

Genomic Tools in Research



Wei Cheng

What I am going to blabla:

- Population Analysis
- Linkage Analysis vs Association Analysis
- GWAS

**P
O
P
U
L
T
I
O
N

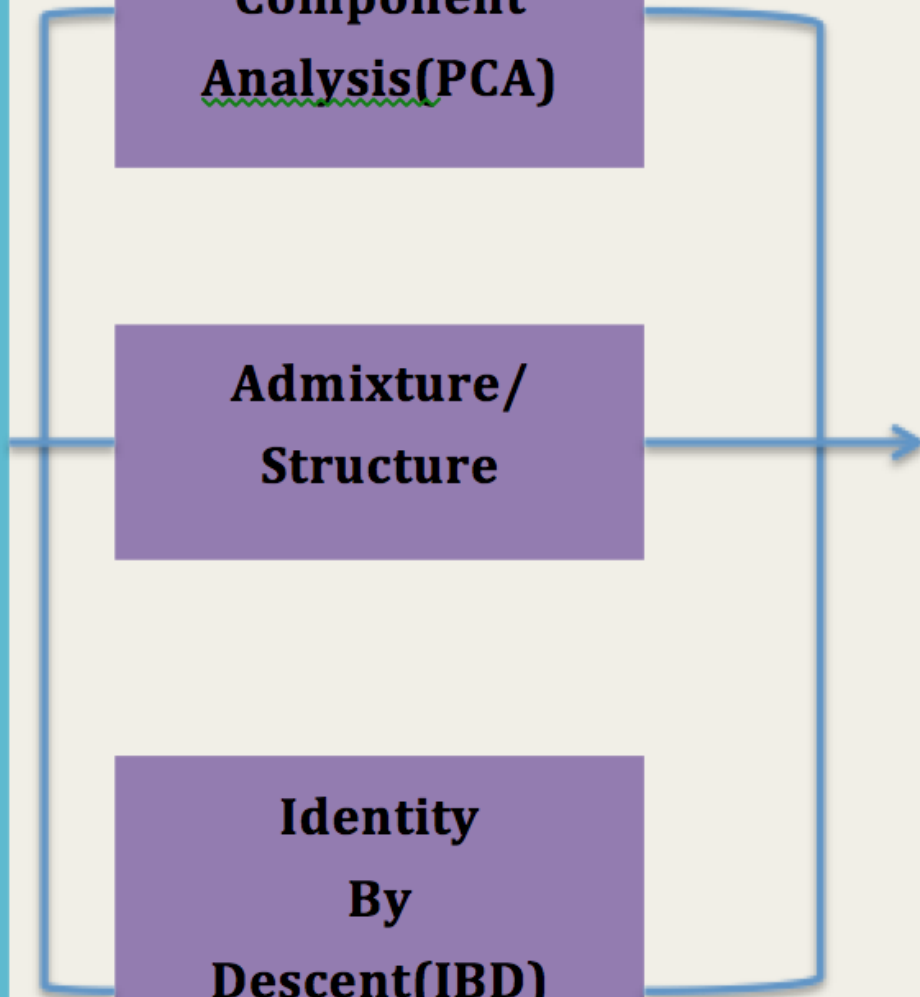
A
N
A
L
Y
S
I
S**

**Principle
Component
Analysis(PCA)**

**Admixture/
Structure**

**Identity
By
Descent(IBD)**

**Population
Stratification
(Relatedness of
Individuals)**



Principle Component Analysis

Definition: Statistical analysis that uses an orthogonal transformation to convert possibly correlated variables into (linear uncorrelated) numerical values called Principle Components.

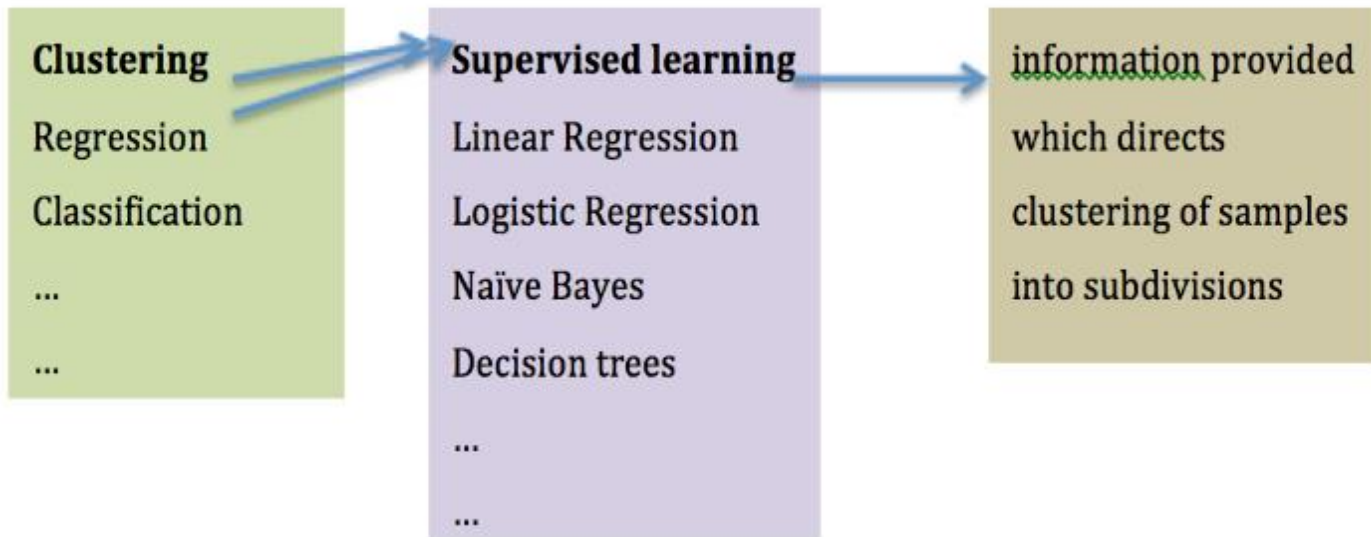
Admixture/Structural/Ancestry Analysis

A method of inferring someone's geographical origins based on an analysis of their genetic ancestry (One of the components of an autosomal DNA test). Admixture calculations offer genetic ancestry analysis for individuals tested for high-density SNP data. It always build ancestry components called cluster.

What is Clustering?

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).

(Machine Learning and Data Mining)



Identity by Descent & Identity by State

Definition: The phenomenon of that two or more than two individuals who share similar nucleotide sequences is identical by state, meanwhile, if they inherit the similar nucleotide sequences from a common ancestor, IBS is identical by descent (IBD).

If this is not clear...

Here is an ancient legend
version...

Suppose you and me have the same mutation, then we are IBS...

if you and me are related, in other words, we share a same ancestor, then we are IBD...

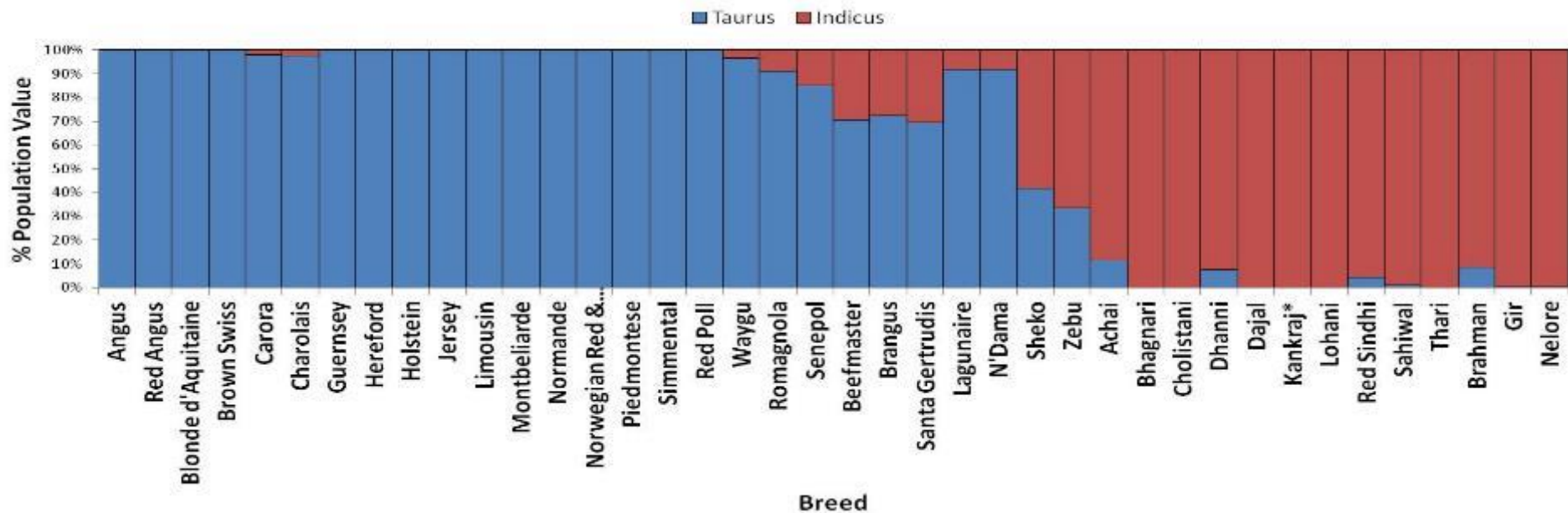
---professor Heather Jay Huson

Exact time: Not clear

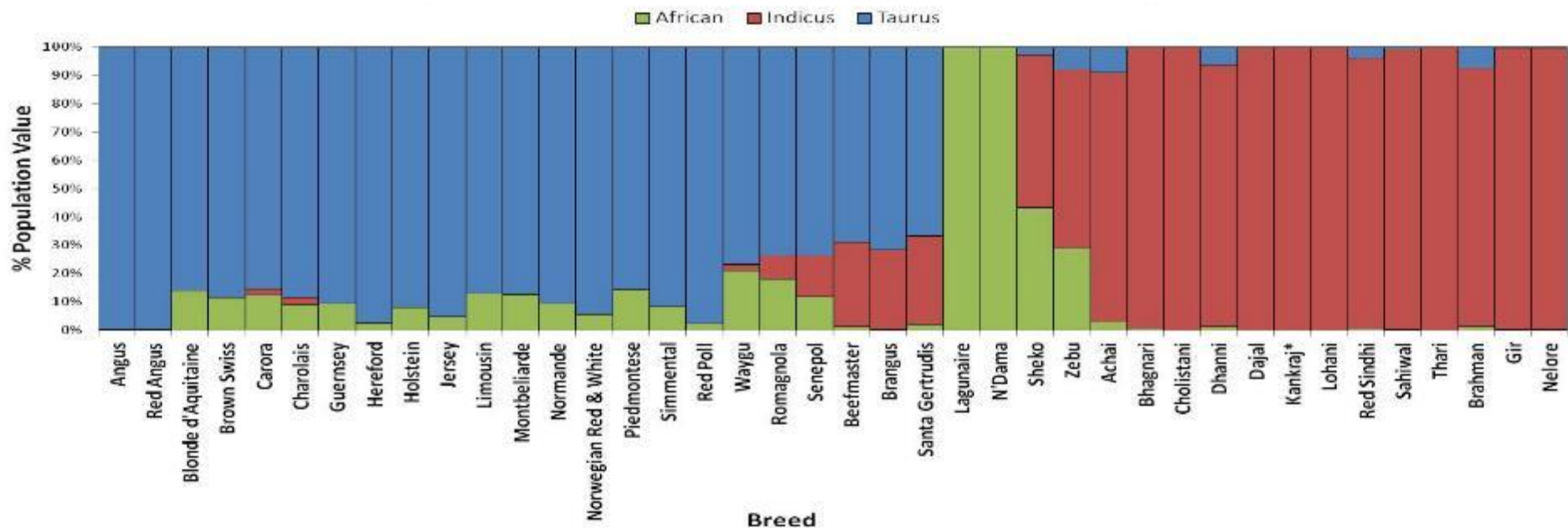
These definitions give you a sense that all of these analysis focus on the relationships of individuals within the whole population.

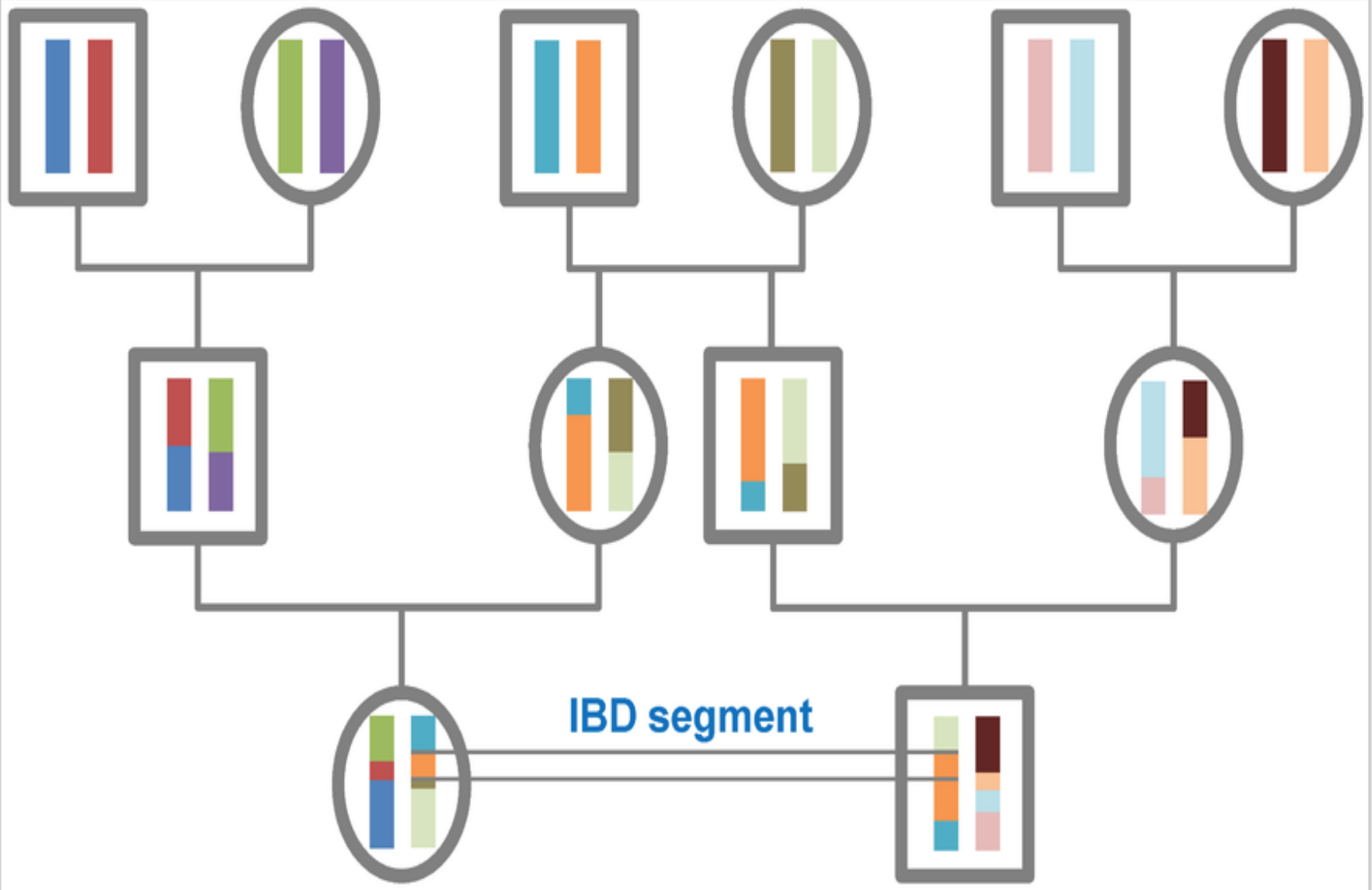
If it's still not clear, let's see some pictures...

Unsupervised Population Clustering of 40 Cattle Breeds; K=2



Unsupervised Population Clustering of 40 Cattle Breeds; K=3





The origin of IBD segments is depicted via a pedigree.



After seeing these
pictures, you guys
should be experts now.
Let's move on...

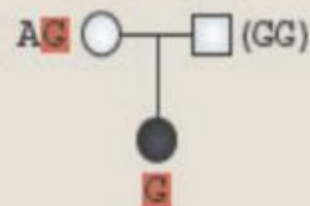
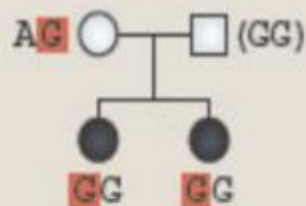
Why bother with population analysis?

Problems such as population Stratification cause errors for linkage analysis and association analysis.

Linkage Analysis vs Association Analysis

Difference:

Linkage is actually looking at physical segments of the genome that are associated with given traits. Association studies go from the other direction, saying, 'given different pieces of the genome, can we then look for different traits that are associated with those different segments of genome?'



Property of mapping approach	Linkage analysis	Association analysis
Data type studied	Relatives	Unrelated or related individuals
Relevant parameter	Recombination fraction	Association statistic
Range of effect detected (linkage or association)	Long (≤ 5 Mb)	Short (≤ 100 kb)
Number of markers required for genome-wide coverage	Moderate (500–1,000)	Large ($> 100,000$)
Statistics used	Cumbersome (requires tailor-made likelihood methods)	Elegant; can use the range of classical statistical tools
Dealing with correlated markers	Pose problems in presence of ungenotyped individuals	Can be handled efficiently
Biological basis of approach	Observe (or infer) recombination in pedigree data	Exploit unobserved recombination events in past generations
Dealing with allelic heterogeneity	Not a problem	Reduces power
Detecting genotyping errors	Potentially detected as Mendelian inconsistencies	Potentially detected only in family data, but not in case–control data
Most suitable application	Rare, dominant traits	Common traits

In the past,
neither of them
were genome
wide because
there wasn't a
technically
feasible or
affordable way to
test the whole
genome at
once...until....



...the invention of the “Chip!”

Not this chip





Now, the most popular way to perform those analysis is by using SNP chips that measure hundreds of thousands of loci spread across the whole genome, thus the name GWAS.

Two most famous companies are producing chips for genotyping

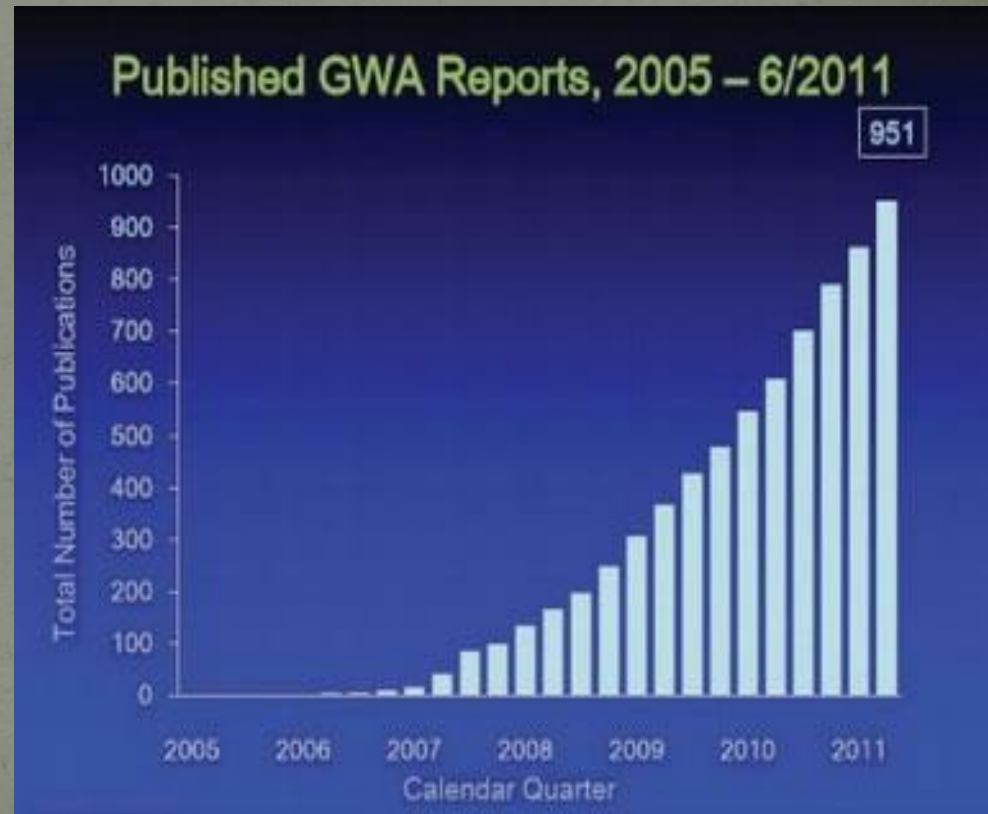


Genome wide associate study(GWAS)

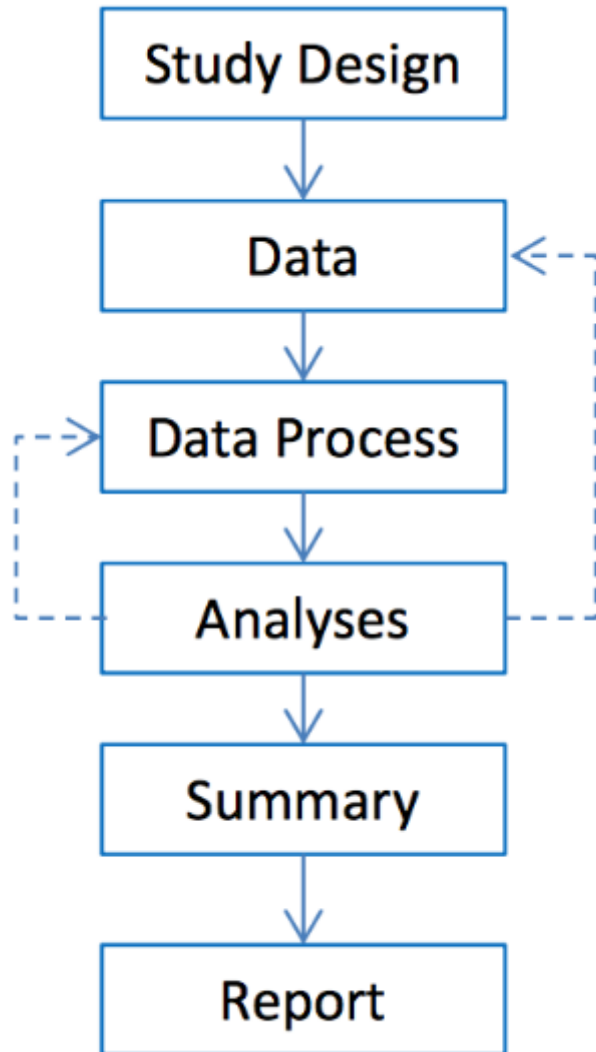
Definition: GWAS is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait

Development of GWAS:

In 2005, Science reported the first successful GWAS which is about investigated patients with age-related molecular degenerations. Then, research of GWAS about obesity, blood pressure, diabetes...were reported...

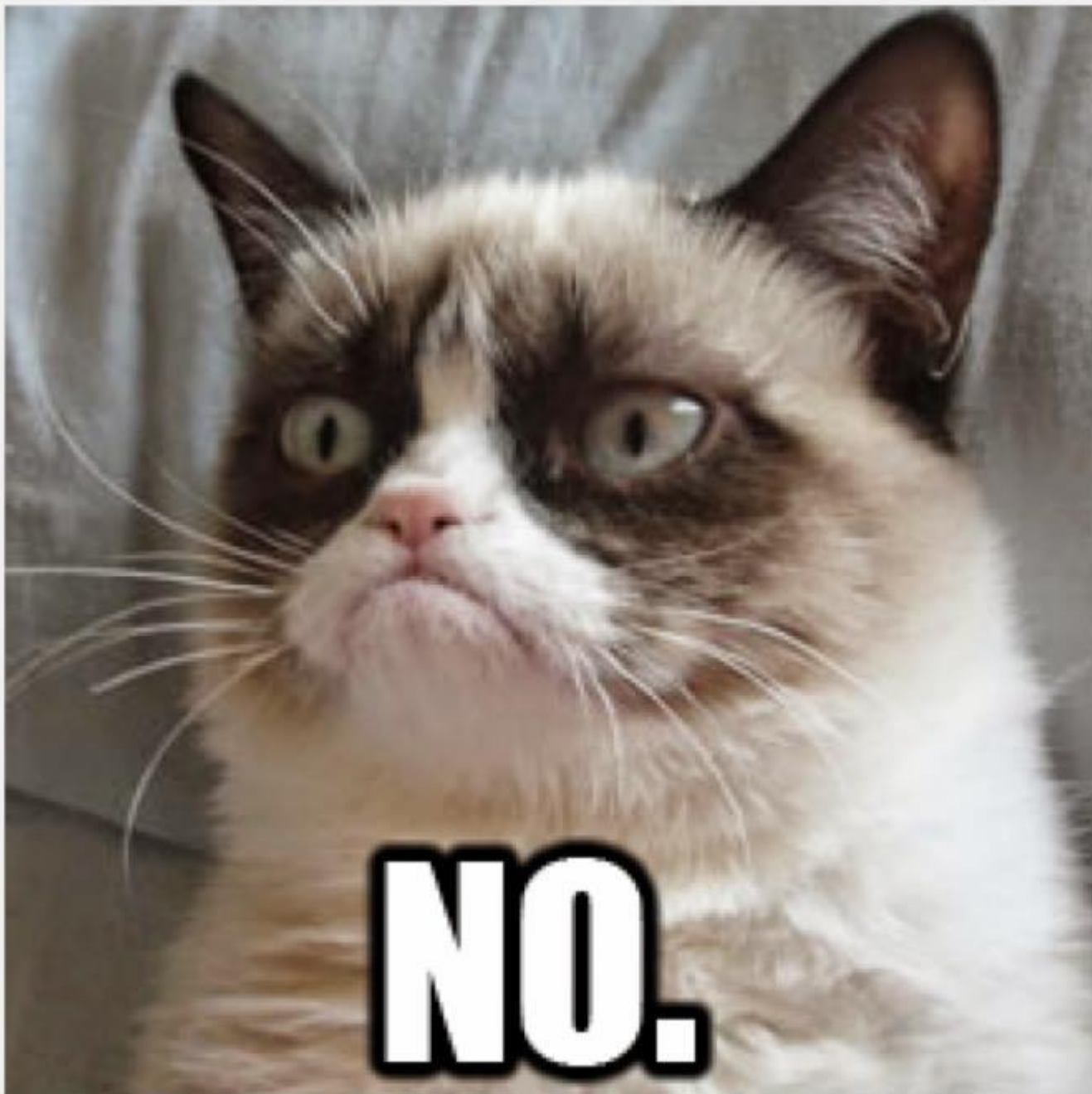


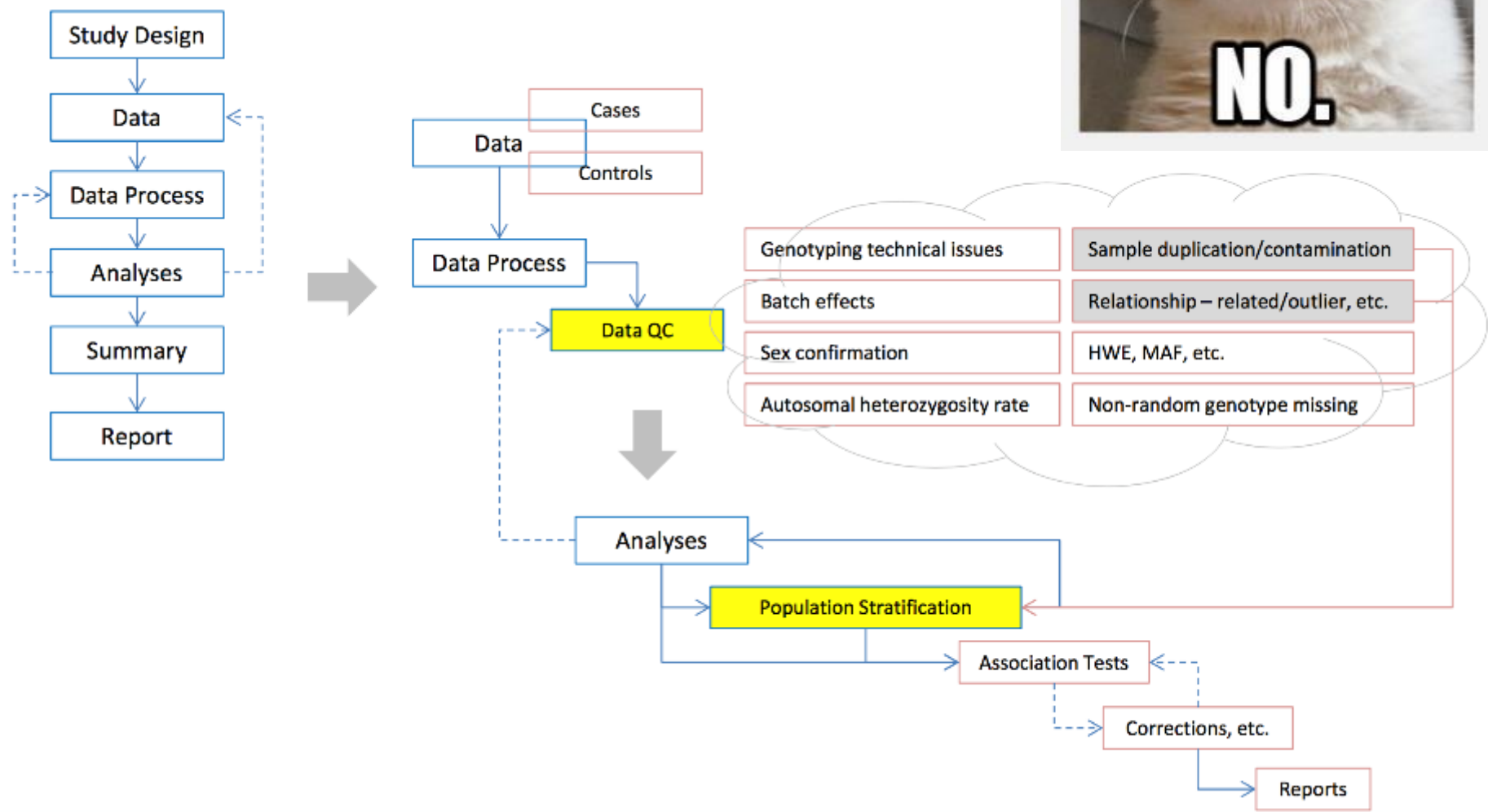
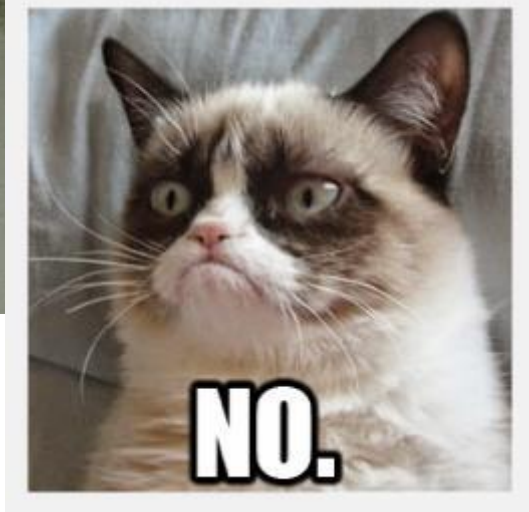
Basic Steps

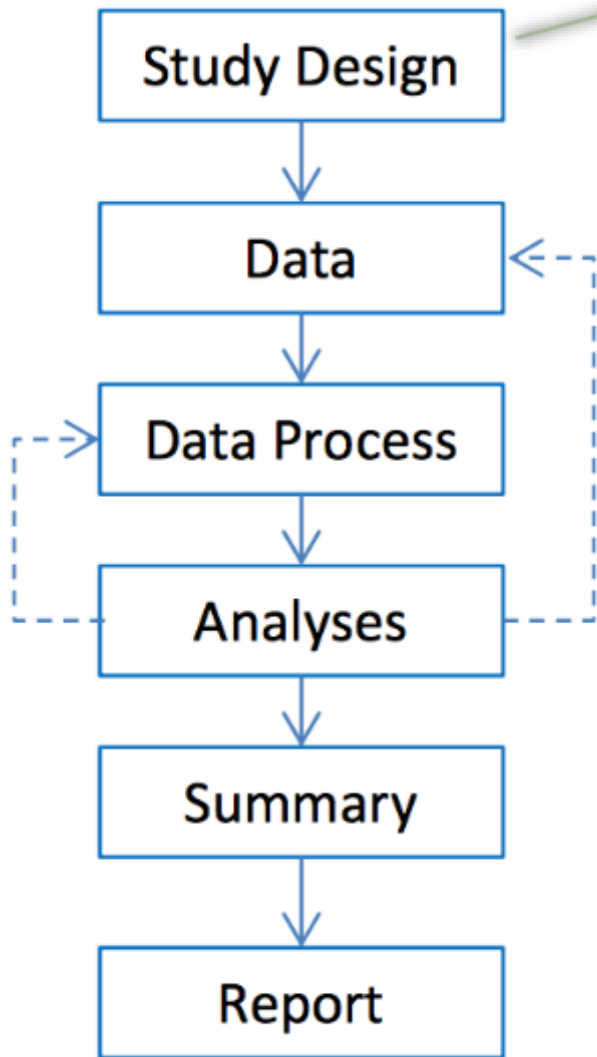


Six Steps, upmost 5 minutes



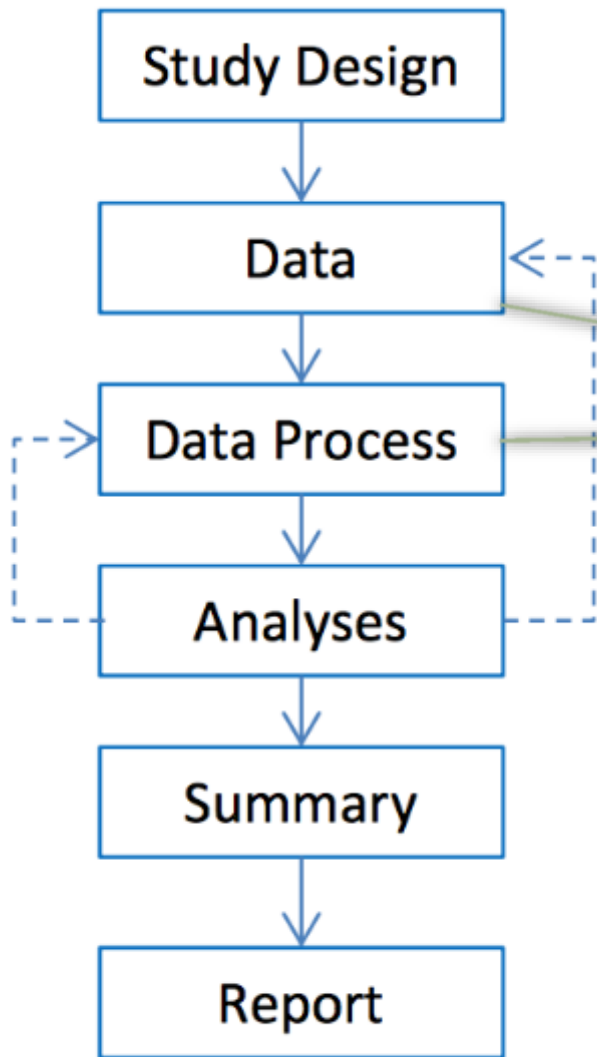






Based on “stage”,
One-stage
design & Two-
stage design or
Multiple-stage
design.

In One-stage design select a large enough samples once, and genotype all of them. In two-stage design or multiple-stage design, we usually select a small sample group for genotyping. Then, we select the SNPs which are obvious significant correlated to the target traits under a loose condition of P-value. After that, we choose the selected SNPs in bigger samples and genotyping. At last, we combine the results of two stages and do the statistics.

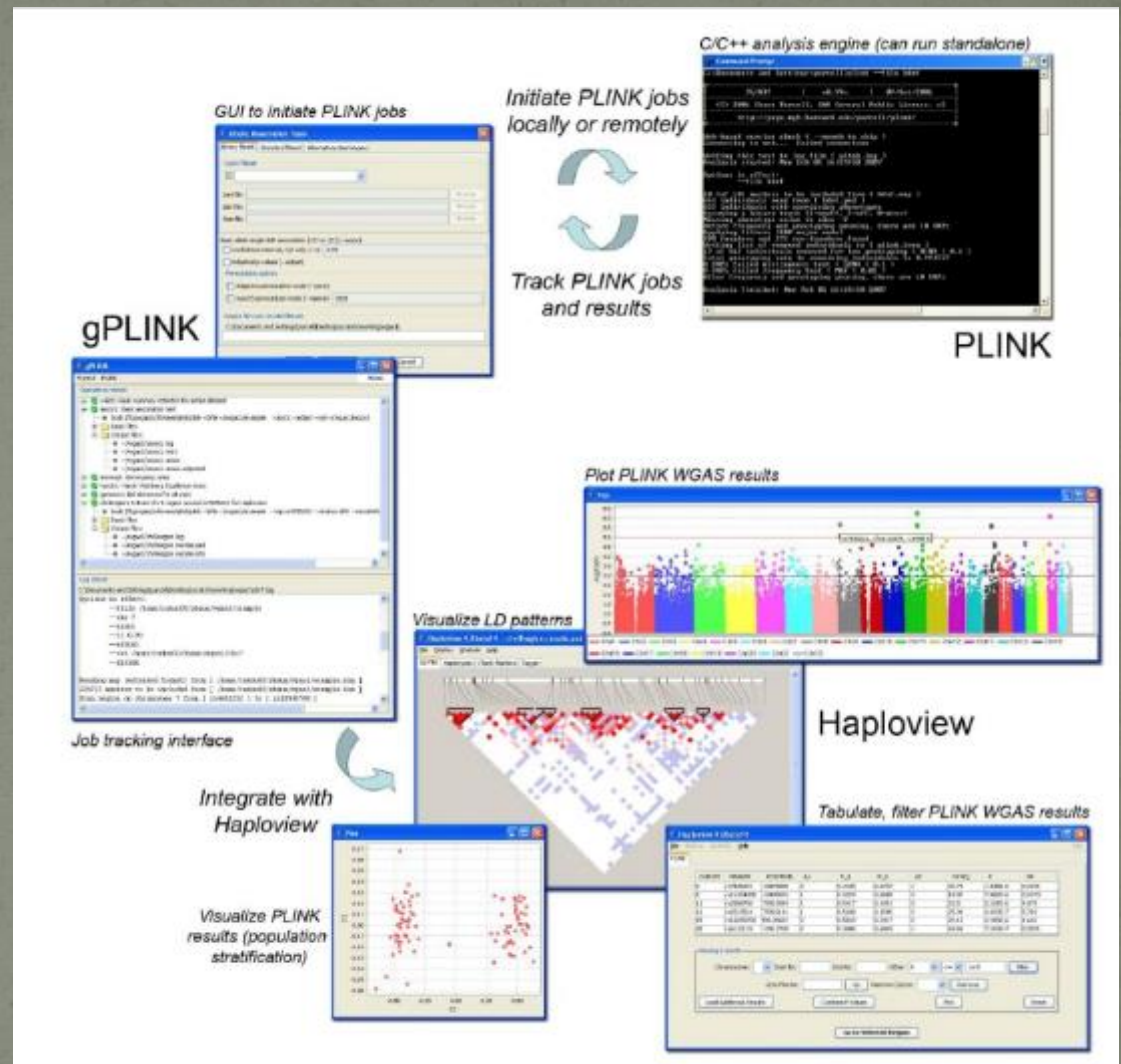


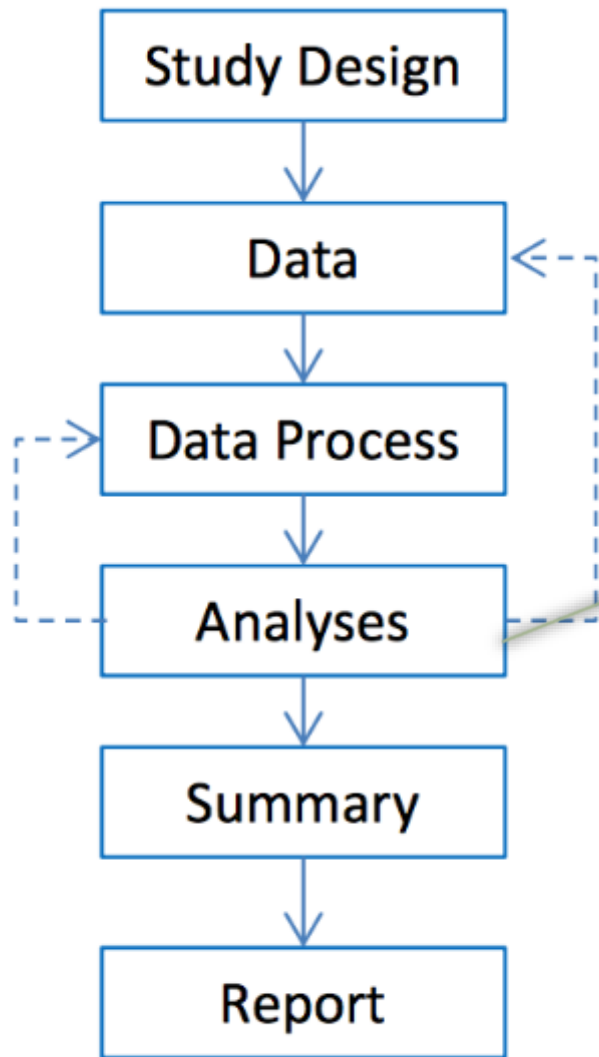
DNA extraction and isolation, genotyping and data review to ensure the high genotyping quality

After genotyping, ----quality control

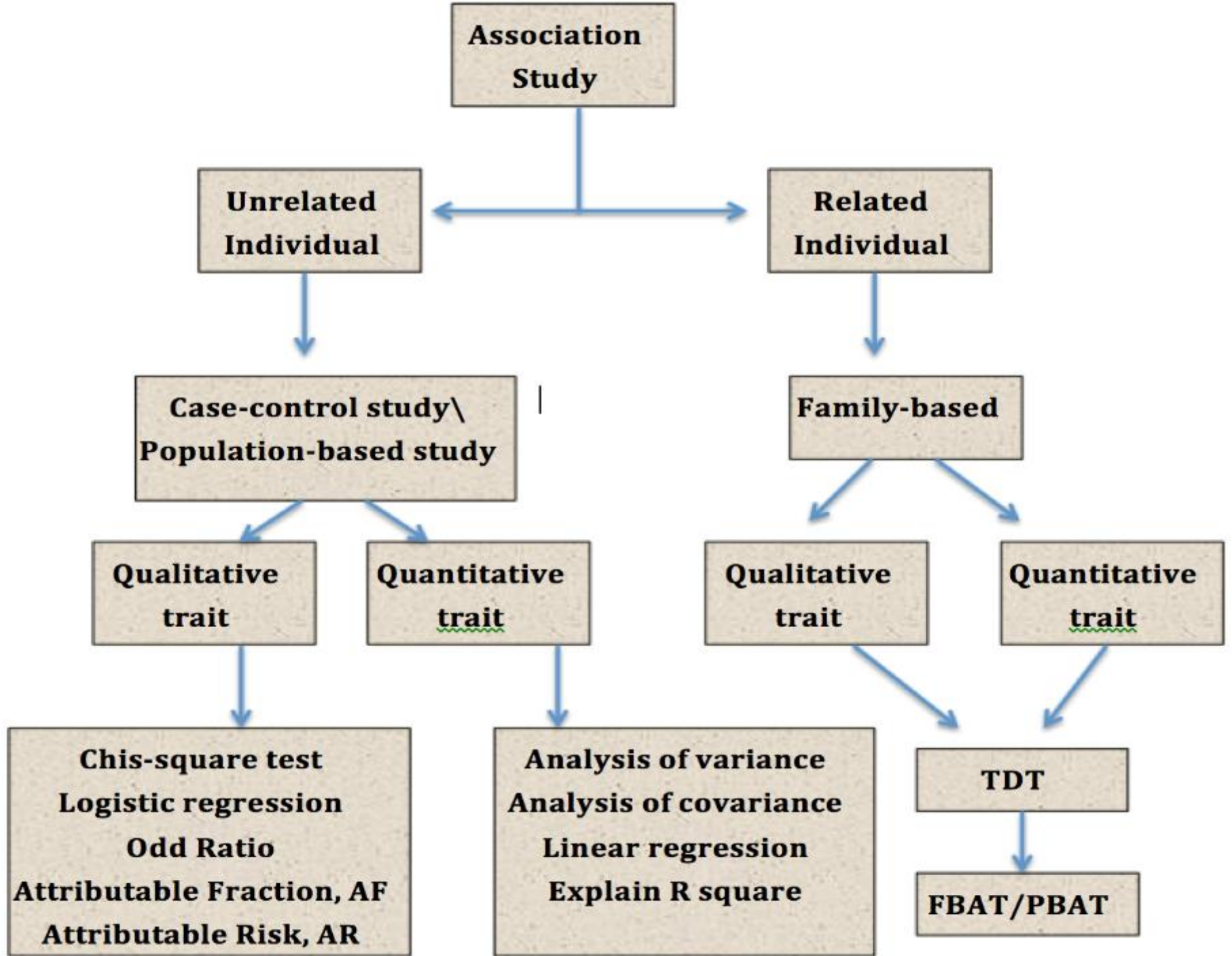
1. Quality control for sampling (Kolmogorov-Smirnov method used to test for normal distribution of data)
2. Genotyping control is basically some primary arrange and analysis for some genotypes after the distinguish.

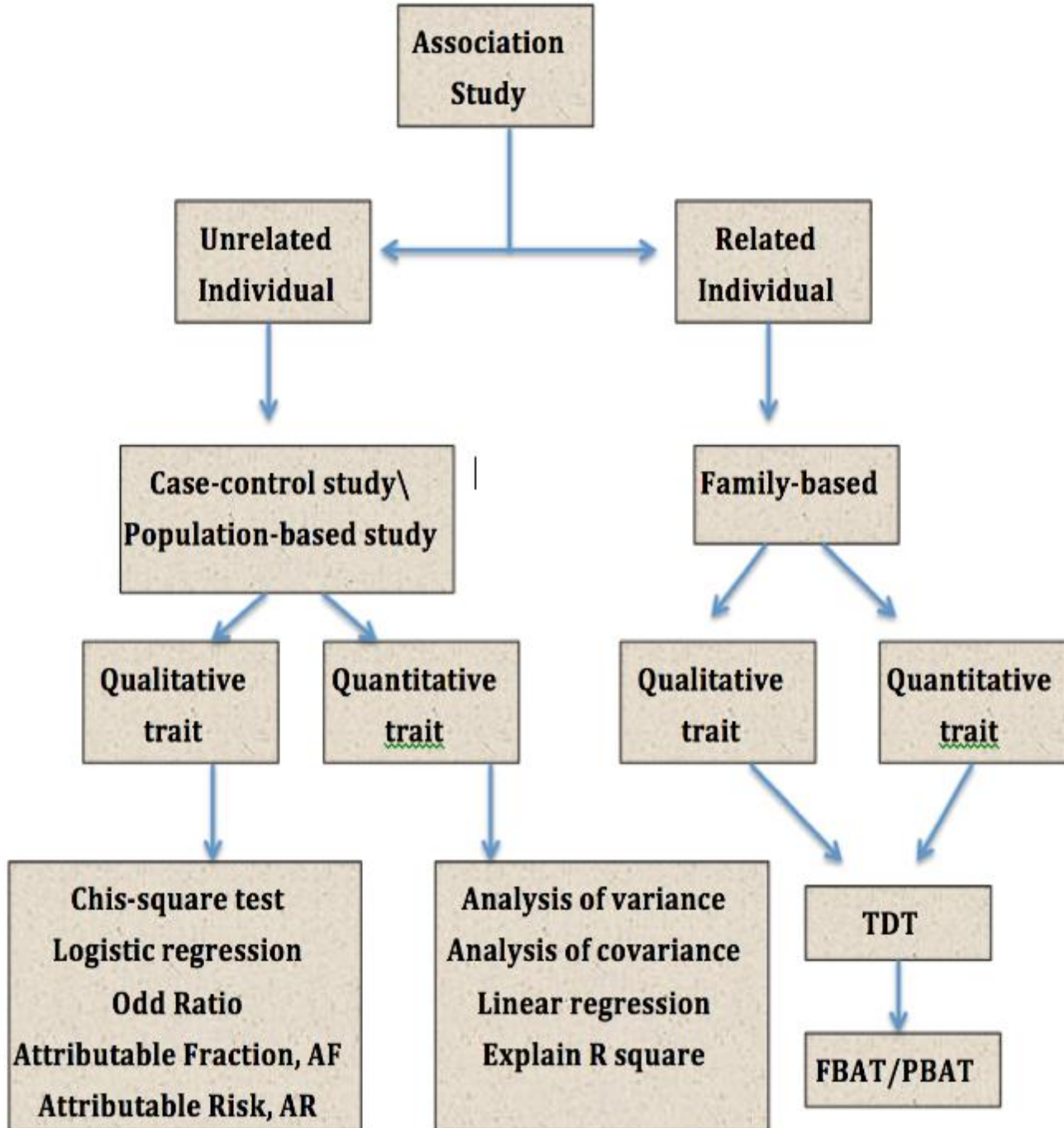
All of them
could be
done in
PLINK!
Check on
our blog





Choose appropriate statistical tests for analysis the relationship between the SNPs and the disease/trait



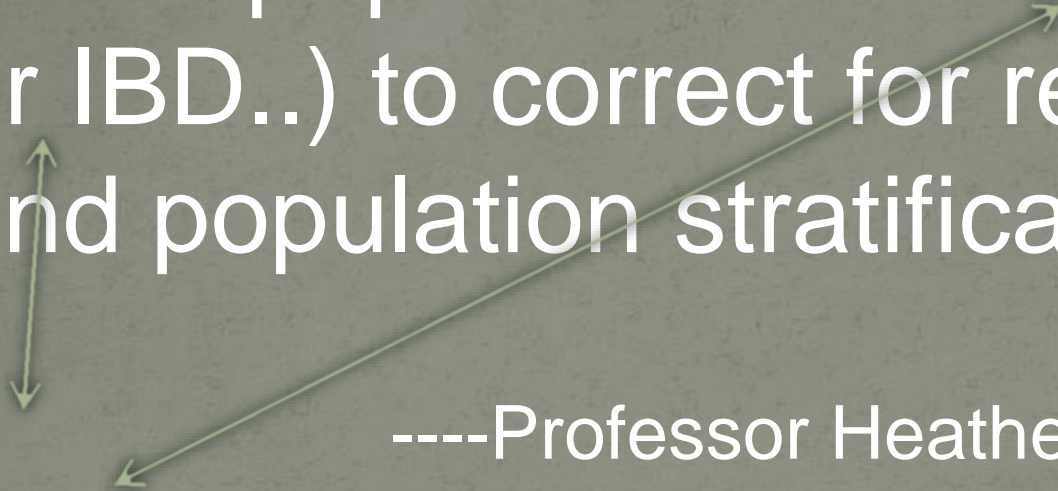


Logistic regression model for qualitative traits.
 Linear regression models for quantitative traits.
 General linear model and mixed linear model, later is better

After seeing this, I need to tell
you another legend story....



When the population is not randomly mating or unrelated, utilize population information (PCA or IBD..) to correct for relatedness and population stratification



----Professor Heather Jay Huson

Population analysis

time: Probably Sunday afternoon

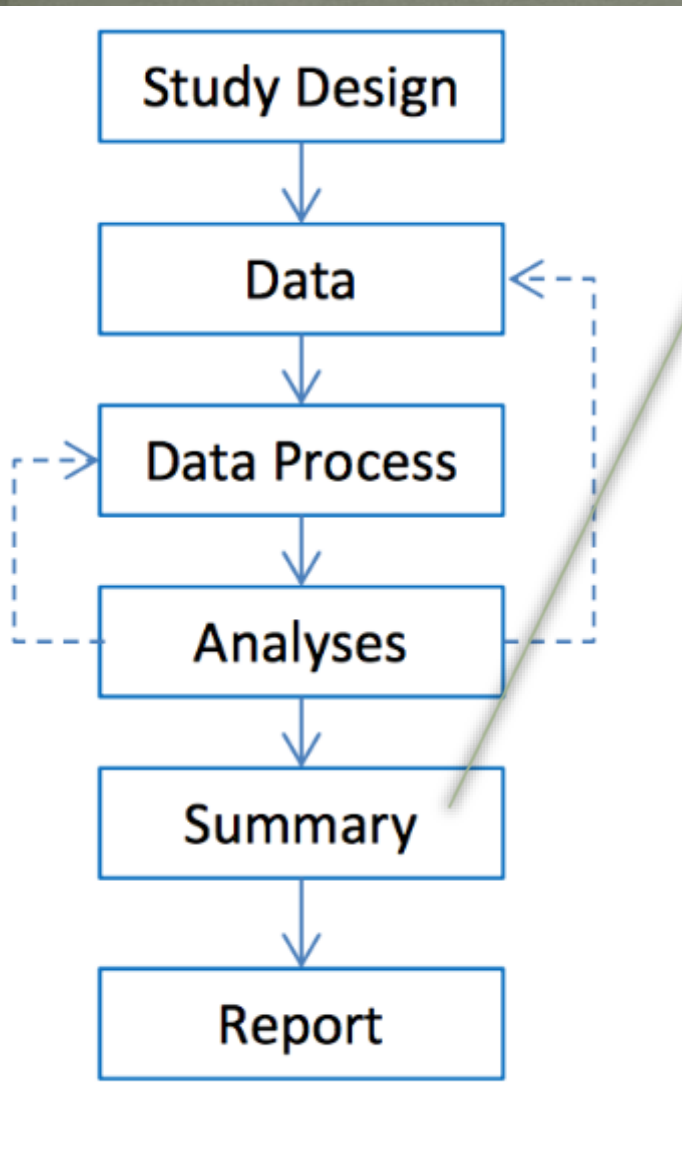
When the population is not randomly mating or unrelated, utilize population information (PCA or IBD..) to correct for relatedness and population stratification



----Professor Heather Jay Huson

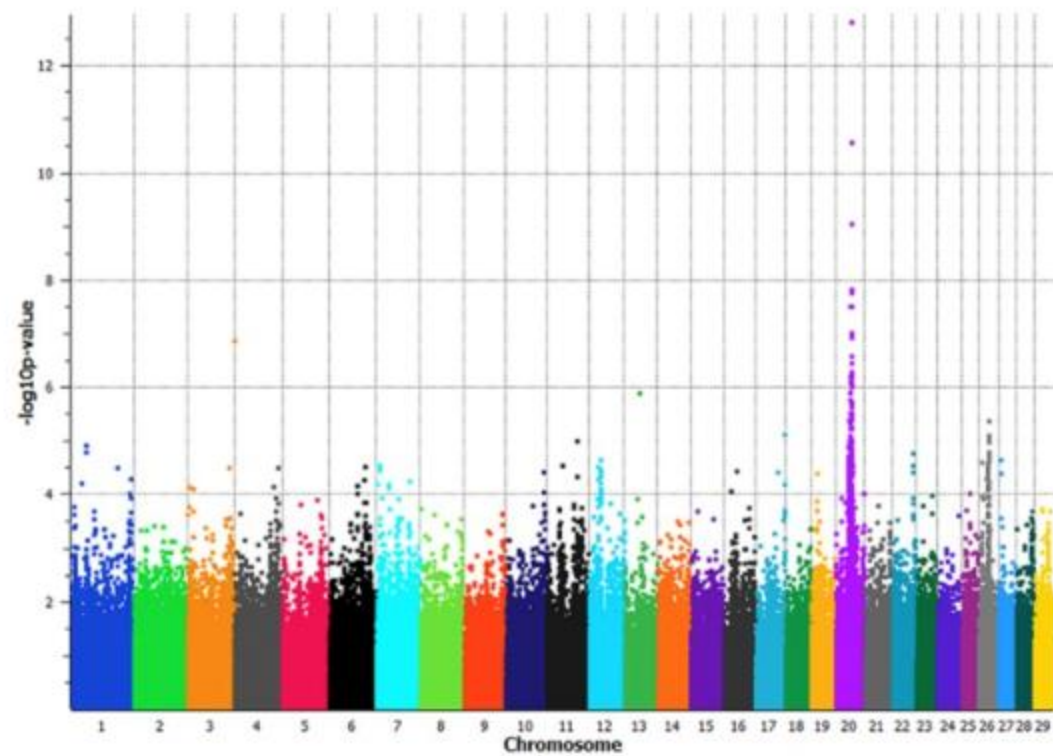
EMMAX
SOFTWARE &
EMMA
ALGORITHM

time: Probably Sunday afternoon

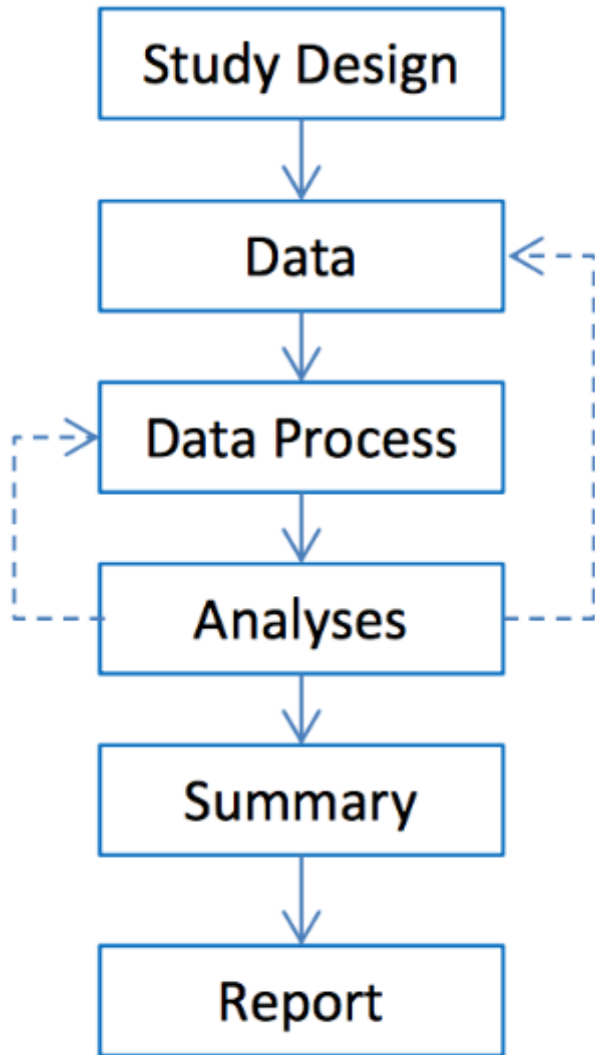


Results and Manhattan plot: Manhattan plot is a scatter plot and is always used to show data with a large number scale project, especially popular for GWAS. More specifically, a GWAS Manhattan plot, genomic coordinates is X-axis, with negative logarithm of the association P-value for each SNP on Y-axis. Thus, a dot on the plot means a SNP.

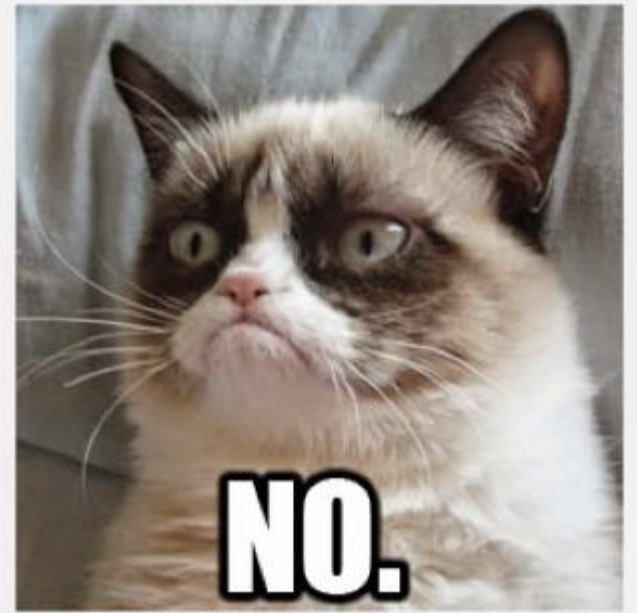
Figure 2



Time to report



One last thing: Replication
of identified associations
in an independent
population samples or
examination of functional
implications
experimentally



All resources on
website, backup
knowledge...



Questions?