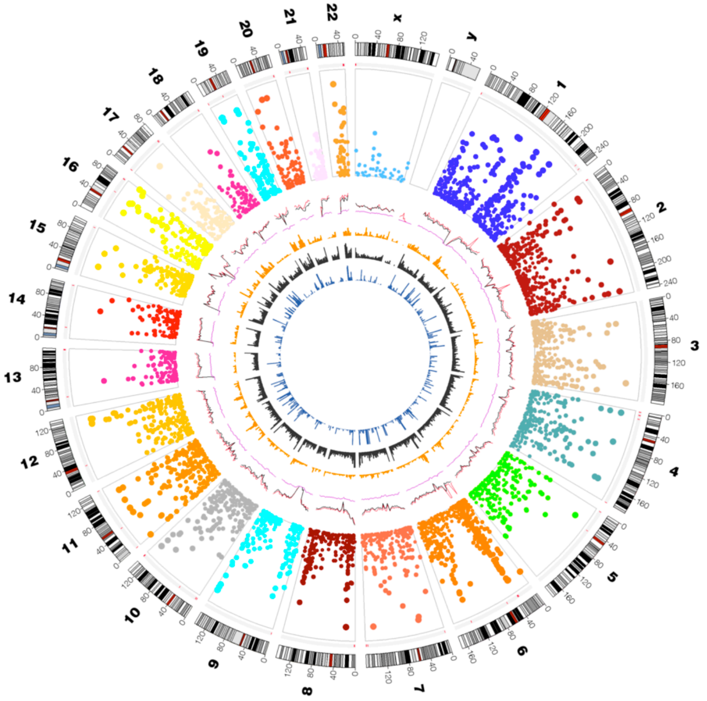
GENOME-WIDE ASSOCIATION STUDIES

Amirtha Ambalavanan, Ph.D.

Post Doctoral Fellow, DBMS

Laboratory of Dr. Qingling Duan
Botterell Hall, room 422
amirtha.ambalavanan@queensu.ca



CAC

- Access to UNIX shell:
  - PC users: download MobaXterm
  - Mac users: download Xquartz

- How to access linux server in CAC
  - ssh -X yourUsername@login.cac.queensu.ca
  - Familiarize yourself with CAC WIKI page: http://cac.queensu.ca/wiki/index.php/Main_Page

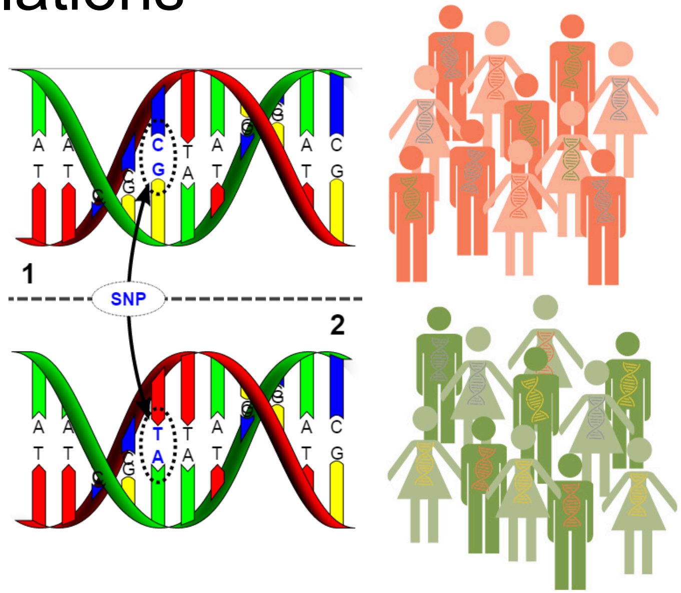
- To transfer files between desktop and CAC account:
  - Use FileZilla

Outline

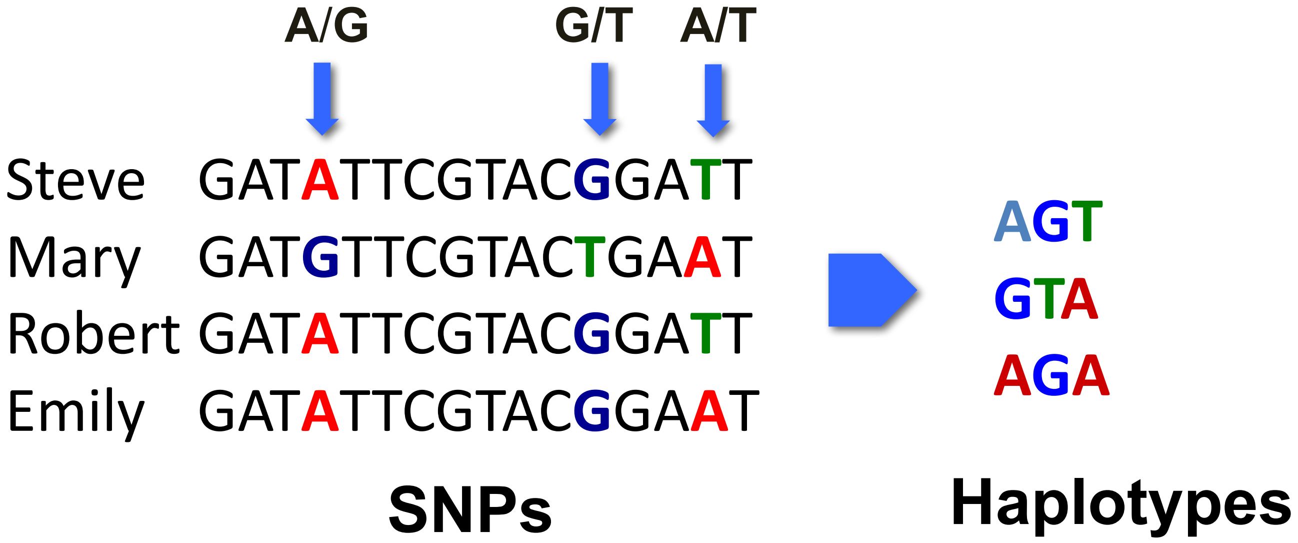
- Introduction to genome-wide association studies (GWAS)
- Key elements of GWAS
- Example of a GWAS
- GWAS Quality Control

Definitions

- Gene – functional unit of DNA that codes for a protein
- Genome – the entirety of an organism’s genetic material
- Genetics – study of heredity
- Genomics - the study of organism’s entire genome
- Genetic association – discern how genetic variations affect traits in populations



Genomics Vocabulary



A/G G/T A/T

Steve GATATTCGTACGGATT

Mary GATGTTCGTACTGAAT

Robert GATATTCGTACGGATT

Emily GATATTCGTACGGAAT

SNPs

AGT

GTA

AGA

Haplotypes

Examples of Genetic Variations

- Single Nucleotide Variations (SNVs) – once every 100-300 bases, polymorphic (SNP) if present in > 1% of population

- Copy Number Variations (CNVs) – structural variations > 1000 bases

- Indels – insertion and deletions

- Microsatellites – DNA motifs consisting of 2-5 nucleotide repeated 5-50 times

Mendelian vs. Complex Traits

Mendelian Disorders

- Rare syndromes (Marfan's disease, cystic fibrosis, sickle cell anemia)
- Single Gene Disorders, high penetrance
- Family-based linkage studies, moderate sample size

Complex Disorders

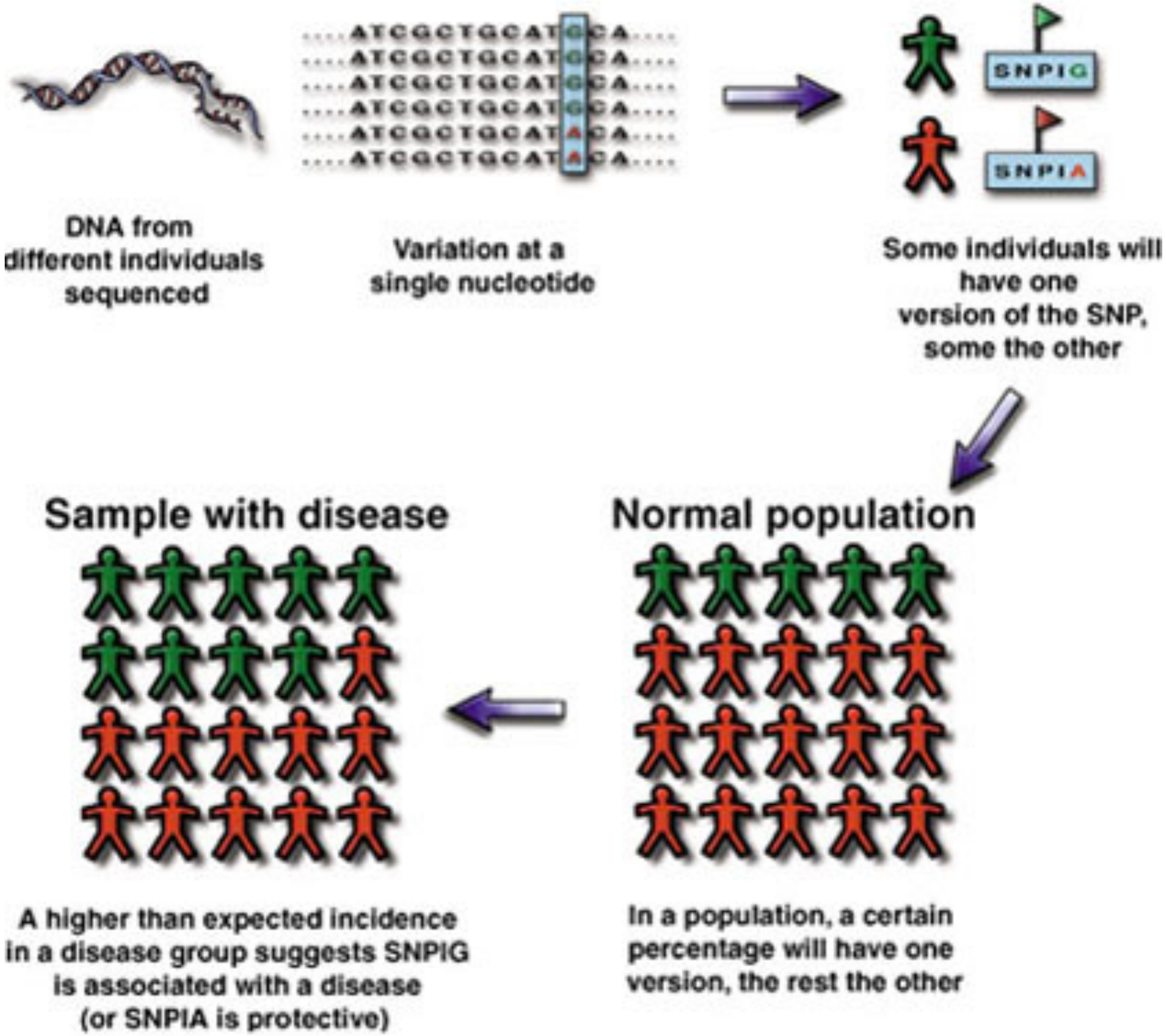
- Common diseases (diabetes, CAD, arthritis, COPD, cancer)
- Multigenic and multifactorial etiology
- Population-based association studies, large sample sizes

Common Disease – Common Variants

The majority of common diseases are strongly influenced by frequent alleles with only moderate effect size.

Eg: Diabetes, high blood pressure, heart disease

Association Studies



DNA from different individuals sequenced

Variation at a single nucleotide

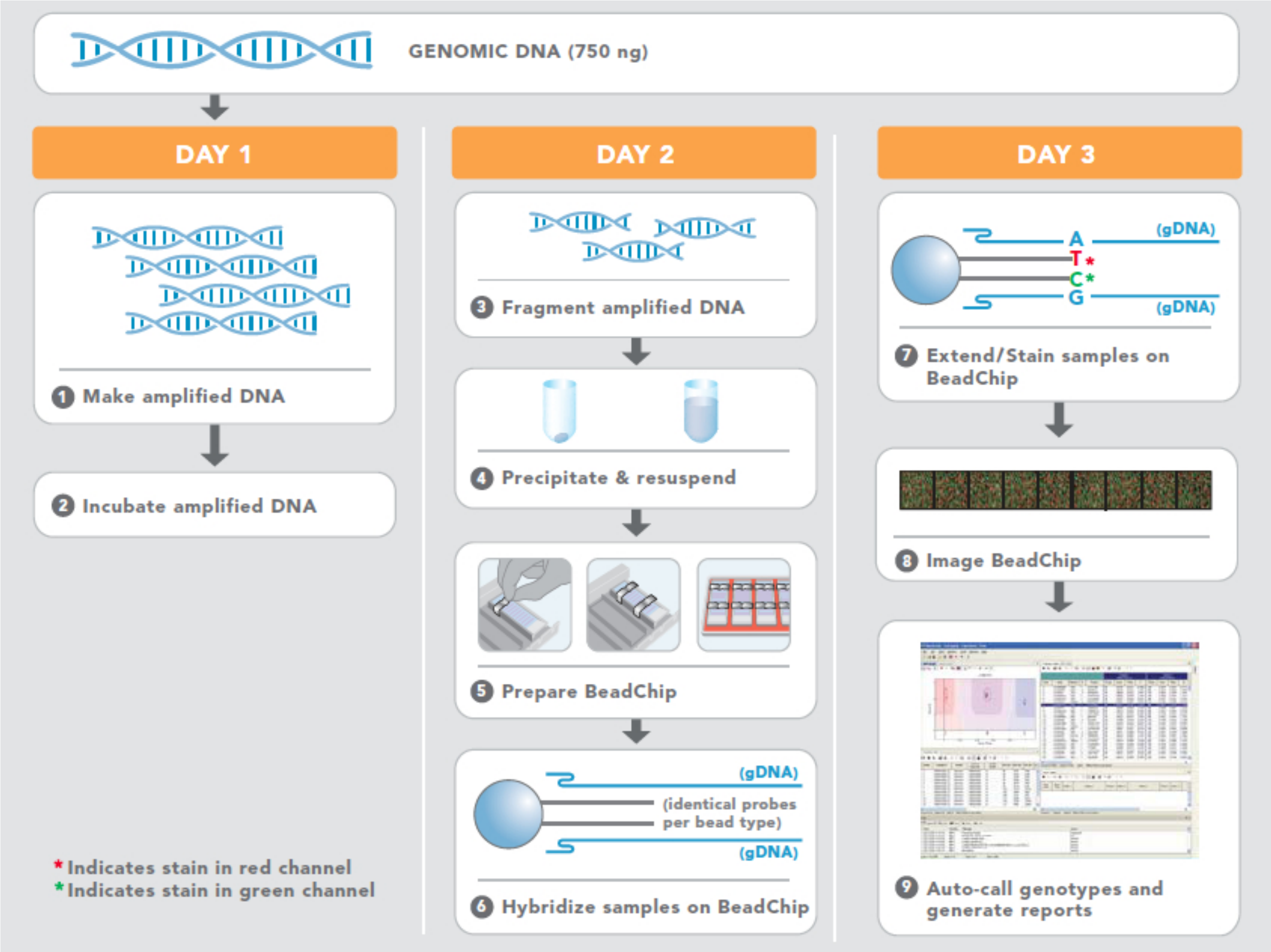
Some individuals will have one version of the SNP, some the other

A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)

In a population, a certain percentage will have one version, the rest the other

© Gibson & Muse, A Primer of Genome Science

Genotyping



GENOMIC DNA (750 ng)

DAY 1

1 Make amplified DNA

2 Incubate amplified DNA

DAY 2

3 Fragment amplified DNA

4 Precipitate & resuspend

5 Prepare BeadChip

6 Hybridize samples on BeadChip

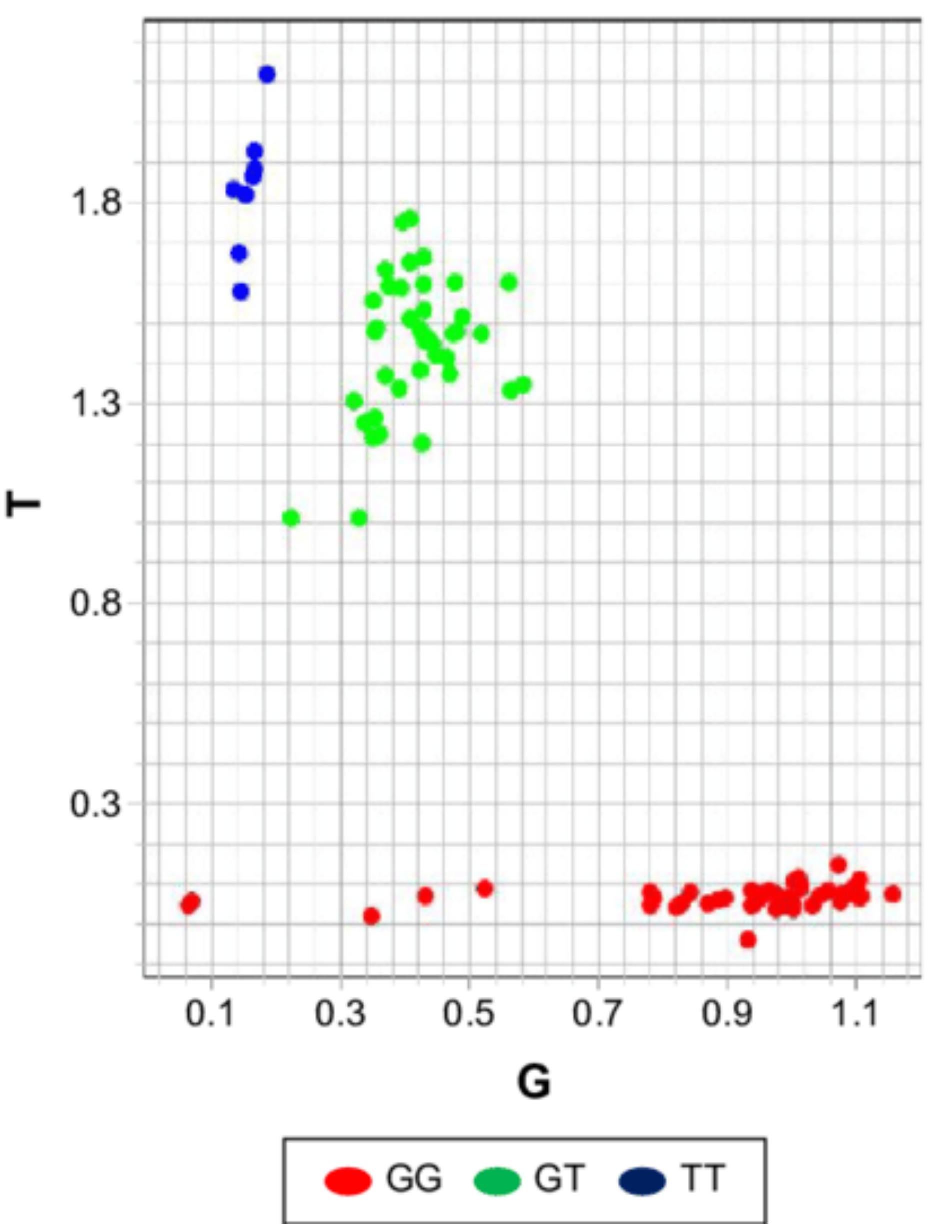
DAY 3

7 Extend/Stain samples on BeadChip

8 Image BeadChip

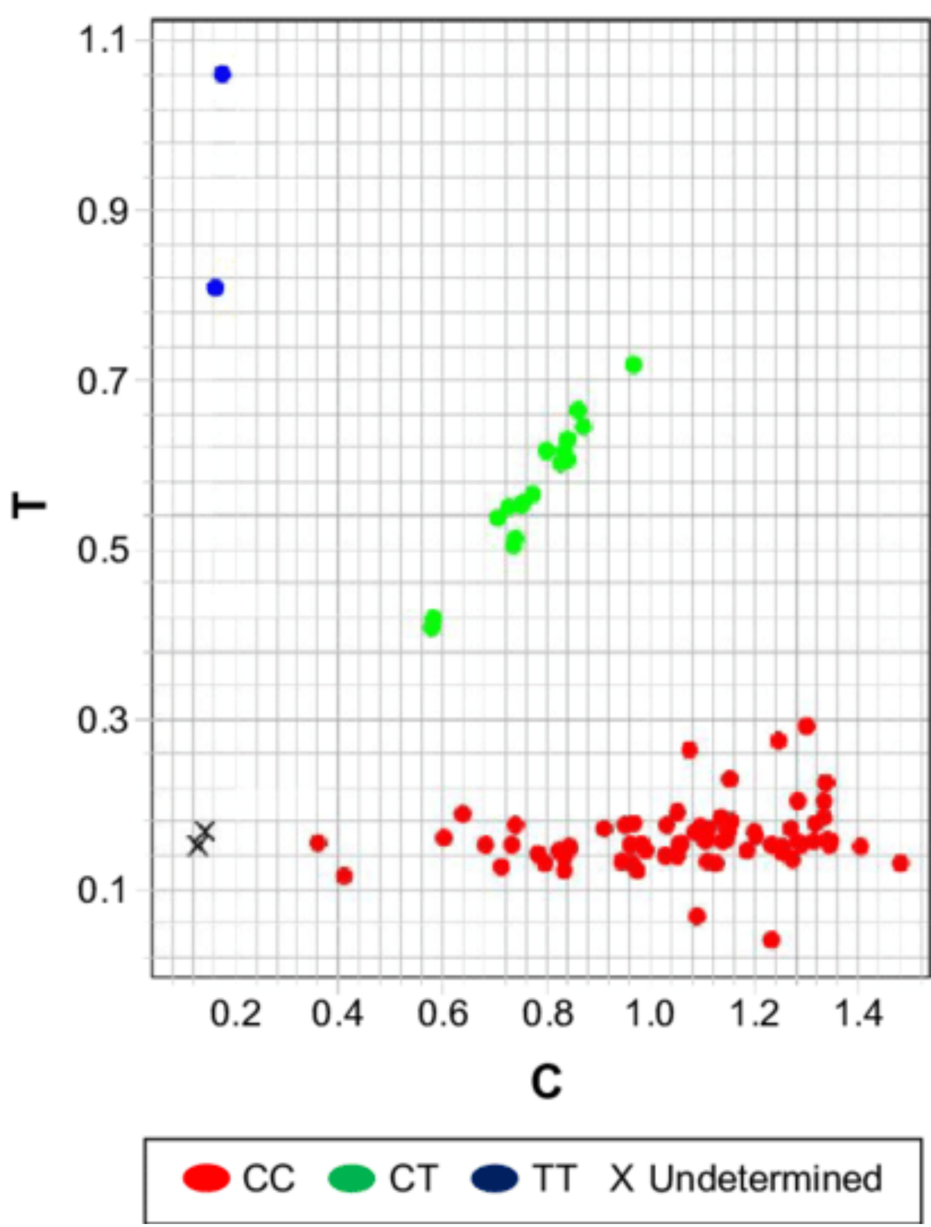
9 Auto-call genotypes and generate reports

* Indicates stain in red channel
* Indicates stain in green channel



rs1801321 (172G>T variant) in *RAD51*

Allele discrimination plot

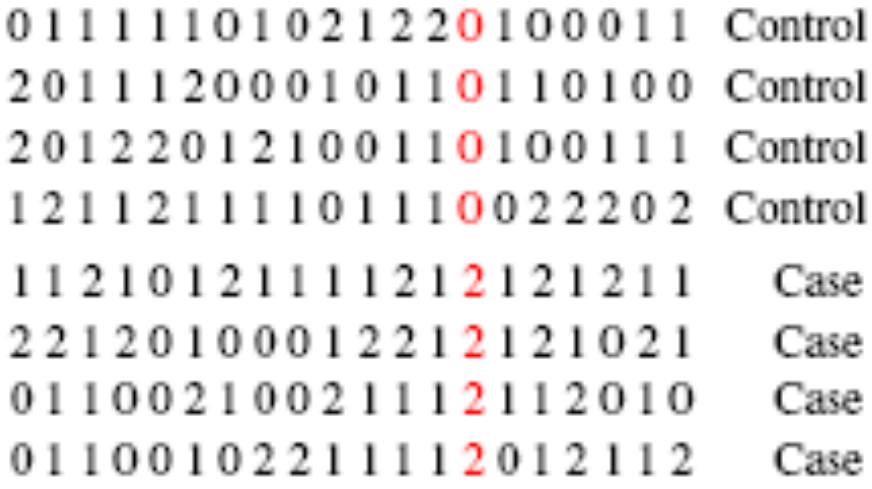


rs2619681 (C>T variant) in *RAD51*

Allele discrimination plot

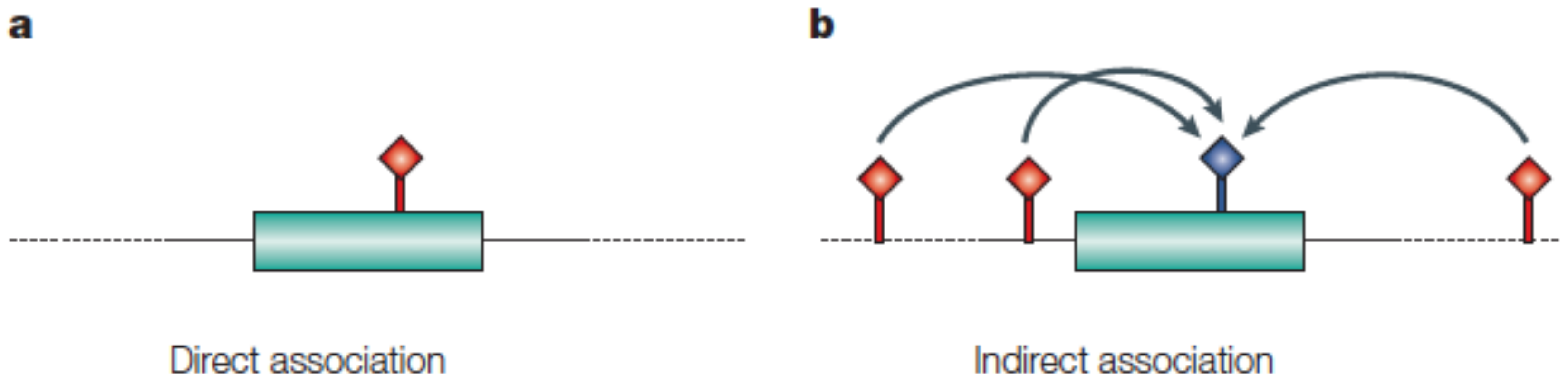
What is a GWAS?

- Genome-Wide Association Study – interrogates the relationship between genome-wide genetic variations and a trait.

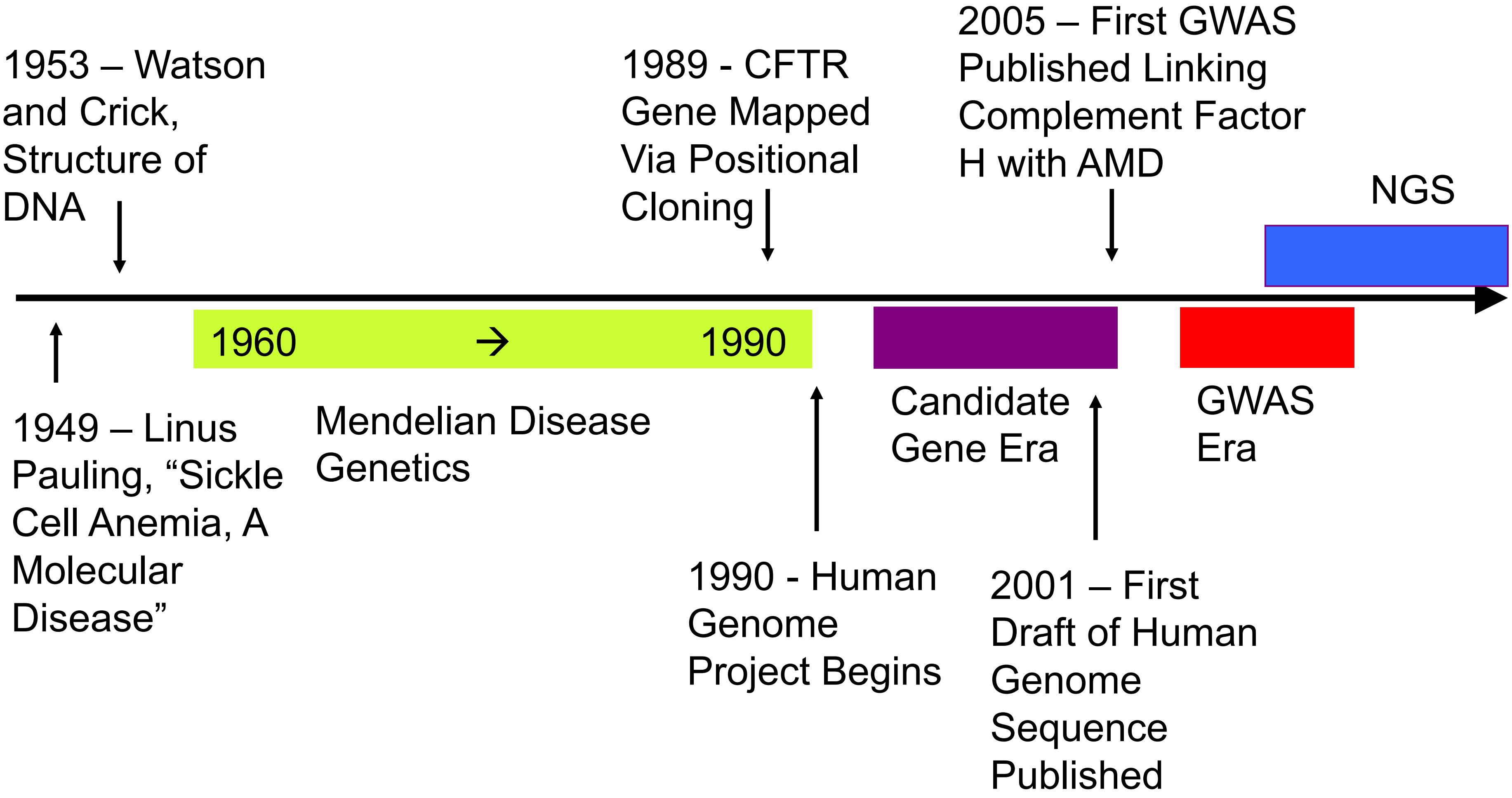


0	1	1	1	1	1	0	1	0	2	1	2	2	0	1	0	0	0	1	1	Control
2	0	1	1	1	2	0	0	0	1	0	1	1	0	1	1	0	1	0	0	Control
2	0	1	2	2	0	1	2	1	0	0	1	1	0	1	0	0	1	1	1	Control
1	2	1	1	2	1	1	1	1	0	1	1	1	0	0	2	2	2	0	2	Control
1	1	2	1	0	1	2	1	1	1	1	2	1	2	1	2	1	2	1	1	Case
2	2	1	2	0	1	0	0	0	1	2	2	1	2	1	2	1	0	2	1	Case
0	1	1	0	0	2	1	0	0	2	1	1	1	2	1	1	2	0	1	0	Case
0	1	1	0	0	1	0	2	2	1	1	1	1	2	0	1	2	1	1	2	Case

Direct and Indirect SNP tests



Genomics Evolution



1953 – Watson and Crick, Structure of DNA

1949 – Linus Pauling, “Sickle Cell Anemia, A Molecular Disease”

1960 → 1990

Mendelian Disease Genetics

1989 - CFTR Gene Mapped Via Positional Cloning

1990 - Human Genome Project Begins

2005 – First GWAS Published Linking Complement Factor H with AMD

Candidate Gene Era

2001 – First Draft of Human Genome Sequence Published

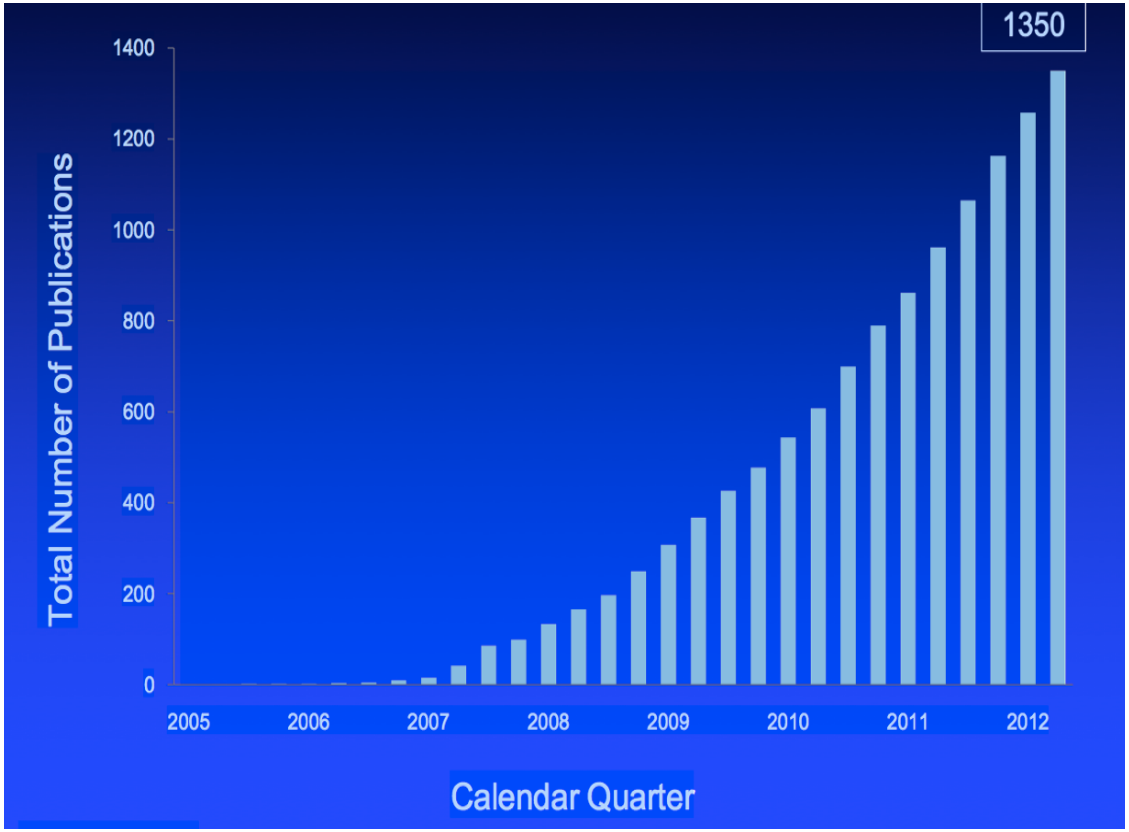
NGS

GWAS Era

GWAS

2005 – 1st GWAS: Age-related macular degeneration

2014 – 14,342 associations



Key Elements of GWAS

- case-control study design
  - potential confounders to analysis (population stratification, ascertainment)
- genome-wide genotyping
  - data management, special programs and computing requirements
  - quality control
- statistical association testing
  - multiple comparisons

GWAS tools

Most popular:

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

Not as popular:

- SNPassoc (Juan R. González 1, et al. Bioinformatics, 2007 23(5):654-655)
- GenABEL (Aulchenko Y.S., Ripke S., Isaacs A., van Duijn C.M. Bioinformatics. 2007, 23(10):1294-6.)

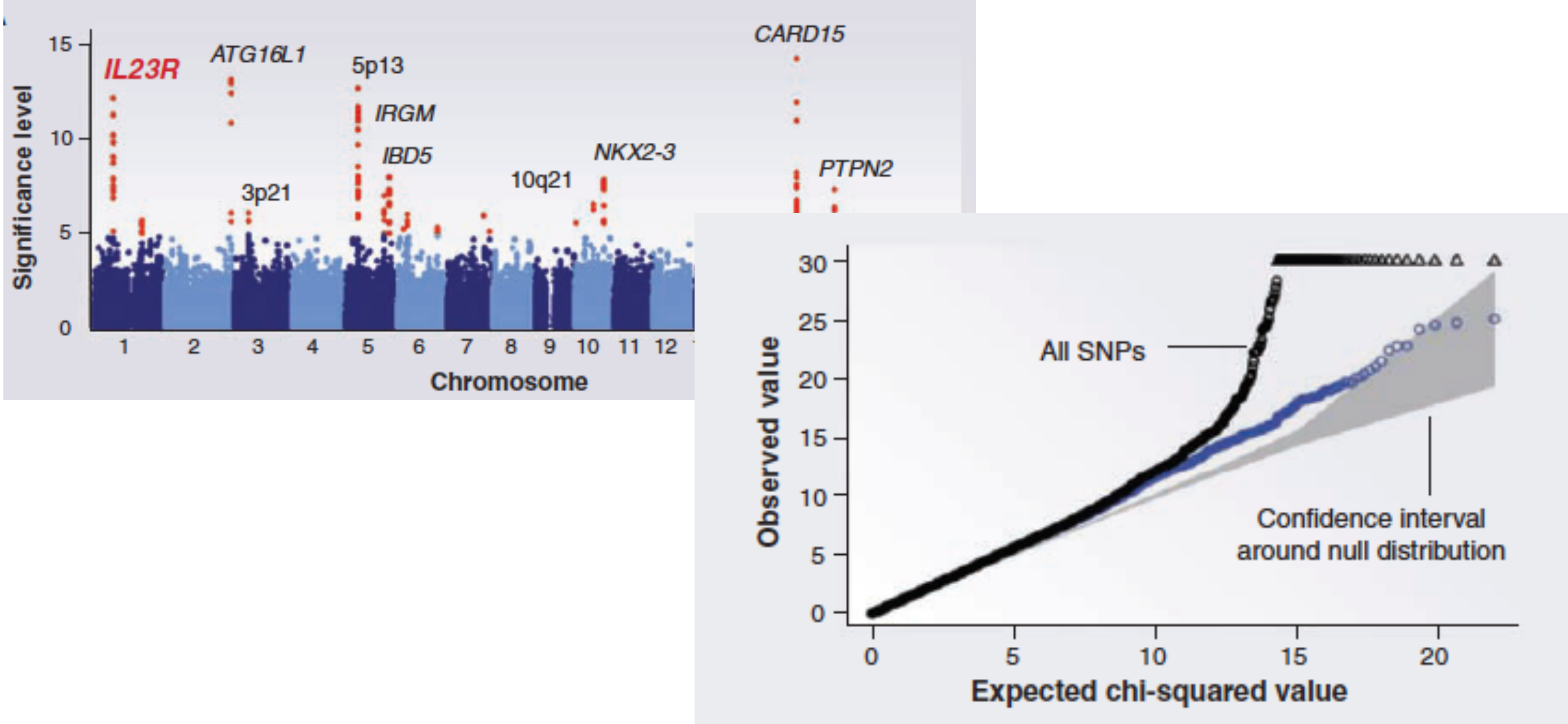
Data Analysis

- Single SNP analysis using pre-specified genetic models
  - 2 x 3 table (2-df)
  - Additive model (1-df), and test for additivity
  - All possible genetic models (recessive, dominant)

	AA	**AB**	**BB**
Affected			
Unaffected			

Visualization of Results

- Manhattan Plots: genome-wide p-values
- QQ Plots: assess bias/significance



False Positives

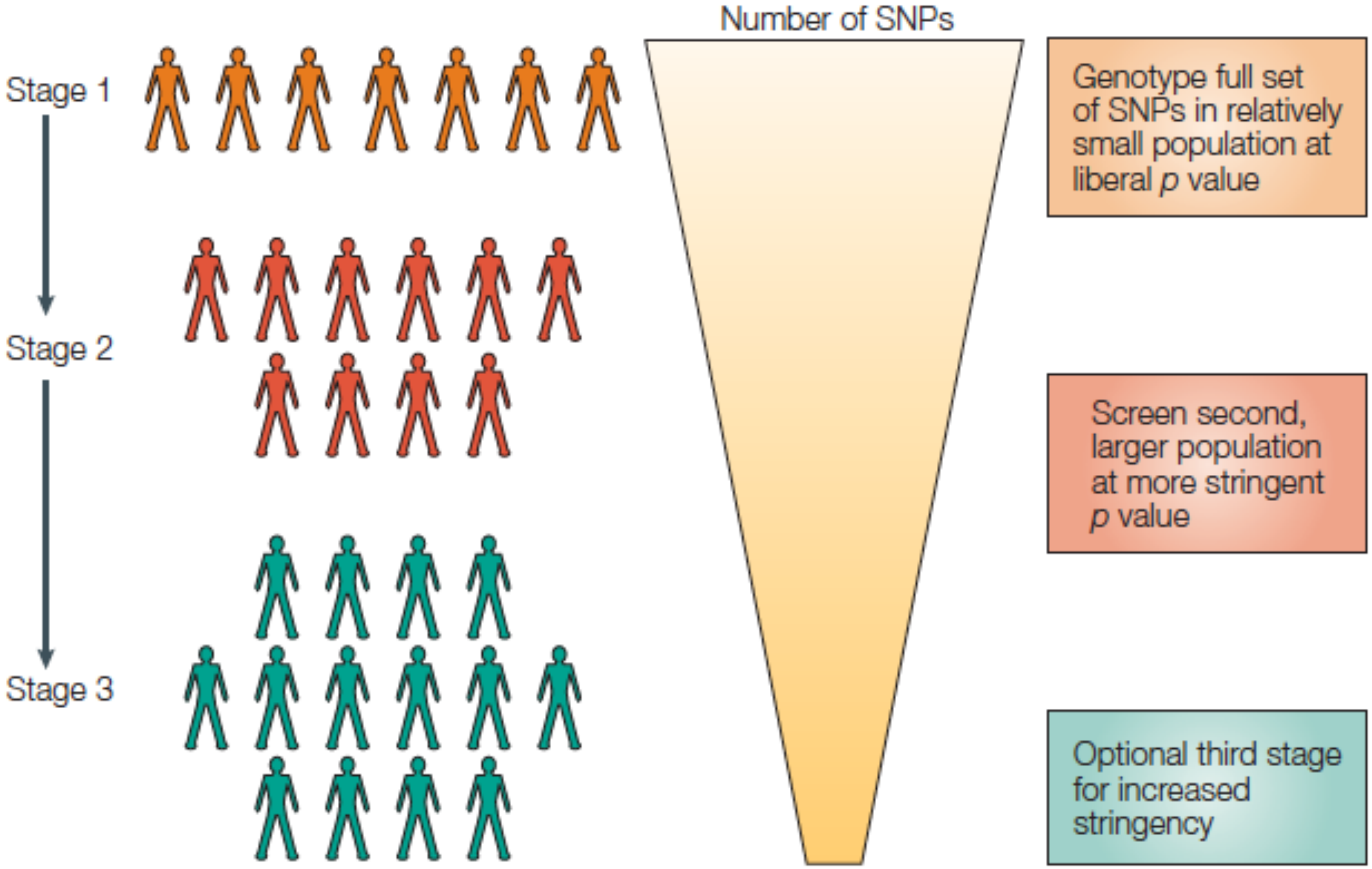
Too many dependent tests: must adjust for number of tests

- **Bonferroni correction**
  - Nominal significance level = study-wide significance / number of tests
  - Nominal significance level = 0.05/500,000 = 10^{-7}
- **Effective number of tests**
  - Take LD into account
- **False discovery rate (FDR)**
  - Expected proportion of false discoveries among all discoveries
  - Offers more power than Bonferroni
  - Holds under weak dependence of the tests

Replication

- The approach may limit the number of false positives
- Confirmation is needed to dissect true from false positives
    - Replication, examine the results from the 2nd stage only
    - Joint analysis, combining data from 1st stage with 2nd stage

Replication

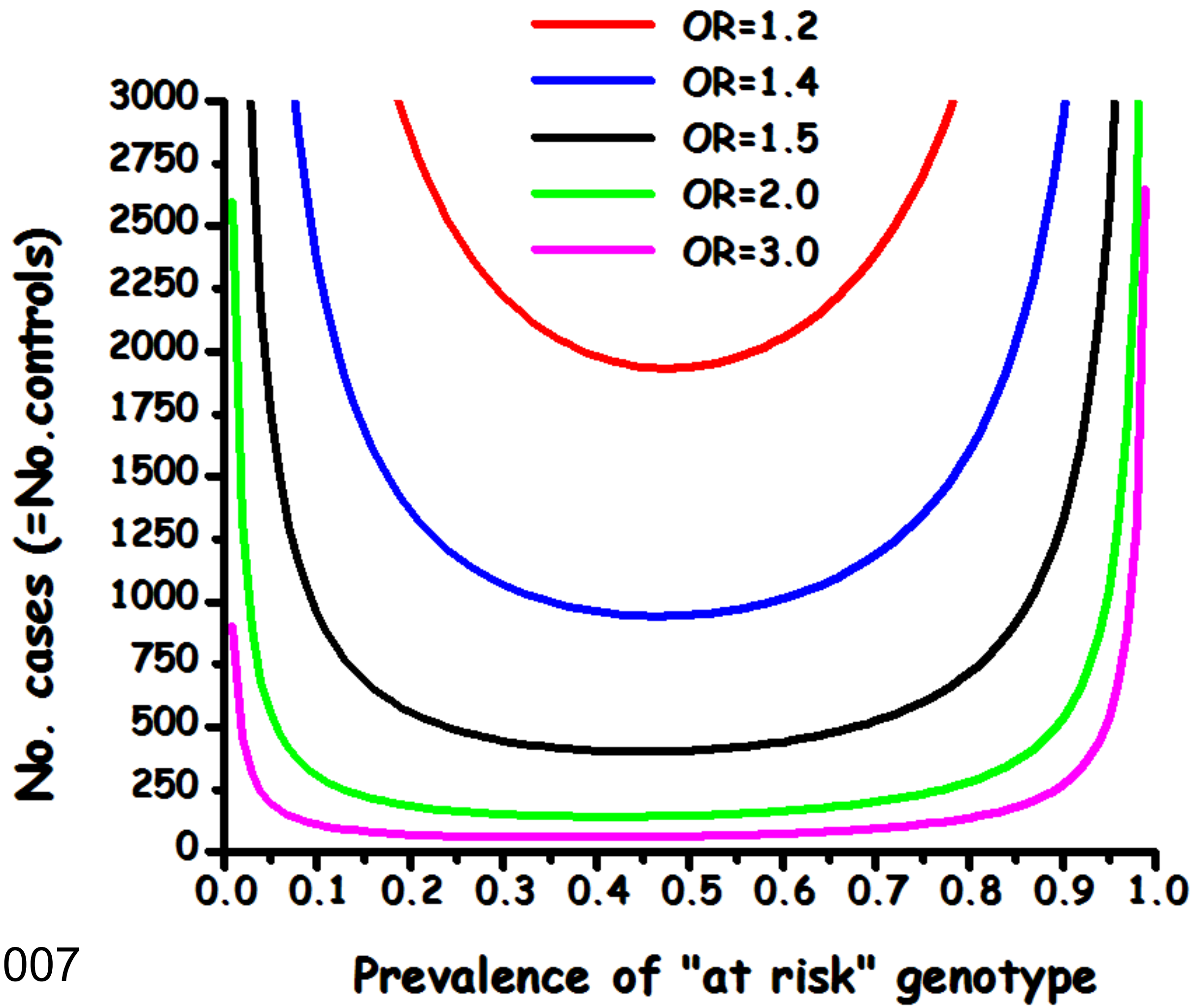


Genotype full set of SNPs in relatively small population at liberal *p* value

Screen second, larger population at more stringent *p* value

Optional third stage for increased stringency

Sample size vs. genotype prevalence and OR

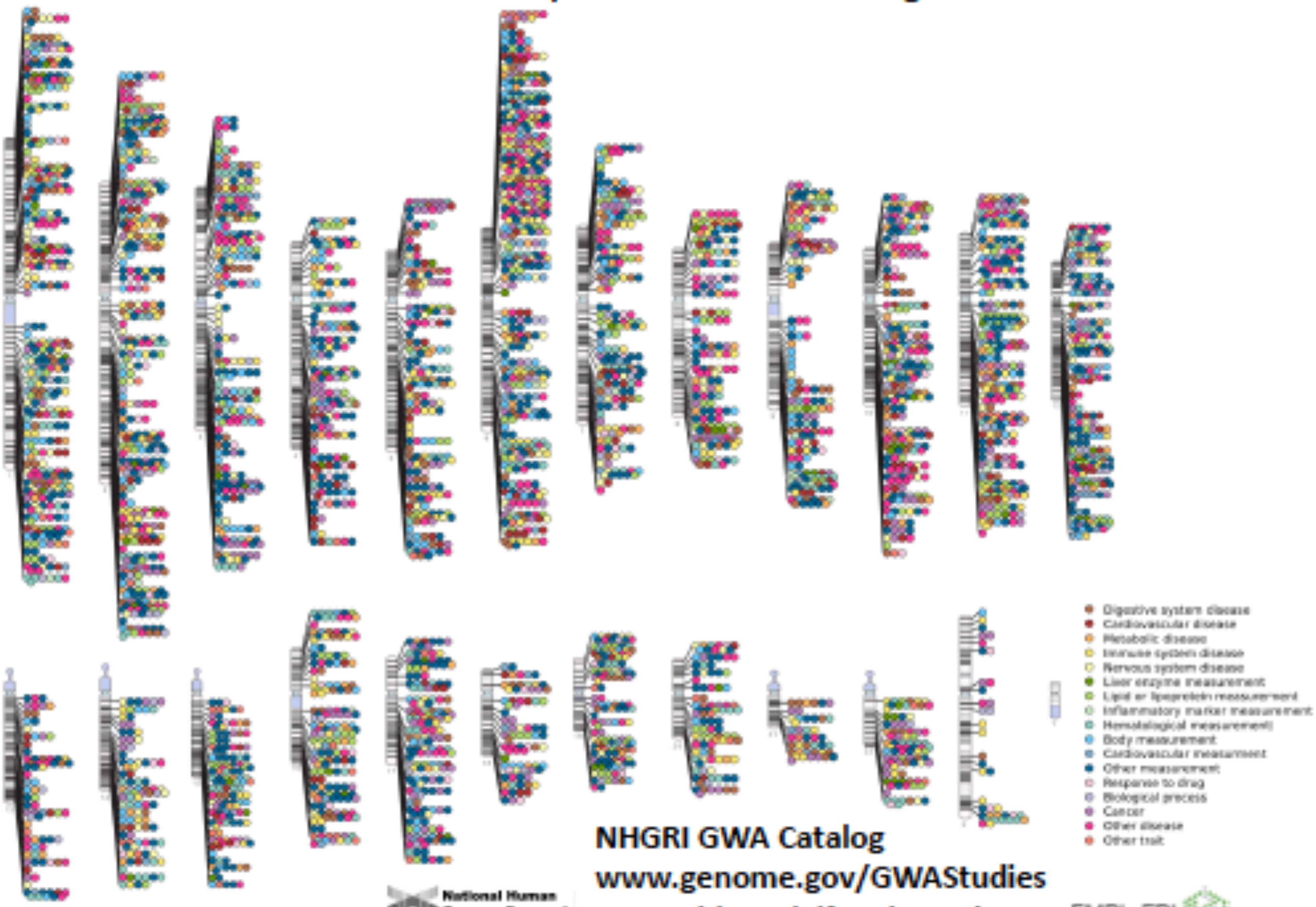


Boffeta, 2007

Case of the Missing Heritability

Environment, Gene-Environment interactions, epistasis, haplotype effects, multifactorial traits, small effects, rare variants, LD, Population Stratification (subtle ancestral differences between case and control groups)

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

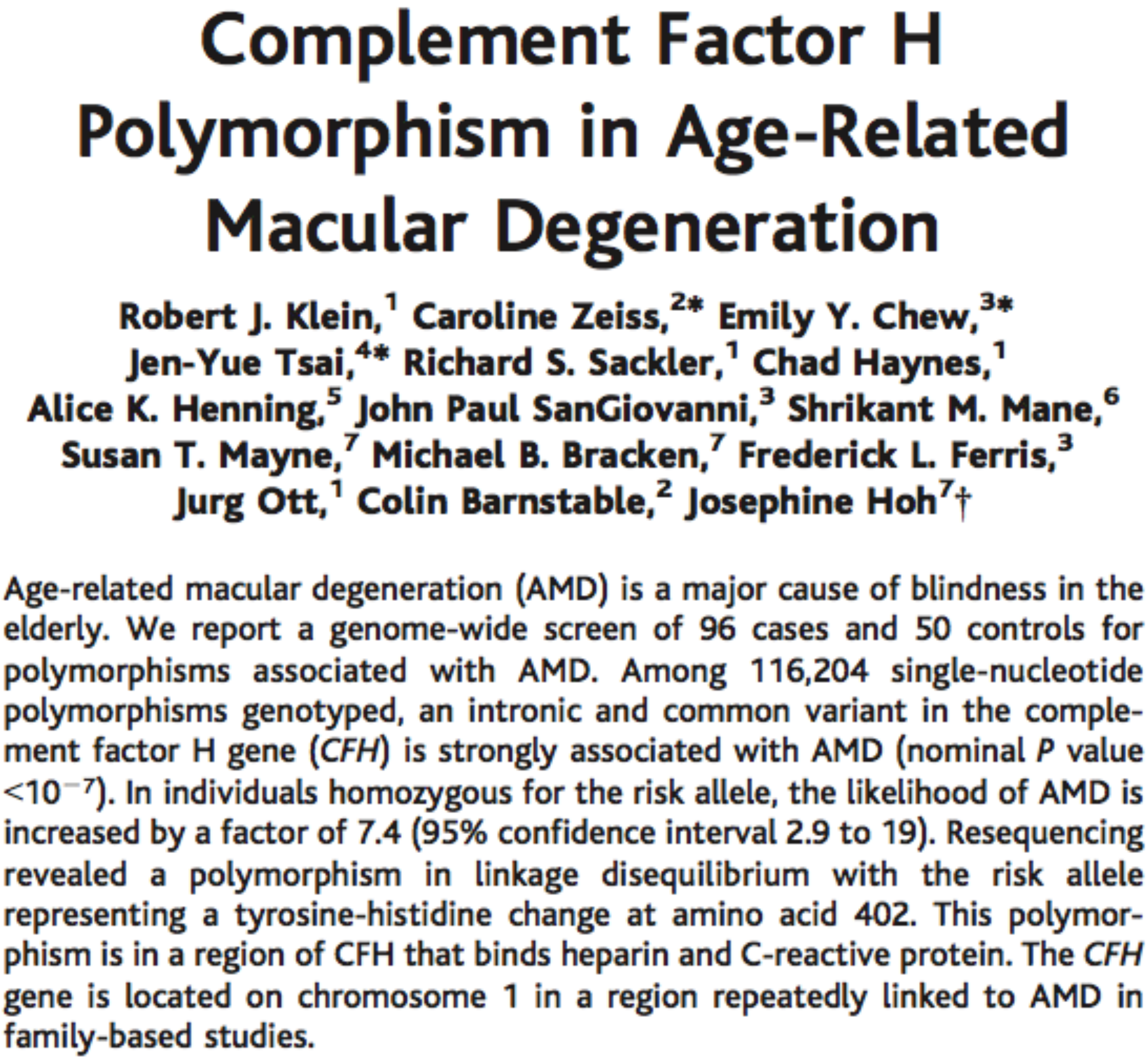


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

EXAMPLE OF GWAS

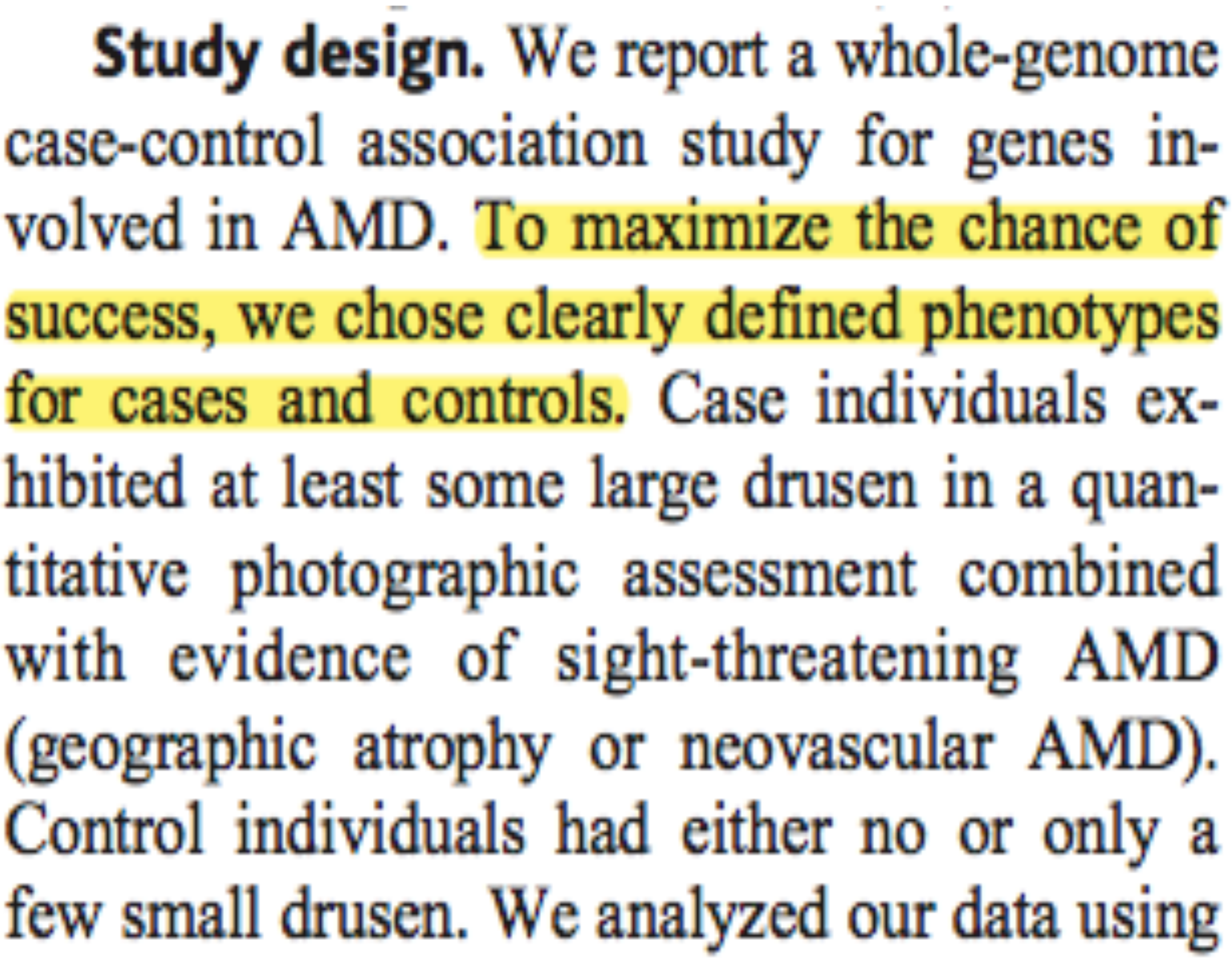


Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

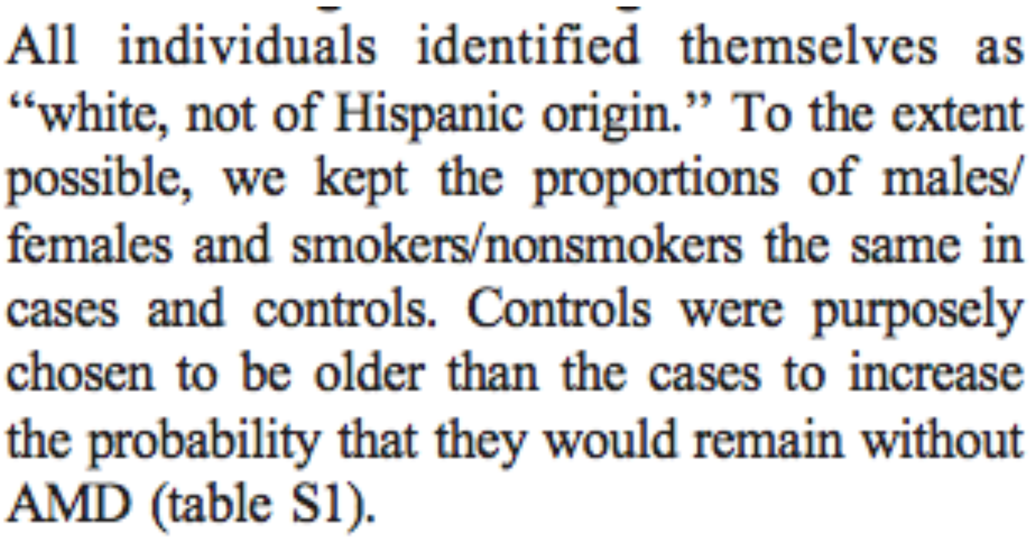
Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

Case-Control Design, Ascertainment



Study design. We report a whole-genome case-control association study for genes involved in AMD. To maximize the chance of success, we chose clearly defined phenotypes for cases and controls. Case individuals exhibited at least some large drusen in a quantitative photographic assessment combined with evidence of sight-threatening AMD (geographic atrophy or neovascular AMD). Control individuals had either no or only a few small drusen. We analyzed our data using

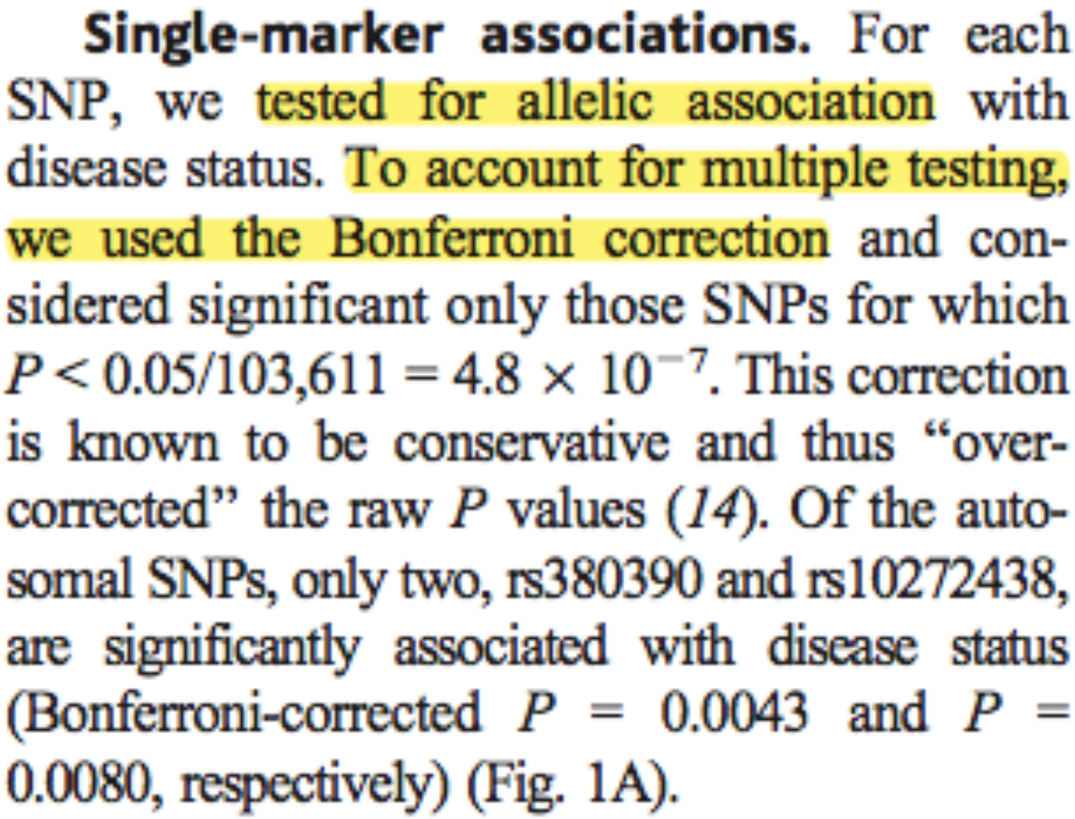
Confounding



All individuals identified themselves as "white, not of Hispanic origin." To the extent possible, we kept the proportions of males/females and smokers/nonsmokers the same in cases and controls. Controls were purposely chosen to be older than the cases to increase the probability that they would remain without AMD (table S1).

- Population Stratification (subtle ancestral differences between case and control groups)
- Traditional confounders (gender, environmental exposures)
- Phenotype misclassification (phenocopies)

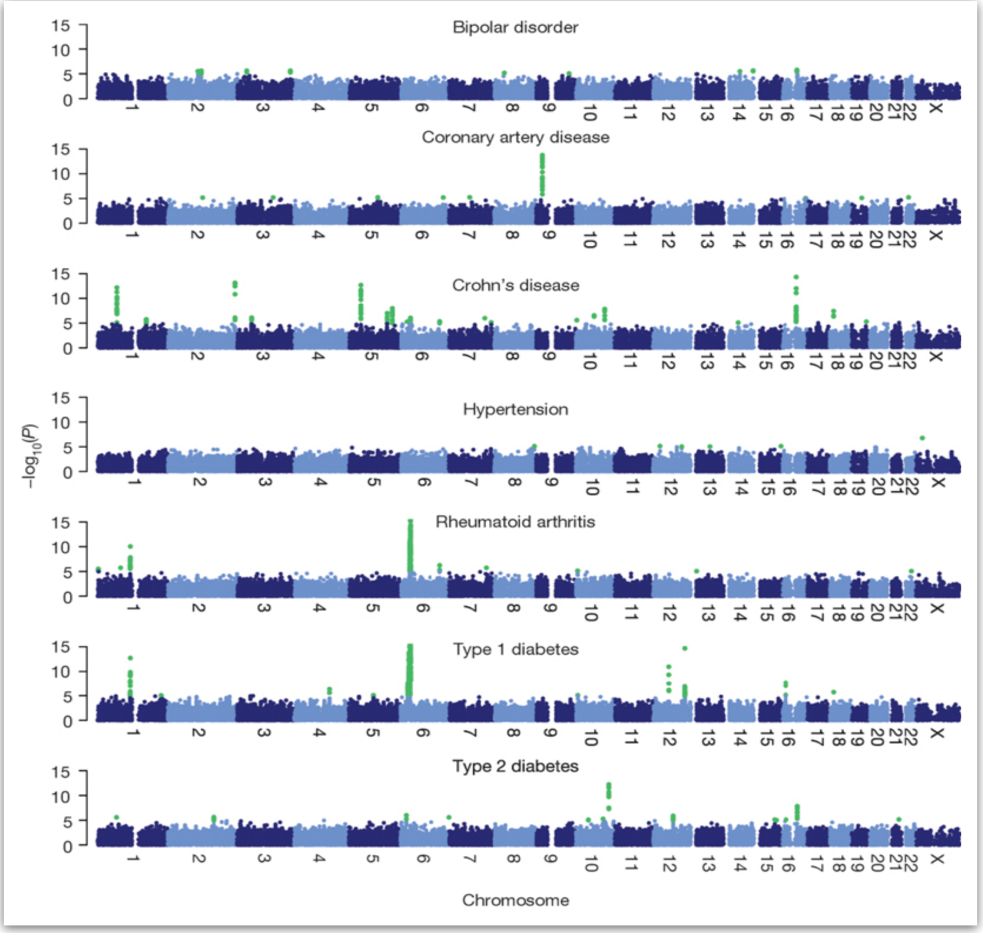
Association Testing

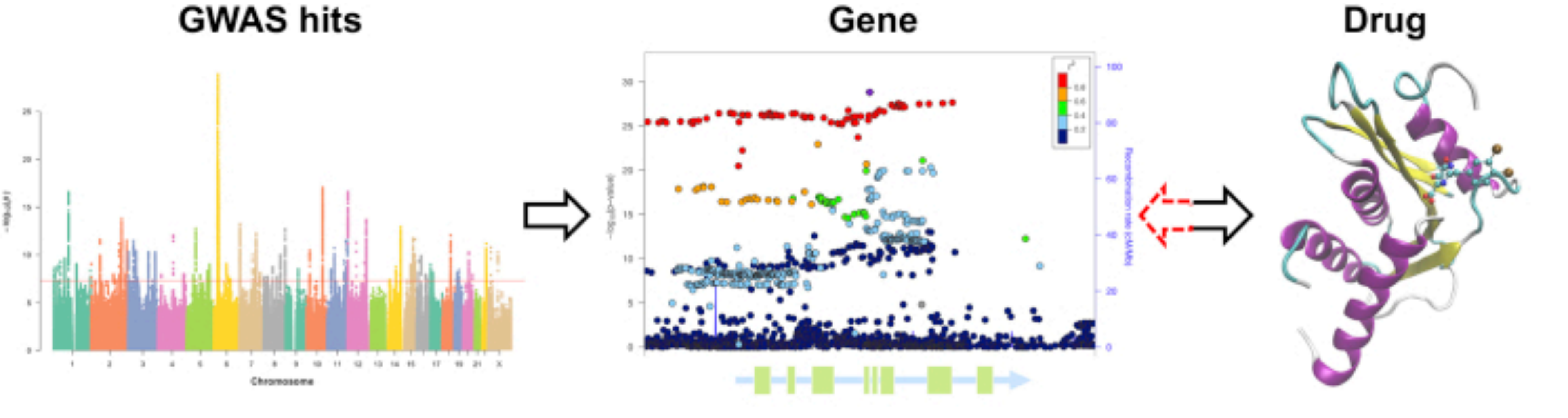


Single-marker associations. For each SNP, we tested for allelic association with disease status. To account for multiple testing, we used the Bonferroni correction and considered significant only those SNPs for which $P < 0.05/103{,}611 = 4.8 \times 10^{-7}$. This correction is known to be conservative and thus “overcorrected” the raw P values (*14*). Of the autosomal SNPs, only two, rs380390 and rs10272438, are significantly associated with disease status (Bonferroni-corrected $P = 0.0043$ and $P = 0.0080$, respectively) (Fig. 1A).

Visualization of Results

- Manhattan Plots
  - genome-wide p-values
- Locus Plots
  - gene-level visualization
- QQ Plots
  - assess bias/significance
- LD Plots
  - visualize local patterns of linkage disequilibrium





GWAS hits

Gene

Drug

Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

QUALITY CONTROL

Table 2: Steps in QC/QA for genome wide association studies as recommended by Laurie et al. (2010) Genet Epi
1. Genotyping batch quality (because all genotypes were re-called together for this study, these QC/QA steps may not be appropriate)
a. Median missing call rate of samples in a batch b. Allelic frequency difference relative to a pool of other batches c. Number of misidentified samples
2. Sample quality
a. Missing call rate over SNPs b. Allelic imbalance measure (“BAlleleFreq”) c. Median genotype confidence score d. Heterozygosity over all SNPs
3. Sample identity
a. Genetic versus annotated gender check b. Planned duplicate sample check c. Relatedness d. Ethnicity
4. Case-control confounding
a. Principal component differences b. Missing call rate differences
5. SNP quality
a. Missing call rate over samples b. Duplicate sample discordance c. Mendelian errors d. Hardy-Weinberg equilibrium e. Minor allele frequency

Subject level QC

SNP level QC

SUBJECT LEVEL QC

Missing genotype calls

The proportion of missing genotype calls for each individual:

- exclude samples that are missing more than 10% of their genotype calls as these are likely to be *low quality DNA* samples with error-ridden genotype calls.

```
$ plink --file GWAS --mind 0.10 --recode --out GWAS2
```

- See file GWAS_clean_mind.log to see how many samples are excluded based on this criteria.

Heterozygosity over all SNPs

Individuals with excessive heterozygosity could represent contamination across samples.

```
$ plink --file GWAS2 --het
```

--het computes observed and expected autosomal homozygous genotype counts for each sample to file plink.het

- reports F coefficient estimates:

[observed hom. count] - [expected count]) / ([total observations] - [expected count]))

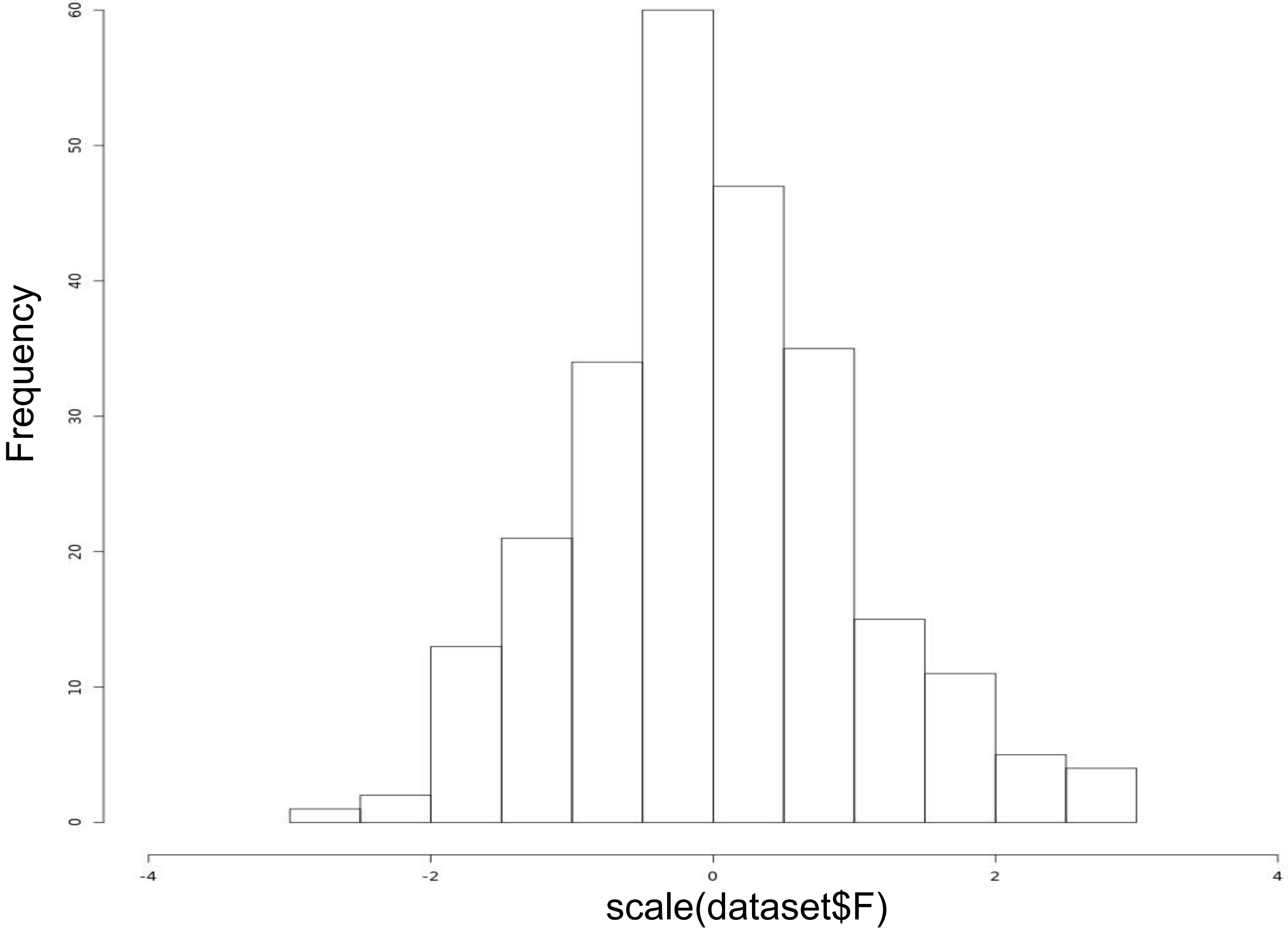
Expected count is based on allele freq.

Plotting heterozygosity

```
>  Dataset <- read.table("plink.het", header=TRUE,
sep="", na.strings="NA", dec=".”, strip.white=TRUE)

> mean(Dataset$F) #F measure of homozygosity
> sd(Dataset$F)
> jpeg("hist.jpeg", height=1000, width=1000)
> hist(scale(Dataset$F), xlim=c(-4,4))
> dev.off()
```

Histogram of Heterozygosity

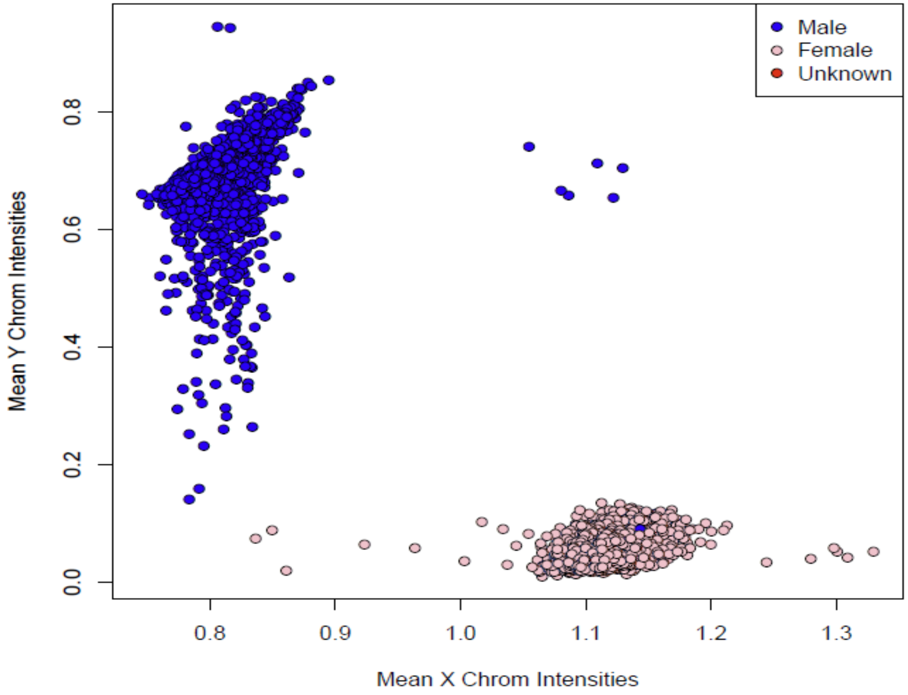


Gender Check

Using SNP genotypes to verify the gender of individuals:

- homozygosity (F) on the X chromosome in each individual: Female if < 0.2, male if > 0.8

```
$ plink --file GWAS2 --check-sex --out GWAS_sex_checking
```



Duplicates

Check if there are any duplicate samples in the dataset:

- Calculate IBS matrix between all members of the study.

```
$plink --file GWAS2 --genome --out duplicates

> dups = read.table(“duplicates.genome”, header = T)
> problem_pairs = dups[which(dups$PI_HAT > 0.4),]
> problem_pairs
```

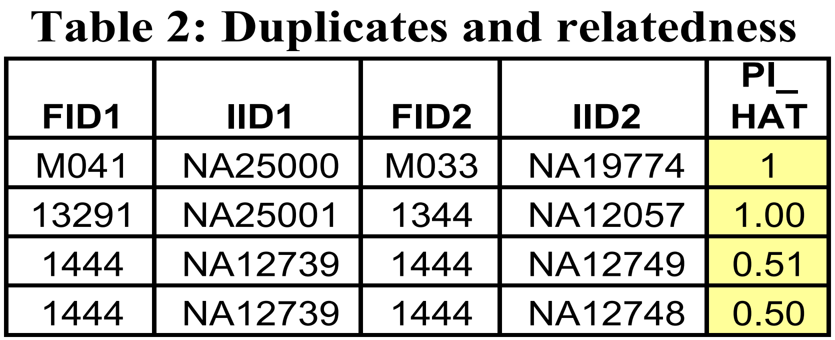
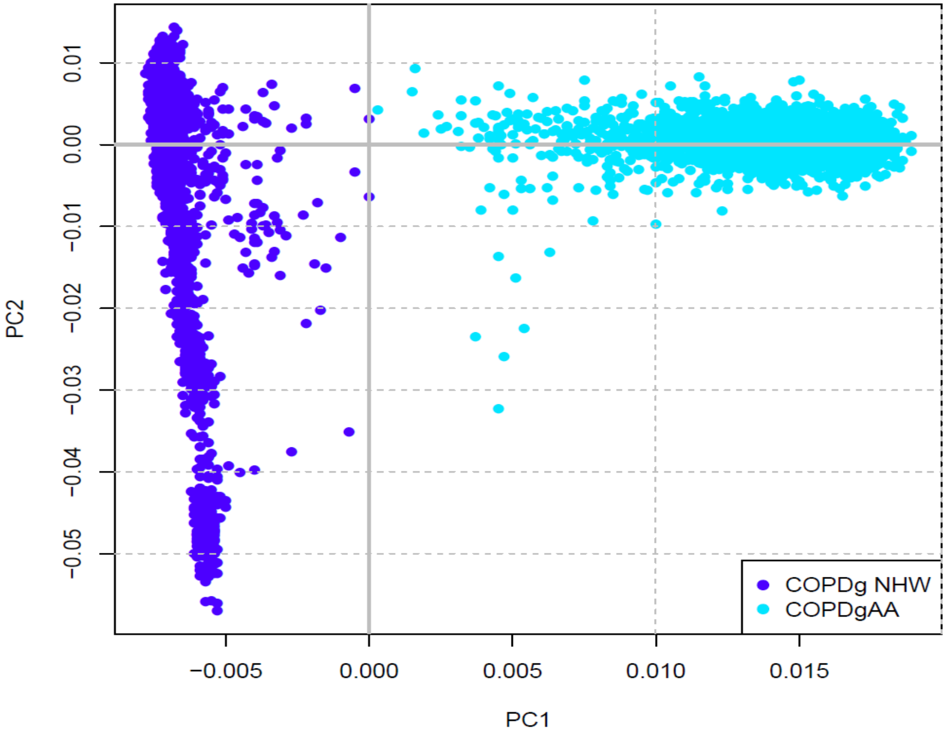


Table 2: Duplicates and relatedness

FID1	IID1	FID2	IID2	PI_HAT
M041	NA25000	M033	NA19774	1
13291	NA25001	1344	NA12057	1.00
1444	NA12739	1444	NA12749	0.51
1444	NA12739	1444	NA12748	0.50

Racial misclassification of individuals

Autosomal SNPs were selected for principal components analysis (PCA) using the following criteria: HWE p-value>0.01, MAF>0.05, and marker represented in HapMap III.



SNP LEVEL QC

Minor Allele Frequency (MAF)

Creating two versions of you dataset:

- One dataset consisting of SNPs with MAF > 0.05 and one with MAF < 0.05.

```
$ plink --file GWAS_clean_mind --maf 0.05 --recode --out MAF_greater_5
$ plink --file GWAS_clean_mind --exclude MAF_greater_5.map –recode --out MAF_less_5
```

Missingness by SNP

- Common SNPs (i.e. MAF≥5%) were flagged if they showed >5% missing calls.
- Less common SNPs (i.e. MAF <5%) were flagged if it had a missing rate >2%.

```
$ plink --file MAF_greater_5 --geno 0.05 --recode --out MAF_greater_5_clean
```

#--geno filters out all variants with missing call rates exceeding the provided value to be removed (similar to --mind for subjects)

Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg Principle is a mathematical model stating that the allele and genotype frequencies within a population will remain constant from generation to generation, in the absence of any other evolutionary influences. This model is only valid under a set of specific conditions:

1. Random mating
2. Infinitely large population
3. No mutations
4. No natural selection
5. No migration (immigration/emigration)

HWE		**Females**	
		A (p)	a (q)
Males	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	Aa (q^2)

$$p^2 + 2pq + q^2 = 1$$

Hardy-Weinberg Equilibrium (HWE)

In reality, this is not the case.

Allele and genotype frequencies change in all human populations worldwide. There are always alleles becoming more common, others becoming less common, some being lost entirely, and brand new alleles being created by mutation.

Therefore, there must be natural evolutionary mechanisms in play, such as:

- natural selection
- genetic drift
- mutations
- gene flow

Hardy-Weinberg Equilibrium (HWE)

- extreme deviations from HWE may be due to genotyping artifacts (p-values $< 10^{-7}$)

Hardy-Weinberg Equilibrium (HWE)

```
$ plink --file GWAS_clean3 --pheno pheno.txt --pheno-name Aff --hardy
```

--hardy writes a list of genotype counts and HW exact test statistics to plink.hwe

Open the file plink.hwe and look for SNPs with p-values of 10-7 or smaller.

```
> hardy = read.table(“plink.hwe”, header = T)
> names(hardy)
> hwe_prob = hardy[which(hardy$P < 0.0000009),]
```

Create a text file called “HWE_out.txt” with the SNPs from hwe_prob.

```
$ plink --file GWAS_clean3 --exclude HWE_out.txt --recode --out GWAS_clean4
```

Or

```
$ plink --file GWAS_clean3 --pheno pheno.txt --pheno-name Aff --hwe
  0.0000009
```

PLINK TUTORIAL