Genome-Wide Association Studies (GWAS) – #1

What is a GWAS?

A genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as height, blood pressure or weight), or the presence or absence of a disease or condition (such as cancer, diabetes or schizophrenia).

GWAS goals

Overall Goal: To identify correlations between a phenotype and whole genome genotypes

Some specific goals:

1. Identify statistical connections between individual genetic variants (or regions) in the genome and the phenotype
   - Drive biological hypothesis generation

2. Generate insights on genetic architecture of phenotype
   - Many genes of small effect?
   - Regions of large effect?

3. Build statistical models to predict phenotype from genotype
   - Predict disease risk from an individual's genome

GWAS Methodology

1. Collect large samples of individuals (n typically in the 10s or 100s of thousands)

2. Measure each individuals for genetic variation throughout the genome

   - Single Nucleotide Polymorphism (SNP)
   - Typicically *m* = 500,000+ SNPS
   - Or whole genome sequencing

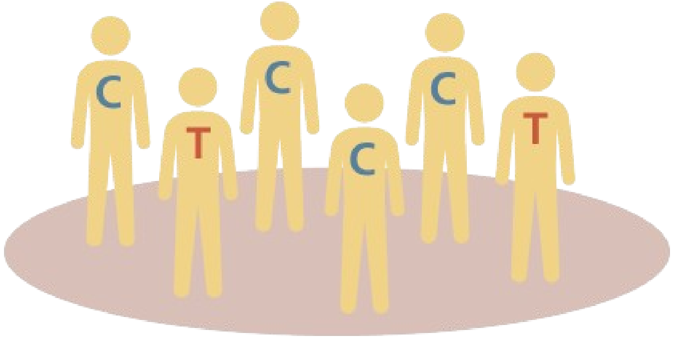
3. Now we can think of our data as X_{n*m} matrix with subjects as rows, SNPs as columns,

   - X_{ij} is in {0,1,2} (genotype at single locus)
   - Also given extra vector Y_n of phenotypes

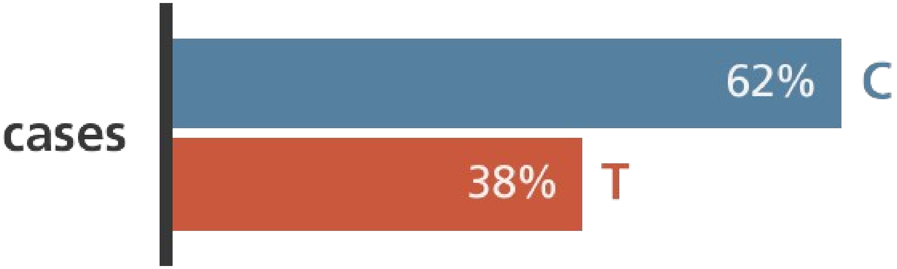
4. Association testing

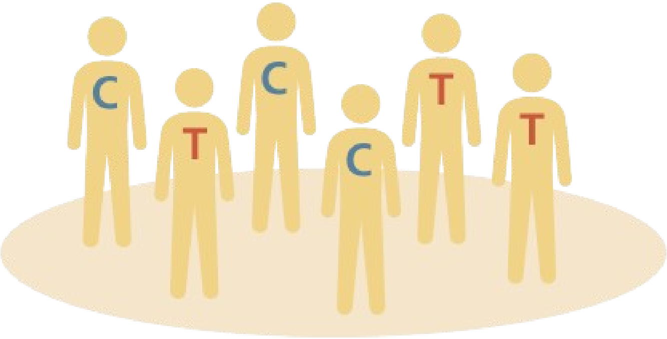
   - Find SNPs (columns in X) that are statistically associated with Y
   - Can be thought of as m separate statistical tests run on this matrix

Single Genotype Example

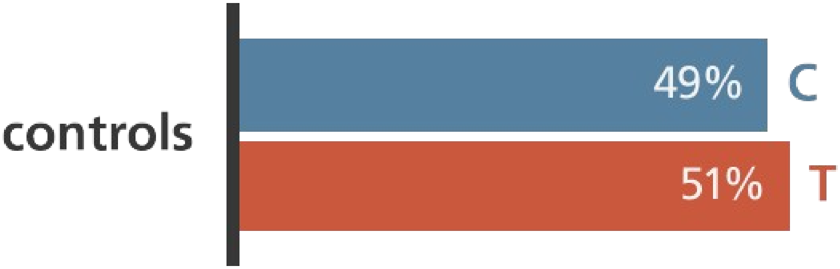


cases (n=1,000)
people with heart disease





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



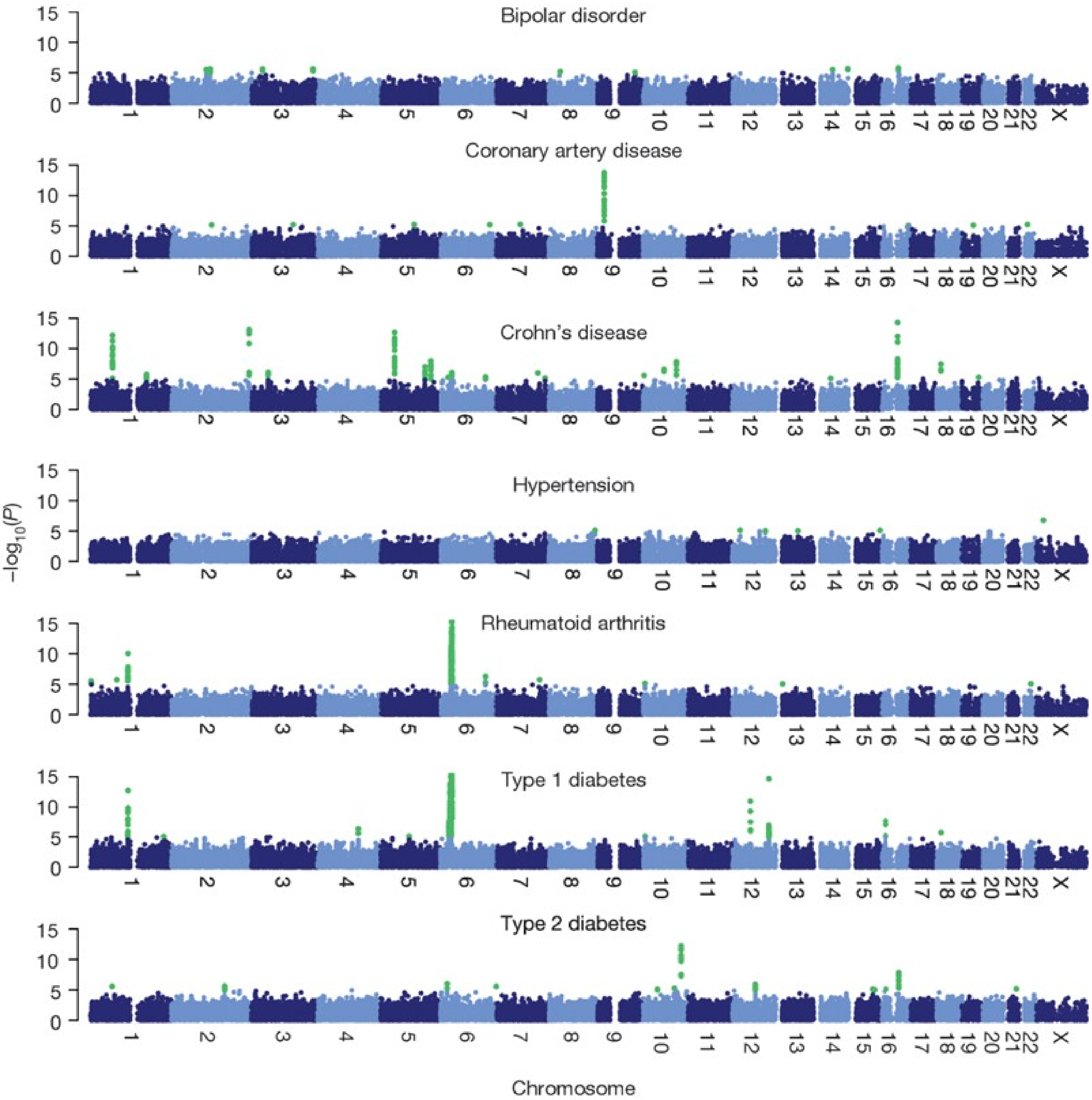
WTCCC

First large
Scale GWAS

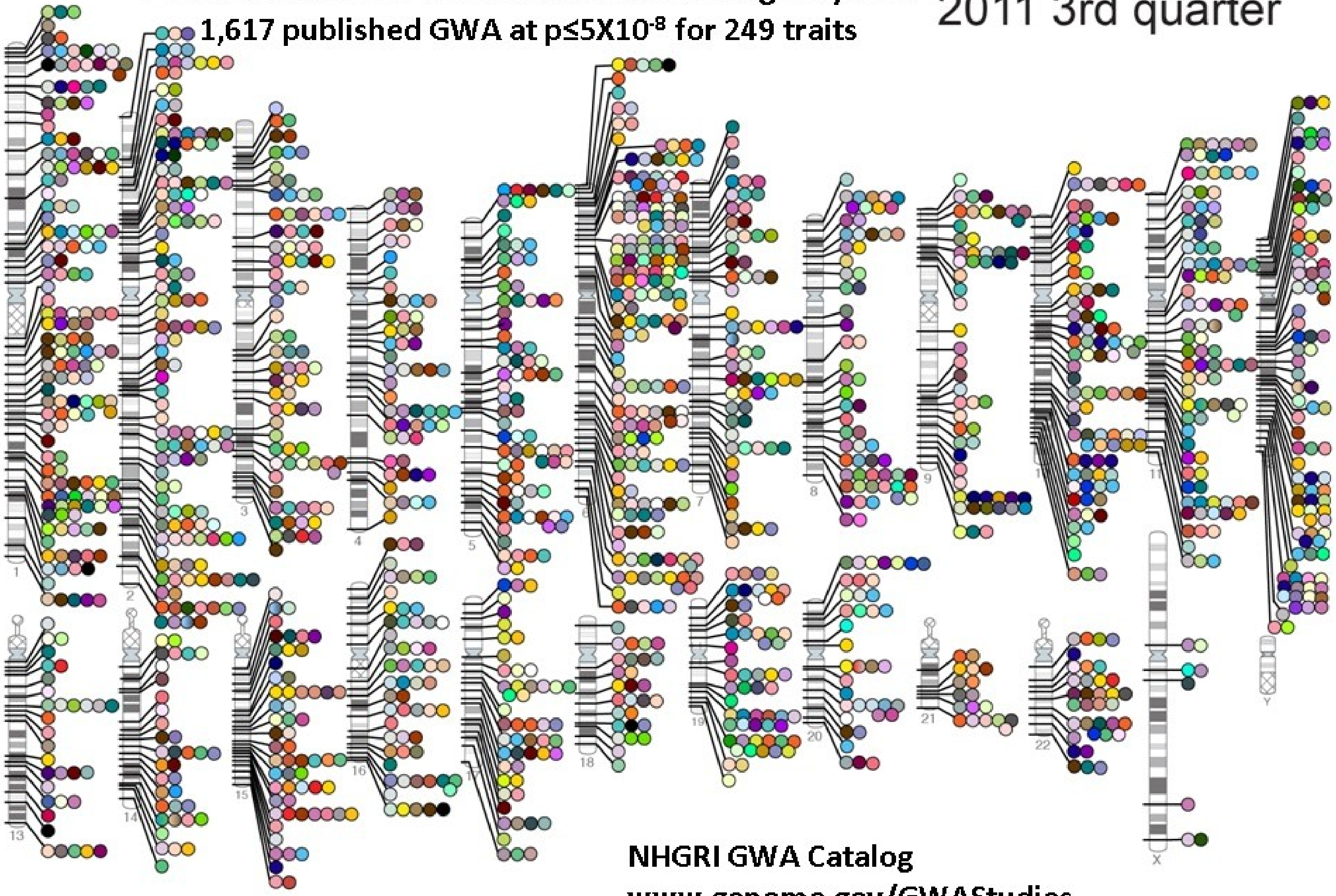
14,000
cases over
7 diseases

3,000 shared
controls

(Nature, 2007)



GWAS Success



Published Genome-Wide Associations through 09/2011

1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

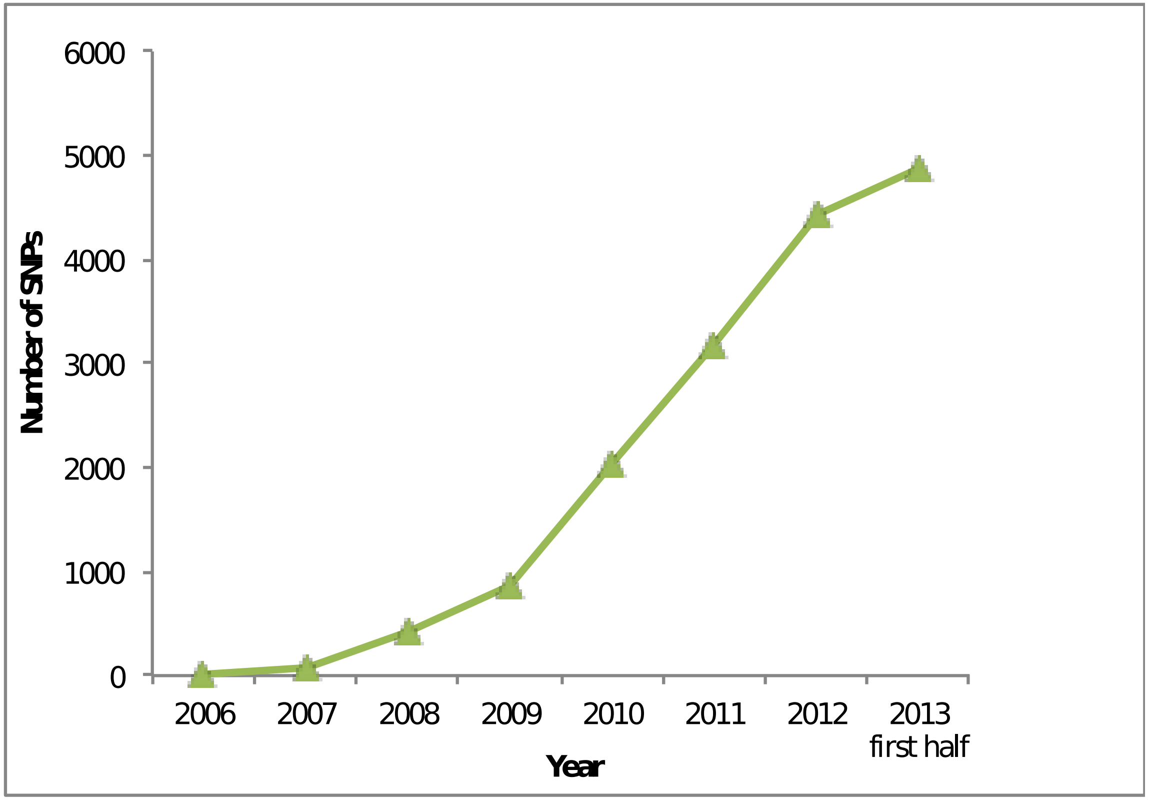
2011 3rd quarter

NHGRI GWA Catalog
www.genome.gov/GWAStudies

GWAS Success

Only 8 genes known for human complex traits until 2002

~5000 genetic variants associated with ~650 traits or diseases by early 2013



That sounds easy…

… why do we have four lectures on this topic?

LOTS can go wrong!

A GWAS performs 100s of thousands or millions of statistical tests and takes the most significant results

Any deviation from underlying assumptions can results in a many false positive results

Most of the time in GWAS is spent in preparing the data to avoid this pitfall

Background Info: Genotyping

An understanding of current genotyping methods in needed to identify problems in the genotype data

Two methods are currently used for genotyping:

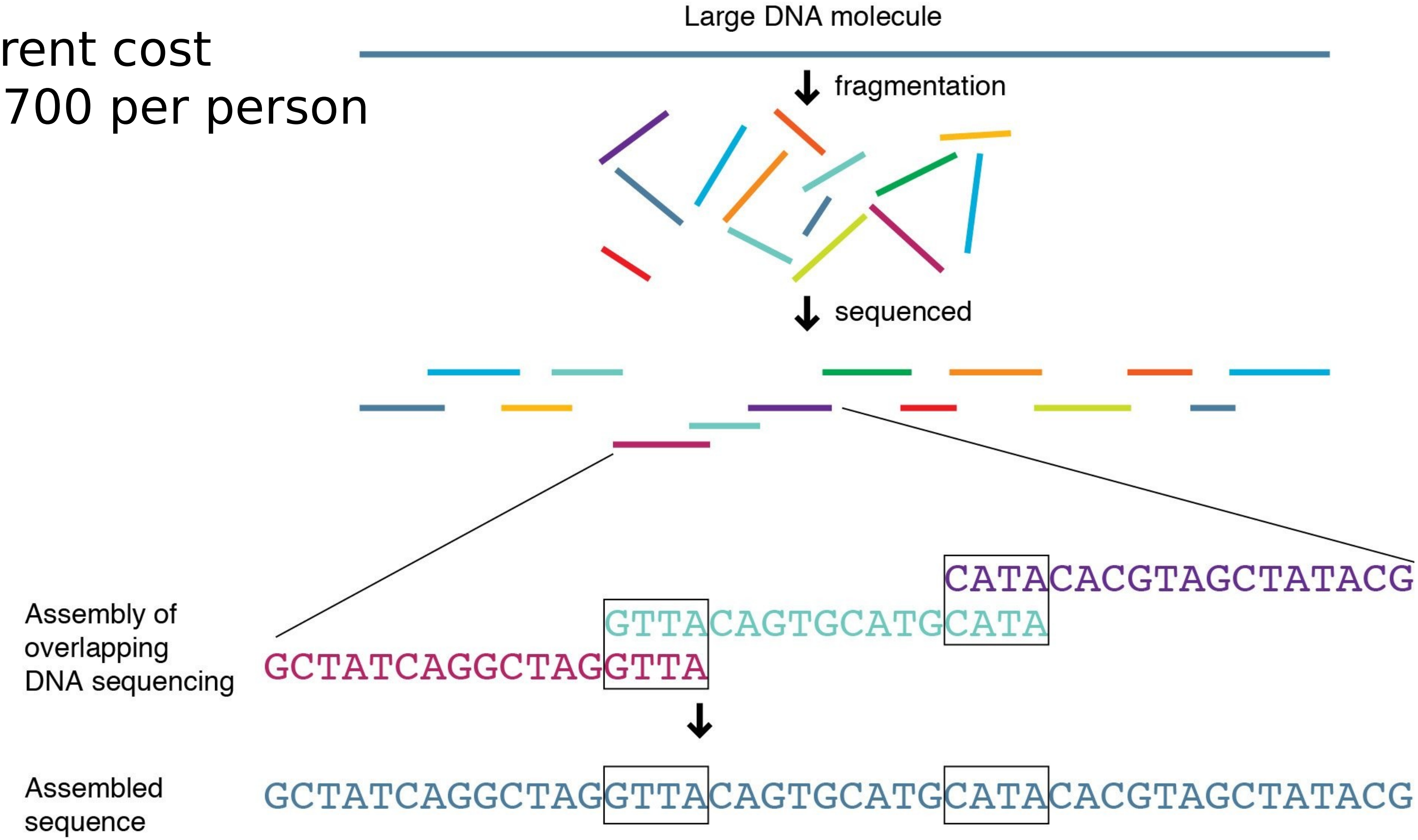
1. Whole Genome Sequencing (WGS)

2. SNP arrays

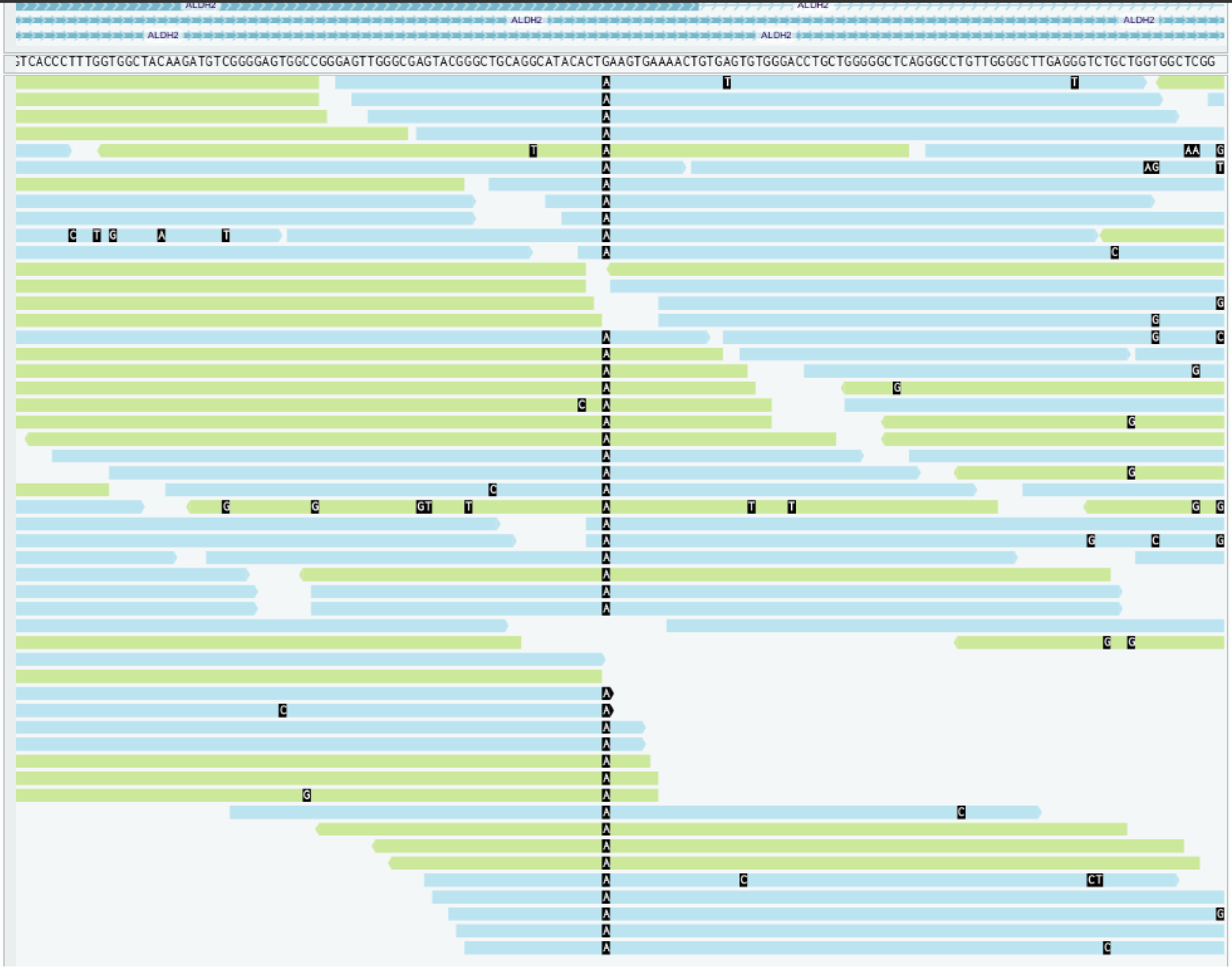
Whole Genome Sequencing

Identifies “all” genetic variation in an individual

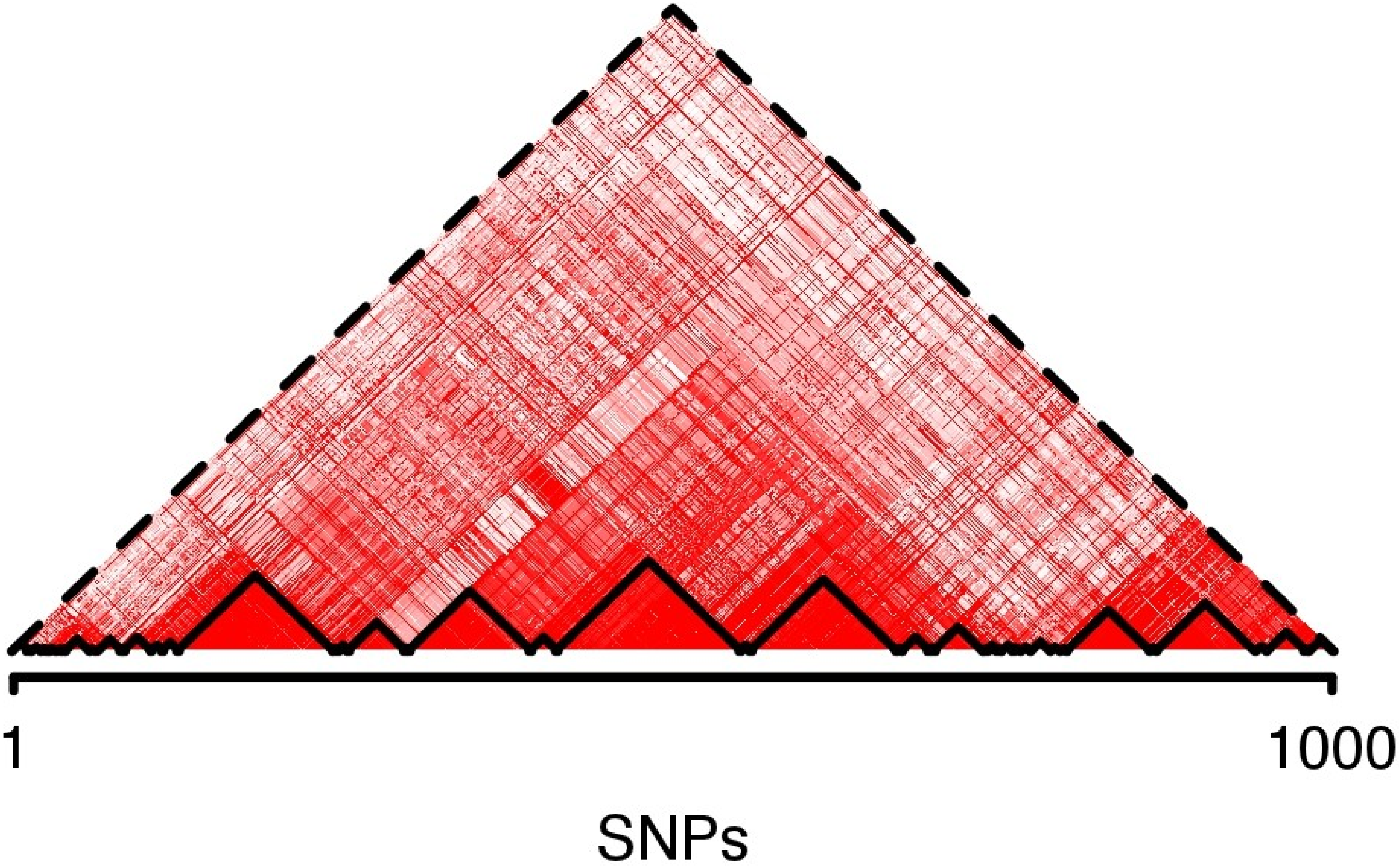
Current cost
~1,700 per person



Whole Genome Sequencing


ALDH2
ALDH2
ALDH2
ALDH2
ALDH2
ALDH2
GTCACCCTTTGGTGGCTACAAGATGTCGGGGAGTGGCCGGGAGTTGGGCGAGTACGGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTGGGACCTGCTGGGGGCTCAGGGCCTGTTGGGGCTTGAGGGTCTGCTGGTGGCTCGG

Linkage Disequilibrium



Linkage Disequilibrium

Blocks of high correlation between genotypes in the population (know as linkage disequilibrium) allow us to capture most of the common genetic variation in the population using a few hundred thousand SNPs

SNP genotyping arrays have been developed to rapidly capture this variation at a lower cost than sequencing

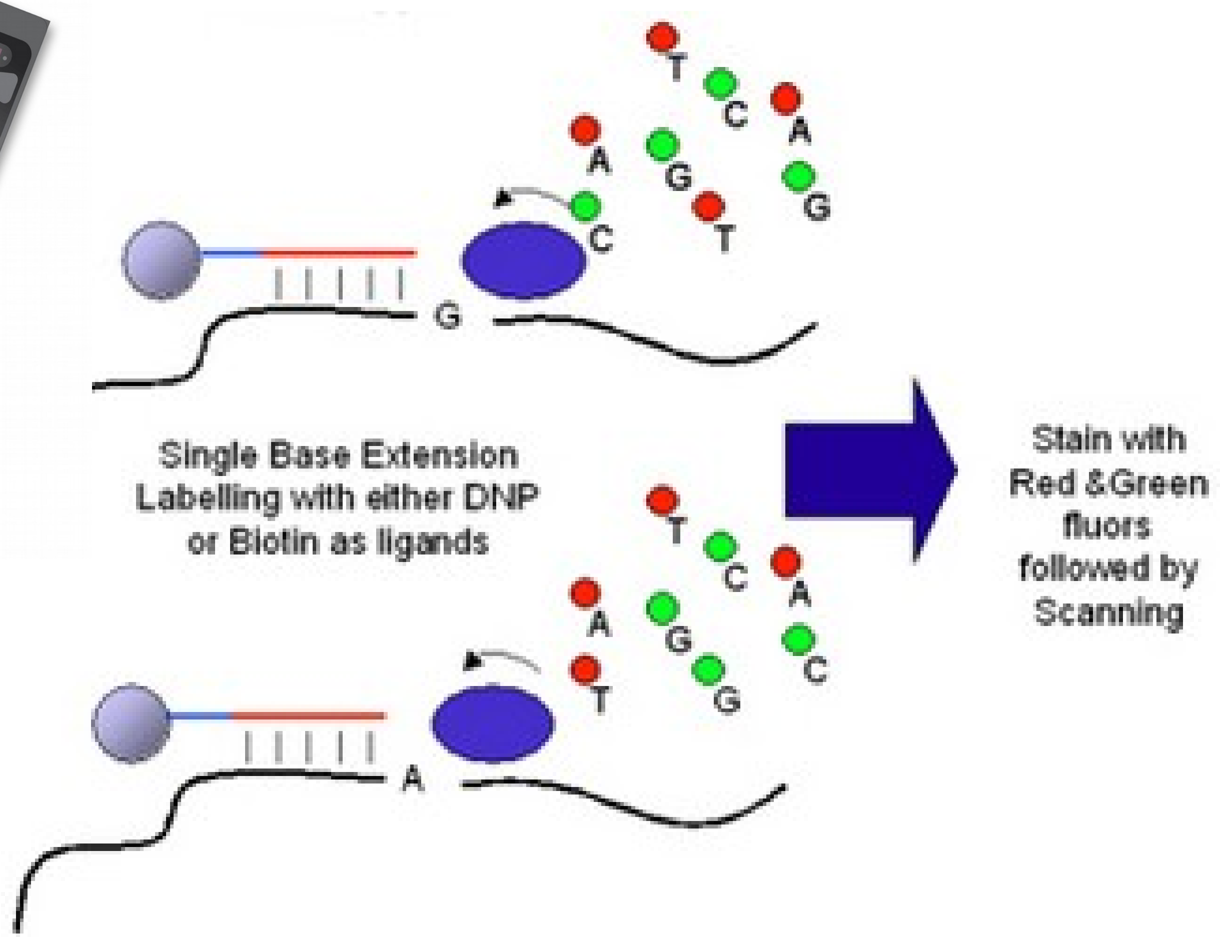
Currently ~$50 - $100 per person

Specific arrays have been designed to cover “important” areas in more depth

- e.g. Immunochip, Psychchip, …

We can recover much of the left out genotypes with imputation (covered later).

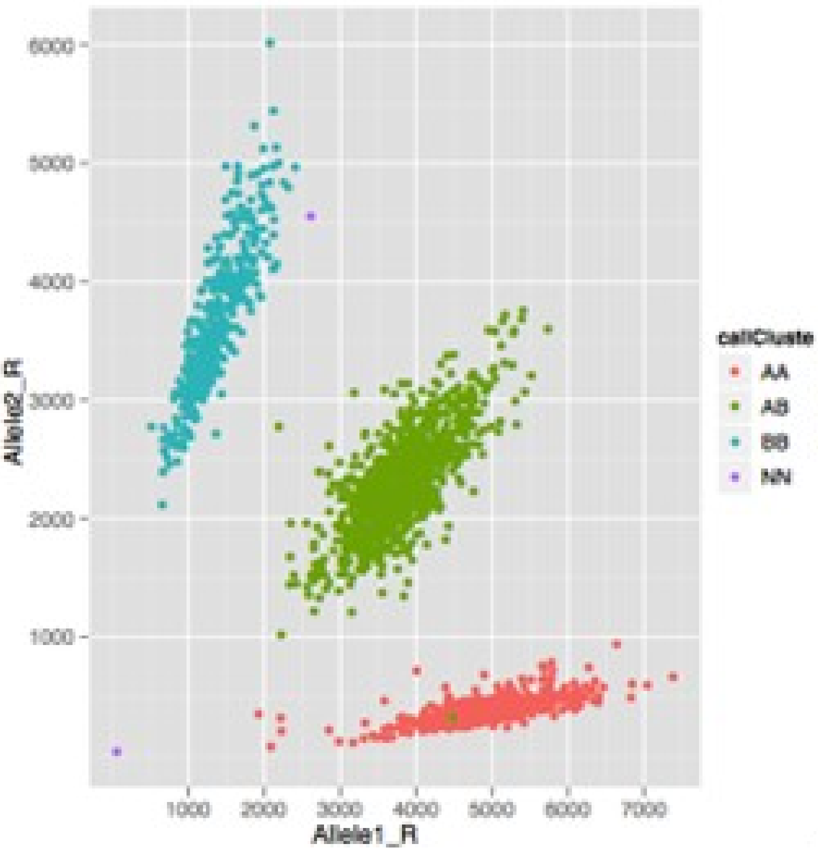
SNP Arrays

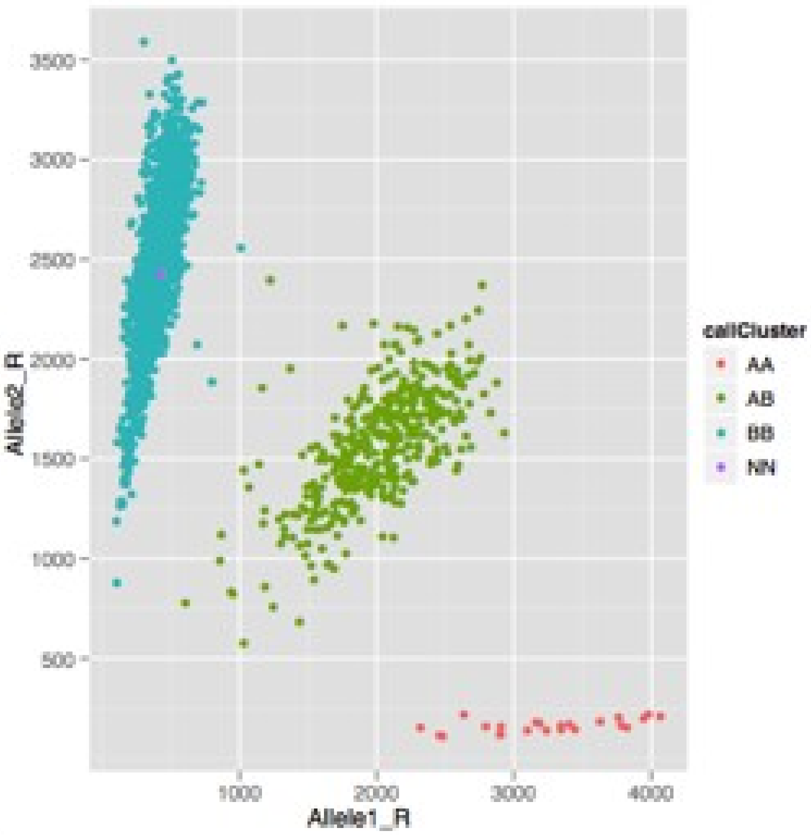


Single Base Extension
Labelling with either DNP
or Biotin as ligands

Stain with
Red &Green
fluors
followed by
Scanning

SNP Arrays


6000
5000
4000
3000
2000
1000
Allele2_R
1000
2000
3000
4000
5000
6000
7000
Allele1_R
callCluste
● AA
● AB
● BB
● NN


3500
3000
2500
2000
1500
1000
500
Allele2_R
1000
2000
3000
4000
Allele1_R
callCluster
● AA
● AB
● BB
● NN

Preparing Genotypic Data

The genotype clean process can be divided into two steps:

1) removing any individuals with poor quality data
2) removing SNP markers that have substandard genotyping performance

Performing the per-individual steps first prevents individuals with poor quality genotypes having an undue influence on the removal of SNP markers in the later step.

Per Individual Quality Control

There are five basic steps to removing bad individuals

1) removal of individuals with excess missing genotypes
2) removal of individuals with outlying homozygosity values
3) remove of samples showing a discordant sex
4) removal of related or duplicate samples, and
5) removal of ancestry outliers

Removal of an individual from the analysis is costly, both in terms of the cost of the genotyping and the time spent preparing the DNA sample

It is important to spend time during the initial study design to ensure to the extent possible that all individuals are from a common ancestral background and that extracted DNA is of high quality

Excess Missing Genotypes

1) removal of individuals with excess missing genotypes

For a SNP array, large numbers of missing SNP calls for an individual indicate:

- intensity measures are failing to fall in any genotype clusters

For sequencing data, large numbers of missing SNP calls indicate:

- low number of reads covering regions of the genome

This can be caused by low quality or concentration of the DNA used for genotyping

Excess Missing Genotypes

Samples with a high missingness rate also tend to have higher genotyping error in the genotypes that are called.

→ Remove any sample with high missingness from further analysis

A threshold in the order of 5% missingness is used to determine which samples need to be removed.

This step is particularly important when using a case-control design, especially when the DNA extraction was performed separately for cases and controls, as differential genotype quality may correlate with disease status and thus introduce a bias to the analysis

Outlying Homozygosity

2) removal of individuals with outlying homozygosity values

The proportion of homozygous (or inversely heterozygous) genotypes across an individual's genome (excluding sex chromosomes) can detect several issues with genotyping

Average heterozygosity correlates with genotype missingness such that samples with high missingness tend to have lower average heterozygosity,

- a reduction in heterozygosity can also reflect inbreeding.

Sample contamination, where multiple samples are accidentally genotyped on a singe array, results in high average heterozygosity.

Outlying Homozygosity

The average value of the proportion of heterozygous genotypes will vary across populations and genotyping platforms and as such the high and low thresholds for sample removal need to be determined by examining the distribution in your cohort.

Discordant Sex

3) remove of samples showing a discordant sex

Determining whether an individual is male or female is straightforward from genotype data

- males only have a single copy of the X chromosome so can not be heterozygous

- the small pseudo-autosomal region at the end of the chromosome may show heterozygosity

- some small number of heterozygotes may be attributable to genotyping errors

Discordant Sex

Females:

- individuals with low heterozygosity across the X chromosome are indicative of a sample mix-up with a male

Males:

- Samples with high heterozyogosity are likely to be females.
- High missingness for X chromosome only can also indicate an incorrectly labelled sample

Provided males and females have been randomly placed on plates for genotyping, patterns of mismatching sex can be used to rectify potential plating errors.

Related/Duplicate Samples

4) removal of related or duplicate samples

Even distantly related samples can bias GWAS results ***if not properly accounted for***.

e.g. if we have two related cases in a case-control analysis, their genotypes being on average more similar to each other than the rest of the cohort will provide a slight bias to the estimate of the allele frequency in cases and its associated standard error

Even this small bias is important when considering the number of statistical tests being performed.

Related/Duplicate Samples

Can detect related individuals by calculating Identity-by-State (IBS) or Identity-by-Descent (IBD)

- IBS measures the average proportion of alleles shared by two individuals across the autosomal genome

- IBD measures the proportion of the genome that is shared between two individuals

IBD = 1	Duplicates or monozygotic (identical) twins
IBD = 0.5	Parent/offspring, siblings
IBD = 0.25	Second degree relatives

For any pair with an IBD > 0.05, remove the one with the lowest genotyping rate

Outlying Ancestry

5) removal of ancestry outliers

Population stratification is the ***major source of bias*** in GWAS, as it is common for disease or quantitative traits to have different frequencies or distributions across populations

Beware the Chopsticks Gene

A case-control study for ability to use chopsticks

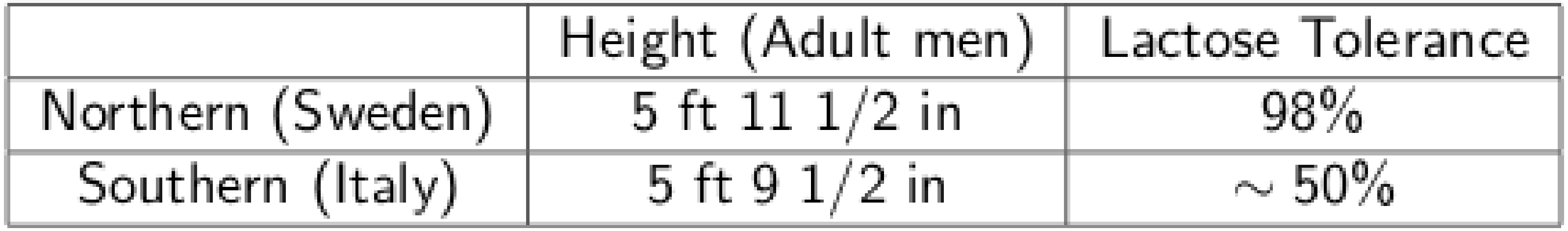
- chopstick usage is very stratified by ancstral group (e.g. Asian vs European)
- A GWAS for chopstick ability using individuals across major population groups would identify many false positives

Outlying Ancestry

Real example:

Campbell et al. Demonstrating stratification in a European American population. *Nature Genetics* (2005) 37:868–72.

Performed GWAS on two groups of individuals of European descent that were discordant for height and identified an association with the LCT (lactase) locus



	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

Outlying Ancestry

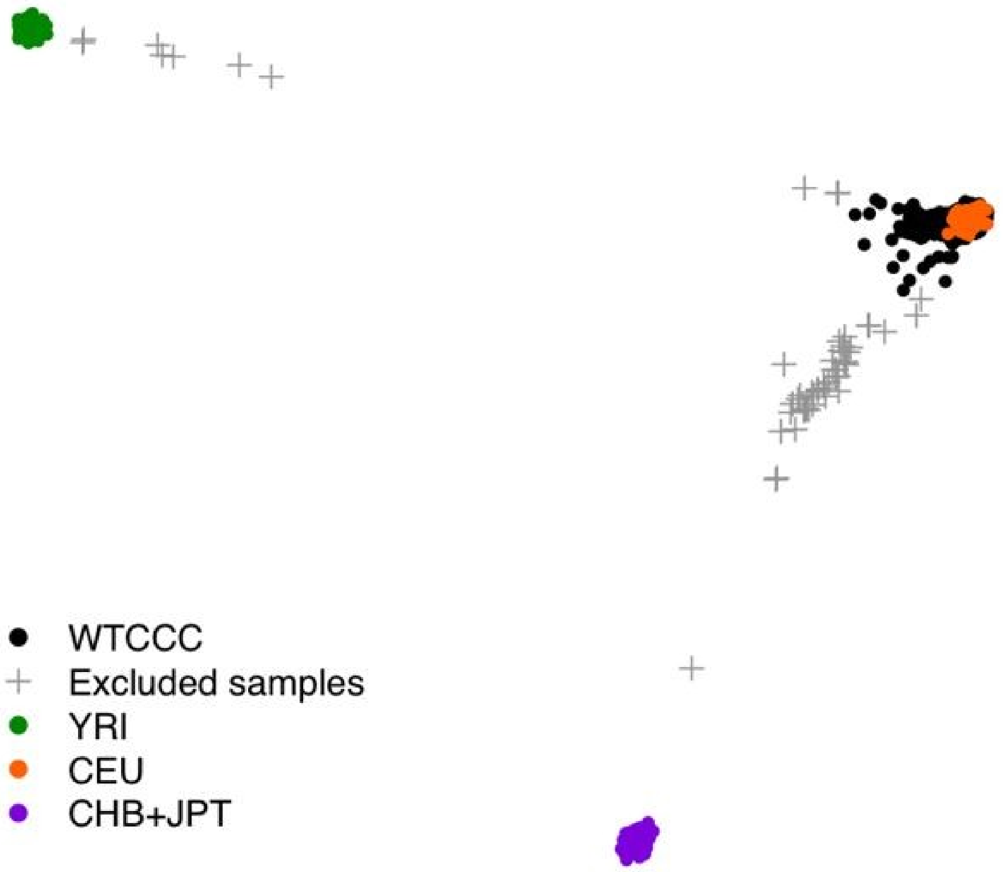
Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among individuals

– Exclude individuals of unexpected ancestry

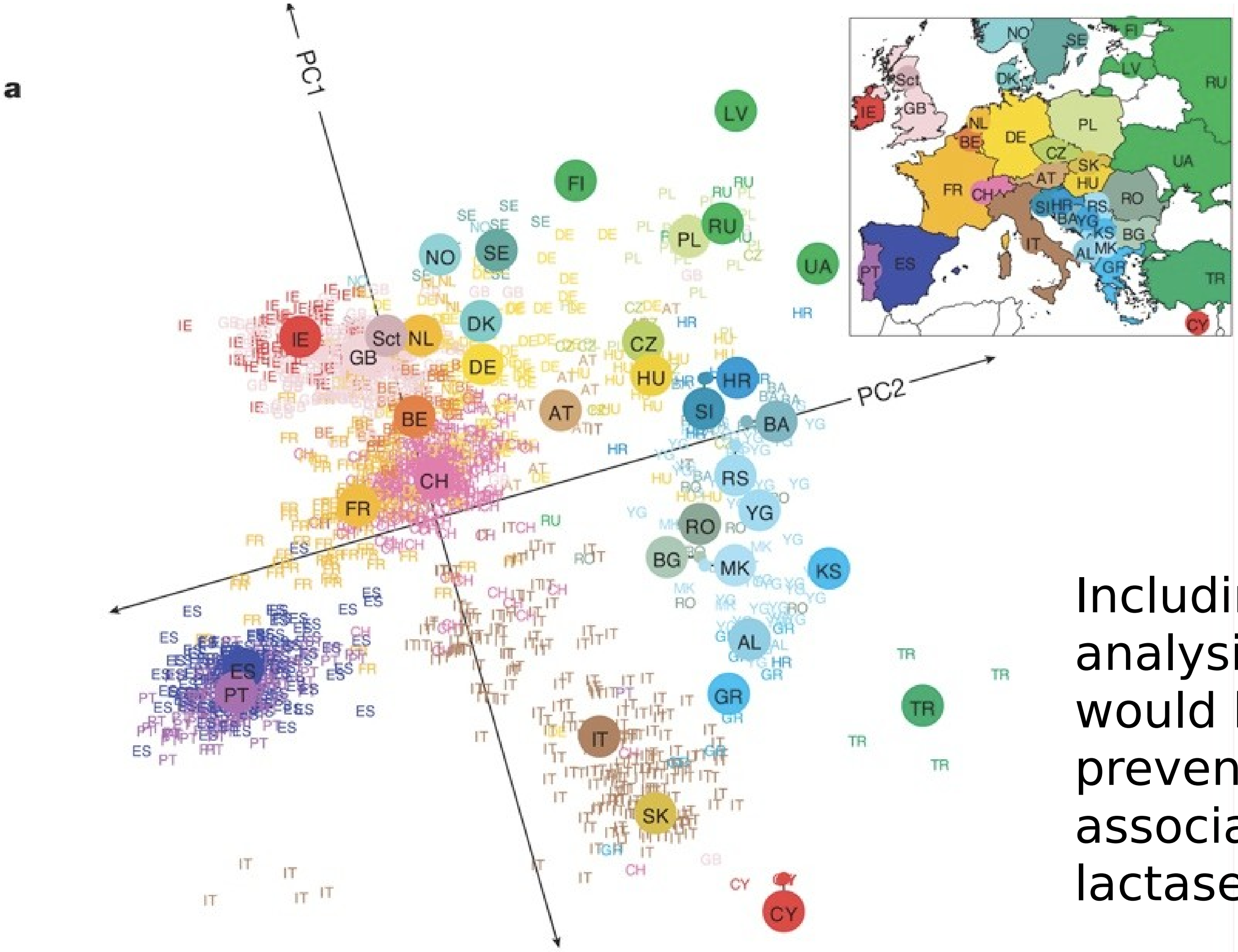
* PC1 separates African vs Asian/European
* PC2 separates Asian vs European

- Use PCs as covariates in the analysis to correct for possible biases induced by sample collection or non-genetic geographical effects on phenotype

Outlying Ancestry



Outlying Ancestry



Including PC1 in analysis of height would have prevented association at lactase gene.