Identifying and Adjusting for Population Structure

Hae Kyung Im, PhD

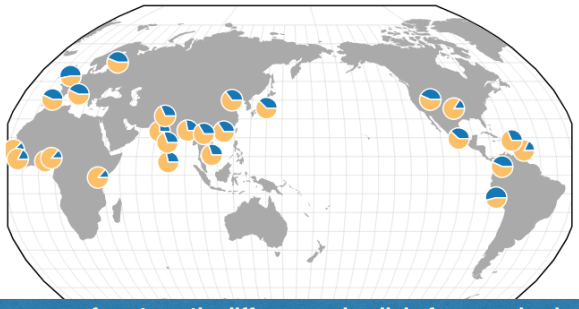
THE UNIVERSITY OF CHICAGO

February 1, 2021

Today's class will focus on how to identify population structure and how to correct false positives that may arise in association studies.

What is Population Structure

chr8:61325684 T/C hg19



Presence of systematic differences in allele frequencies between subsets of individuals due to different ancestries

Marcus & Novembre (2016) Visualizing the Geography of Genetic Variants. Bioinformatics

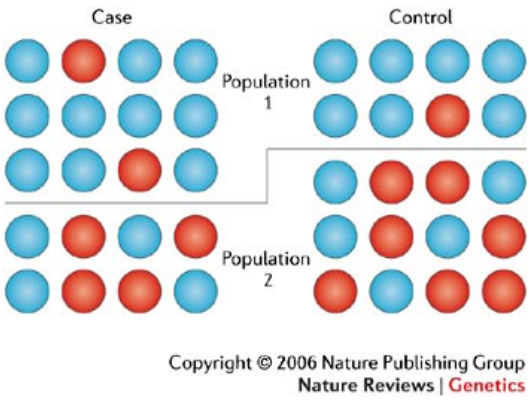
Population structure refers to systematic differences in allele frequencies in different regions of the world due to different ancestries. This figure shows the allele frequencies of one variant

Population Structure Manifests in

- Departure from Hardy Weinberg Equilibrium
- Reduced heterozygosity (fraction of markers that are called heterozygous) due to population structure (Wahlund effect)
- Patterns in principal components analysis of genetic data
- Spurious associations leading to false positives

Population structures manisfets in departure from hardy weinberg equilibrium, reduced heterozygosity, patterns in principal component analysis of genetic data, spurious associations leading to false positives

Spurious Association Due to Population Structure



Spurious associations can result due to different composition of populations among cases and controls as shown in this example. Here, population 1 is overrepresented among cases while population 2 is overrepresented among controls. The blue variant is more common in population 1 and the analysis will suggest that the blue allele increases the risk of disease, even if there is no effect on the disease.

HapMap Project

nature

Feature | Published: 18 December 2003

The International HapMap Project

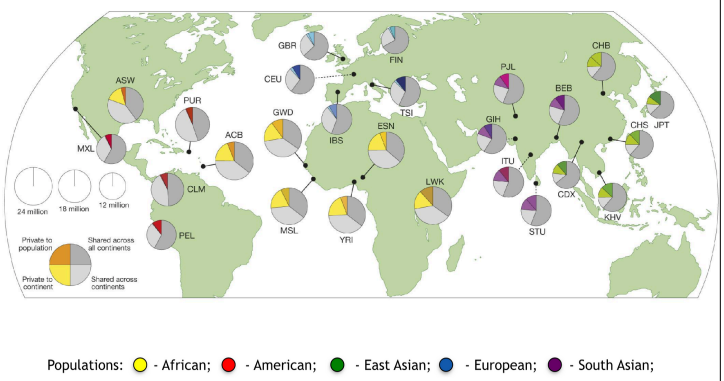
†The International HapMap Consortium

Nature **426**, 789–796(2003) | Cite this article

An international project to create a haplotype map of the human genome

The international HapMap project recruited individuals who consented to have their genotypes publicly available with the goal of mapping haplotypes of the human genome.

1000 Genomes Project



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

Auton, A., Altshuler, D. M., Durbin, R. M., Chakravarti, A., Clark, A. G., Donnelly, P., et al. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. http://doi.org/10.1038/nature15393

6

The HapMap project evolved into the 1000 Genomes project, which aimed to sequence 1000 individuals from around the globe including African, American (native), East Asian, European, and South Asian populations. As of 2020, these resources are still being heavily used to understand the human genetic diversity: https://twitter.com/JeffreyMKidd/status/1222532448744923136

Newly sequenced 1000 genomes to high coverage https://twitter.com/1000genomes/status/1294222026769604608

Hardy Weinberg Equilibrium in Multi-ancestry Samples

HapMap Phase 3 Populations

ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

HapMap Phase 3 Populations

```{r}
popinfo = read_tsv("relationships_w_pops_051208.txt")

## What's population composition?
popinfo %>% count(population)
```

R Console | tbl_df 11 x 2

population <chr>	n <int>
ASW	90
CEU	180
CHB	90
CHD	100
GIH	100
JPT	91
LWK	100
MEX	90
MKK	180
TSI	100

1–10 of 11 rows Previous 1 2 Next

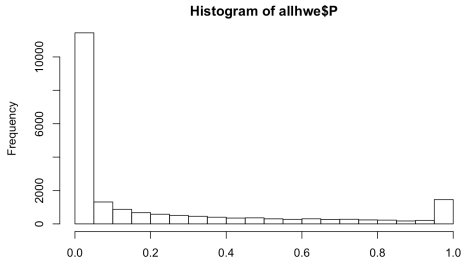
https://hakyimlab.github.io/hgen471/L6-population-structure.html

we can check the number of individuals in each of the HapMap 3 populations.

HWE with Mixed Population

```{r }
## what happens if we calculate HWE with this mixed population?
system("./plink --bfile hapmapch22 --hardy --out out/allhwe")
allhwe = read.table("out/allhwe.hwe",header=TRUE,as.is=TRUE)
hist(allhwe$P)
qqunif(allhwe$P,main='HWE HapMap3 All Pop')
```

R Console



https://hakyimlab.github.io/hgen471/L6-population-structure.html

Plink will test the Hardy Weinberg Equilibrium (HWE) of all variants in the genotype file. This is the departure from the expected counts: n*p^2, n*2*p*(1-p), n*(1-p) for aa, aA, and AA where

a = minor allele

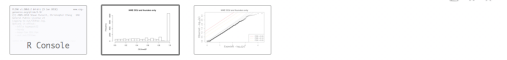
n = number of individuals

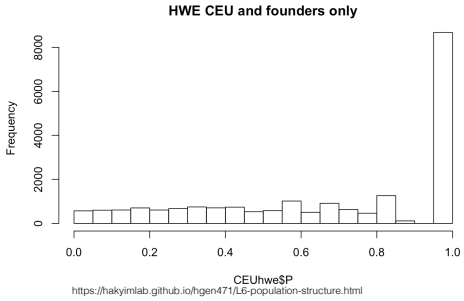
p = minor allele frequency

The concentration of small p-values (peak near 0) indicates that the majority of variants do not follow HWE.

HWE with European population only

```{r}
## what if we calculate with single population?
popinfo %>% filter(population=="CEU") %>% write_tsv(path="CEU.fam")
system("./plink --bfile hapmapch22 --hardy --keep-fam CEU.fam --out out/CEUhwe")
CEUhwe = read.table("out/CEUhwe.hwe",header=TRUE,as.is=TRUE)
hist(CEUhwe$P,main="HWE CEU and founders only")
qqunif(CEUhwe$P,main="HWE CEU and founders only")
```





https://hakyimlab.github.io/hgen471/L6-population-structure.html

11

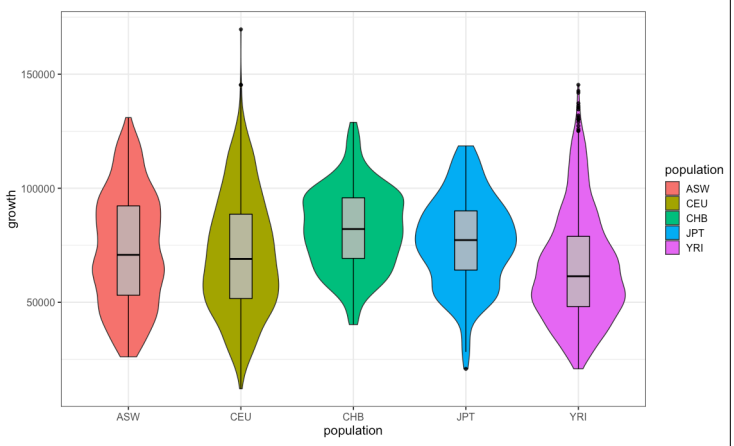
When we test HWE using only one population, here CEU/Europeans, we no longer see the peak at 0.

The peak at 1 is an artifact, probably due to pre-selection of variants that are in HWE.

GWAS in Multi-ethnic Samples

We will see next how population structure can inflate association statistics.

Example: Growth Phenotype by Population



H. K. Im et al, "Mixed effects modeling of proliferation rates in cell-based models: consequence for pharmacogenomics and cancer.," PLoS Genetics, 2012.

13

Here we are showing the proliferation rate of the lymphoblastoid cell lines of HapMap individuals. Notice the difference in mean among populations. This is likely to cause spurious associations since SNPs with different allele frequencies among populations will correlate with these differences leading to false positives.

Example: Growth Phenotype by Population

```
lm(formula = growth ~ population, data = igrowth)

Residuals:
   Min     1Q Median     3Q    Max
-58821 -18093  -2242  15896  98760

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    73080.8      938.2  77.894  < 2e-16 ***
populationCEU  -2190.1     1175.4  -1.863   0.0625 .
populationCHB   9053.1     2043.9   4.429 9.73e-06 ***
populationJPT   3476.8     2034.8   1.709   0.0876 .
populationYRI  -7985.2     1137.2  -7.022 2.61e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24160 on 3591 degrees of freedom
  (130 observations deleted due to missingness)
Multiple R-squared:  0.0345,    Adjusted R-squared:  0.03342
F-statistic: 32.08 on 4 and 3591 DF,  p-value: < 2.2e-16
```

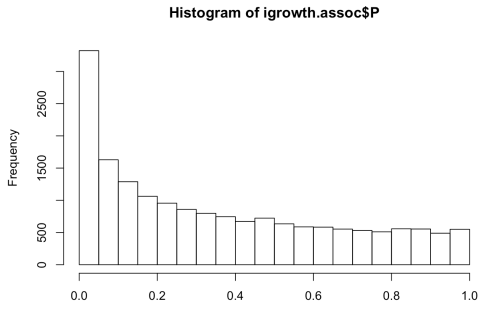
https://hakyimlab.github.io/hgen471/L6-population-structure.html

14

Linear regression of growth phenotype on population yields significant differences in mean growth between populations. ASW is here the reference to which the other populations are being compared too. (Default in R is to order factors by alphabetic order.)

Populations Differences Lead to Inflation of Small P-values

```
system(glue::glue("~/bin/plink --bfile {work.dir}hapmapch22 --linear --pheno {work.dir}igrowth.pheno --pheno-name growth --maf 0.05 --out {work.dir}output/igrowth"))
igrowth.assoc = read.table(glue::glue("{work.dir}output/igrowth.assoc.linear"),header=T,as.is=T)
hist(igrowth.assoc$P)
```



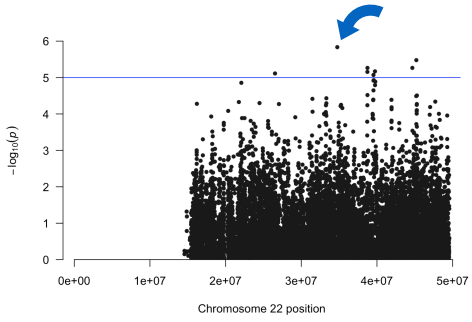
https://hakyimlab.github.io/hgen471/L6-population-structure.html

15

Here we show the histogram of p-values of the association between growth phenotype and about 20K SNPs in chr 22. A concentration of small p-values suggests that either causal variants are associated with proliferation rate or that they are inflated by population structure.

Populations Differences Lead to Inflation of Small P-values

```
manhattan(igrowth.assoc, chr="CHR", bp="BP", snp="SNP", p="P" )
```



https://hakyimlab.github.io/hgen471/L6-population-structure.html

16

Manhattan plot shows no genome-wide significant hit but if we take into account that we have only 20,649 variants tested here so Bonferroni corrected threshold would be 10^{-5.6}

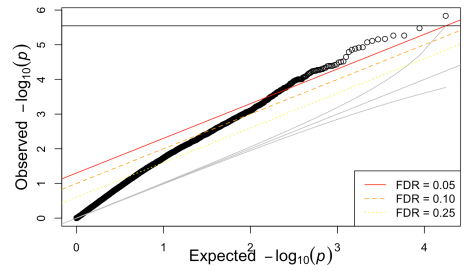
--

> 0.05/20649
[1] 2.421425e-06
> -log10(0.05/20649)
[1] 5.615929

--

Populations Differences Lead to Inflation of Small P-values

```
qqunif(igrowth.assoc$P)
```



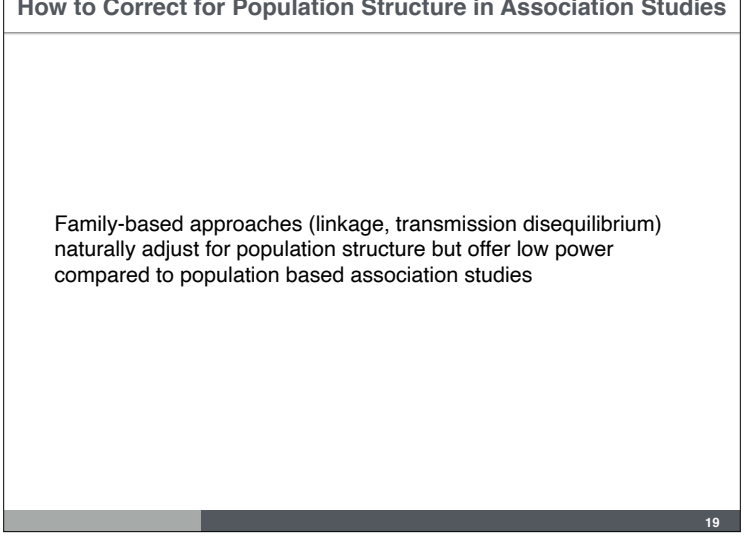
https://hakyimlab.github.io/hgen471/L6-population-structure.html

17

In general (exceptions to be discussed later), well behaved qq-plots follow the identity line (representing the variants that are true nulls, no relationship between the variant and the phenotype) and tend to depart significantly from the identity line for a smaller set of variants. In this qq plot, the variants depart from the identity line from the very beginning, indicating some level of associations for most variants. This is not impossible (highly polygenic traits where all are causal) but more likely to indicate inflation of the qq-plot due to population stratification. Most variants have differences between populations, and that difference is being used to "explain" the mean

differences in proliferation rate.

How to Correct for Population Structure



In general, it is hard to get family studies with very large sample sizes so we will look for other ways to account for population structure.

How to Correct for Population Structure in Association Studies

1. Correcting with genomic control (Devlin and Roeder 1999)

2. Inferring the latent sub-populations (Pritchard et al 2000)
   Fit association in each population separately and combine

3. Adjusting for principal components
   (Patterson 2006, Novembre 2008, Price et al 2010)

4. Mixed effects modeling (EMMAX, Kang et al 2010)

1. Genomic control method by Devlin and Roeder is a simple approach broadly used. As sample sizes increase, we will see that we need to revisit this approach.
2. If we know the subpopulations, we can run the GWAS within each and meta-analyze the results (more on this next lecture). Even if we don't know the subpopulations a priori, if they are distinct enough we may be able to identify them and run GWAS within each latent sub-population (principal component analysis will help for this).
3. The most common approach used now is using principal components are covariates.
4. Mixed effects modeling is another approach.

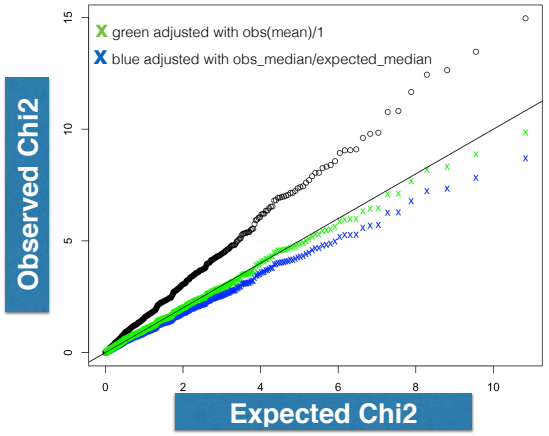
---

1. Genomic Control (Devlin and Roeder, 1999)

- **assumption**: the effects of population stratification and cryptic relatedness are constant across the genome
  - the test statistics distributed λ * Chi2
- estimate λ using the
  - mean of test statistics, or
  - median/qchisq(0.5)=0.456
- procedure works fairly well for λ close to 1 (1 - 1.1)
- λ < 1.05 considered acceptable inflation

As seen in the growth phenotype example, population stratification can lead to inflation of false positives. More small p-values than we would expect (uniformly distributed).

1. Genomic Control - Adjustment



22

Let's simulate a really dumb case control study that has all cases from one population and all controls from another population. Here all SNPs that have different frequencies between populations will be significantly associated with case-control status. Chi2 of these associations are shown in this figure. If we apply regression, Chi2 stat = (effect size / standard error)^2. Green dots correspond to corrected chi2 with genomic control. Genomic control here is λ = mean(obs chi2 stat) / mean(expected chi2 stat). One can also adjust with medians, to keep it robust to outliers (true signals with large chi2 stat could skew the mean and over-correct the statistic)

Function to Calculate Genomic Control

```
## Reads data
S <- read.table(input, header = FALSE)

if (stat_type == "Z")
   z = S[, 1]

if (stat_type == "CHISQ")
   z = sqrt(S[, 1])

if (stat_type == "PVAL")
   z = qnorm(S[, 1] / 2)

## calculates lambda
lambda = round(median(z^2) / 0.4549, 3)

lambda
```

```
Where 0.4549 is the median of a
chi2 r.v. with 1 df
qchisq(0.5, 1)
[1] 0.4549364
```

23

here is a function that will calculate the genomic control (λ) given Z scores, Chi2, or pvalues in a data frame called input.

2. Infer Latent Population Structure

- Infer latent population structure
  - e.g. STRUCTURE Pritchard et al
- Perform association in each sub-population and aggregate using meta-analysis

We can also infer the latent subpopulations with methods such as STRUCTURE or others, and perform association within each subpopulation and then meta-analyze to get the full population association statistics.

3. Adjust with Principal Components

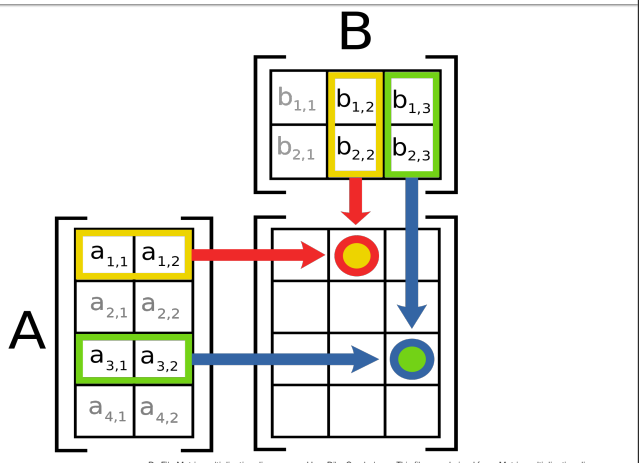
- This is currently the most common approach

Review
Matrix Algebra

Matrix Algebra

	Example
Addition	$\begin{bmatrix} 1 & 3 & 1 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 5 \\ 7 & 5 & 0 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 & 1+5 \\ 1+7 & 0+5 & 0+0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 6 \\ 8 & 5 & 0 \end{bmatrix}$
Scalar Multiplication	$2 \cdot \begin{bmatrix} 1 & 8 & -3 \\ 4 & -2 & 5 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 & 2 \cdot 8 & 2 \cdot -3 \\ 2 \cdot 4 & 2 \cdot -2 & 2 \cdot 5 \end{bmatrix} = \begin{bmatrix} 2 & 16 & -6 \\ 8 & -4 & 10 \end{bmatrix}$
Transposition	$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$

Matrix Multiplication



Matrix Multiplication

$$(A * B)_{i,j} = \sum_k A_{i,k} B_{k,j}$$

order of matrices matters

$$(A * B)_{i,j} \neq (B * A)_{i,j}$$

Einstein notation: repeated index gets summed over

$$A_{i,k} B_{k,j} \rightarrow \sum_k A_{i,k} B_{k,j}$$

Matrix Form of System of Linear Equations

$$\mathbf{Ax} = \mathbf{b}$$

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n &= b_1 \\ &\vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n &= b_m \end{aligned}$$

https://en.wikipedia.org/wiki/Matrix_(mathematics)

Derive Linear Regression Solution with Matrix Notation

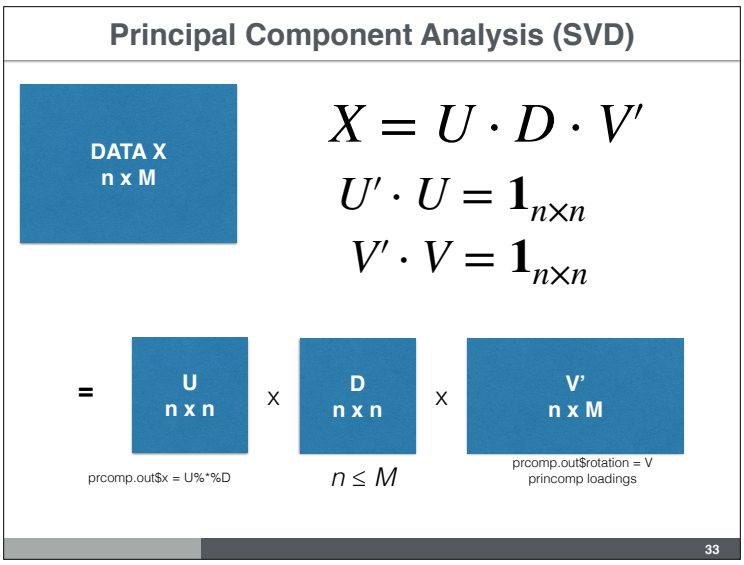
Let's assume we have demeaned and divided by sd Y's and X's

Write $Y = X \beta + \varepsilon$

Check dimensions

$X'Y = X'X \beta + X' \varepsilon$

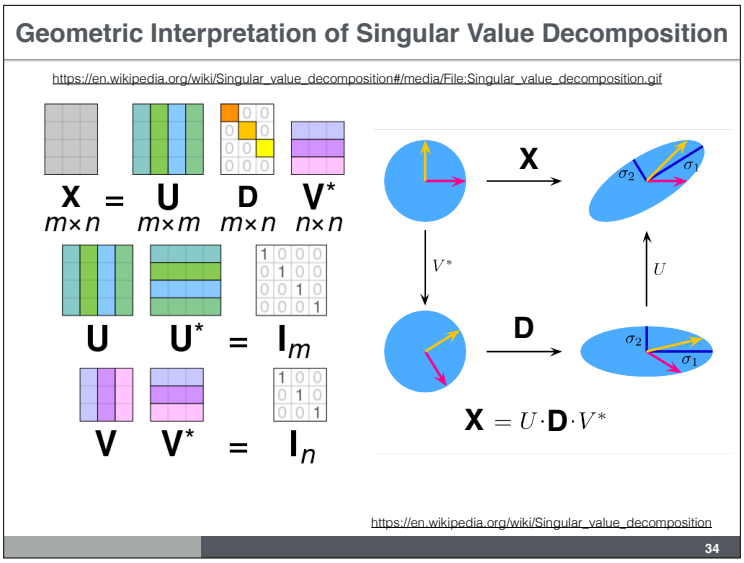
Principal Component Analysis



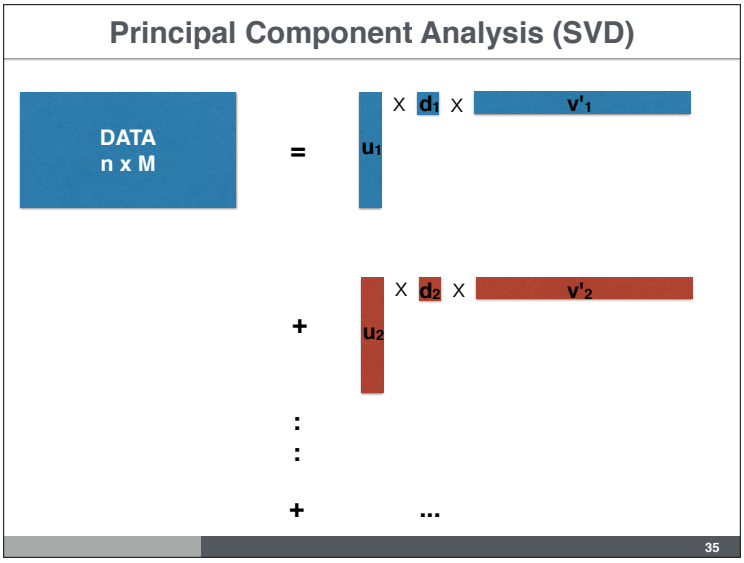
Principal component analysis of a data matrix can be performed using a method called singular value decomposition (SVD). SVD will find the main axis where the data varies.

U has left singular vectors as columns

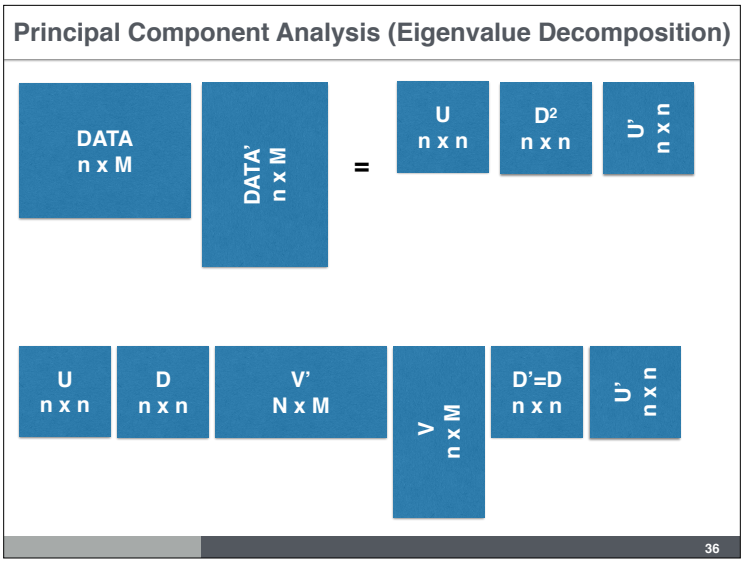
V has right singular vectors as columns



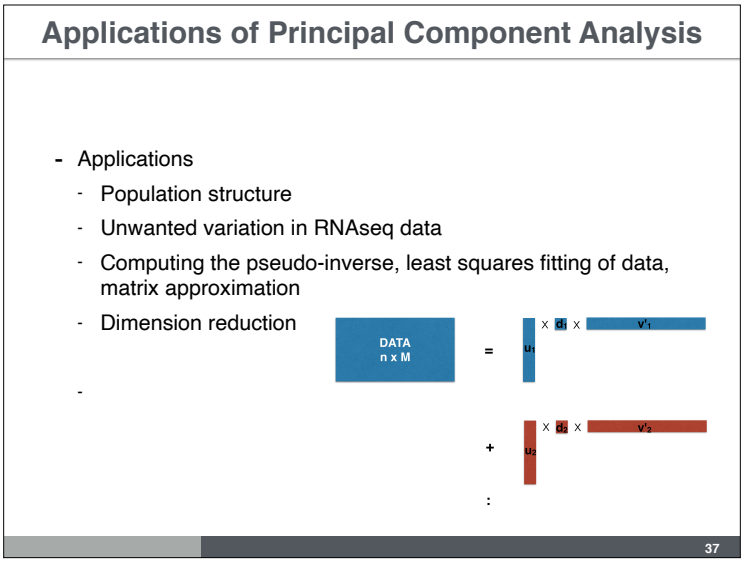
Geometric interpretation of SVD. U and V are vectors in a coordinate systems that are intrinsically attached to every data matrix. It is quite remarkable that any matrix can be decomposed into this product of three matrices, with the geometric interpretation that applying X to a vector is equivalent to applying a rotation (V*) followed by strecthing/shrinking, a final rotation (U).



This equivalent representation of the SVD makes evident the application to latent factor identification and dimension reduction.



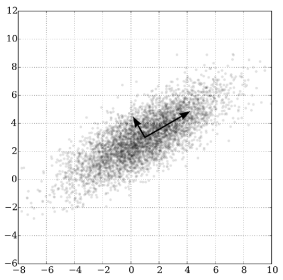
Connection between eigenvalue decomposition and singular value decomposition. R implements principal components analysis both ways. SVD based one is more numerically stable.



Principal component analysis can be applied in multiple cases. They can help us tease out population structure, find effects of unknown variation in expression or other omics data, it helps in reducing the dimension of the data. It is also very helpful in making computationally more stable algorithms.

PCA: Axes of Maximum Variation

- Eigenvalues are usually ordered by the value
- Top eigenvectors represent axes of maximum variation in the data

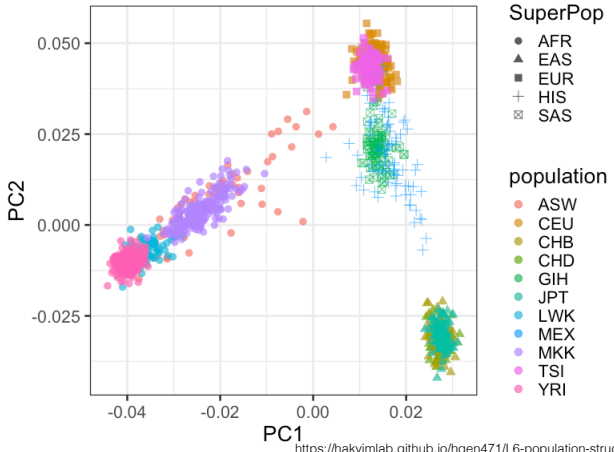


https://en.wikipedia.org/wiki/Principal_component_analysis#/media/File:GaussianScatterPCA.svg

38

In this two-dimensional dataset (in general we would be dealing with tens of thousands or millions of features) the axis rotated by 30 degrees is a good approximation to the data and represents the main axis of variation. Extend this idea when you have millions of axis and just a few were most of the action (variation) happens.

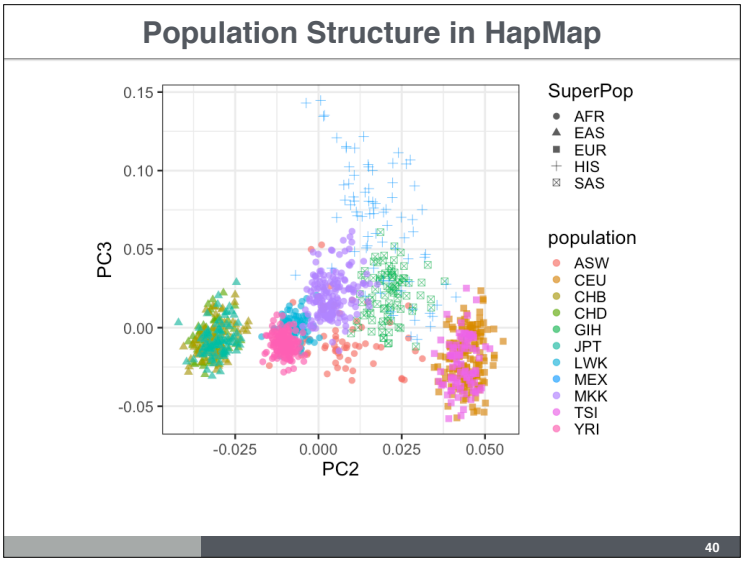
Population Structure in HapMap



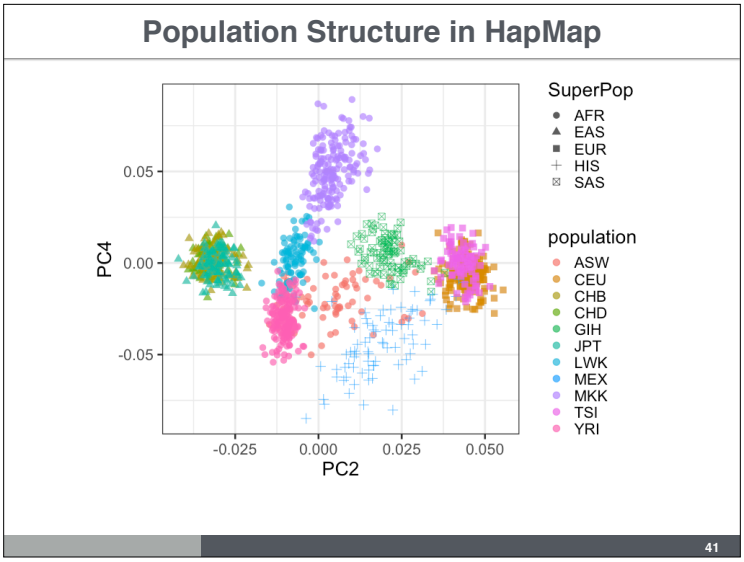
https://hakyimlab.github.io/hgen471/L6-population-structure.html

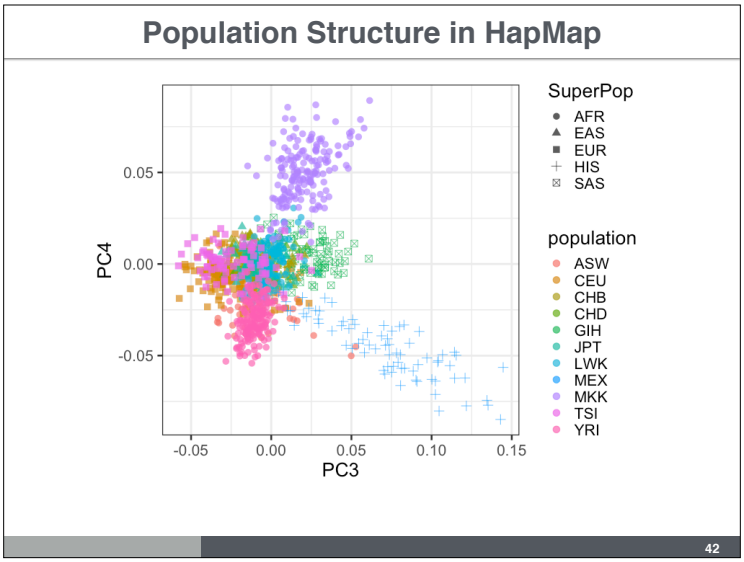
39

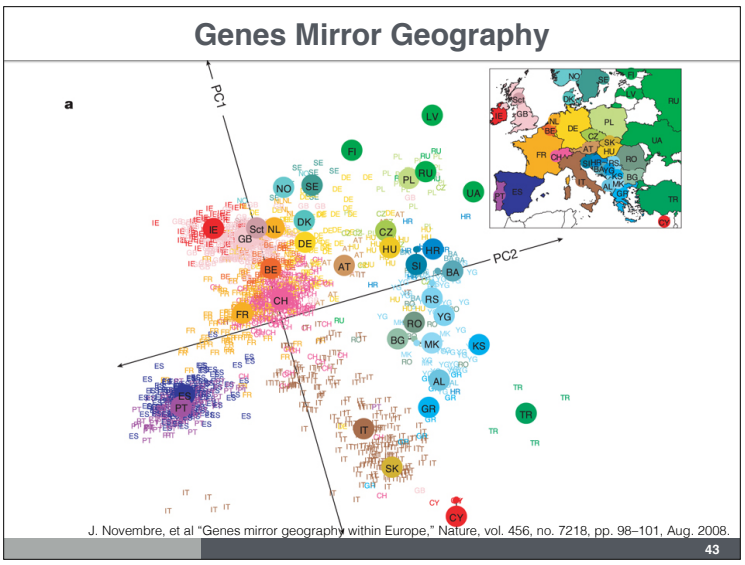
You can check out the linked html where the code is shown to generate the principal components of the genotype matrix from the HapMap project. Each population is represented with different colors. PC1 separates the African populations (circle) from the European (square), Asian (triangles), Hispanic (+), and South Asians (crossed square). PC2 separates the European and Asian populations further.



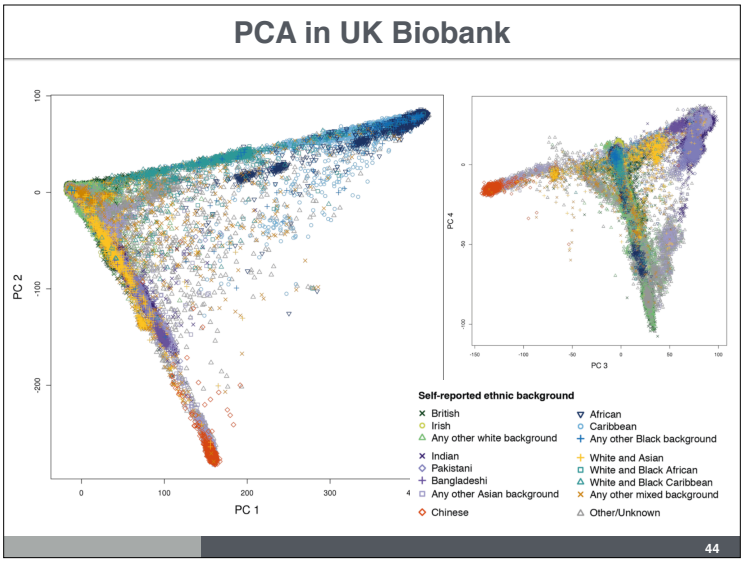
Subsequence PCs helps further distinguish different populations.







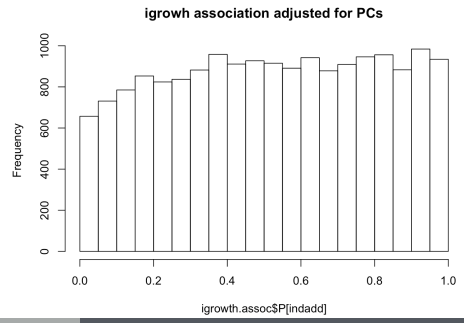
This is probably one of the most used figures in genetic talks. The first two principal components of a matrix 1500 European individuals genotypes are shown. When each individual is colored by the color of the country of origin, the map of Europe emerges. This is a remarkable consequence of the genetic similarity tied to geographic proximity.



Here are the two principal components of the UK Biobank genotype data. Special methods had to be developed to be able to compute these principal components given the sheer size of the data with 500,000 individuals.

Growth GWAS Adjusted with PCs

```
system(glue::glue("~/bin/plink --bfile {work.dir}hapmapch22 --linear --pheno {work.dir}igrowt
h.pheno --pheno-name growth --covar {work.dir}output/pca.eigenvec --covar-number 1-4 --maf 0.
05 --out {work.dir}output/igrowth-adjPC"))
igrowth.assoc = read.table(glue::glue("{work.dir}output/igrowth-adjPC.assoc.linear"),header
=T,as.is=T)
indadd = igrowth.assoc$TEST=="ADD"
titulo = "igrowh association adjusted for PCs"
hist(igrowth.assoc$P[indadd],main=titulo)
```



Here we adjust the regression adding the first 4 principal components as covariates. How many to use depends on the application, sample size, etc. In UK Biobank some people recommend using 14 but there is no consensus. It's always recommended to use sensitivity analysis. Try different numbers of PC's. Don't do p-hacking, that is do not choose the number of PC's that give you significant association for a cherry picked variant.