Population structure and stratification

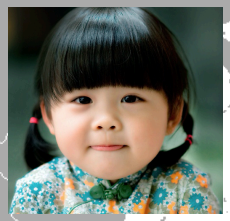
Kridsadakorn Chaichoompu
GIGA-Medical Genomics (BIO3)
University of Liege

Population Structure

- Population genetics is a subfield of genetics that deals with genetic differences within and between populations, and is a part of evolutionary biology. Studies in this branch of biology examine such phenomena as adaptation, speciation, and population structure.
- Population stratification is the presence of a systematic difference in allele frequencies between subpopulations in a population possibly due to different ancestry, especially in the context of association studies. Population stratification is also referred to as population structure, in this context.


Diversity
• Human
• Plants
• Animals
• Bacteria
• etc

Human Diversity



How to group people?

Countries

Languages

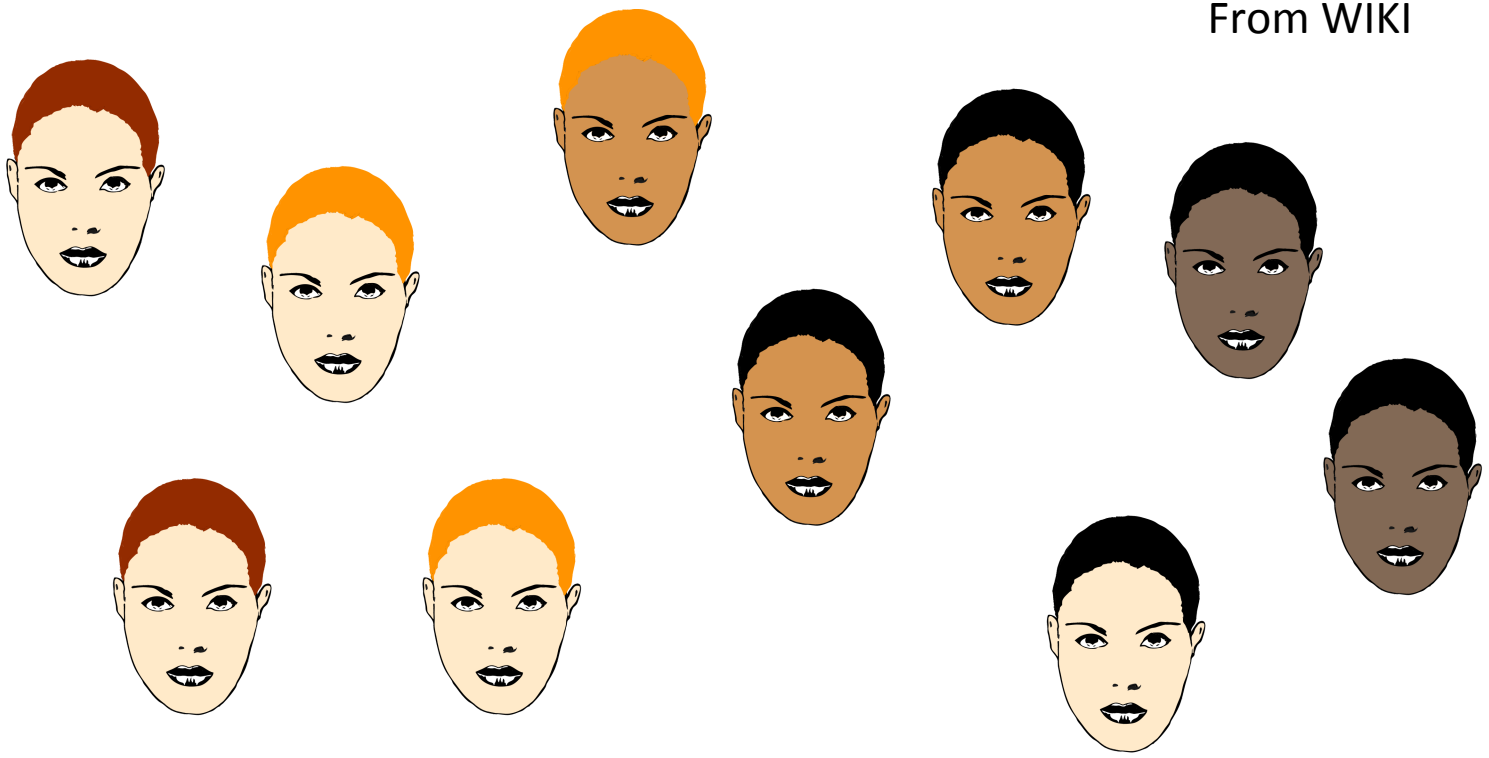
Physical appearances: Hair colors, Eye colors, Skin colors

Diversity in Population

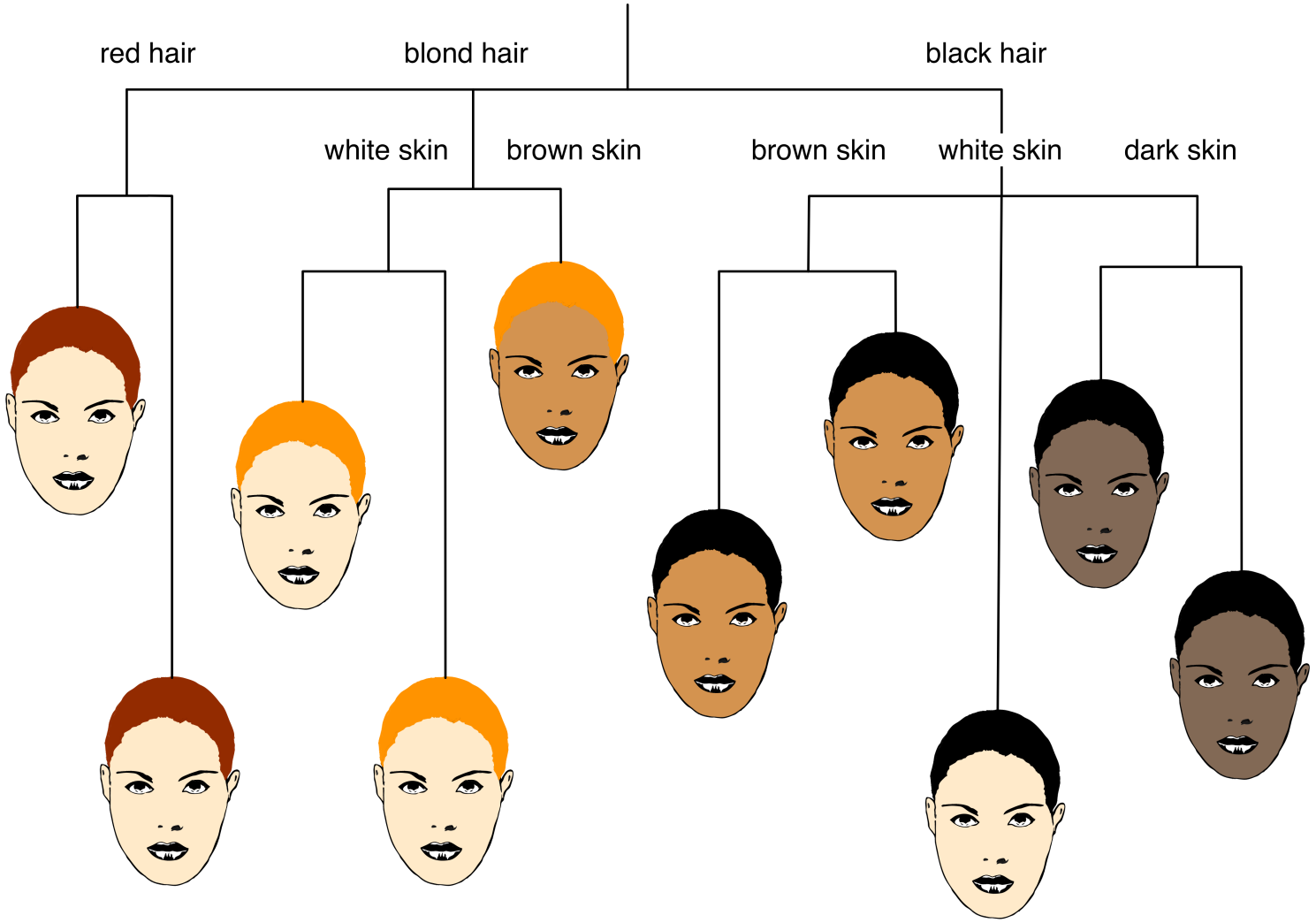
Languages?
Example: Belgium

- Dutch 59%
- French 41%
- German ?%

From WIKI

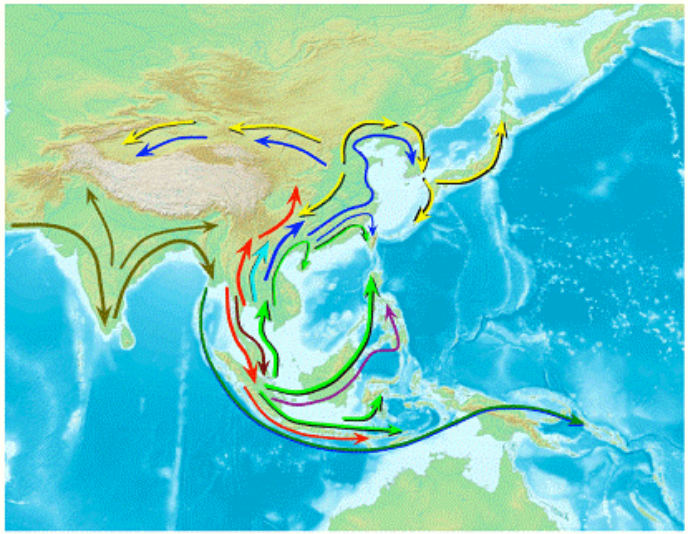


Population clustering



What is the benefit of knowing Population Substructures?

- Population evolution
- Population ancestry
- Population migration
- Population based analysis
- Subgroup of patients

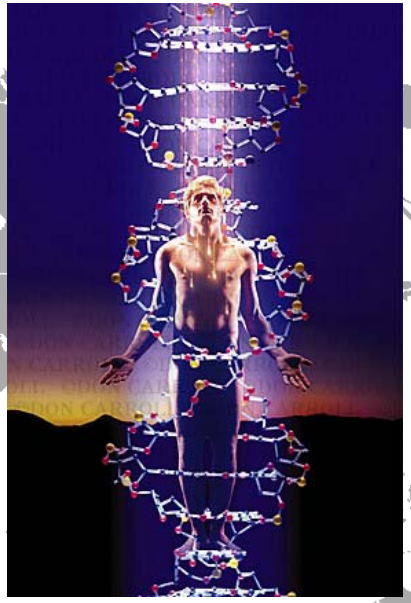


Indigenous populations

- 54,794 SNPs
- 1,928 individuals
- 73 Asian and 2 non-Asian populations

Mapping Human Genetic Diversity and tracing the genetic origins of Asian populations
The HUGO Pan-Asian SNP Consortium
Science, October 2009

DNA: the blueprint of our lives



PROPER DRUGS AND TREATMENT

HAPMAP Project

feature

The International HapMap Project

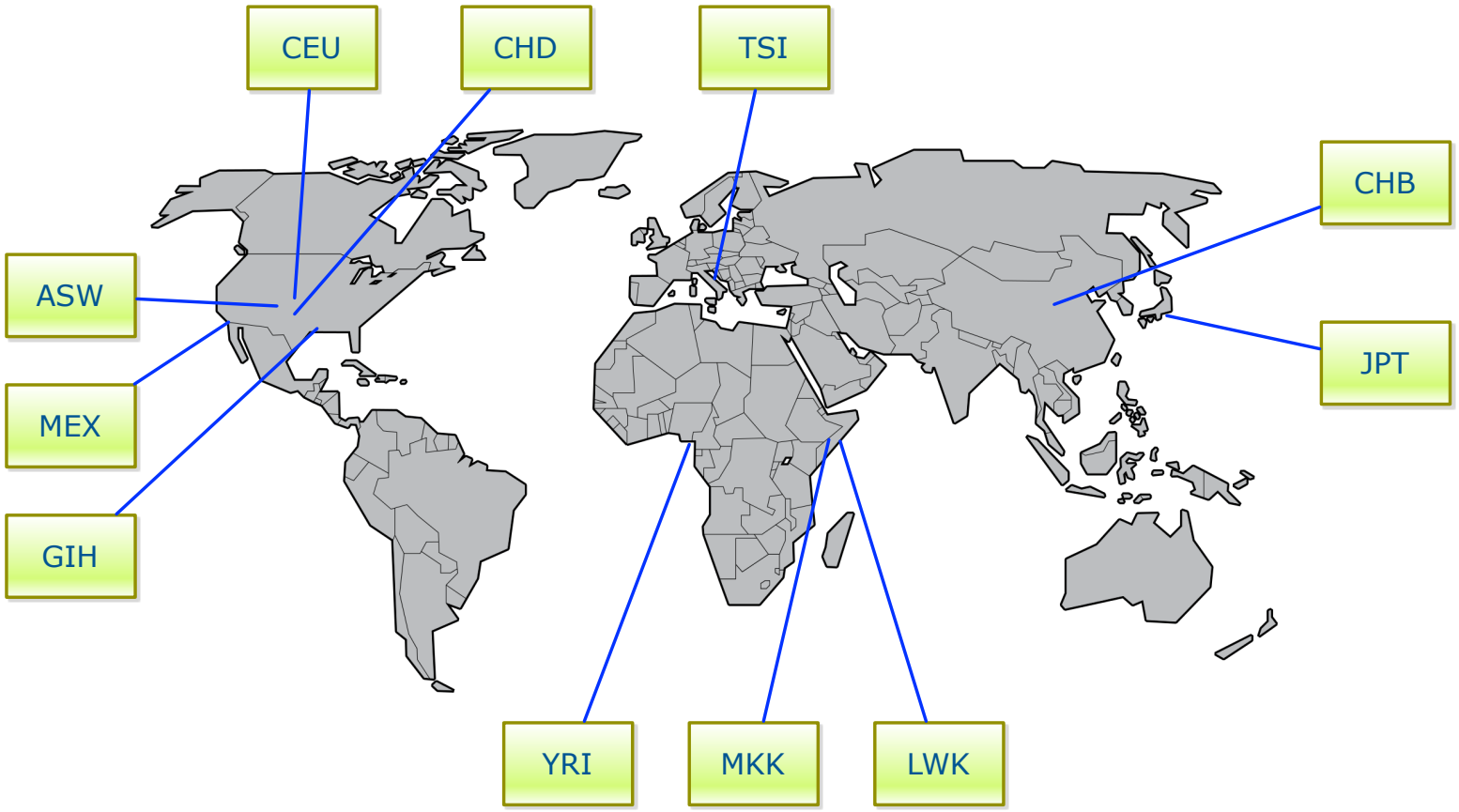
The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.

https://www.genome.gov/10001688/international-hapmap-project/

HAPMAP samples



ftp://ftp.ncbi.nlm.nih.gov/hapmap

1000 Genomes project

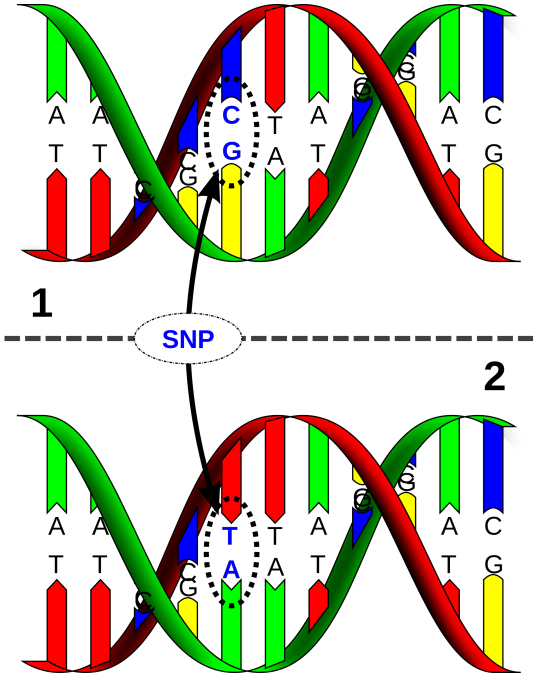
IGSR and the 1000 Genomes Project

Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

Clustering using genetic profile

- Single Nucleotide Polymorphisms (SNPs) are commonly used to capture variations between populations.
- Small scale: small subsets of SNPs or ancestry-informative markers (AIM)
- Genome-wide scale: 600K – 4M SNPs



Quality Control

- Missing data
- Linkage Disequilibrium (LD) pruning
- Hardy-Weinberg Equilibrium (HWE)
- Minor allele frequency (MAF) filtering

Suggestion: use PLINK

http://pngu.mgh.harvard.edu/~purcell/plink/

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs).

PCA in R

- prcomp(x, retx = TRUE, center = TRUE, scale. = FALSE, tol = NULL, ...)
- princomp(formula, data = NULL, subset, na.action, ...)
- eigen(x, symmetric, only.values = FALSE, EISPACK = FALSE)
- svd(x, nu = min(n, p), nv = min(n, p), LINPACK = FALSE)

library(rARPACK)

- svds(A, k, nu = k, nv = k, opts = list(), ...)
- eigs(A, k, which = "LM", sigma = NULL, opts = list(), ...)

snpStats – Bioconductor Package

- http://www.bioconductor.org/packages/release/bioc/html/snpStats.html

Usually, principal components analysis is carried out by calculating the eigenvalues and eigenvectors of the correlation matrix. With N cases and P variables, if we write X for the $N \times P$ matrix which has been standardised so that columns have zero mean and unit standard deviation, we find the eigenvalues and eigenvectors of the $P \times P$ matrix $X^{\mathrm{T}}.X$ (which is N or $(N-1)$ times the correlation matrix depending on which denominator was used when calculating standard deviations). The first eigenvector gives the loadings of each variable in the first principal component, the second eigenvector gives the loadings in the second component, and so on. Writing the first C component loadings as columns of the $P \times C$ matrix B, the $N \times C$ matrix of subjects' principal component scores, S, is obtained by applying the factor loadings to the original data matrix, *i.e.* $S = X.B$. The sum of squares and products matrix, $S^{\mathrm{T}}.S = D$, is diagonal with elements equal to the first C eigenvalues of the $X^{\mathrm{T}}.X$ matrix, so that the variances of the principal components can obtained by dividing the eigenvalues by N or $(N-1)$.

PCA for GWAS

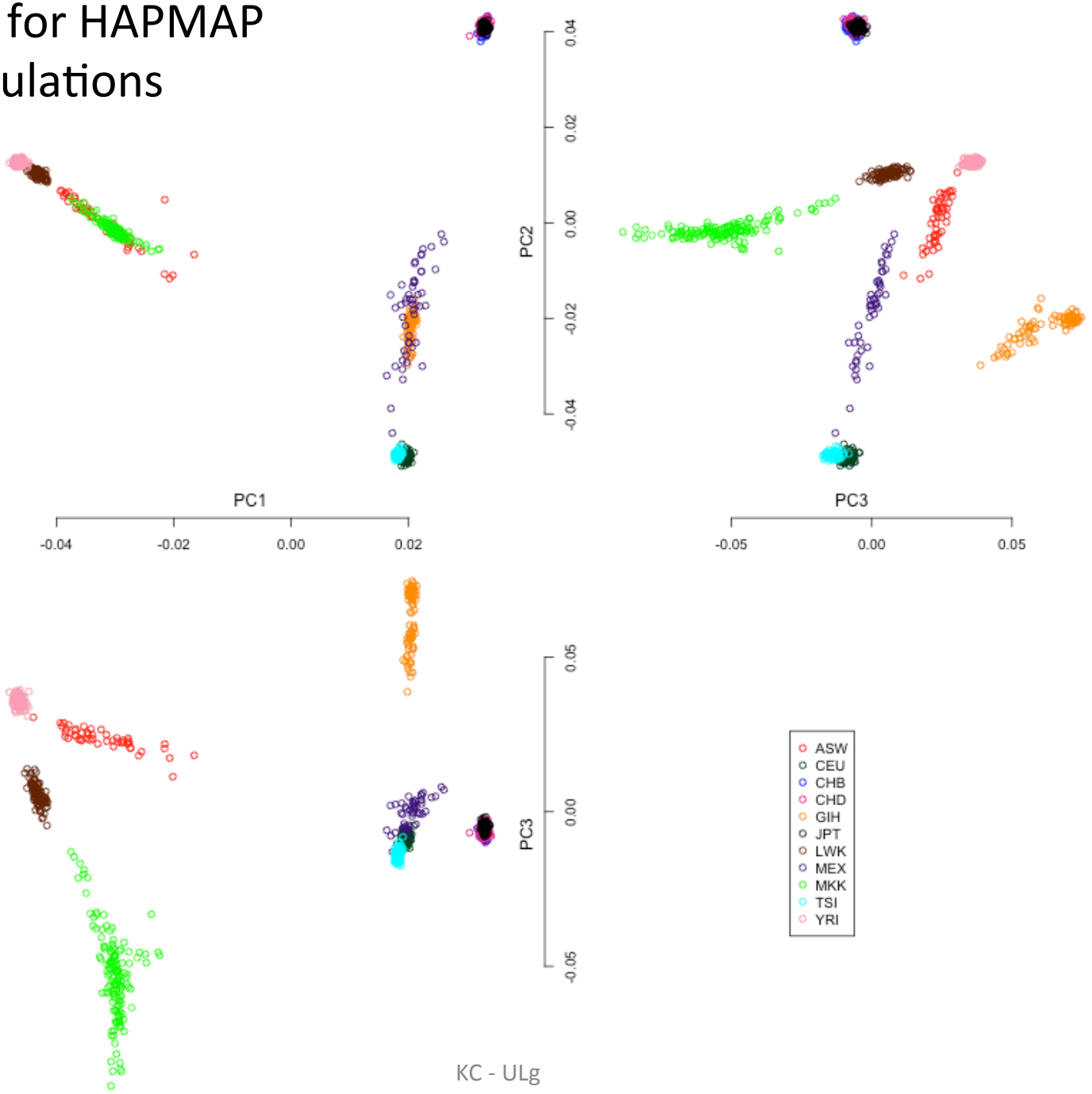
nature genetics

Principal components analysis corrects for stratification in genome-wide association studies

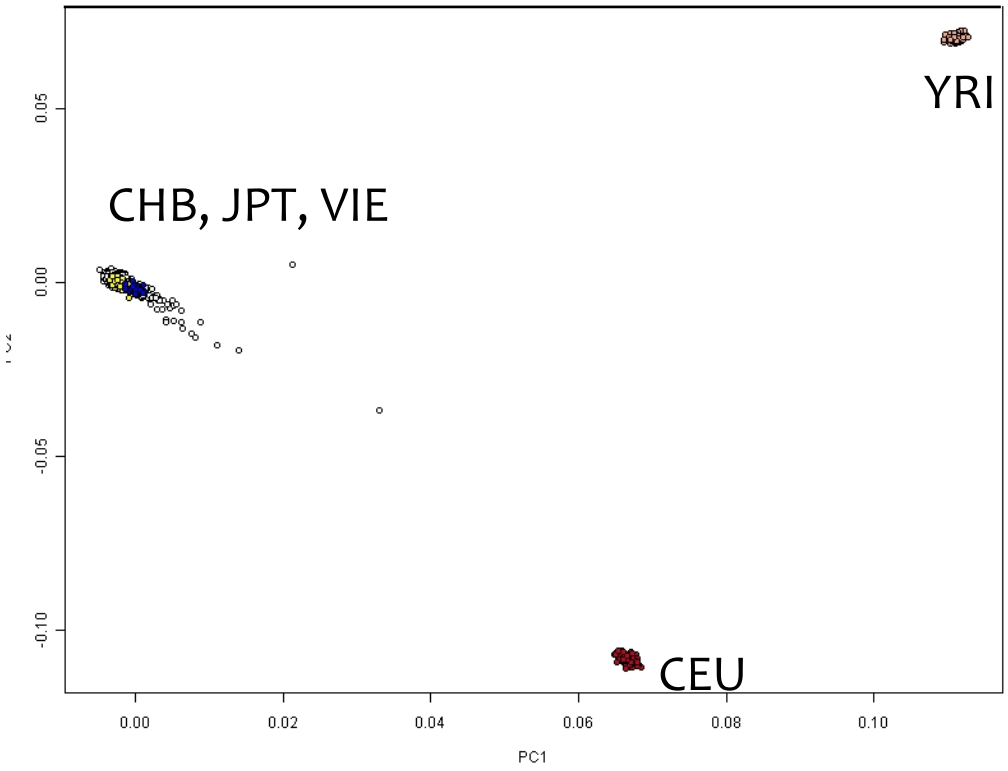
Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

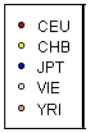
Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

PCA plot for HAPMAP Populations



The importance of substructures



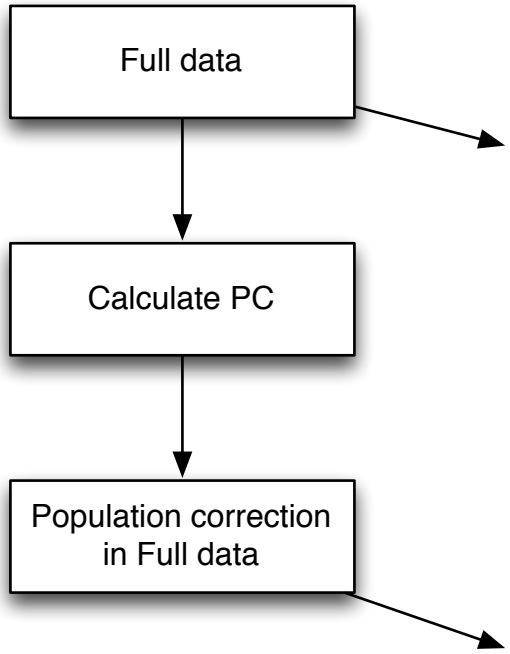


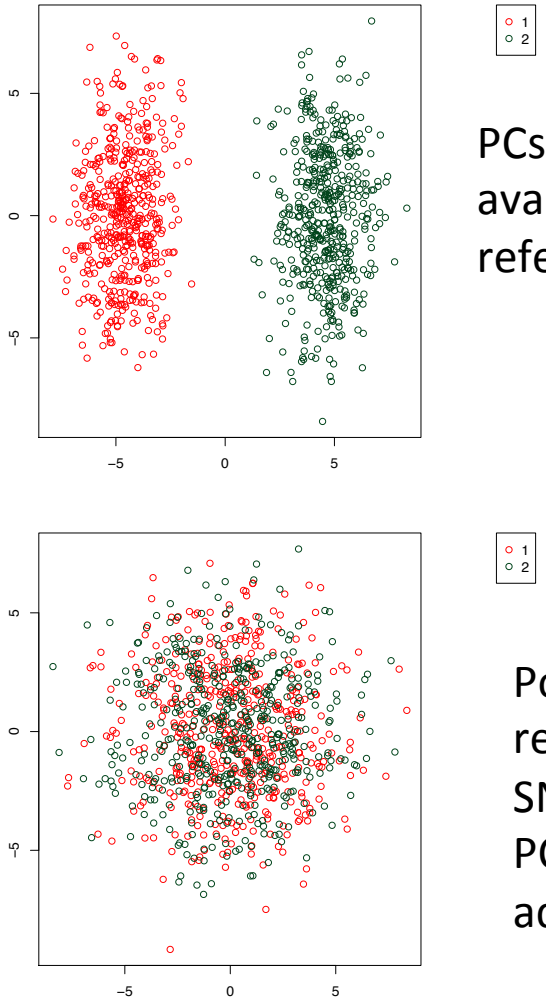
Genome-wide association study for Dengue shock syndrome

Chiea Chuen Khor et al.
Nature Genetics 2012

- 657,366 SNPs
- 4,028 individuals from Vietnam

Population correction using linear model





PCs were calculated from all available data (2 populations), referred to as “Pooled PCs”

Population Correction: PCs regressed out from original SNPs.
PCs were calculated from adjusted SNPs.

Linear Regression in R

Linear models

lm(formula, data, subset, ...)

Example in help page:

```
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2, 10, 20, labels = c("Ctl","Trt"))
weight <- c(ctl, trt)
lm.D9 <- lm(weight ~ group)
plot(lm.D9)
```

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

Generalized Linear Models - GLM

glm(formula, family = gaussian, data, weights, ...)

Example from help page:

```
counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
print(d.AD <- data.frame(treatment, outcome, counts))
glm.D93 <- glm(counts ~ outcome + treatment, family =
poisson())
```

http://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html

Models for GLM

glm(formula, family=familytype(link=linkfunction), data=)

Family	**Default Link Function**
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

http://www.statmethods.net/advstats/glm.html

GWAS with regression

- Linear model:

  plink --bfile mydata --linear

- Logistic model:

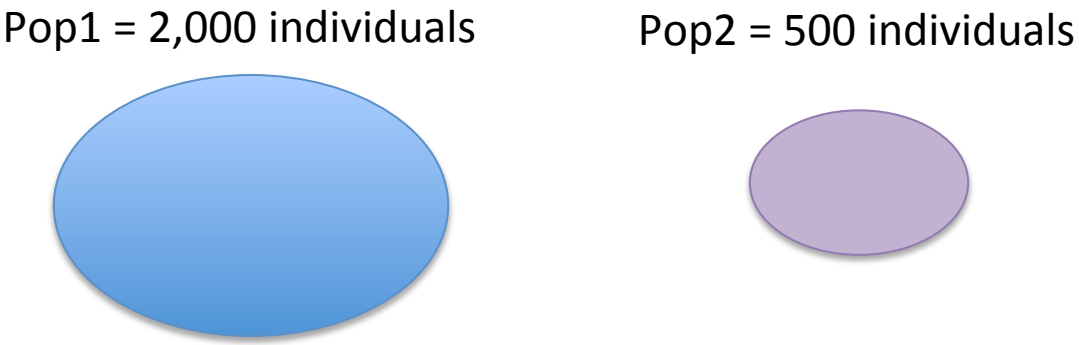
  plink --bfile mydaya –logistic

- Check:

http://zzz.bwh.harvard.edu/plink/anal.shtml#glm

Fixation index (F_{ST})

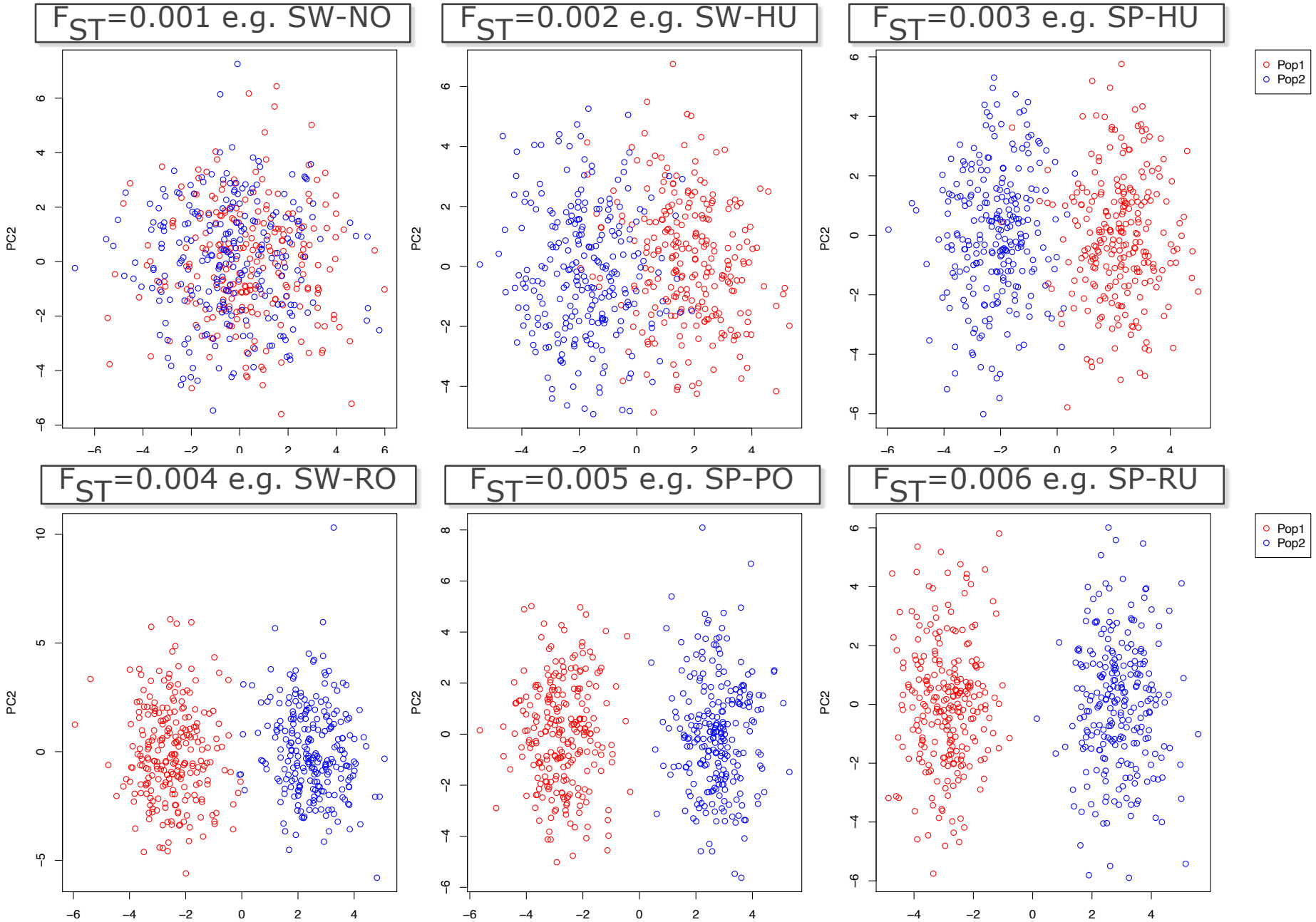
- F_{ST} can be used to describe a distance among population.
- F_{ST} can be biased due to the allele frequencies and the number of independent SNPs.



F_{ST} among European populations

	Sp	*Fr*	*Be*	*UK*	*Sw*	*No*	*Ge*	*Ro*	*Cz*	*Sl*	*Hu*	*Po*	*Ru*	*CEU*	*CHB*	*JPT*
Fr	0.0008															
Be	0.0015	0.0002														
UK	0.0024	0.0006	0.0005													
Sw	0.0047	0.0023	0.0018	0.0013												
No	0.0047	0.0024	0.0019	0.0014	0.0010											
Ge	0.0025	0.0008	0.0005	0.0006	0.0011	0.0016										
Ro	0.0023	0.0017	0.0018	0.0028	0.0041	0.0044	0.0016									
Cz	0.0033	0.0016	0.0013	0.0014	0.0016	0.0024	0.0003	0.0016								
Sl	0.0034	0.0017	0.0015	0.0017	0.0019	0.0026	0.0005	0.0014	0.0001							
Hu	0.0030	0.0015	0.0013	0.0016	0.0020	0.0026	0.0004	0.0011	0.0001	0.0001						
Po	0.0053	0.0032	0.0028	0.0027	0.0023	0.0034	0.0012	0.0028	0.0004	0.0004	0.0006					
Ru	0.0059	0.0037	0.0034	0.0032	0.0025	0.0036	0.0016	0.0030	0.0008	0.0007	0.0009	0.0003				
CEU	0.0026	0.0008	0.0005	0.0002	0.0011	0.0012	0.0006	0.0028	0.0014	0.0016	0.0016	0.0026	0.0031			
CHB	0.1096	0.1094	0.1093	0.1096	0.1073	0.1081	0.1085	0.1047	0.1080	0.1069	0.1058	0.1086	0.1036	0.1095		
JPT	0.1118	0.1116	0.1114	0.1117	0.1095	0.1103	0.1107	0.1068	0.1102	0.1091	0.1079	0.1108	0.1057	0.1117	0.0069	
YRI	0.1460	0.1493	0.1496	0.1513	0.1524	0.1531	0.1502	0.1463	0.1503	0.1498	0.1490	0.1520	0.1504	0.1510	0.1901	0.1918

Simon et al. 2008



To understand F_{ST}, here are simulated data using Balding method and the examples of EU populations as reported in (Simon et al. 2008)

F_{ST} – R Packages

Package 'PopGenome'

May 4, 2015

Type Package
Title An Efficient Swiss Army Knife for Population Genomic Analyses
Version 2.1.6
Date 2015-05-1

Package 'hierfstat'

December 4, 2015

Version 0.04-22
Date 2015-11-24
Title Estimation and Tests of Hierarchical F-Statistics

Package 'StAMPP'

July 6, 2015

Type Package
Title Statistical Analysis of Mixed Ploidy Populations
Depends R (>= 2.14.0), pegas
Imports parallel, doParallel, foreach, adegenet, methods, utils
Version 1.4
Date 2015-06-30

Estimating F_{ST}

Method

Estimating and interpreting F_{ST}: The impact of rare variants

Gaurav Bhatia,[1,2,6,7] Nick Patterson,[2,6,7] Sriram Sankararaman,[2,3] and Alkes L. Price[2,4,5,7]

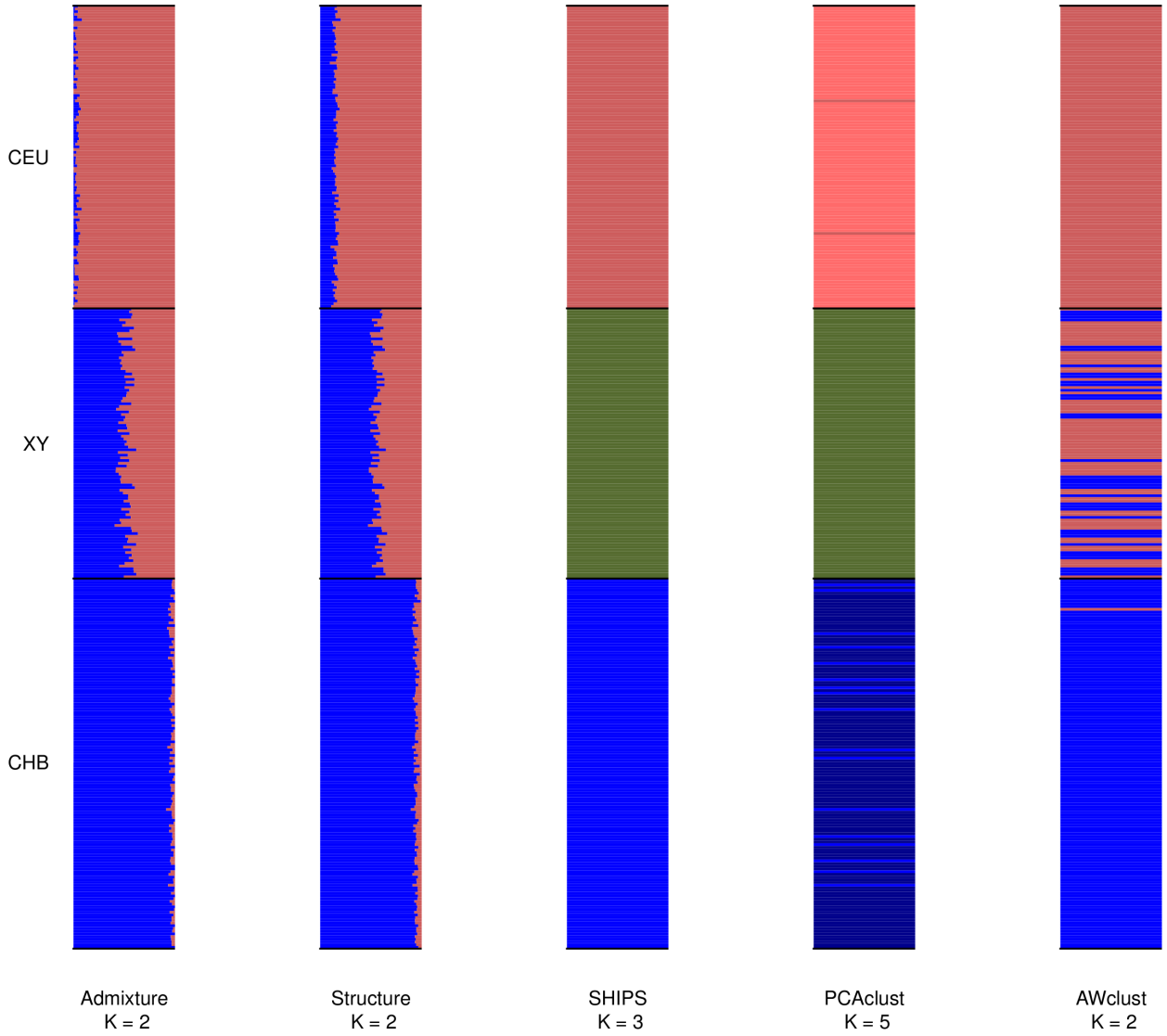
[1]Harvard–Massachusetts Institute of Technology (MIT), Division of Health, Science, and Technology, Cambridge, Massachusetts 02139, USA; [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [4]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA; [5]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA

In a pair of seminal papers, Sewall Wright and Gustave Malécot introduced F_{ST} as a measure of structure in natural populations. In the decades that followed, a number of papers provided differing definitions, estimation methods, and interpretations beyond Wright's. While this diversity in methods has enabled many studies in genetics, it has also introduced confusion regarding how to estimate F_{ST} from available data. Considering this confusion, wide variation in published estimates of F_{ST} for pairs of HapMap populations is a cause for concern. These estimates changed—in some cases more than twofold—when comparing estimates from genotyping arrays to those from sequence data. Indeed, changes in F_{ST} from sequencing data might be expected due to population genetic factors affecting rare variants. While rare variants do influence the result, we show that this is largely through differences in estimation methods. Correcting for this yields estimates of F_{ST} that are much more concordant between sequence and genotype data. These differences relate to three specific issues: (1) estimating F_{ST} for a single SNP, (2) combining estimates of F_{ST} across multiple SNPs, and (3) selecting the set of SNPs used in the computation. Changes in each of these aspects of estimation may result in F_{ST} estimates that are highly divergent from one another. Here, we clarify these issues and propose solutions.

Genetic Admixture

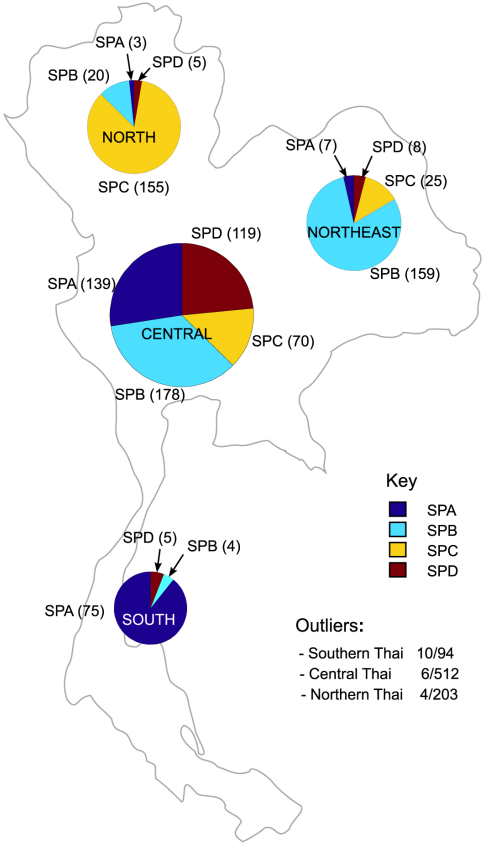
Genetic admixture occurs when two or more previously isolated populations begin interbreeding. Admixture results in the introduction of new genetic lineages into a population. It has been known to slow local adaptation by introducing foreign, unadapted genotypes (known as gene swamping). It also prevents speciation by homogenizing populations.

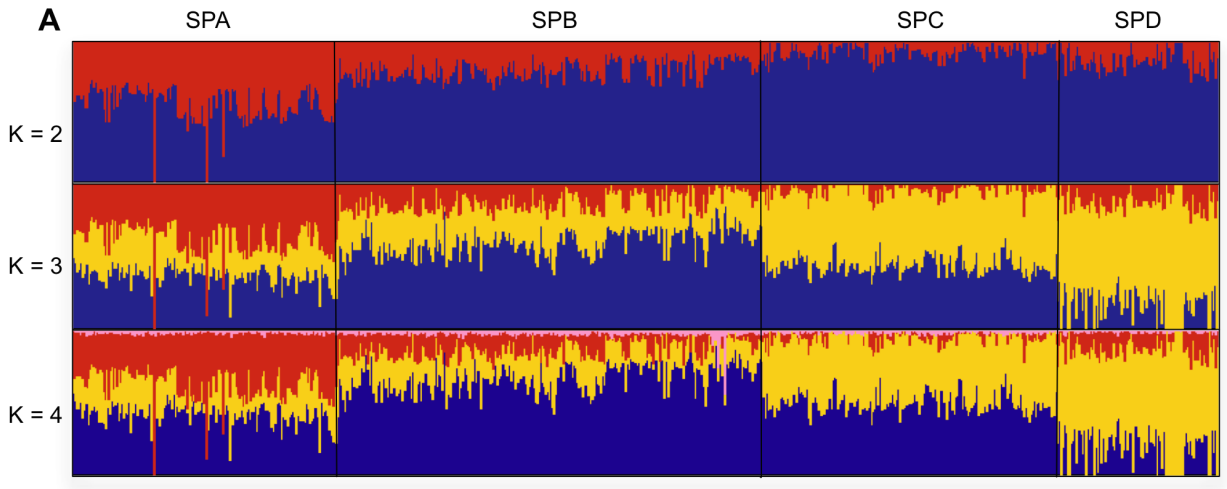
Tools for Admixture profiling



Bouaziz 2012

Thai population





Thai population genetic structure
Wangkumhang, P et al. PLoS One, 2013