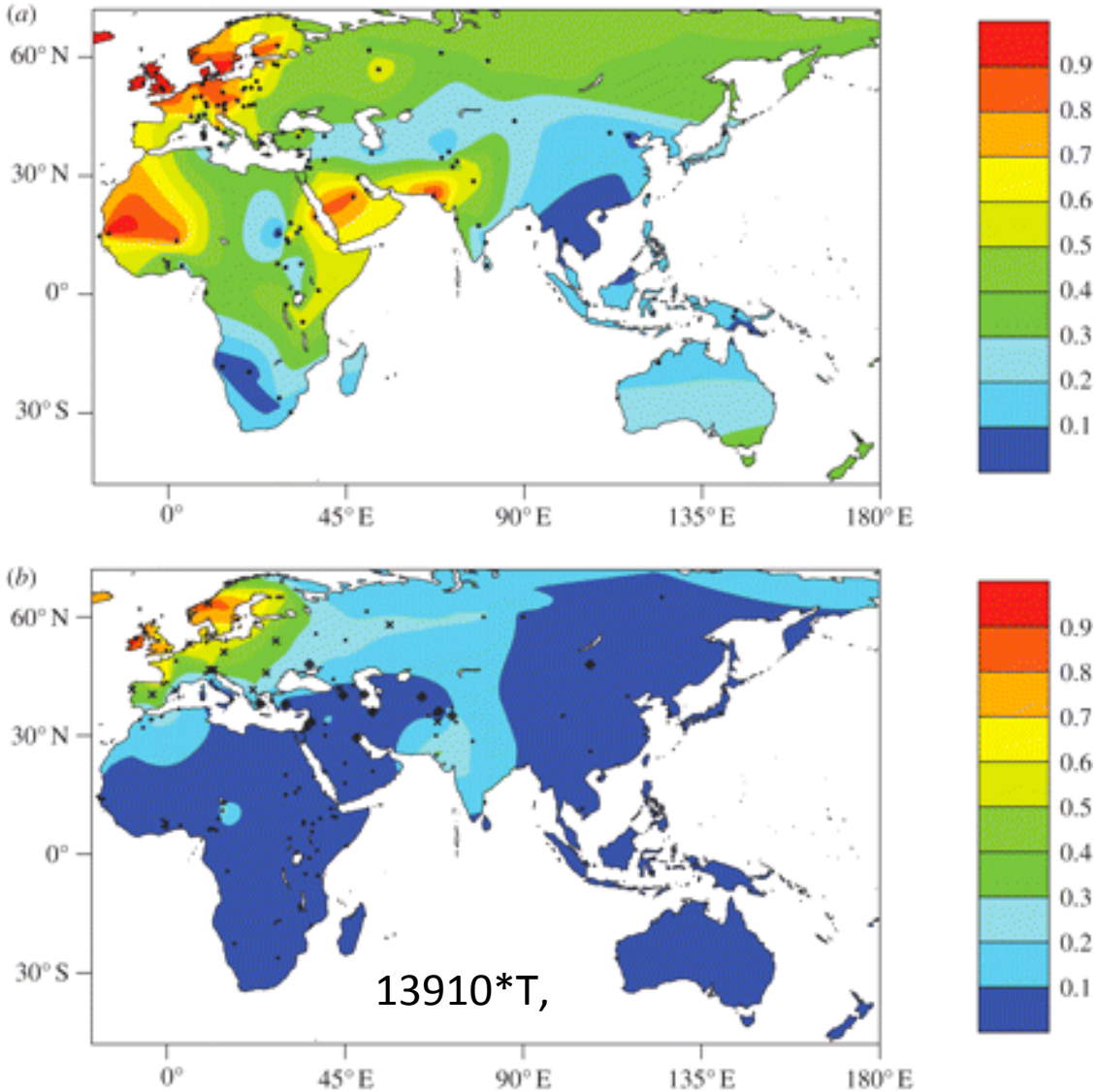
Change in human diversity with migration. The number of different colors (representing amount of genetic variation) reduces as explorers move into newer territories

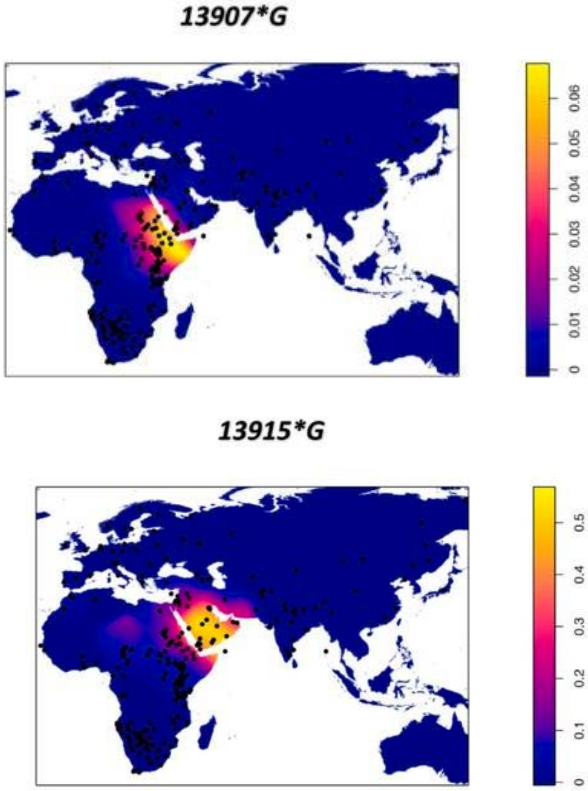
.

(From: https://blogs.plos.org/)

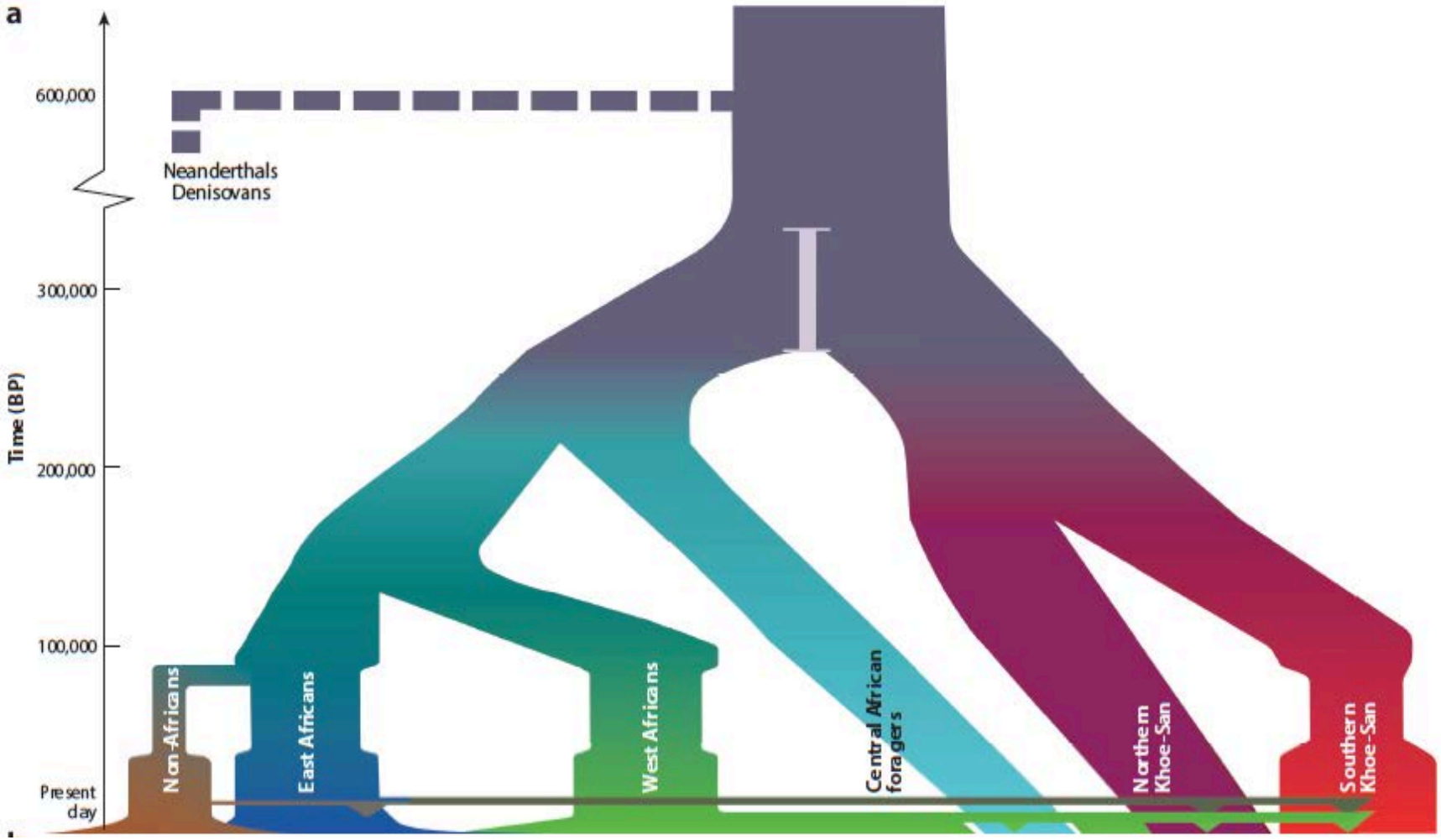
Lactase Persistence

http://www.ucl.ac.uk/mace-lab/resources/glad

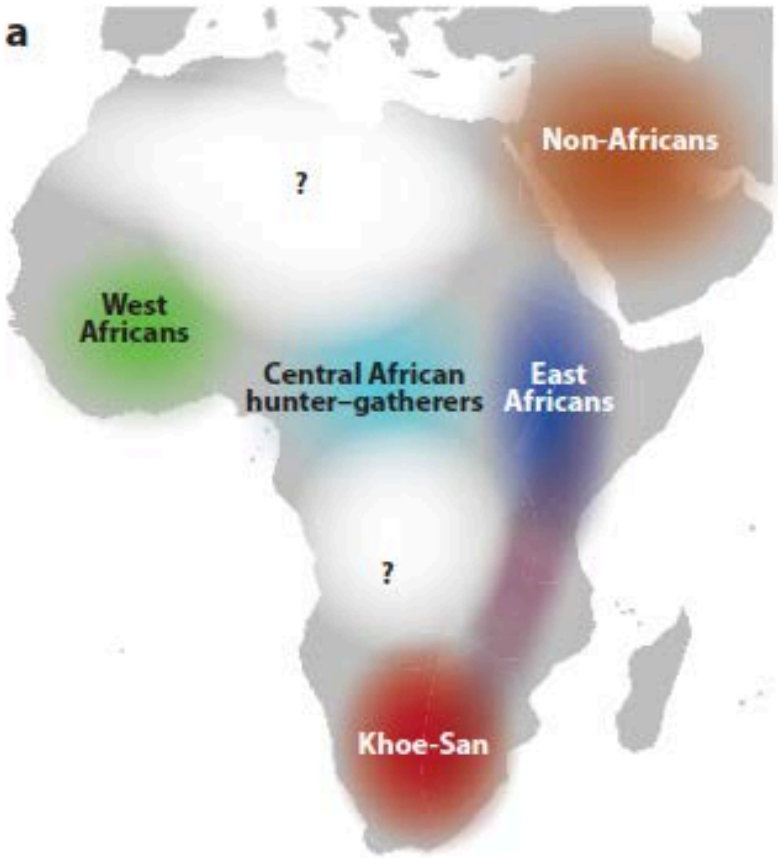


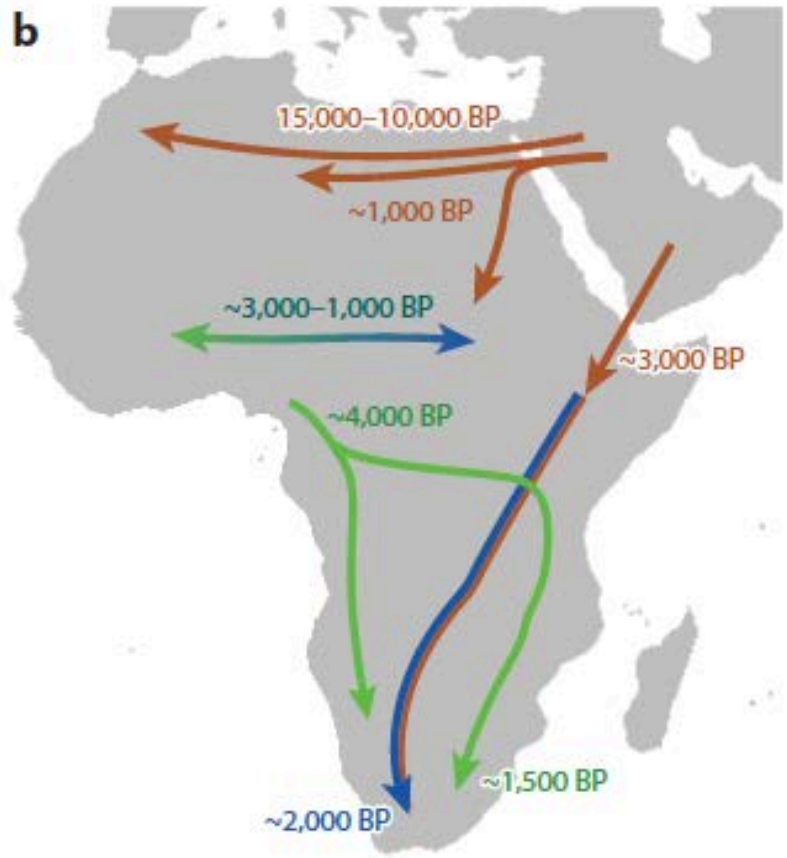


Gerbault et al. 2014



Schlebusch C and Jakobsson M, 2018

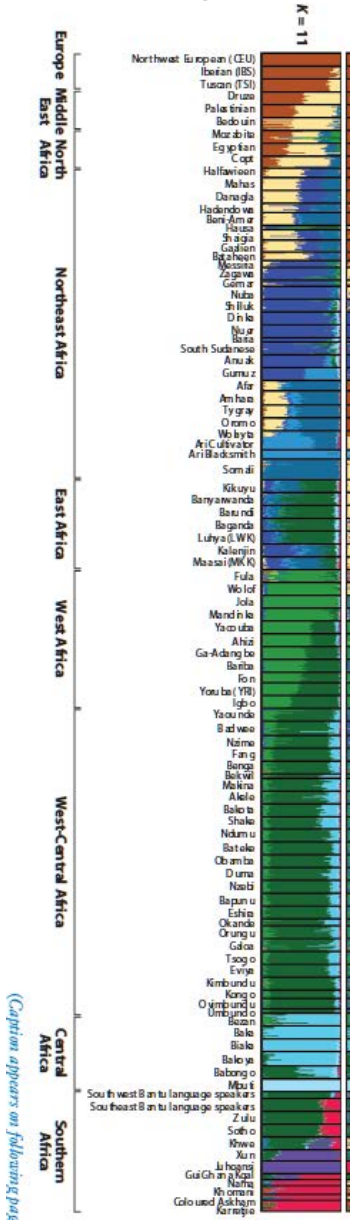




Map of major pre-farming population stratification across the African continent, showing non-Africans

Migration of herders and farmers since Holocene

Ethnolinguistic composition of present day Africa



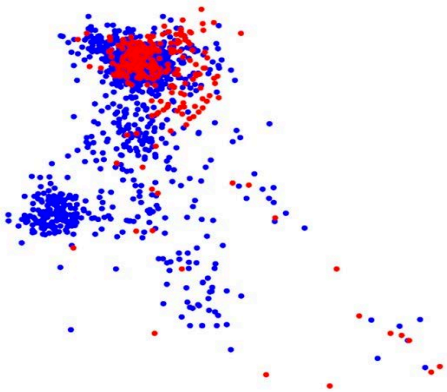
Cryptic relatedness

- Cryptic relatedness refers to the idea of presence of relatives in a set of ostensibly unrelated individuals in a case-control association study.

- An essential assumption for a GWAS is the independence of subject genomes. Cryptic relatedness violates this assumption and could, thereby, confound the inference of an association study.

- Factors that give rise to cryptic relatedness
  - Assortative mating
  - Effective population sizes /Recent bottleneck
  - Sampling biases

Section III.
How to detect PS and CR?

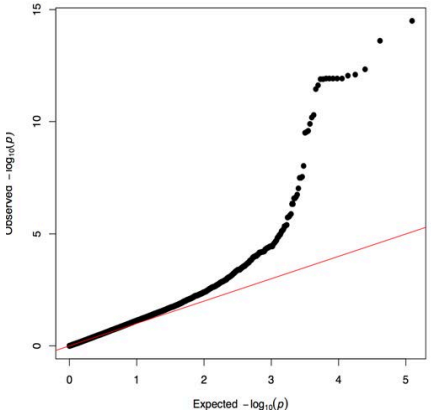


Pre association

- PCA based approach
- IBD based approach

Post association

- Genomic control
- QQ plots



Other approaches

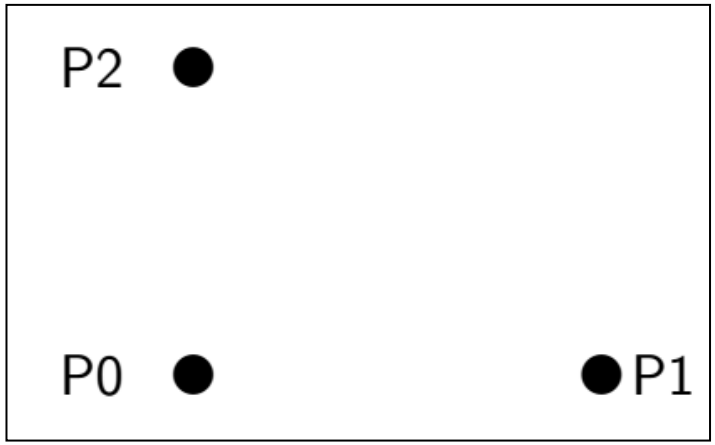
Principal Component Analysis

	1 2	3 4	5 6	7 8
P0:	AA	AC	AT	AA
P1:	TT	AA	TT	AA
P2:	AA	AC	TT	TT

Genotype information

Distance:

- P0, P1: 4
- P0, P2: 3
- P1, P2: 5



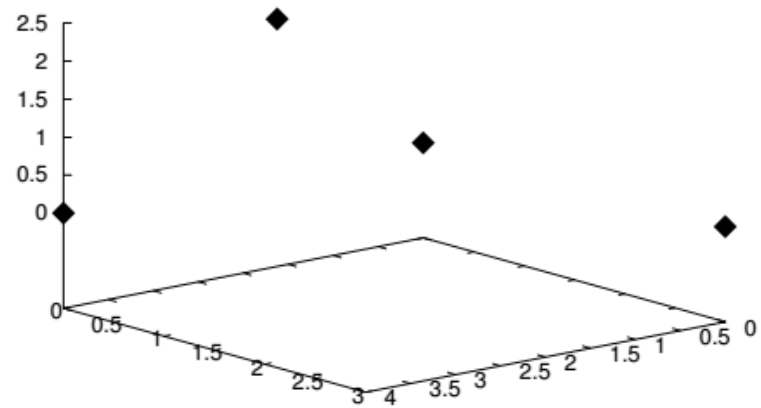
Genotype information:

```
P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
P3: AT CC TT AA
```

Distance:

	P0	P1	P2	P3
P0	0	4	3	3
P1		0	5	3
P2			0	4

Can only be embedded in 3D space



P0: (0,0,0) P1: (4,0,0)
P2: (0,3,0) P3=(2,0.33,2.21)

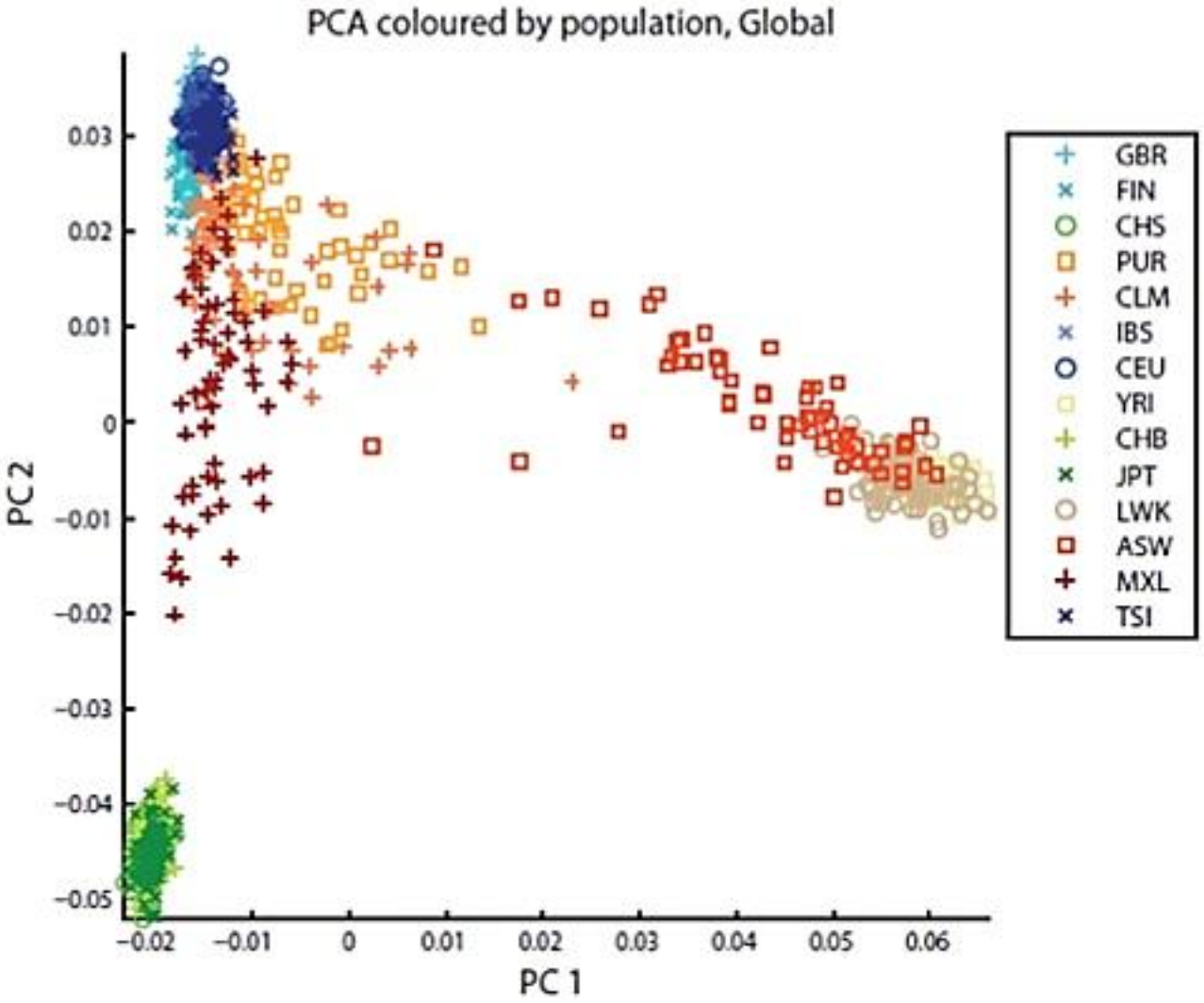
PCA

Uses a method known as *eigendecomposition* – in which it takes distance matrix and produces:

- **Eigenvalues**: λ_i is relative importance of dimension *i*
- **Eigenvectors**: v_i coordinates of each individual in the i-th dimension.

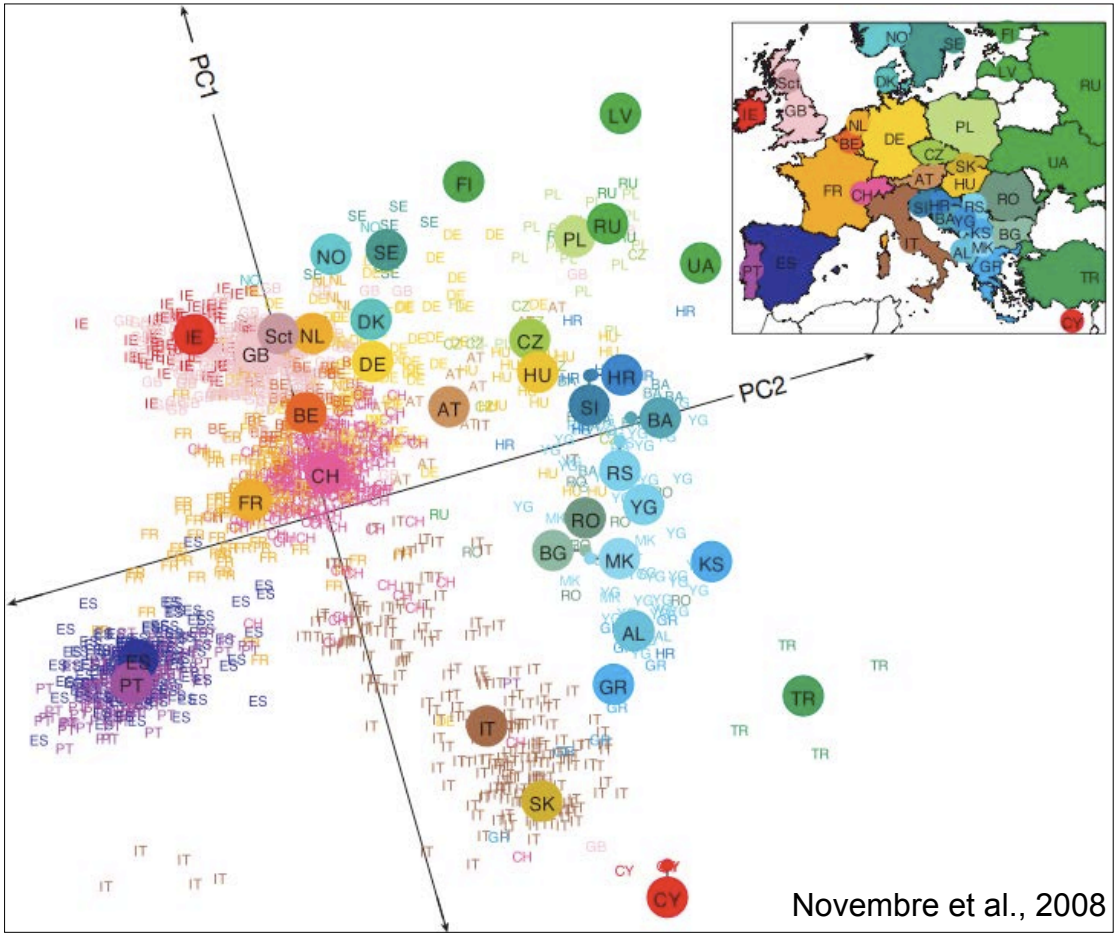
- Preserve relative distance between individuals.
- Number of dimensions/components are reduced.
- Components independent of each other.
- Ordered by importance.

Example of PCA



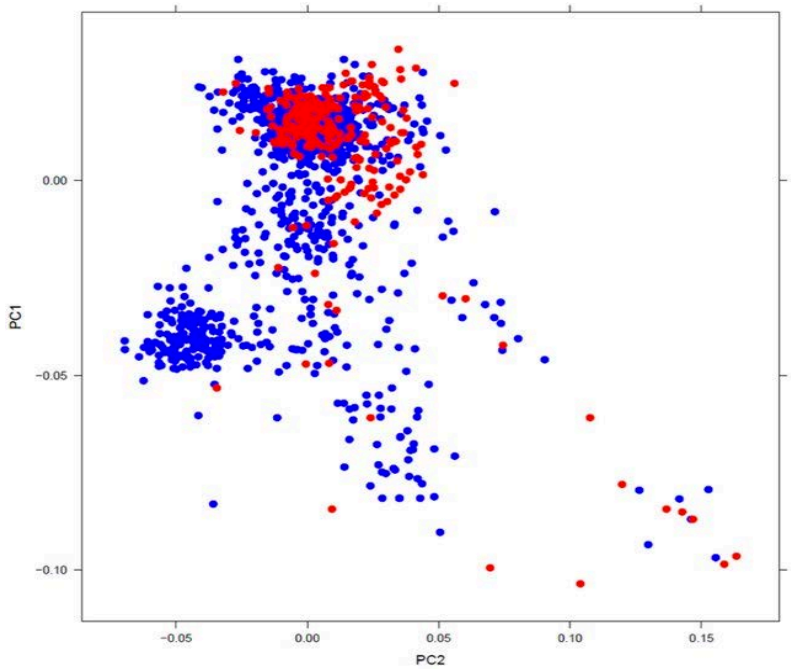
1000 Genomes, 2012

To what extent can population structure be detected using PCA?

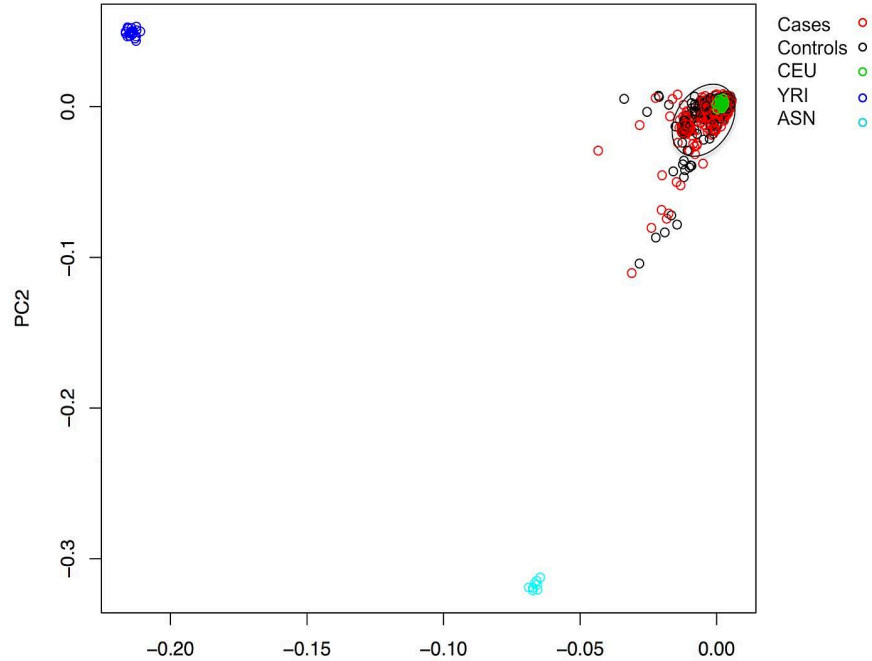


Novembre et al., 2008

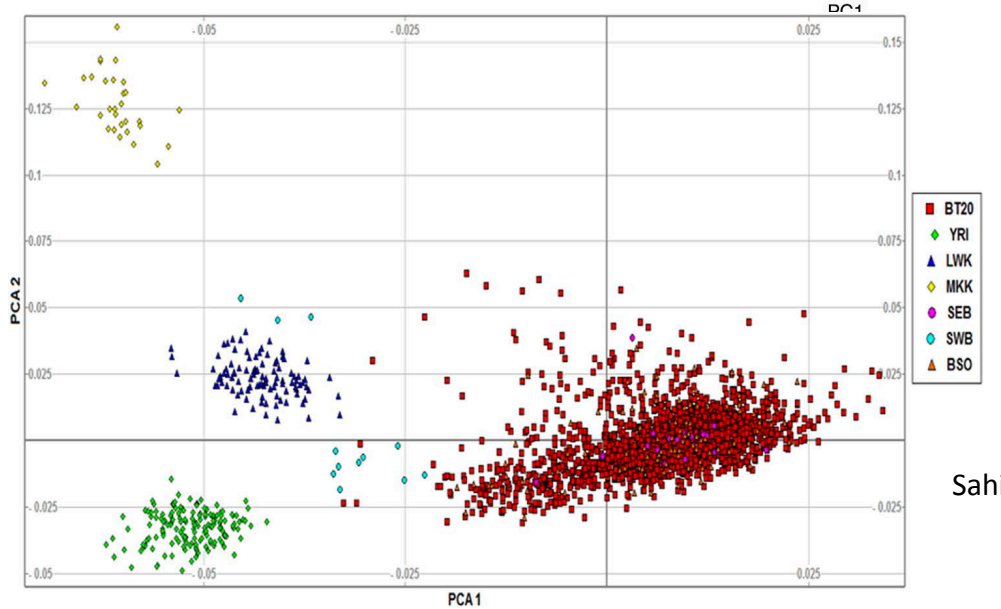
A geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans



Behr et al. 2017



Hulur et al. 2017



Sahibdeen et al. 2017

PCA is easy

it is a advisable to perform a LD based pruning before running the PCA

```
Running PCA

plink --bfile mydata
--pca --out mypca
```

creates the file mypca.eigenvec and mypca.eigenval, containing the eigenvector and values respectively.

Home

Genesis PCA and Admixture Plot Viewer

Wits Bionformatics, Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg

Genesis can be used to display screen and publication quality pictures of population PCA and admixture charts

Why use Genesis?

Genesis takes the output of popular programs such as Admixture and EIGENSTRAT and produces good quality pictures, which the user can interactively change. There are first class tools that can be used to create good quality pictures, but they require expertise to use and best used when one already knows exactly what the output should look like. In practice there is a huge need for an interactive tool. Which PCs display interesting data can be interactively explored. Which colours are best to use is not just an aesthetic problem: in some cases a set of colours works well but with other data the same colours doesn't because the colours don't clearly contrast with a new position of the objects being drawn. There may be a need to rearrange the labelling or the data. We want to make the fonts as big as possible, but what is "big as possible" depends on the quantity and arrangement of data. Often when displaying admixture charts, multiple charts are shown in one diagram, we need to to keep consistency of colours and may want to play with the ordering of data.

We see the need for an interactive tool that can be used to explore possibilities and produce good quality data. Although tools like Distruct and R are more flexible and produce very high quality pictures, Genesis is interactive and requires much less expertise to use.

Requirements

Genesis requires Java 1.8 with SWT libraries installed. Genesis runs on Windows, Linux and MacOS X. For Mac OS X, X11 must be installed. (Download XQuartz here)

Quick start

On Windows and Linux, the program should be run as

```
java -jar Genesis.jar
```

On Mac OS X, X11 must be installed and the program should be run as

```
java -XstartOnFirstThread -jar Genesis.jar
```

Some sample data files can be found here

Downloads

- Download executable and/or source code from here

Latest version:0.2.6b
November 2015

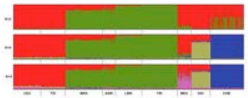
Documentation

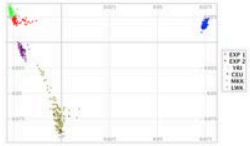
The manual is available as

- a PDF file
- html

The main distribution includes documentation in PDF, HTML and info format. Amateur videos showing use of Genesis for PCA and admixture are also available.

Click on the pictures below for examples.



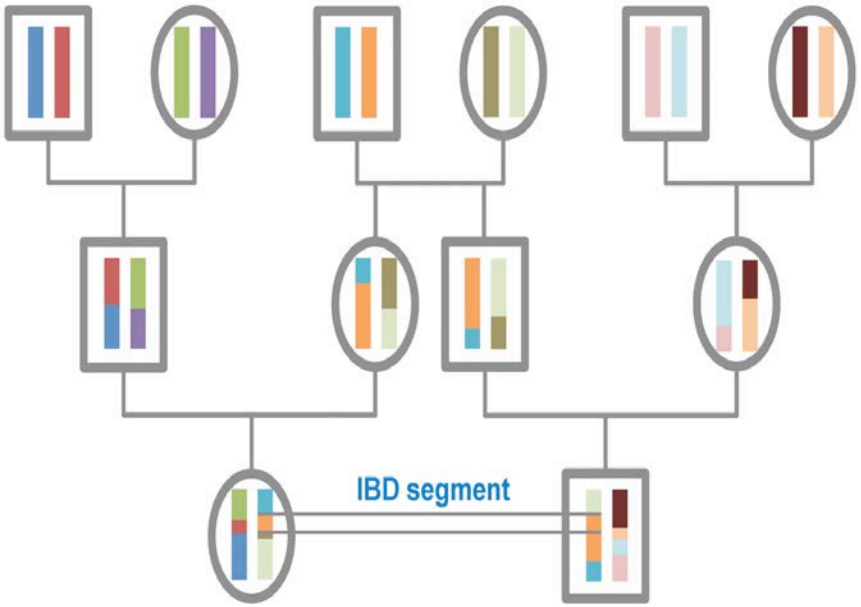


http://www.bioinf.wits.ac.za/software/genesis/

Assessing Structure Due to Relatedness

- Presence of duplicate and related individuals in the dataset may **introduce bias** and cause **genotypes in families** to be over-represented.

- To identify duplicate and related individuals, a metric (identity by descent, IBD) is calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosomes)



- The degree of **recent shared ancestry for a pair of individuals** (identity by descent, IBD) can be estimated using genome-wide IBS data using Plink. (IBD shown as pi_hat in plink)

```
CALCULATING IBD
plink –bfile example--genome --out example
```

- The expectation is that :
  IBD = 1 for duplicates or monozygotic twins
  IBD = 0.5 for first-degree relatives,
  IBD = 0.25 for second-degree relatives
  IBD = 0.125 for third-degree relatives

- The IBS method works best when only **independent SNPs** are included in the analysis. Independent SNP set for IBS calculation is generally prepared by removing regions of extended LD and pruning the remaining regions so that no pair of SNPs within a given window (say, 50kb) is correlated .

Genomic control

- In the presence of population structure, the chi-squared statistic X^2 is inflated by a constant inflation factor lambda ,

- Lambda is defined as the empirical median of L unrelated statistics divided by the expected median under the null distribution.

- In theory lambda should be equal to 1 in a homogeneous population. So a **value greater than one** implies **population structure**.

- PLINK estimates this value as GIF, while running association test. GIF greater than **1.05** often indicates the presence of a structure.

$$\lambda = \frac{\text{median}(X_1^2, X_2^2, \ldots, X_L^2,)}{0.456},$$

```
Reading map (extended format) from [ xxx.bim ]
.......

....
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is
1.15865
Mean chi-squared statistic is 1.2009
Correcting for 61997 tests
Writing multiple-test corrected significance values to
[ xxx_assoc.assoc.adjusted ]
```

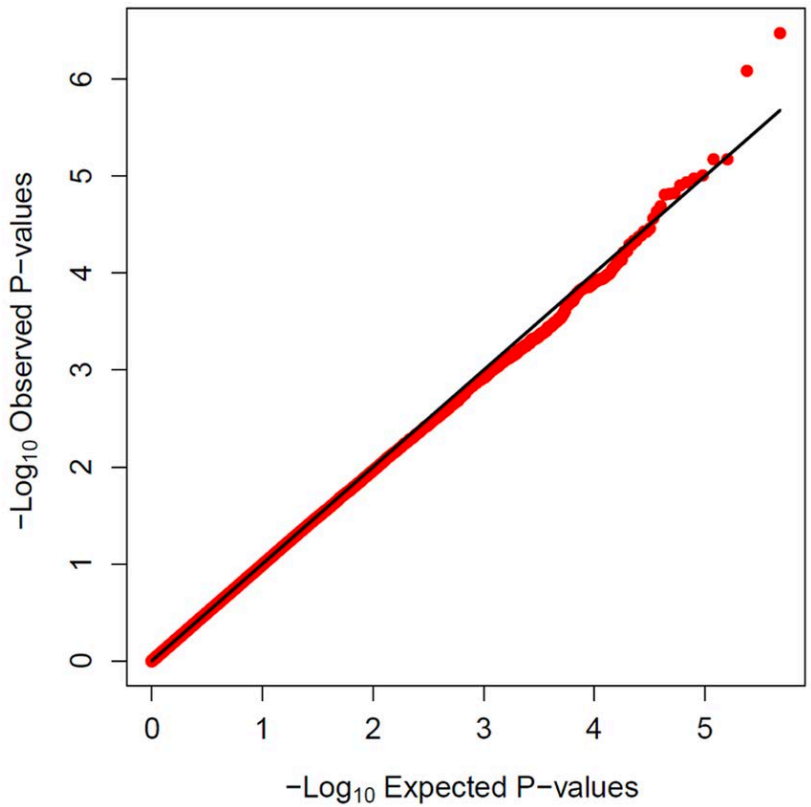
- Other factors could also cause GIF to have higher values (Clayton et al. 2005)

Q-Q plots

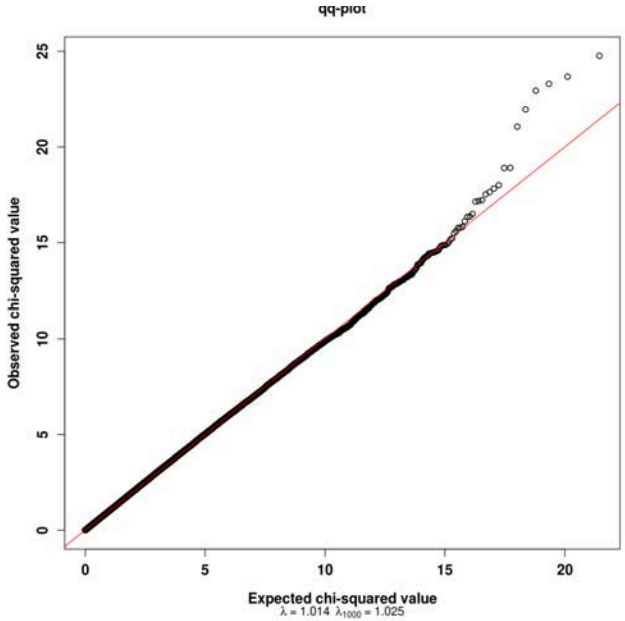
- Quantile-quantile (Q-Q) plots could detect the existence of population structure.

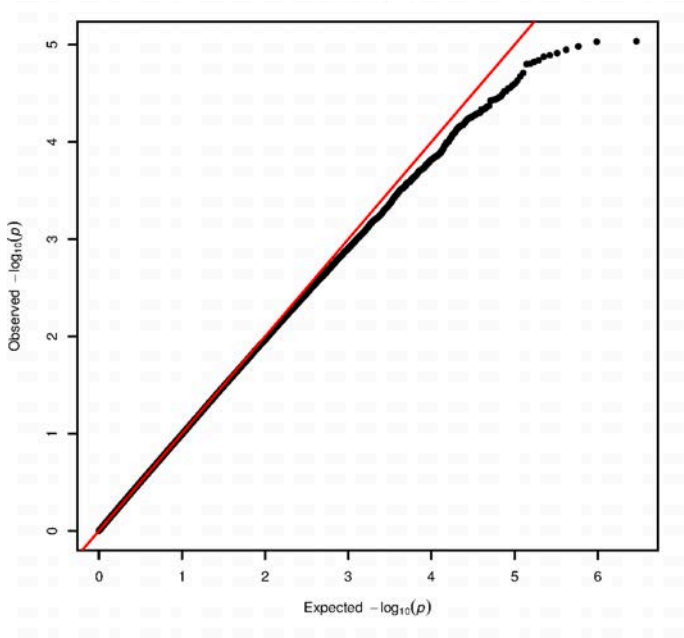
- The Q-Q plot is constructed as a scatter plot of the observed ranked P-values from the largest to smallest against the theoretical values under the null hypothesis of no association.

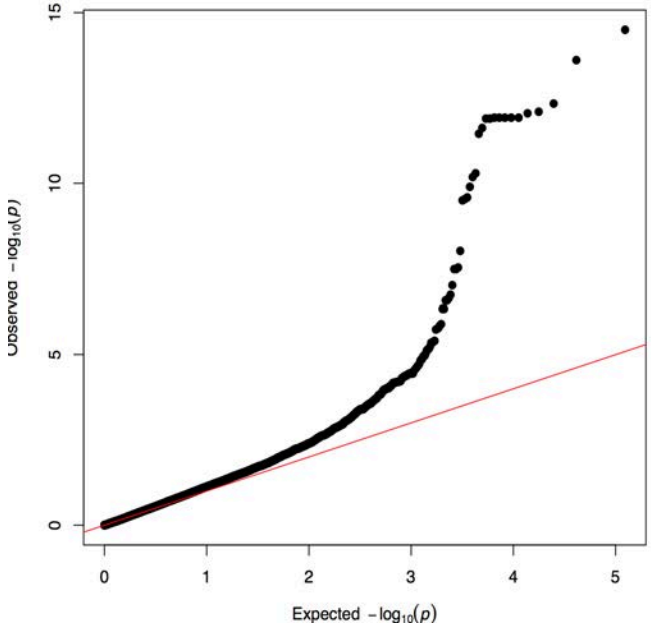
- If the statistics come from null distribution, the plot should go along the diagonal linearly. Large/Early deviation from the diagonal indicates **population structure**.




THE GOOD
THE BAD
and THE UGLY


qq-plot
Observed chi-squared value
Expected chi-squared value
λ = 1.014 λ1000 = 1.025


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

Other methods

Liu *et al. BMC Bioinformatics* 2013, **14**:132
http://www.biomedcentral.com/1471-2105/14/132

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Robust methods for population stratification in genome wide association studies

2015

RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness

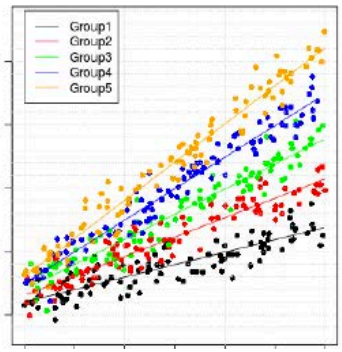
Matthew P. Conomos,[1] Michael B. Miller,[2] and Timothy A. Thornton[1]*

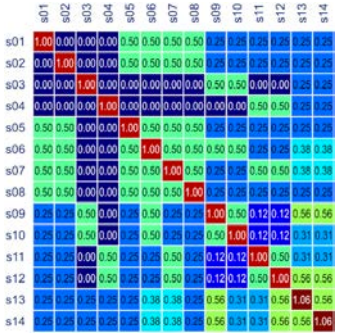
2015

PLOS | ONE

RESEARCH ARTICLE

Using Network Methodology to Infer Population Substructure

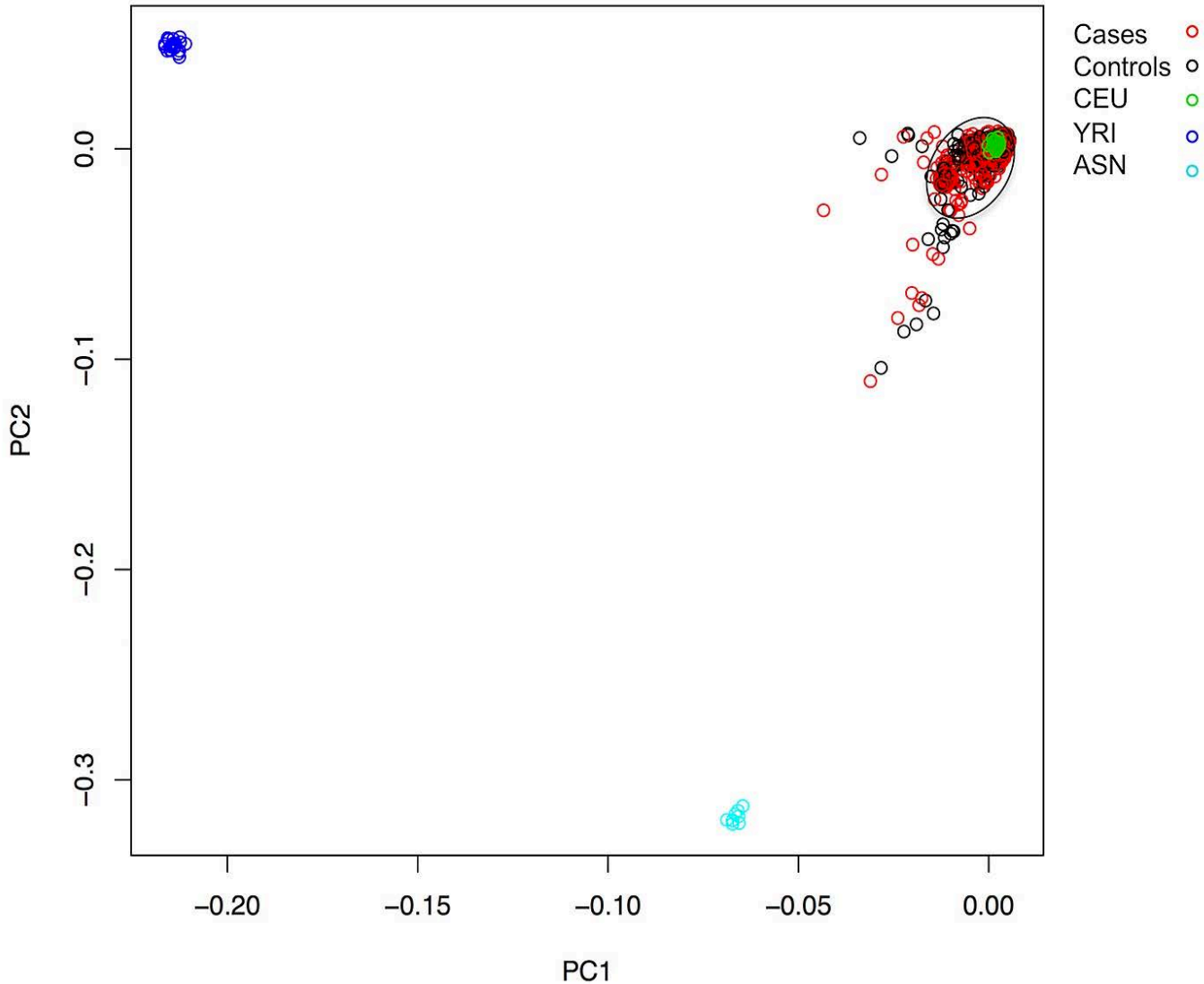




Section IV.
Correcting for Population Structure and Relatedness

- PCA based
- GIF based correction
- Removing related individuals
- LMMs
- PCA and Kinship matrix as covariates
- Pros and cons of various approaches

Using PCA to remove outliers

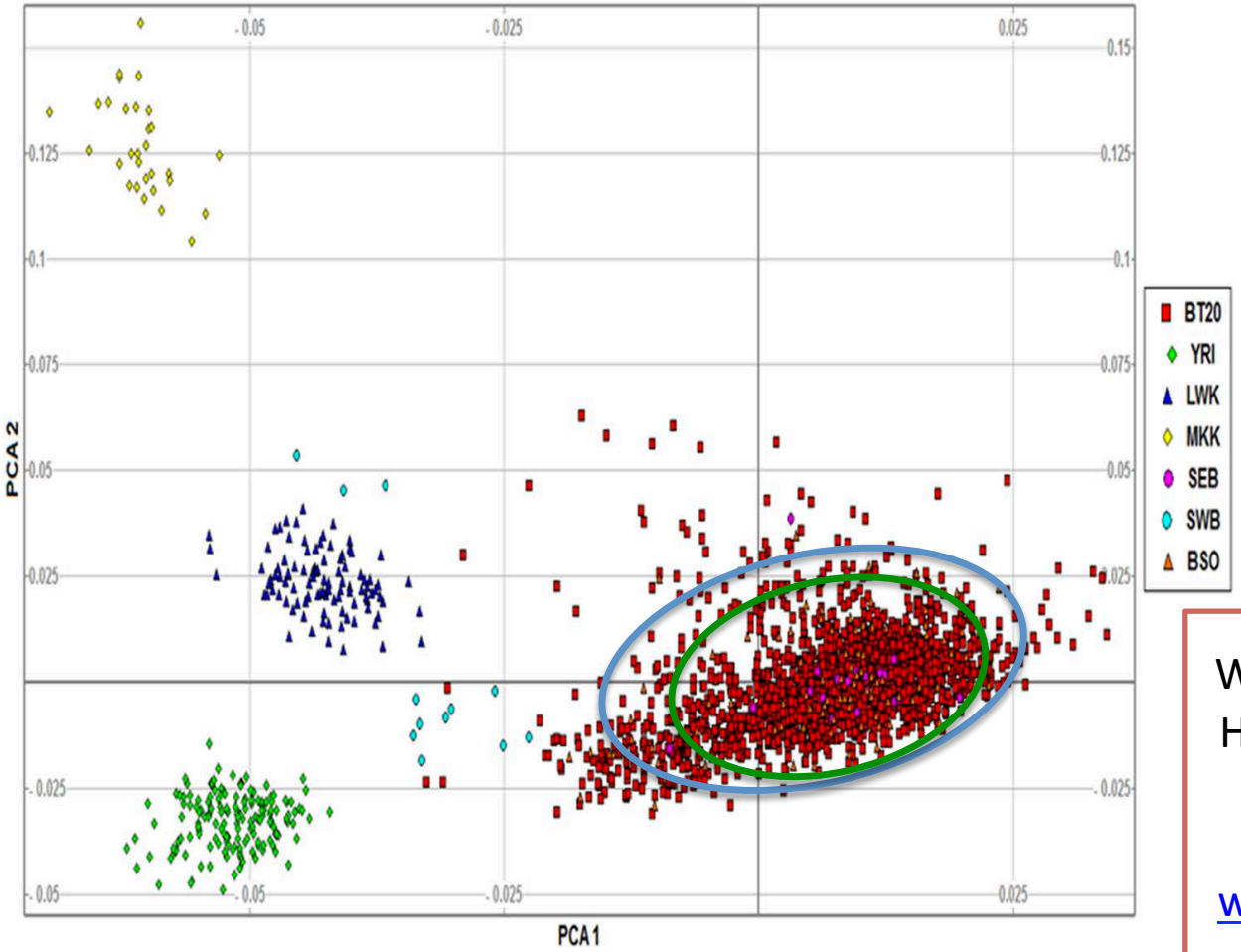


Draw a boundary containing the core of cases and control.

Exclude individuals outside the boundary

Hulur et al. 2017

Using PCA to remove outliers



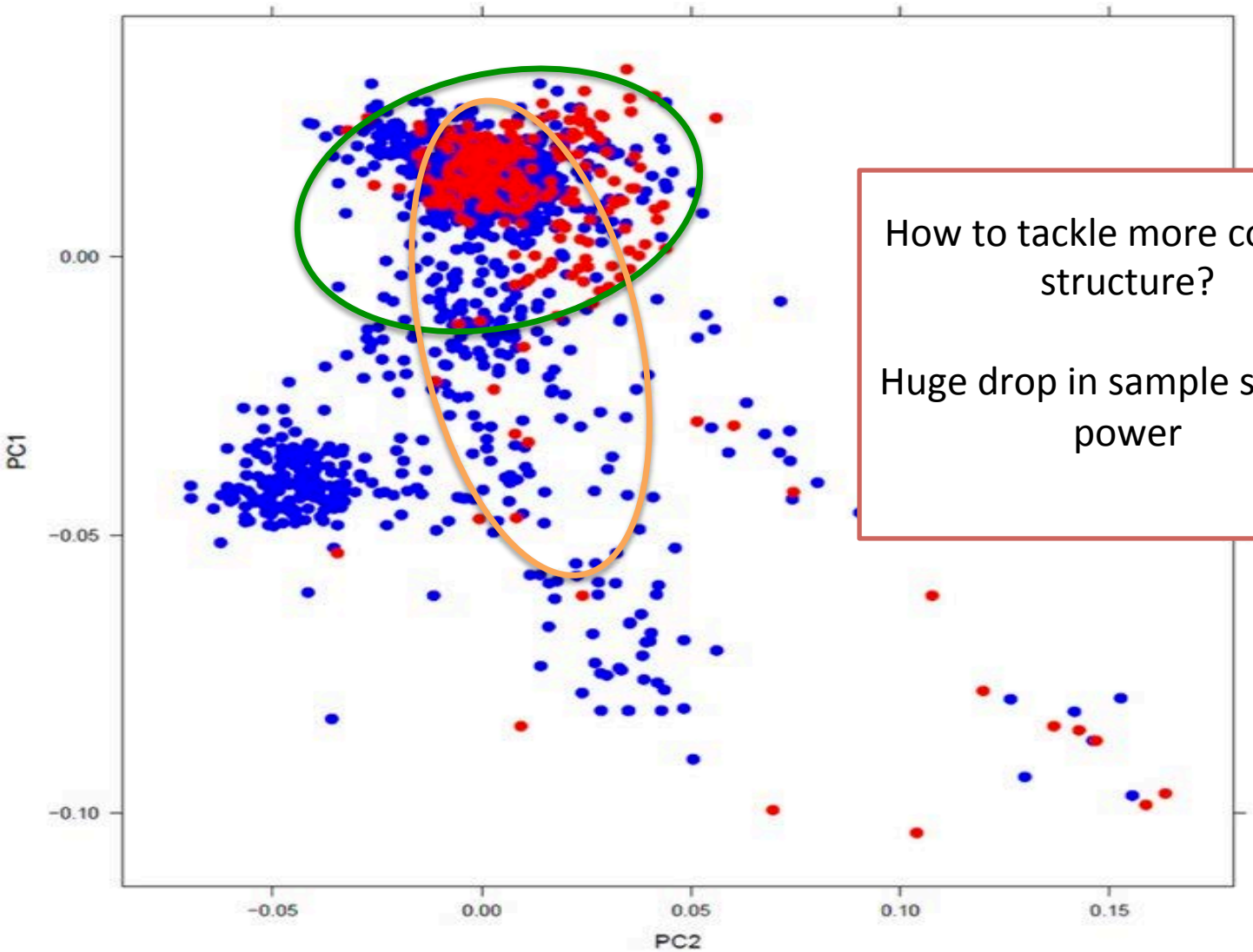
Sahibdeen et al. 2018

Which ellipse is better?
How many to exclude?

http://www.bioinf.wits.ac.za/software/poputils/

Is the correction enough?

Using PCA to remove outliers



How to tackle more complex structure?

Huge drop in sample size and power

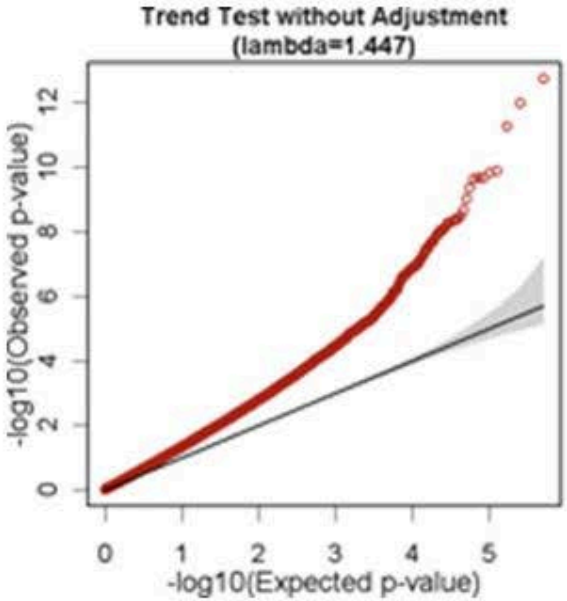
Behr et al. 2017

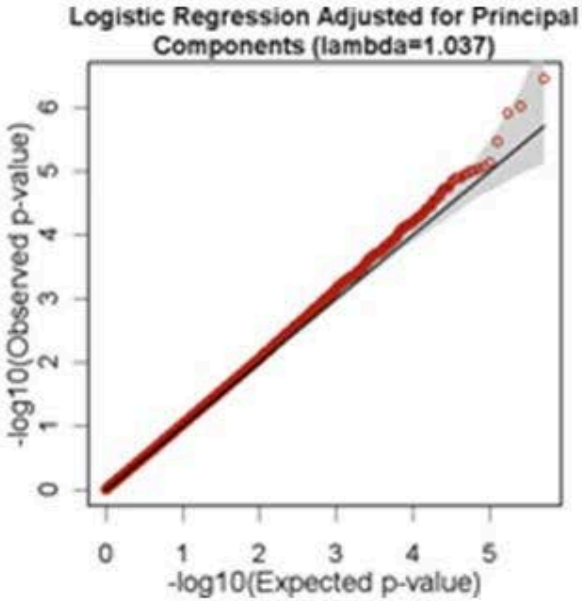
PCA as covariates

- The top *n* principal components, selected on the basis of distribution of the values are incorporated into the regression testing model as covariates. Depending on the dataset - *n* usually ranges from 3 to 10.

- Let $Z1_j$, $Z2_j$, … denote the eigenvectors for *j*th person

$$g(E(Y|X)) = a + bX + c1Z1 + c2Z2 + \ldots$$





Removing related individuals

```
CALCULATING IBD
plink –bfile example--genome --out example
```

- The expectation is that :
  IBD = 1 for duplicates or monozygotic twins
  IBD = 0.5 for first-degree relatives,
  IBD = 0.25 for second-degree relatives
  IBD = 0.125 for third-degree relatives

- Genotyping error, LD and population structure cause variation around these theoretical values and it is typical to remove one individual from each pair with an IBD > 0.1875 (halfway between the expected IBD for third- and second-degree relatives).

- In some cases this may lead to loss of lot many individuals.
- There might still be more distant relatedness which is not addressed at higher cutoffs but could lead to spurious associations

Genomic control based correction

The GC based correction implemented in PLINK can be accessed using --**g**c flag in addition to the --adjust flag, while running association.

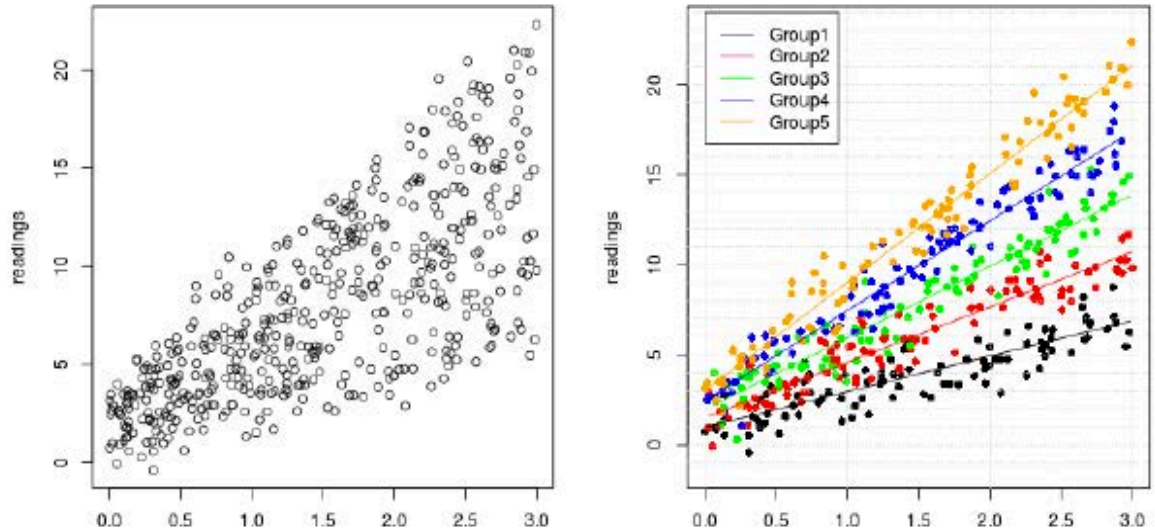
```
plink --bfile --assoc --gc --adjust --out results_with_gc
```

$$\chi^2_{fair} = \chi^2_{biased}/\lambda$$

Plink provides GC adjusted P-Values

- Implementation is difficult for non-additive model
- The correction applies to all the variants, however, some SNPs exhibit more differences in their allele frequencies than others; thus, the uniform adjustment is inappropriate and leads to a loss of power.
- Threshold for applying GC is empirical.

$$phenotype \sim \beta \times genotype + \beta_1 \times covariates + \beta_2 \times structure + \epsilon$$



- Covariates: sex, batch effects, chip effects
- Structure: villages or subpopulations: linear mixed models can model the intra-group effect

Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly used when there is non independence in the data, such as arises from a hierarchical structure.

LMM can address both PS and Relatedness

Random effects: account for familial relatedness

Grand Mean

Marker effect

$$Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + Line_i + \varepsilon_i$$

Phenotype of ith individual

Fixed effects: account for population structure

Observed SNP alleles of ith individual

Random error term

- $(Line_1, \ldots, Line_n) \sim \mathrm{MVN}(\mathbf{0}, 2K\sigma_G^2)$
- K = kinship matrix
- $\varepsilon_i \sim \text{i.i.d. } \mathrm{N}(0, \sigma_E^2)$

Measures relatedness between individuals

Yu et al. (2006)

Software

- **GEMMA**
- **EMMAX**
- **BOLT-LMM**
- **FaST-LMM**
- **GRAMMAR**

While these tools are primarily targeted for quantitative traits, These can be also applied to analyze binary traits, by **treating them as quantitative traits**. However, using these for **unbalanced case-control** might provide false associations.

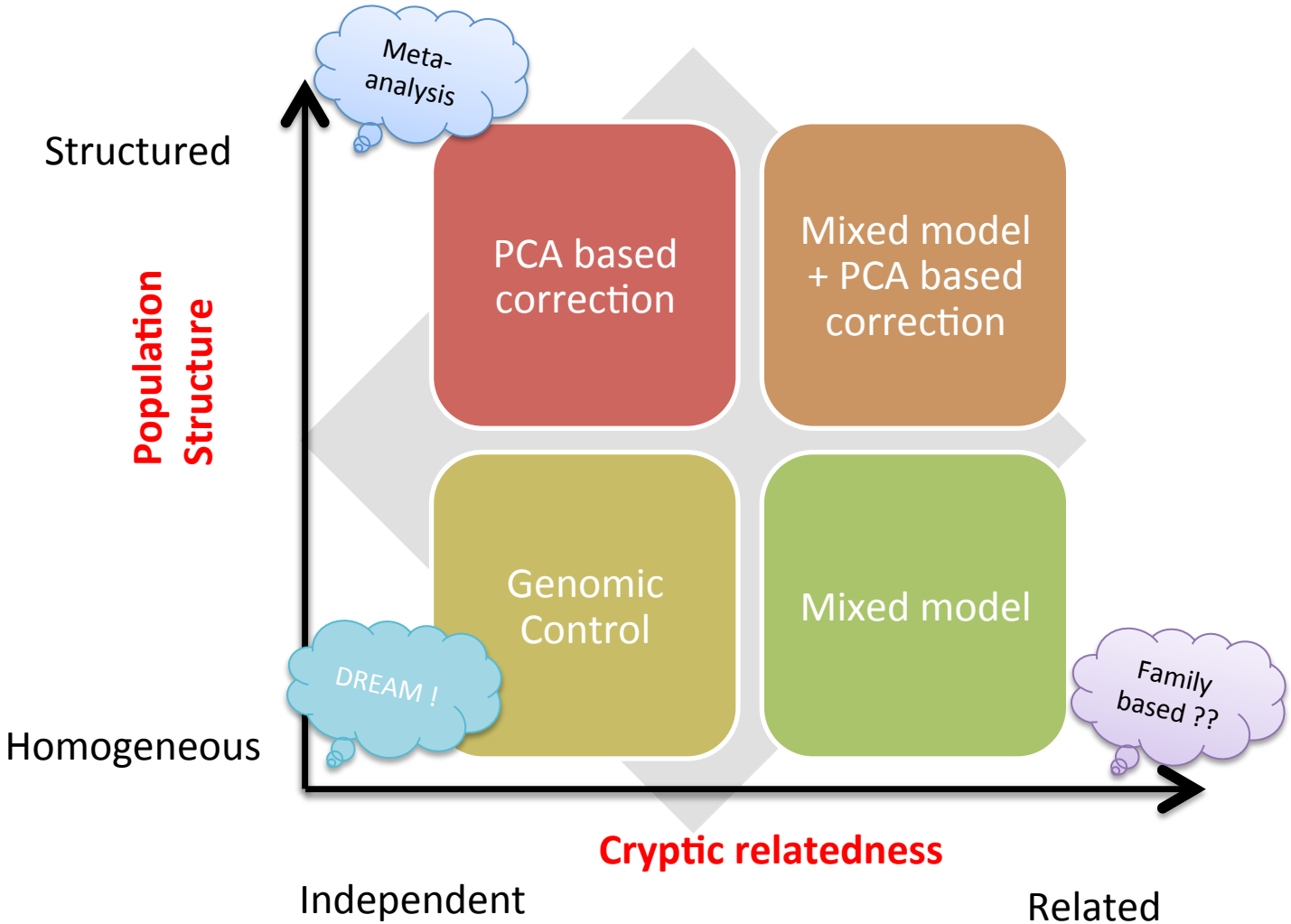
Special modification for case-control using LMM such as **liability-threshold mixed linear model** (LTMLM) are also available.

Table 1 Comparison of fast mixed-model association methods that model all SNPs

Method[a]	Requires $O(MN^2)$ time	Avoids proximal contamination	Models non-infinitesimal genetic architecture
EMMAX (ref. 3)	X		
FaST-LMM (ref. 5)	X[b]	X	
FaST-LMM-Select (refs. 9,11,15)	X[b]	X	X[c]
GEMMA (ref. 6)	X		
GRAMMAR-Gamma (ref. 10)	X[d]		
GCTA-LOCO (ref. 12)	X	X	
BOLT-LMM		X	X

BOLT-LMM only recommended for sample sizes **>5000.**

Which correction to use?



- Population structure often confounds inference and needs to be addressed for a successful GWAS.

- There are a variety of computational approaches to enable meaningful association studies to be conducted in a dataset with considerable structure and/or relatedness.

- Given the diversity of African populations and the history of migration and admixture, observing population structure in samples from most geographies is unsurprising. Do not be happy with P-values unless you have seen the Q-Q plot!

- Different approaches for correction are better suited for different scenarios. Choose your approach judiciously.

- Most of the real life analysis (especially those based on Imputed data) are computationally intensive and require nuanced interpretation.

- Population structure often confounds inference and needs to be addressed for a successful GWAS.

- There are a variety of computational approaches to enable meaningful association studies to be conducted in a dataset with considerable structure and/or relatedness.

- Given the diversity of African populations and the history of migration and admixture, observing population structure in samples from most geographies is unsurprising. Do not be happy with P-values unless you have seen the Q-Q plot!

- Different approaches for correction are better suited for different scenarios. Choose your approach judiciously.

- Most of the real life analysis (especially those based on Imputed data) are computationally intensive and require nuanced interpretation.