Population stratification and Epistasis using PLINK

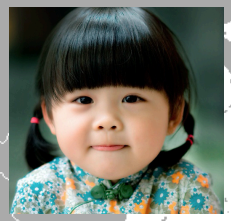
Kridsadakorn Chaichoompu
GIGA-Medical Genomics (BIO3)
University of Liege

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

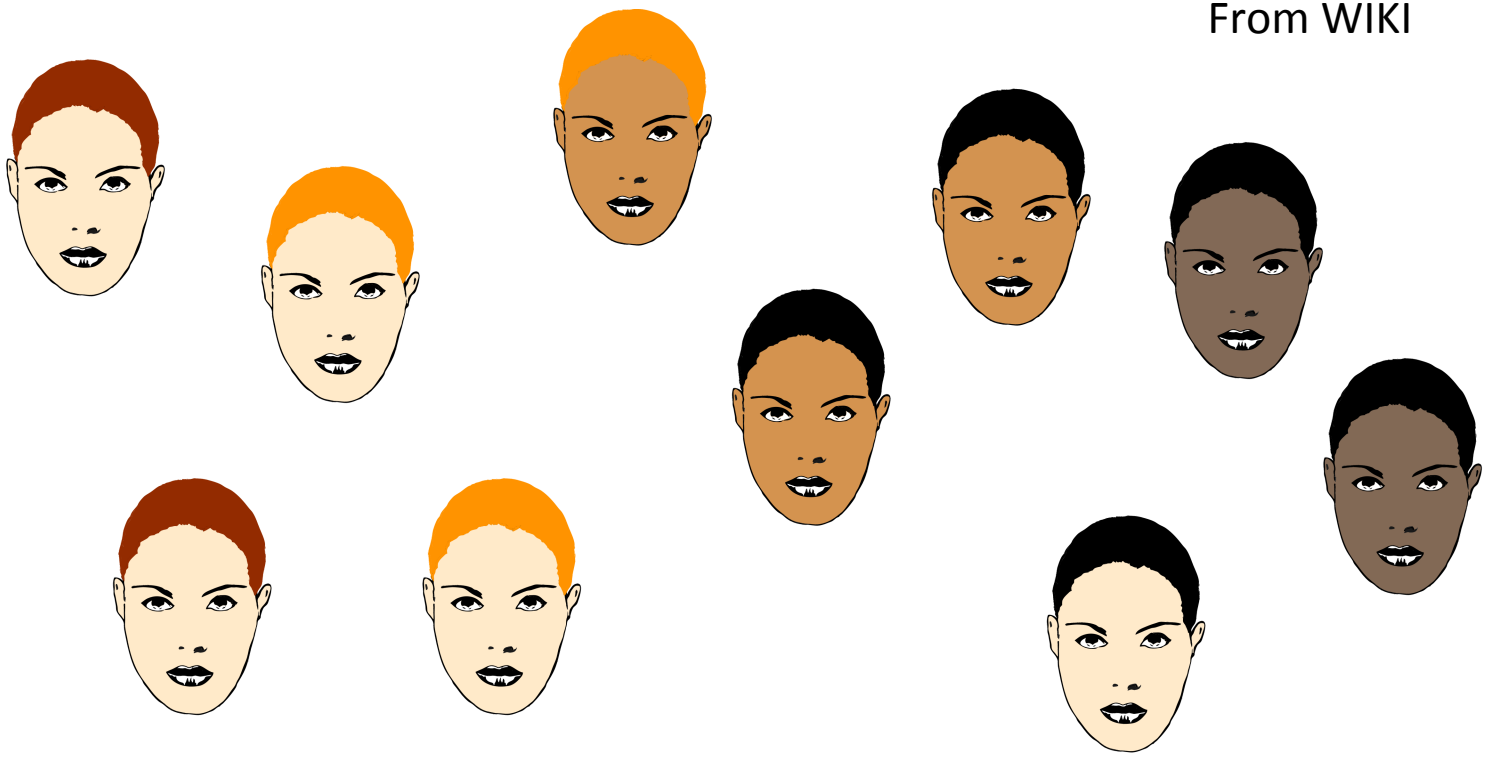
Physical appearances: Hair colors, Eye colors, Skin colors

Diversity in Population

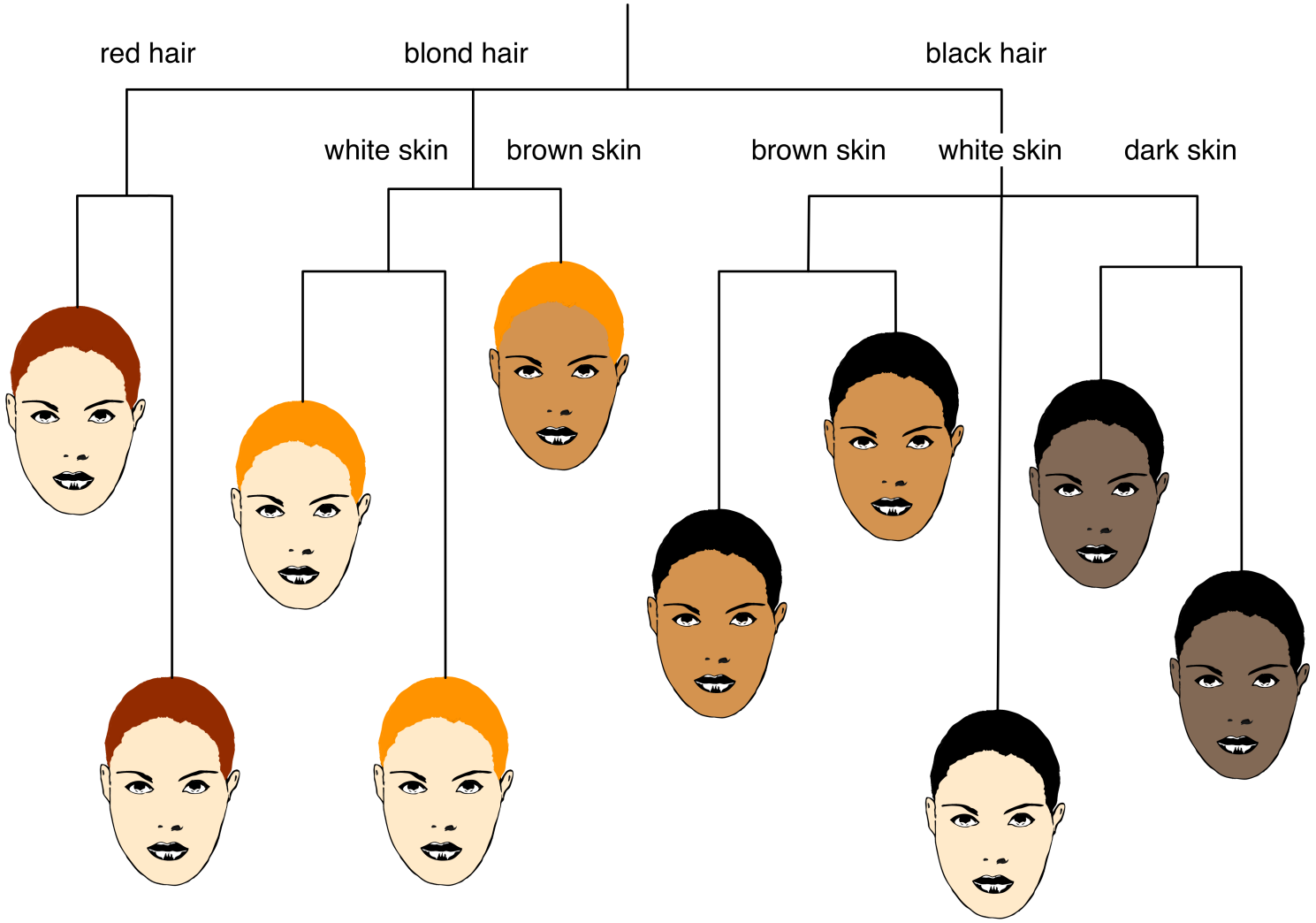
Languages?
Example: Belgium

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

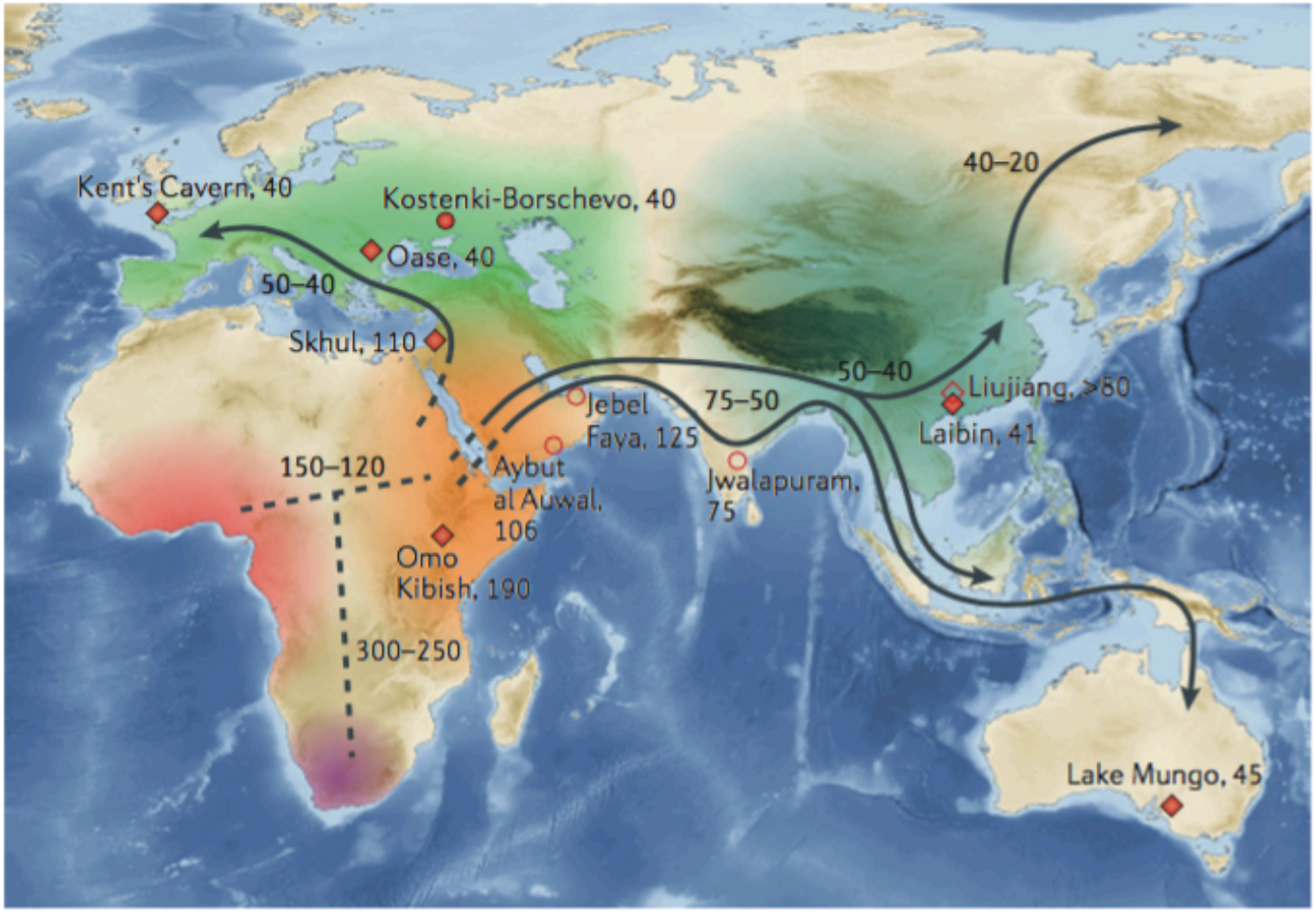
From WIKI



Population clustering

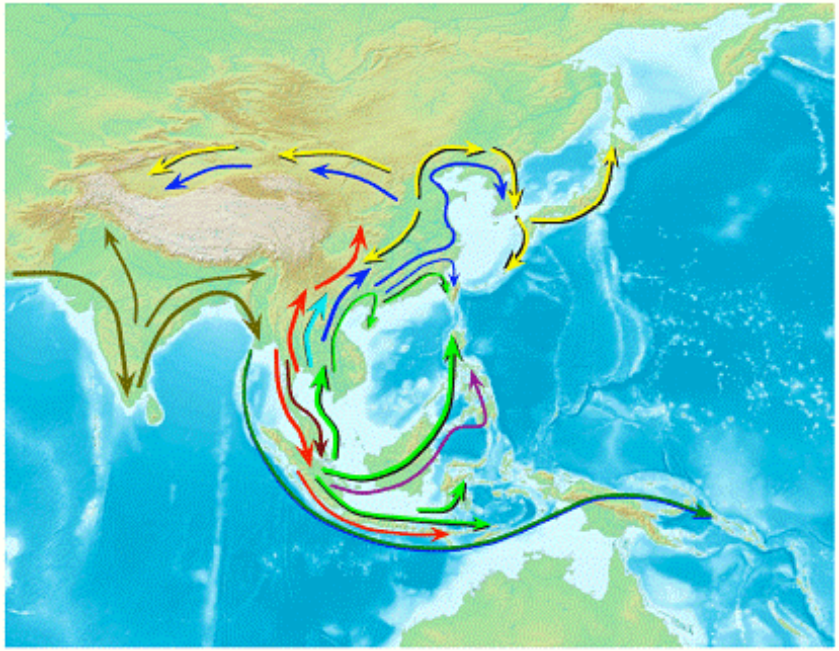


What causes population structure? World-wide migration



(Scally, 2012)

Migration within region: East Asia

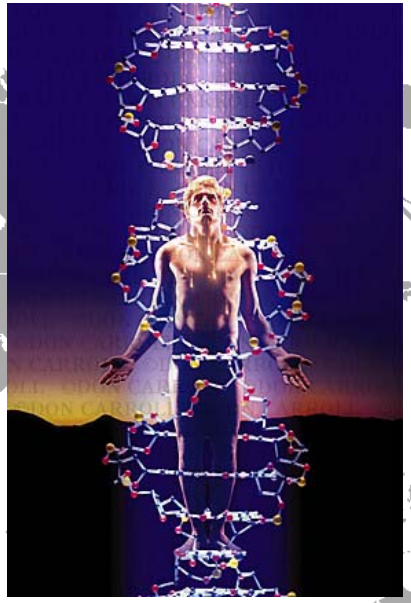


Indigenous populations

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

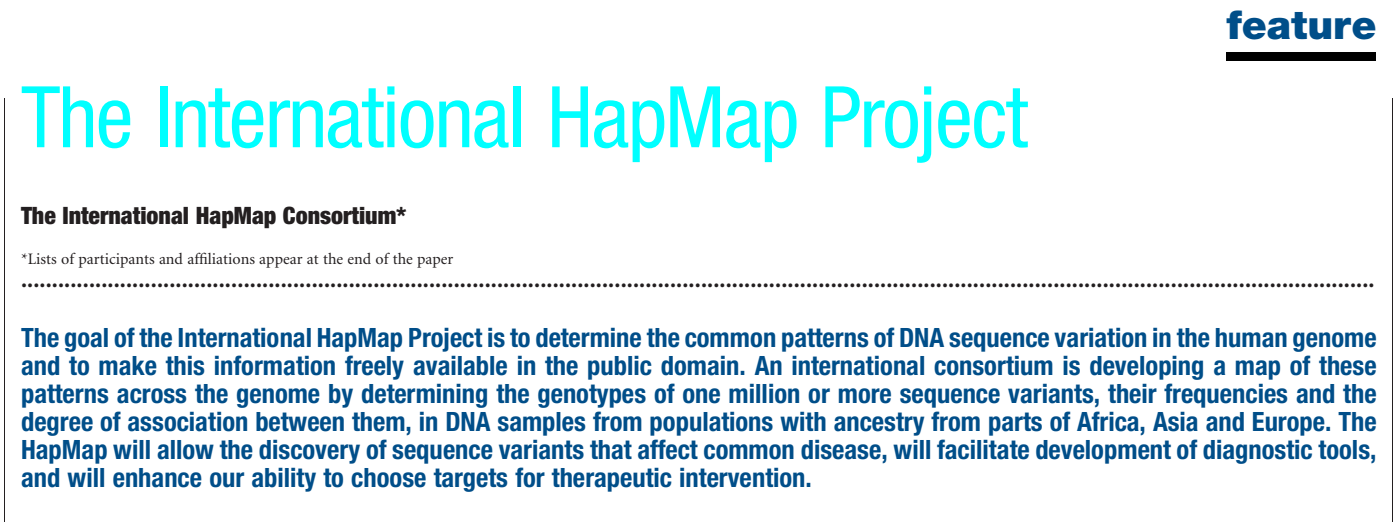
Mapping Human Genetic Diversity and tracing the genetic origins of Asian populations
The HUGO Pan-Asian SNP Consortium
Science, October 2009

DNA: the blueprint of our lives



PROPER DRUGS AND TREATMENT

HAPMAP Project



feature

The International HapMap Project

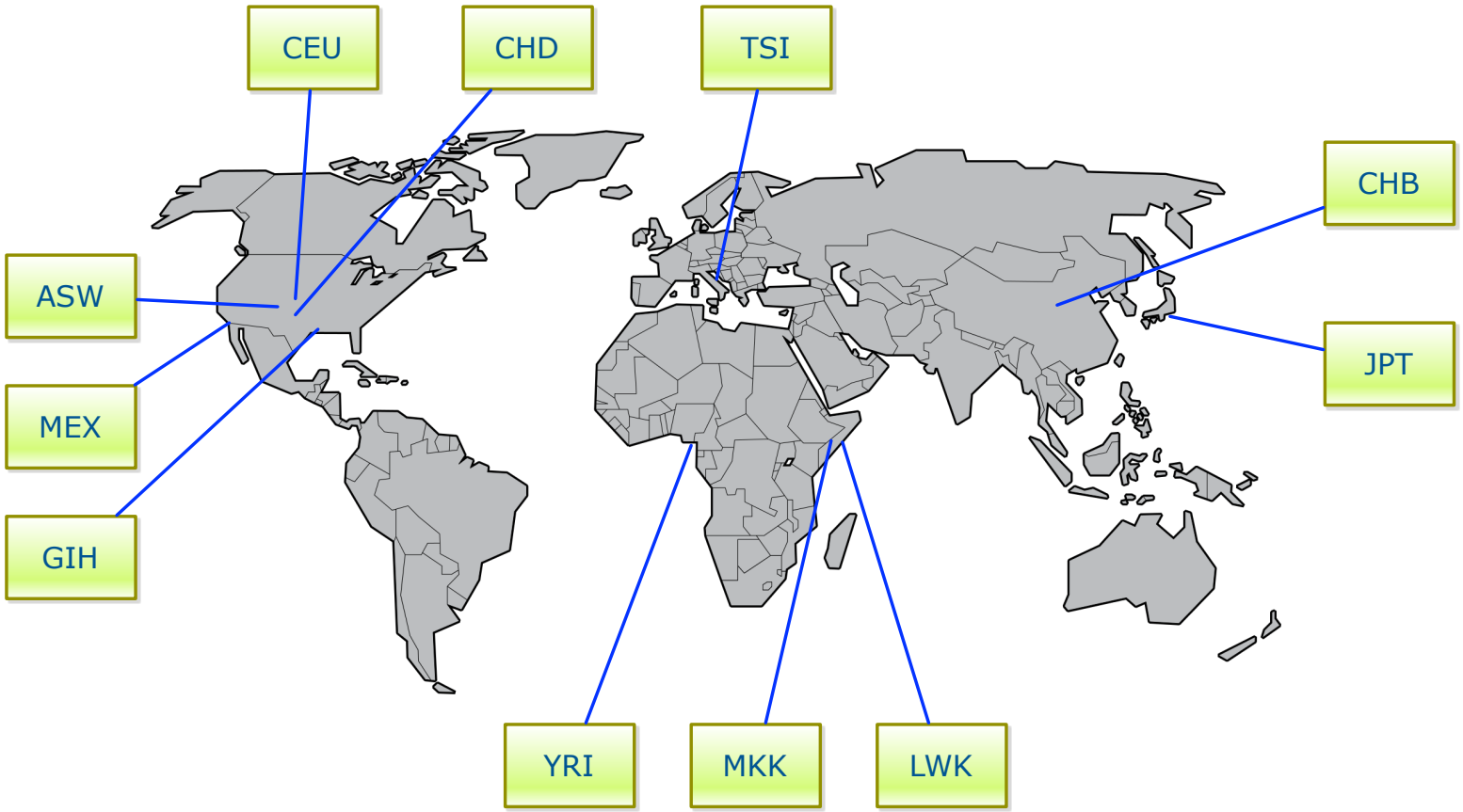
The International HapMap Consortium*

*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

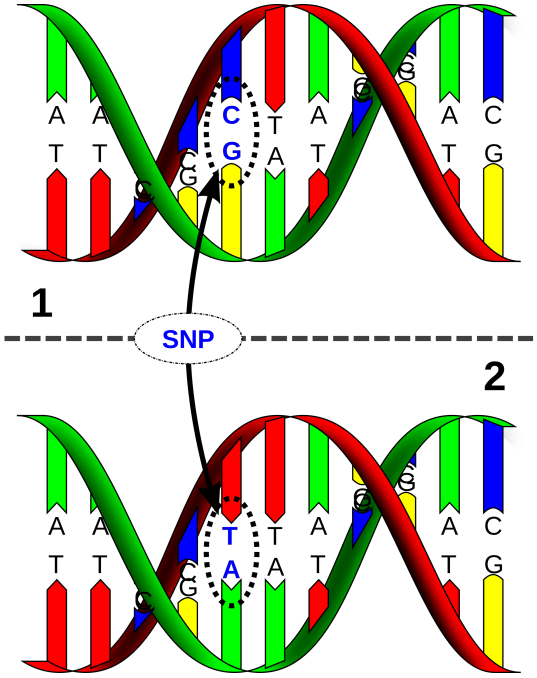
IGSR and the 1000 Genomes Project

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



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

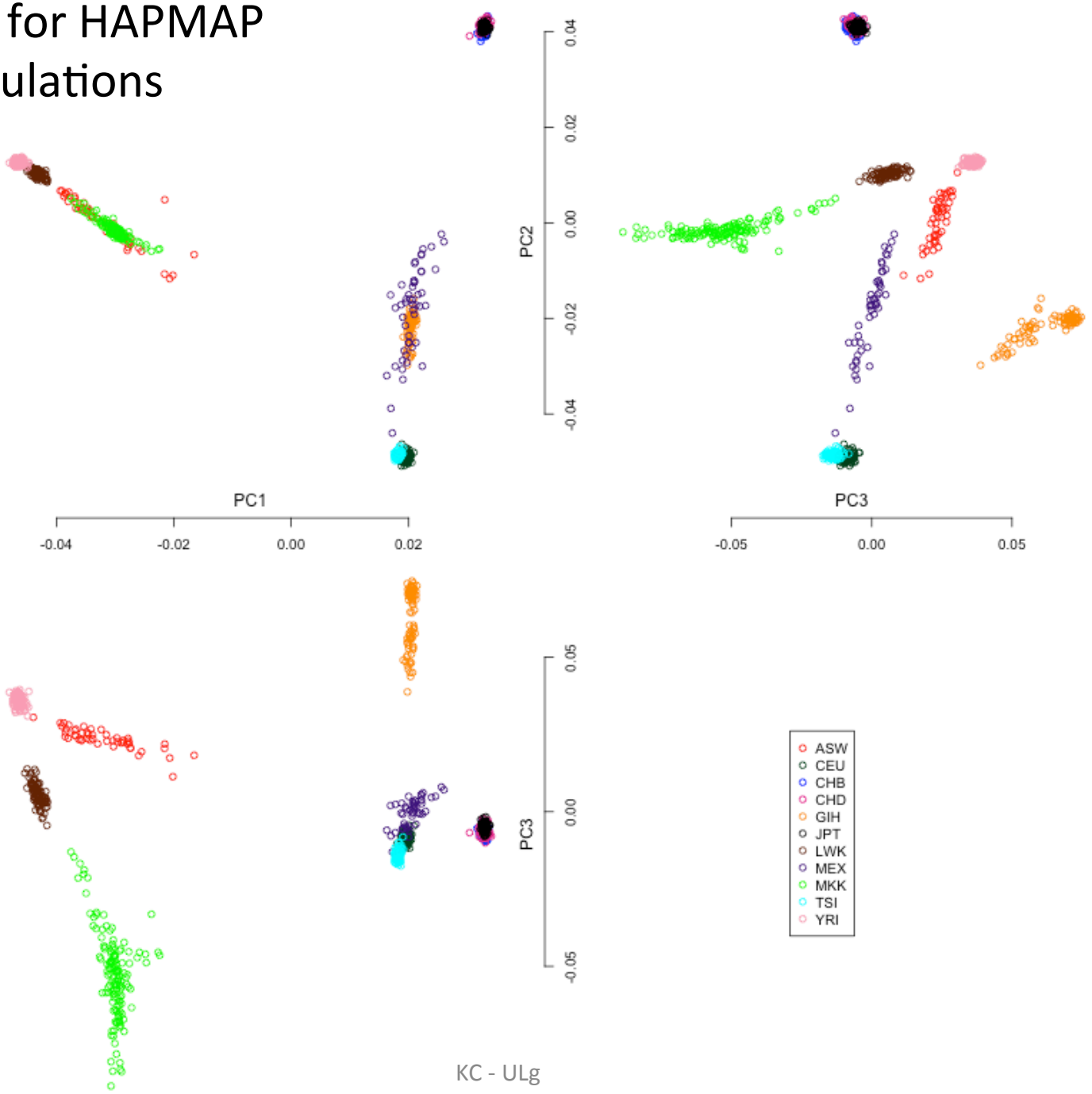
nature genetics

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

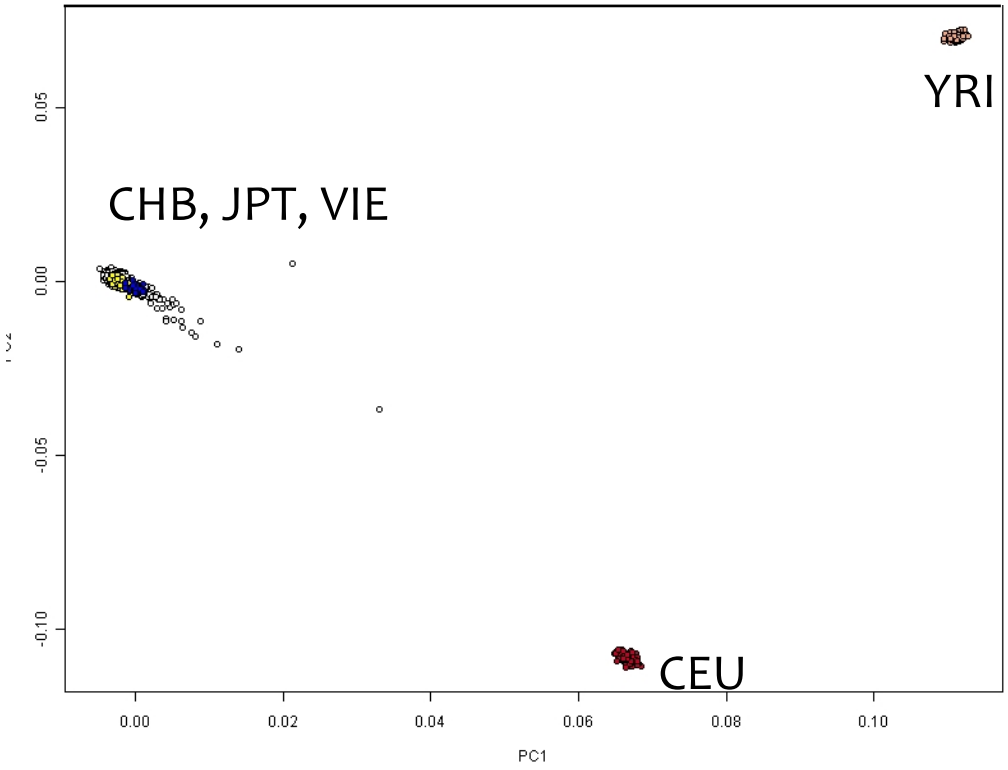
Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

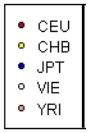
Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method 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



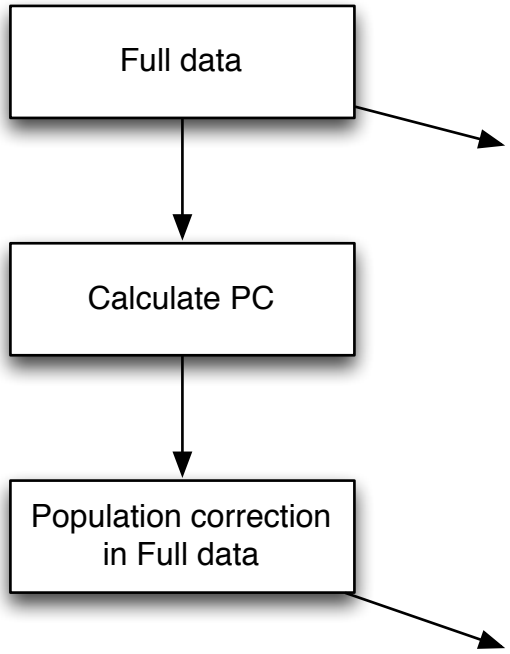


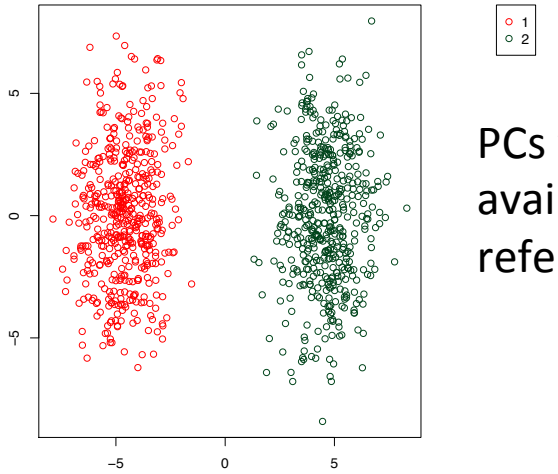
Genome-wide association study for Dengue shock syndrome

Chiea Chuen Khor et al.
Nature Genetics 2012

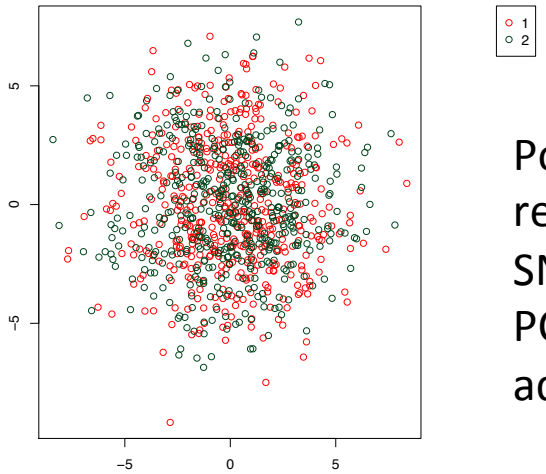
- 657,366 SNPs
- 4,028 individuals from Vietnam

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

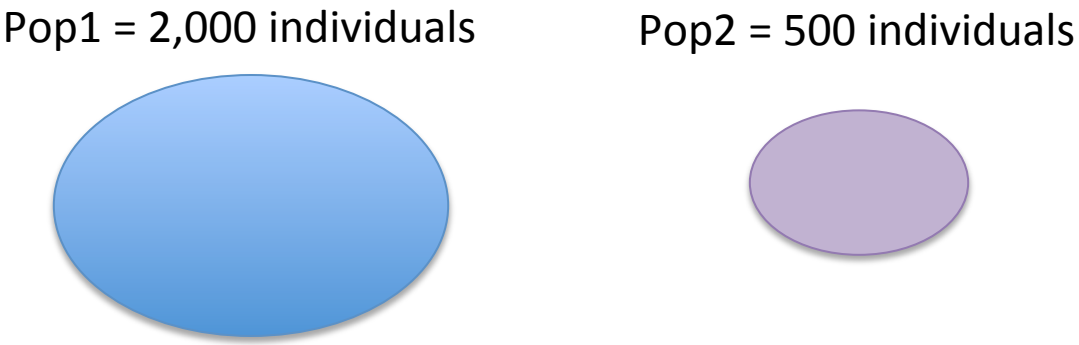
glm(formula, family=familytype(link=linkfunction), data=)

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

Fixation index (F_{ST})

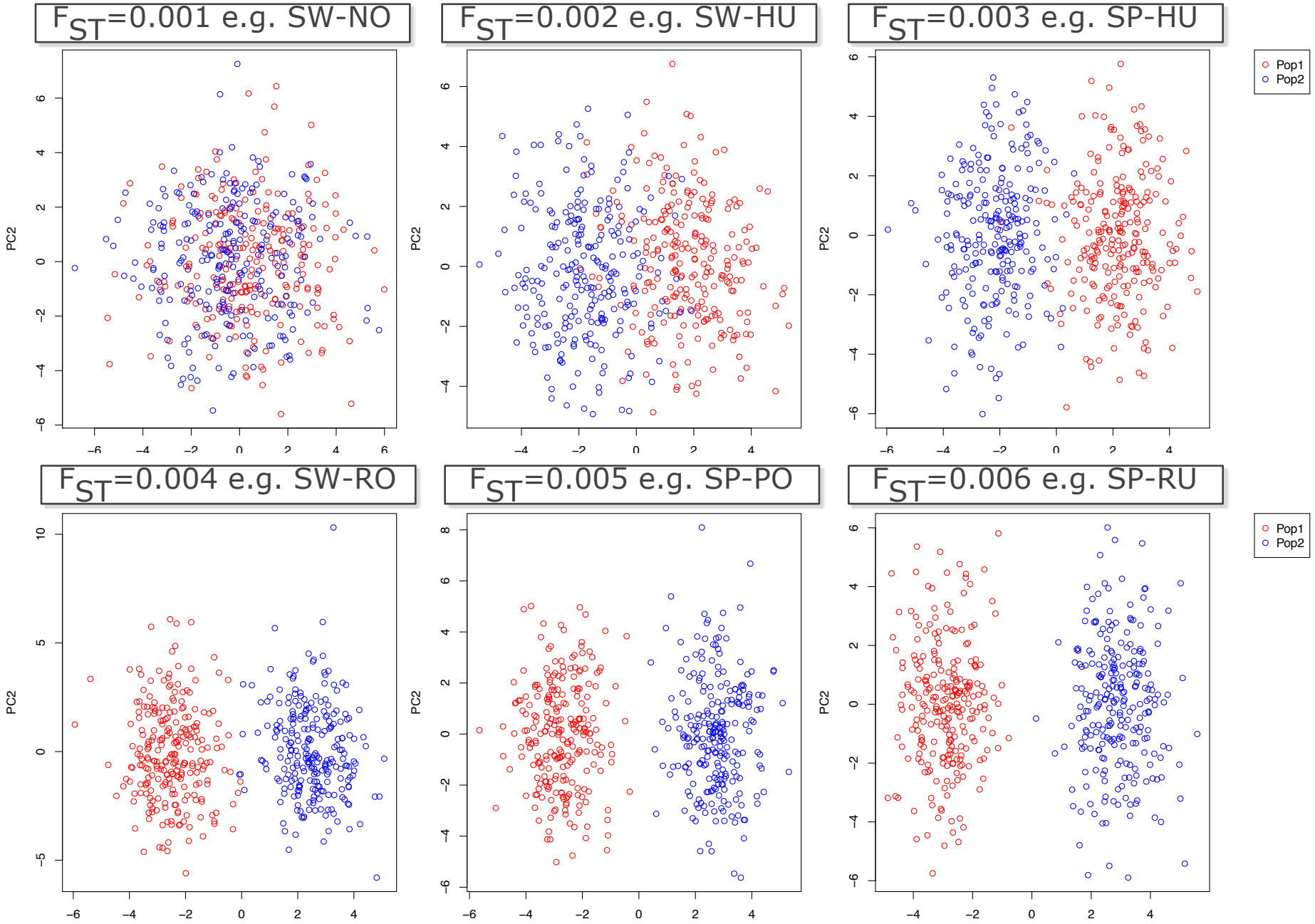
- F_{ST} can be used to describe a distance among population.
- F_{ST} can be biased due to the allele frequencies and the number of independent SNPs.



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]

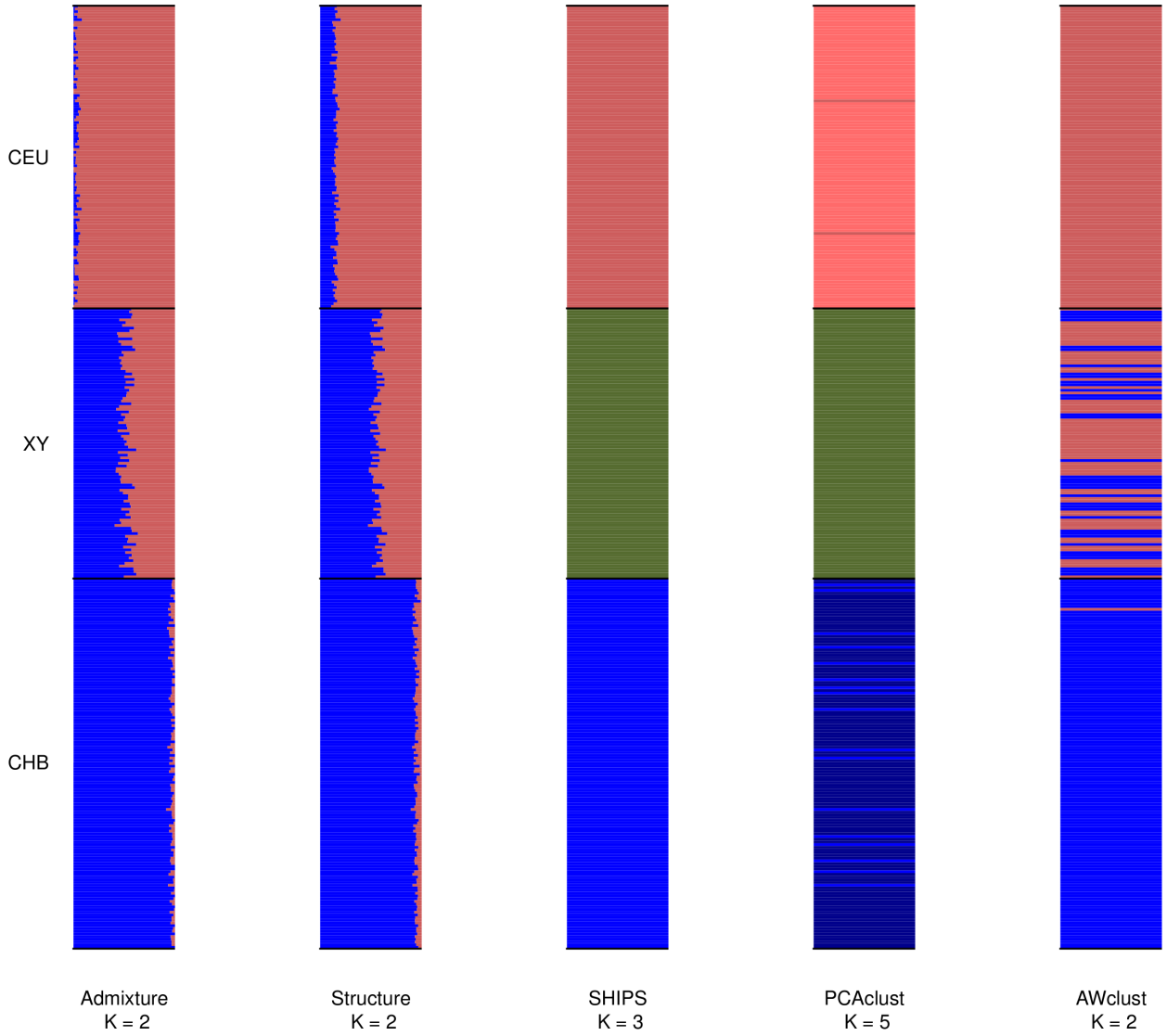
[1]Harvard–Massachusetts Institute of Technology (MIT), Division of Health, Science, and Technology, Cambridge, Massachusetts 02139, USA; [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [4]Department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

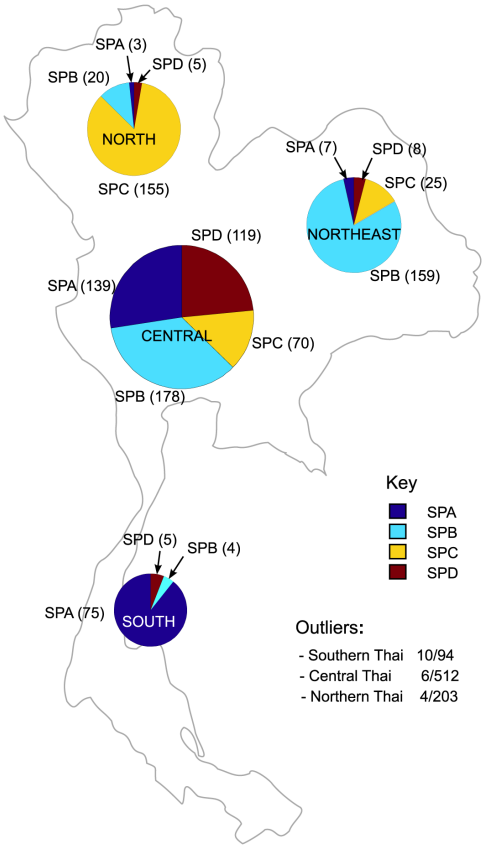
Genetic admixture occurs when two or more previously isolated populations begin interbreeding. Admixture results in the introduction of new genetic lineages into a population. It has been known to slow local adaptation by introducing foreign, unadapted genotypes (known as gene swamping). It also prevents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

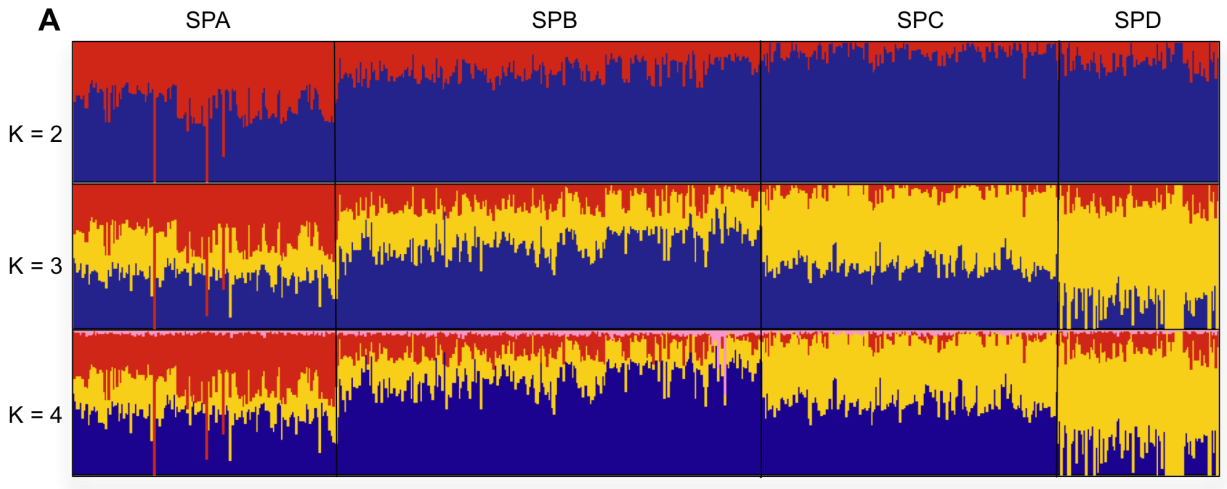
Tools for Admixture profiling



Bouaziz 2012

Thai population





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

PLINK: Why PLINK?

- PLINK is a whole genome association analysis software, and it is FREE!

  http://pngu.mgh.harvard.edu/~purcell/plink/
- PLINK has a well-documented manual to explain all features
- PLINK is available for Linux, Mac OS, and MS-DOS
- PLINK has 2 versions, the stable version (1.07) and the beta version (1.9)
  - PLINK 1.9 works much faster than 1.07
  - PLINK 1.9 has many new features
- gPLINK is the other version of PLINK that provides graphical user interface. Please be aware that using PLINK for a while genome analysis usually takes a long time, it is better to use a command-line version

PLINK: File Formats

PLINK mainly supports 3 types of formats

- Standard text format (PED and MAP) Note that all files must have the same name, otherwise we need to clearly indicate by using *--ped* and *–map*

  plink --file test

- Binary format (BED, BIM, and FAM)

  plink --bfile test

- Transposed text format (TPED, and TFAM) Note that all files must have the same name, otherwise we need to clearly indicate by using *--tped* and *--tfam*

  plink --tfile test

Format conversion

- To convert or to indicate output as text format (PED and MAP

  plink --file test --recode --out test_ped
- To convert or to indicate output as Binary format (BED, BIM, and FAM)

  plink --file test --make-bed --out test_bin
- To convert or to indicate output as Transposed text format (TPED, and TFAM)

  plink --file test --transpose --recode --out test_tp
- Alternatively, it is possible to recode data as 1/2 encoding

  plink --file test --recode12 --out test_12
- To convert to additive encoding

  plink --file test --recodeAD --out test_12
- It is possible to switch between A,T,G,C encoding to 1,2,3,4 encoding by using --allele1234 or --alleleACGT vice versa

Data manipulation: SNPs (1/3)

To get a set of SNPs, you can specify a single SNP and, optionally, also ask for all SNPs in the surrounding region, with the --window option:

plink --bfile mydata --snp rs652423 --window 20

which extracts only SNPs within +/- 20kb of rs652423 based on multiple SNPs and ranges (--snps)

The --snps command will accept a comma-delimited list of SNPs, including ranges based on physical position. For example,

plink --bfile mydata --snps rs273744-rs89883,rs12345-rs67890,rs999,rs222

Based on physical position (--from-kb, etc)

plink --bfile mydata --chr 2 --from-kb 5000 --to-kb 10000

to select all SNPs within this 5000kb region on chromosome 2.

Data manipulation: SNPs (2/3)

To merge more than two standard and/or binary filesets, it is often more convenient to specify a single file that contains a list of PED/MAP

For example, consider we had 4 PED/MAP filesets (labelled fA.* through fD.*) and 4 binary filesets, labelled fE.* through fH.*).

Then using the command:

plink --file fA --merge-list allfiles.txt --make-bed --out mynewdata

Data manipulation: SNPs (3/3)

To exclude some sets of SNPs

plink --file data --exclude mysnps.txt

where the file mysnps.txt is, as for the --extract command, just a list of SNPs, one per line.

Data manipulation: individuals (1/3)

To get a set of individuals

```
plink --file data --keep mylist.txt
```

where the file mylist.txt is, as for the --remove command, just a list of Family ID / Individual ID pairs, one set per line, i.e. one person per line. (fields can occur after the 2nd column but they will be ignored -- i.e. you could use a FAM file as the parameter of the --keep command, or have comments in the file. For example

```
F101   1
F1001  2_B
F3033  1_A  Drop this individual because of consent issues
F4442  22
```

Data manipulation: individuals (2/3)

To exclude a set of individuals

plink --file data --remove mylist.txt

where the file mylist.txt is, as for the --keep command, just a list of Family ID / Individual ID pairs, one set per line, i.e. one person per line (although, as for --keep, fields after the 2nd column are allowed but they will be ignored).

Data manipulation: individuals (3/3)

Filter some individuals

```
plink --file data --filter myfile.raw 1 --freq
```

implies a file myfile.raw exists which has a similar format to phenotype and cluster files: that is, the first two columns are family and individual IDs; the third column is expected to be a numeric value (although the file can have more than 3 columns), and only individuals who have a value of 1 for this would be included in any subsequent analysis or file generation procedure. e.g. if myfile.raw were

```
F1 I1 2
F2 I1 7
F3 I1 1
F3 I2 1
F3 I3 3
```

Because filtering on cases or controls, or on sex, or on position within the family, will be common operations, there are some shortcut options that can be used instead of --filter. These are:

```
--filter-cases
--filter-controls
--filter-males
--filter-females
--filter-founders
--filter-nonfounders
```

Quality control processes

- Missing genotype
- Hardy-Weinberg Equilibrium
- Minor Allele frequency
- Linkage disequilibrium pruning

Missing rate per person

The initial step in all data analysis is to exclude individuals with too much missing genotype data. This option is set as follows:

```
plink --file mydata --mind 0.1
```

which means exclude with more than 10% missing genotypes. A line in the terminal output will appear, indicating how many individuals were removed due to low genotyping. If any individuals were removed, a file called

```
plink.irem
```

will be created, listing the Family and Individual IDs of these removed individuals. Any subsequent analysis also specifeid on the same command line will be performed without these individuals.

Missing rate per SNP

Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate, with the --geno option: the default is to include all SNPS (i.e. --geno 1).

To include only SNPs with a 90% genotyping rate (10% missing) use

```
plink --file mydata --geno 0.1
```

As with the --maf option, these counts are calculated after removing individuals with high missing genotype rates.

Hardy-Weinberg Equilibrium

To exclude markers that failure the Hardy-Weinberg test at a specified significance threshold, use the option:

plink --file mydata --hwe 0.001

By default this filter uses an exact test. The standard asymptotic (1 df genotypic chi-squared test) can be requested with the --hwe2 option instead of --hwe.

The following output will appear in the console window and in plink.log, detailing how many SNPs failed the Hardy-Weinberg test, for the sample as a whole, and (when PLINK has detected a disease phenotype) for cases and controls separately:

Writing Hardy-Weinberg tests (founders-only) to [plink.hwe]

30 markers failed HWE test (p <= 0.05) and have been excluded

34 markers failed HWE test in cases

30 markers failed HWE test in controls

This test will only be based on founders (if family-based data are being analysed) unless the --nonfounders option is also specified.

Minor Allele frequency

Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency):

```
plink --file mydata --maf 0.05
```

means only include SNPs with MAF >= 0.05. The default value is 0.01. This quantity is based only on founders (i.e. individuals for whom the paternal and maternal individual codes and both 0).

This option is appropriately counts alleles for X and Y chromosome SNPs.

Linkage disequilibrium pruning (1/2)

Sometimes it is useful to generate a pruned subset of SNPs that are in approximate linkage equilibrium with each other. This can be achieved via two commands: --indep which prunes based on the variance inflation factor (VIF), which recursively removes SNPs within a sliding window; second, --indep-pairwise which is similar, except it is based only on pairwise genotypic correlation.

The VIF pruning routine is performed:

plink --file data --indep 50 5 2

will create files

plink.prune.in

plink.prune.out

Each is a simlpe list of SNP IDs; both these files can subsequently be specified as the argument for a --extract or --exclude command.

The parameters for --indep are: window size in SNPs (e.g. 50), the number of SNPs to shift the window at each step (e.g. 5), the VIF threshold. The VIF is 1/(1-R^2) where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously. That is, this considers the correlations between SNPs but also between linear combinations of SNPs.

Linkage disequilibrium pruning (2/2)

The second procedure is performed:

plink --file data --indep-pairwise 50 5 0.5

This generates the same output files as the first option; the only difference is that a simple pairwise threshold is used. The first two parameters (50 and 5) are the same as above (window size and step); the third parameter represents the r^2 threshold.

To give a concrete example: the command above that specifies 50 5 0.5 would a) consider a window of 50 SNPs, b) calculate LD between each pair of SNPs in the window, b) remove one of a pair of SNPs if the LD is greater than 0.5, c) shift the window 5 SNPs forward and repeat the procedure.

To make a new, pruned file, then use something like (in this example, we also convert the standard PED fileset to a binary one):

plink --file data --extract plink.prune.in --make-bed --out pruneddata

PCA using PLINK

PLINK 1.9 provides two dimension reduction routines: --pca, for principal components analysis (PCA) based on the variance-standardized relationship matrix. Top principal components are generally used as covariates in association analysis regressions to help correct for population stratification.

plink --file data --pca {count} <header> <tabs> <var-wts>

By default, --pca extracts the top 20 principal components of the variance-standardized relationship matrix.

You can change the number by passing a numeric parameter {count}.

Eigenvectors are written to plink.eigenvec, and top eigenvalues are written to plink.eigenval.

The 'header' modifier adds a header line to the .eigenvec file(s), and the 'tabs' modifier makes the .eigenvec file(s) tab- instead of space-delimited.

You can request variant weights with the 'var-wts' modifier

Epistasis using PLINK

--fast-epistasis <boost | joint-effects | no-ueki> <case-only> <set-by-set | set-by-all>

--fast-epistasis starts an imprecise but fast scan for epistasis based on inspection of 3x3 joint genotype count tables.

Linear/logistic regression-based test

--epistasis <set-by-set | set-by-all>

--epistasis, for a quantitative trait, uses linear regression to fit the model

$Y = \beta 0 + \beta 1gA + \beta 2gB + \beta 3gAgB$

for each inspected variant pair (A, B), where gA and gB are allele counts; then the β3 coefficients are tested for significance

https://www.cog-genomics.org/plink/1.9/epistasis

Covariates

To do population correction, you can use EigenVectors as covariates.

--covar [filename] <keep-pheno-on-missing-cov>

--covar-name [column ID(s)/range(s)...]

--covar-number [column number(s)/range(s)...]

--covar designates the file to load covariates from. The file contains FID and IID in first two columns and covariates in remaining columns.

--covar-name lets you specify a subset of covariates to load, by column name; separate multiple column names with spaces or commas, and use dashes to designate ranges. (Spaces are not permitted immediately before or after a range-denoting dash.)

--covar-number lets you use column numbers instead.

For example, if the first row of the covariate file is

FID IID SITE AGE DOB BMI ETH SMOKE STATUS ALC

then the following two expressions have the same effect:

--covar-name AGE, BMI-SMOKE, ALC

--covar-number 2, 4-6, 8