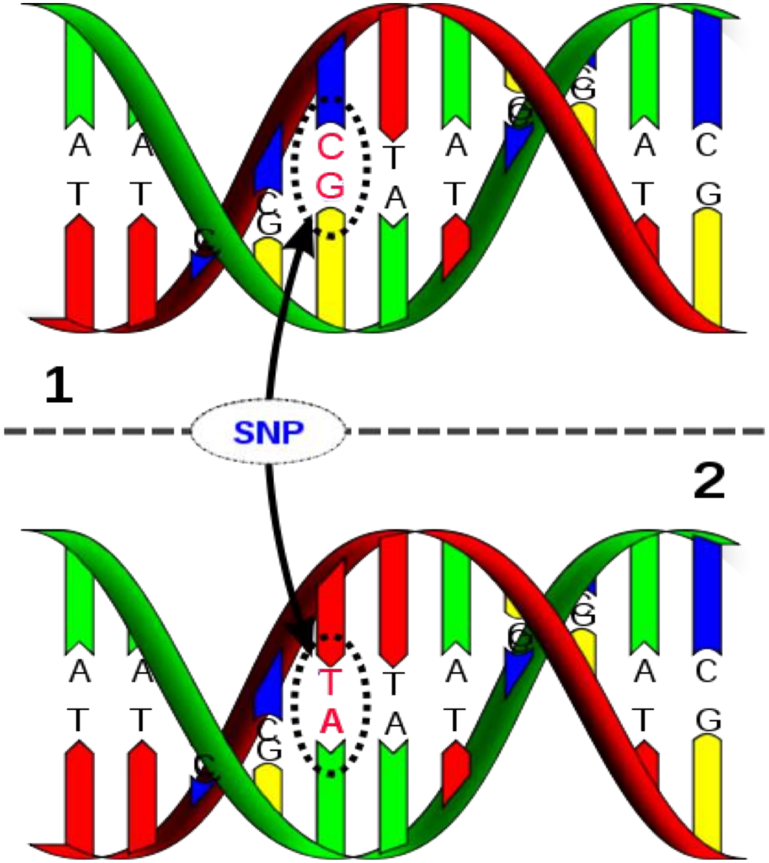
Genome-wide Association Study

02/11/2019

Qi Yan
Department of Pediatrics
Children’s Hospital of Pittsburgh of UPMC
University of Pittsburgh

Human Genome and Single Nucleotide Polymorphisms (SNPs)



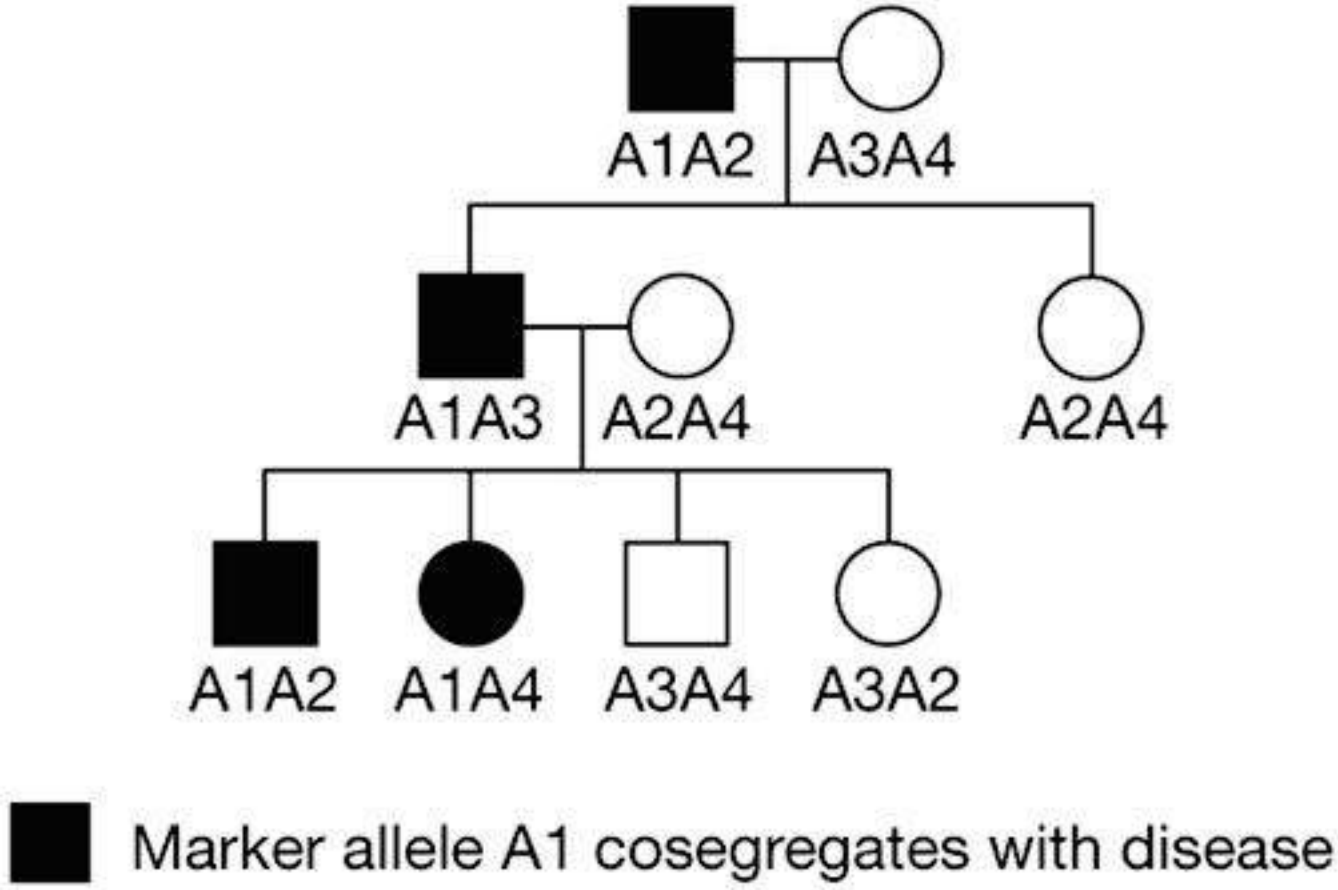
- 23 chromosome pairs
- 3 billion bases

- A single nucleotide change between pairs of chromosomes

- E.g.

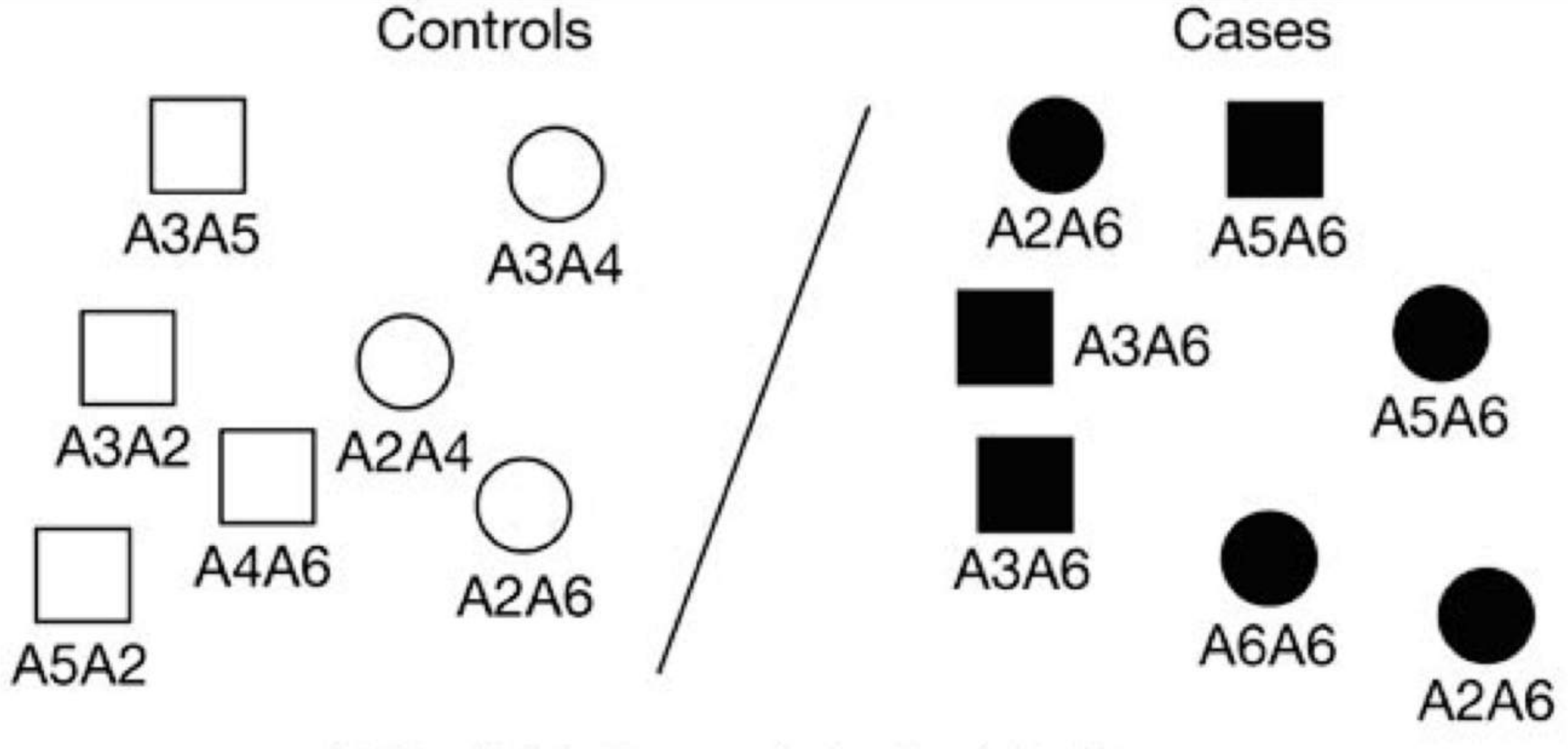
Haplotype1: AAGG**G**ATCCAC
Haplotype2: AAGG**A**ATCCAC

Where are Genes?



Kullo et al. *Nature Clinical Practice Cardiovascular Medicine* (2007)

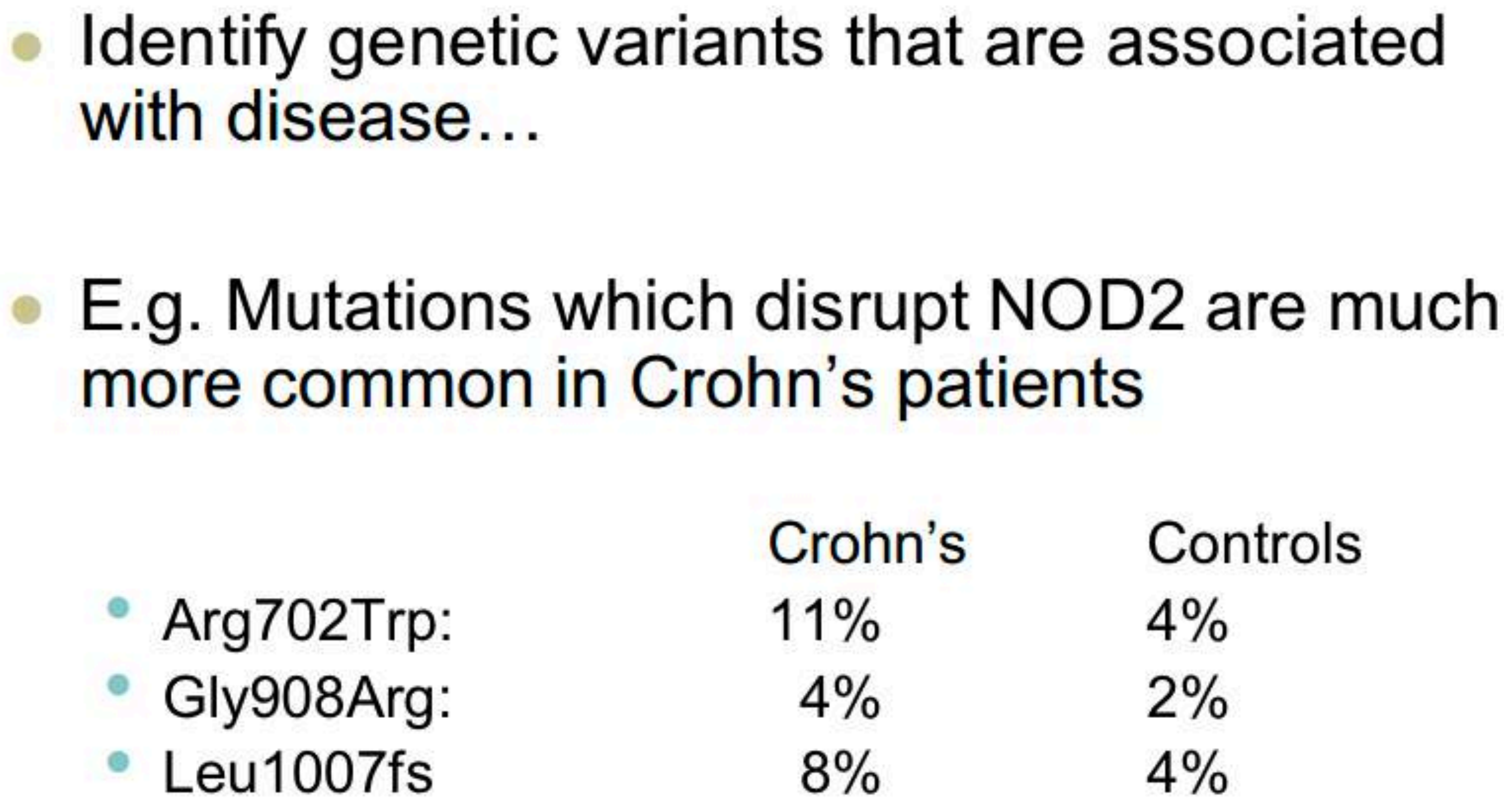
Association Study in Population



Allele A6 is ‘associated’ with disease

Kullo et al. *Nature Clinical Practice Cardiovascular Medicine* (2007)

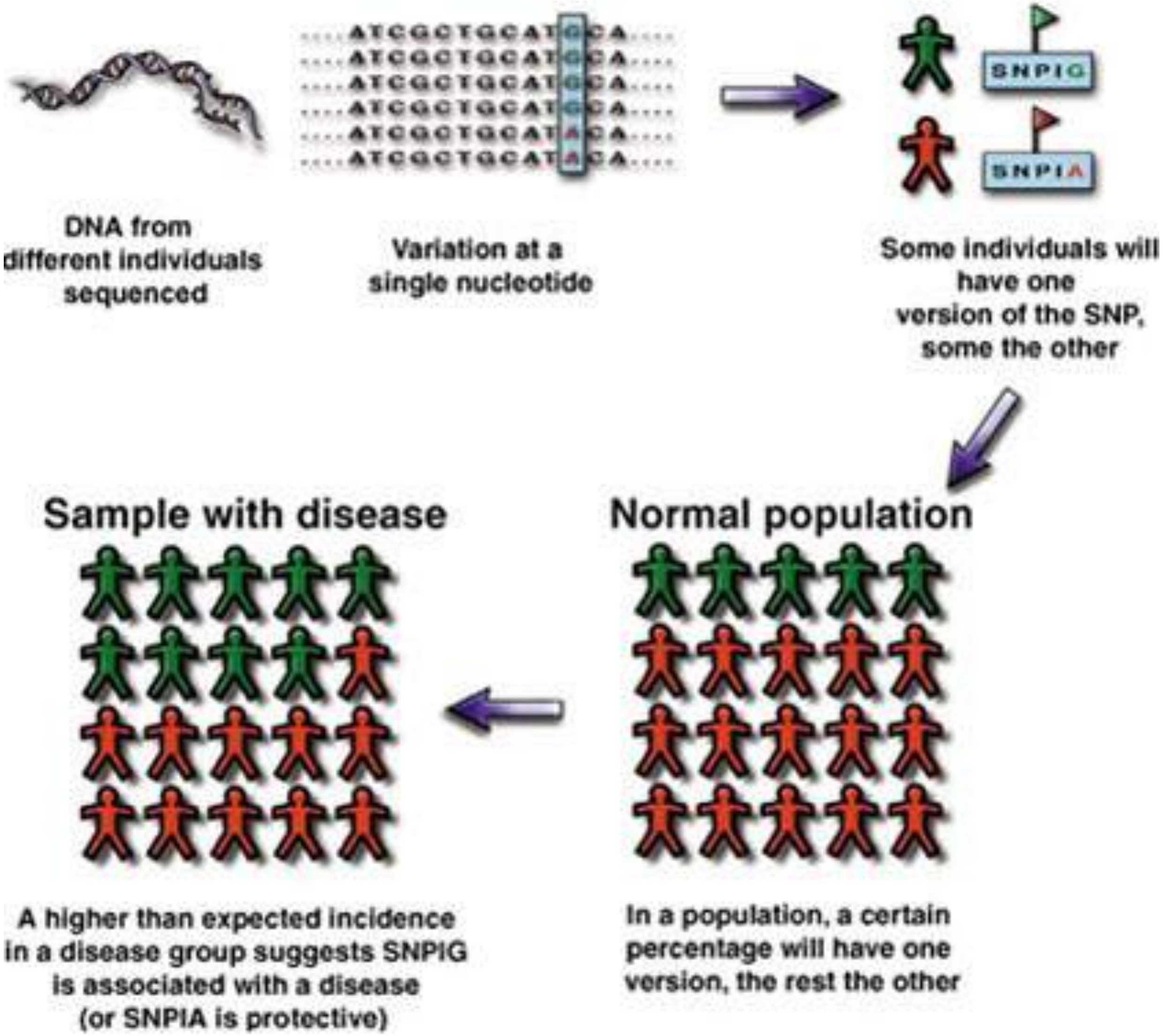
What are Genes?



- Identify genetic variants that are associated with disease…
- E.g. Mutations which disrupt NOD2 are much more common in Crohn's patients

	Crohn's	Controls
Arg702Trp:	11%	4%
Gly908Arg:	4%	2%
Leu1007fs	8%	4%

Using SNPs to Track Predisposition to Disease



DNA from different individuals sequenced

Variation at a single nucleotide

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

Normal population

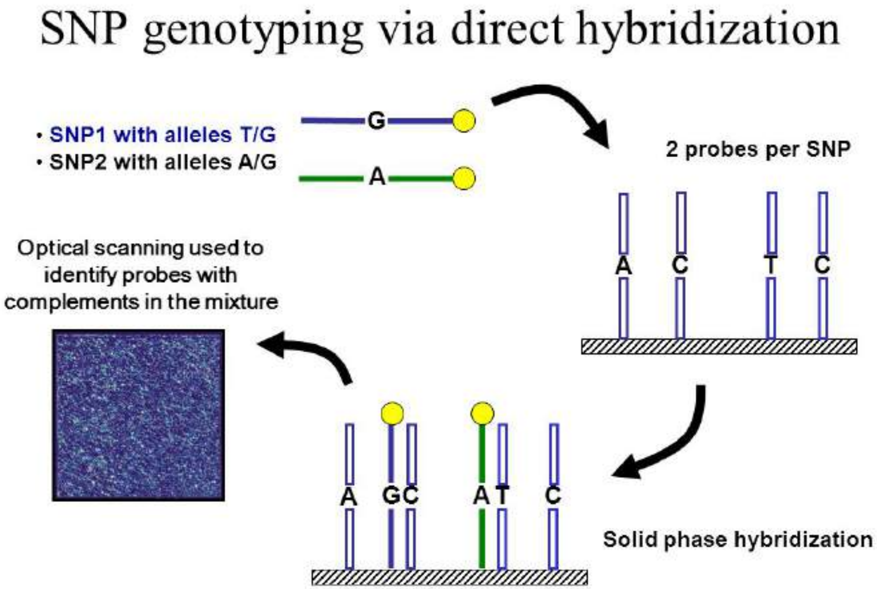
In a population, a certain percentage will have one version, the rest the other

Sample with disease

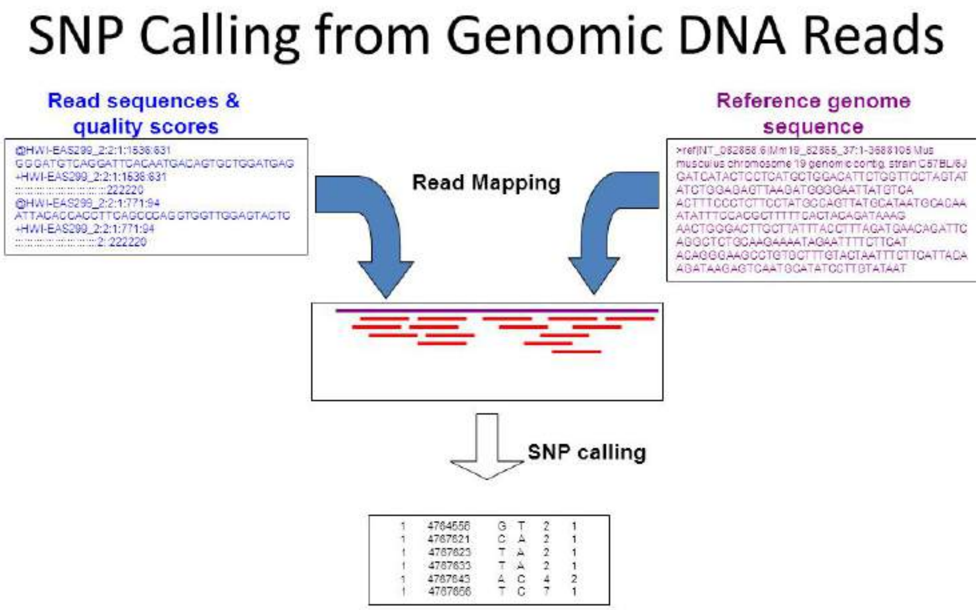
A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)

© Gibson & Muse, A Primer of Genome Science

SNP Data


SNP genotyping via direct hybridization
• SNP1 with alleles T/G
• SNP2 with alleles A/G
G
A
2 probes per SNP
A
C
T
C
Optical scanning used to identify probes with complements in the mixture
A
GC
AT
C
Solid phase hybridization

www.cd-genomics.com


SNP Calling from Genomic DNA Reads
Read sequences & quality scores
Reference genome sequence
Read Mapping
SNP calling

```
@HWI-EAS299_2:2:1:1536:631
GGGATGTCAGGATTCACAATGACAGTGCTGGATGAG
+HWI-EAS299_2:2:1:1536:631
::::::::::::::::::::::::::222220
@HWI-EAS299_2:2:1:771:94
ATTACACCACCTTCAGCCCAGGTGGTTGGAGTACTC
+HWI-EAS299_2:2:1:771:94
::::::::::::::::::::::2:222220
```

```
>ref|NT_082868.6|Mm19_82865_37:1-3688105 Mus
musculus chromosome 19 genomic contig, strain C57BL/6J
GATCATACTCCTCATGCTGGACATTCTGGTTCCTAGTAT
ATCTGGAGAGTTAAGATGGGGAATTATGTCA
ACTTTCCCTCTTCCTATGCCAGTTATGCATAATGCACAA
ATATTTCCACGCTTTTTCACTACAGATAAAG
AACTGGGACTTGCTTATTTACCTTTAGATGAACAGATTC
AGGCTCTGCAAGAAAATAGAATTTTCTTCAT
ACAGGGAAGCCTGTGCTTTGTACTAATTTCTTCATTACA
AGATAAGAGTCAATGCATATCCTTGTATAAT
```

```
1    4764558    G  T   2   1
1    4767821    C  A   2   1
1    4767823    T  A   2   1
1    4767833    T  A   2   1
1    4767843    A  C   4   2
1    4767856    T  C   7   1
```

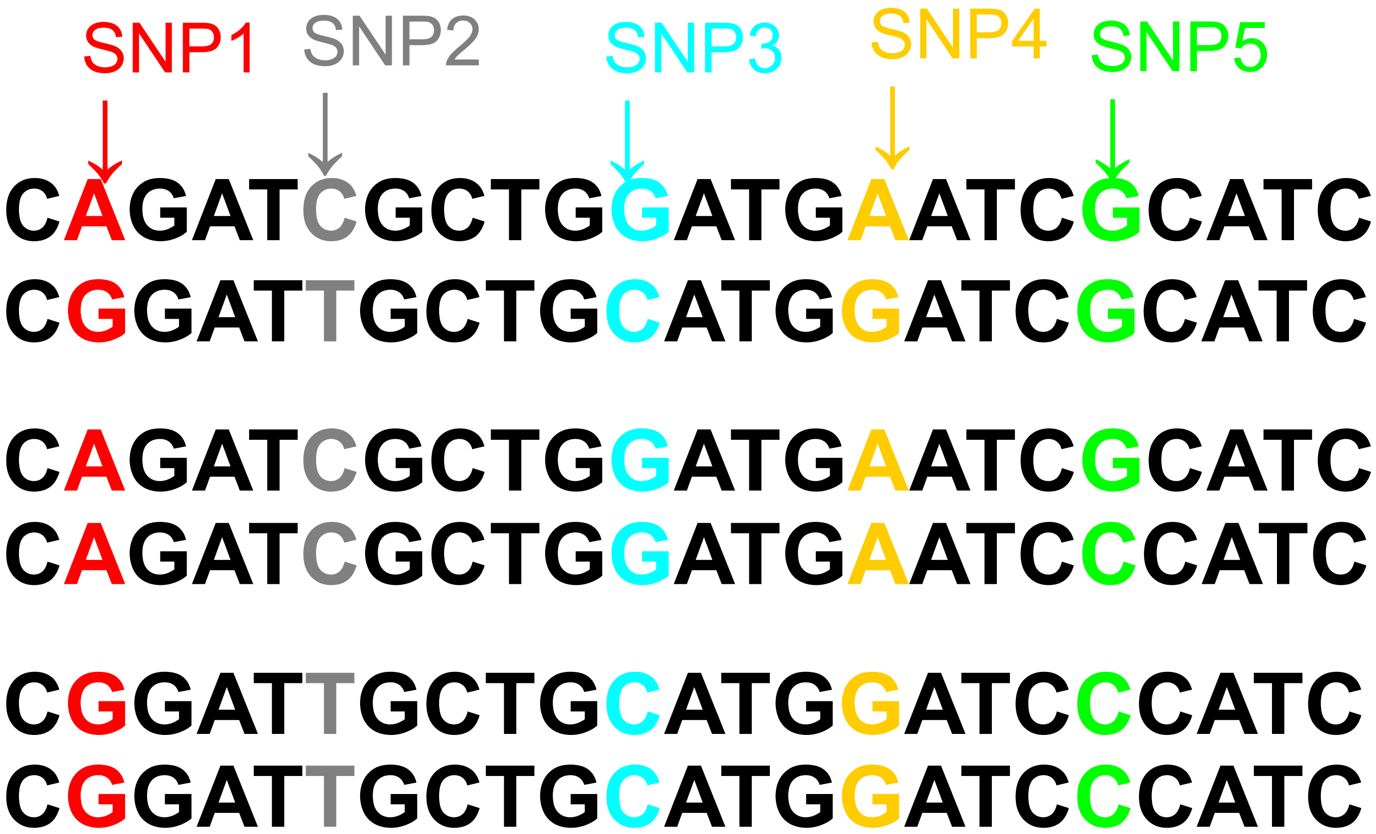
BIBM 2011 Tutorial

	SNP1	SNP2	SNP3	…
Sample1	A/A	G/G	G/G	…
Sample2	A/A	A/G	G/G	…
Sample3	A/C	G/G	G/G	…
…	…	…	…	…

OR

	SNP1	SNP2	SNP3	…
Sample1	0	2	2	…
Sample2	0	1	2	…
Sample3	1	2	2	…
…	…	…	…	…

SNPs in Population



CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

Association Study in Case Control Samples

SNP1 SNP2 SNP3 SNP4 SNP5

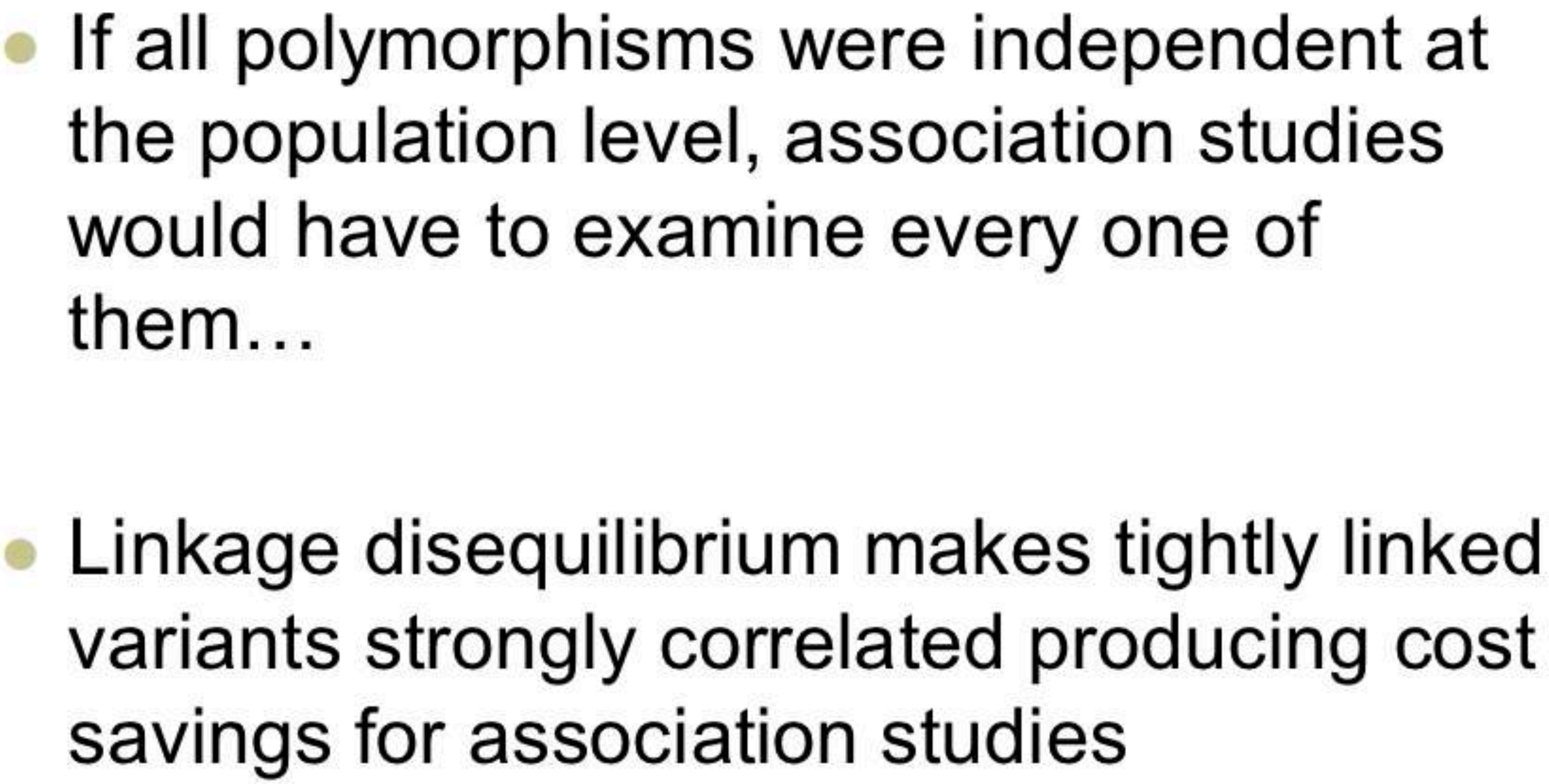
Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

Association Studies and Linkage Disequilibrium



- If all polymorphisms were independent at the population level, association studies would have to examine every one of them…

- Linkage disequilibrium makes tightly linked variants strongly correlated producing cost savings for association studies

Linkage Disequilibrium (LD)

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

$$p_{AB} = p_A p_B$$
$$p_{Ab} = p_A p_b = p_A(1 - p_B)$$
$$p_{aB} = p_a p_B = (1 - p_A) p_B$$
$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B)$$

$$D_{AB} = p_{AB} - p_A p_B$$
$$p_{AB} = p_A p_B + D_{AB}$$
$$p_{Ab} = p_A p_b - D_{AB}$$
$$p_{aB} = p_a p_B - D_{AB}$$
$$p_{ab} = p_a p_b + D_{AB}$$

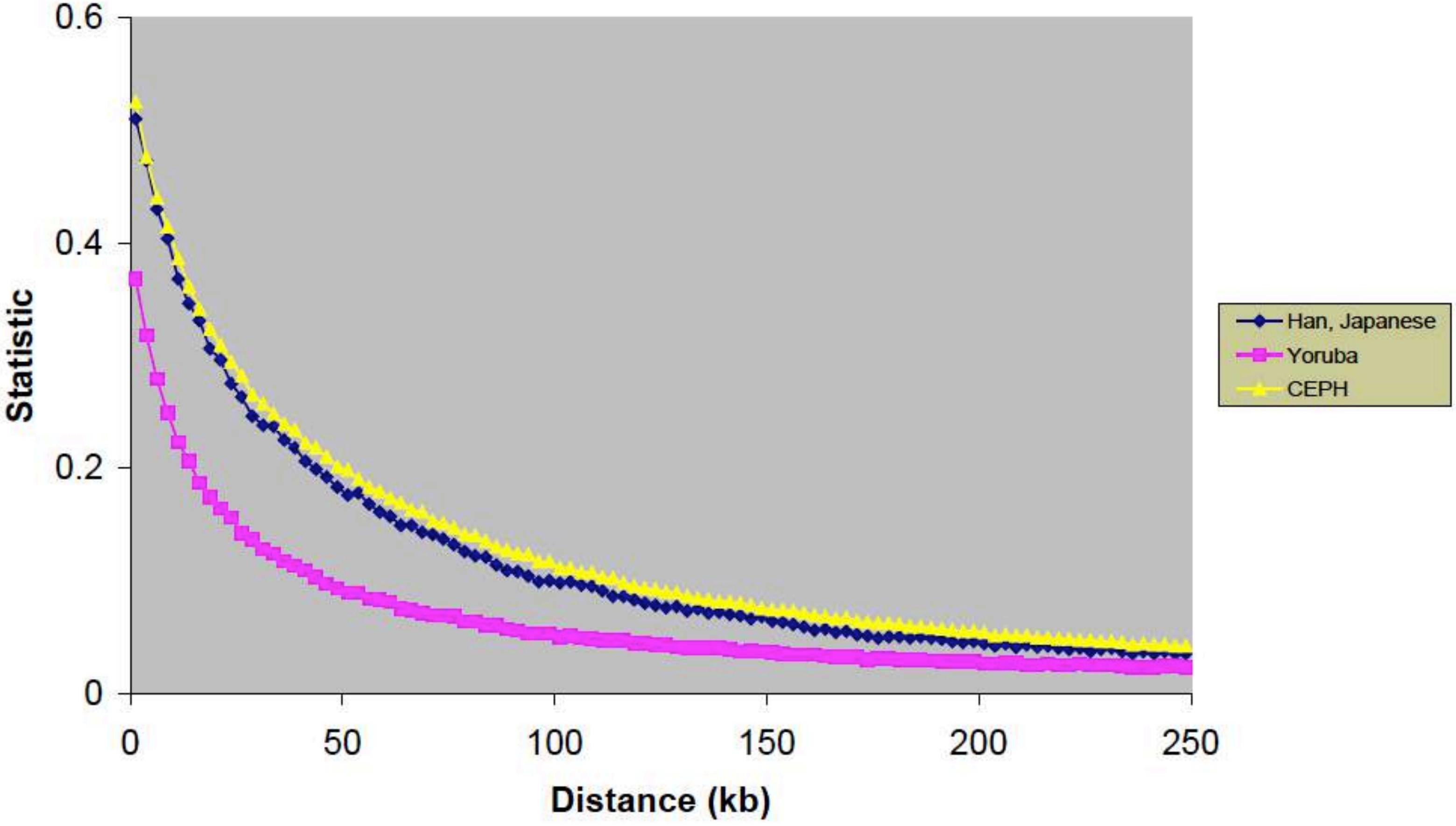
$\Downarrow$

$$D_{AB} = p_{AB} p_{ab} - p_{Ab} p_{aB}$$

$$\Delta^2 = \frac{D_{AB}^2}{p_A(1 - p_A) p_B(1 - p_B)}$$

- Ranges between 0 and 1
  - 1 when the two markers provide identical information
  - 0 when they are in perfect equilibrium

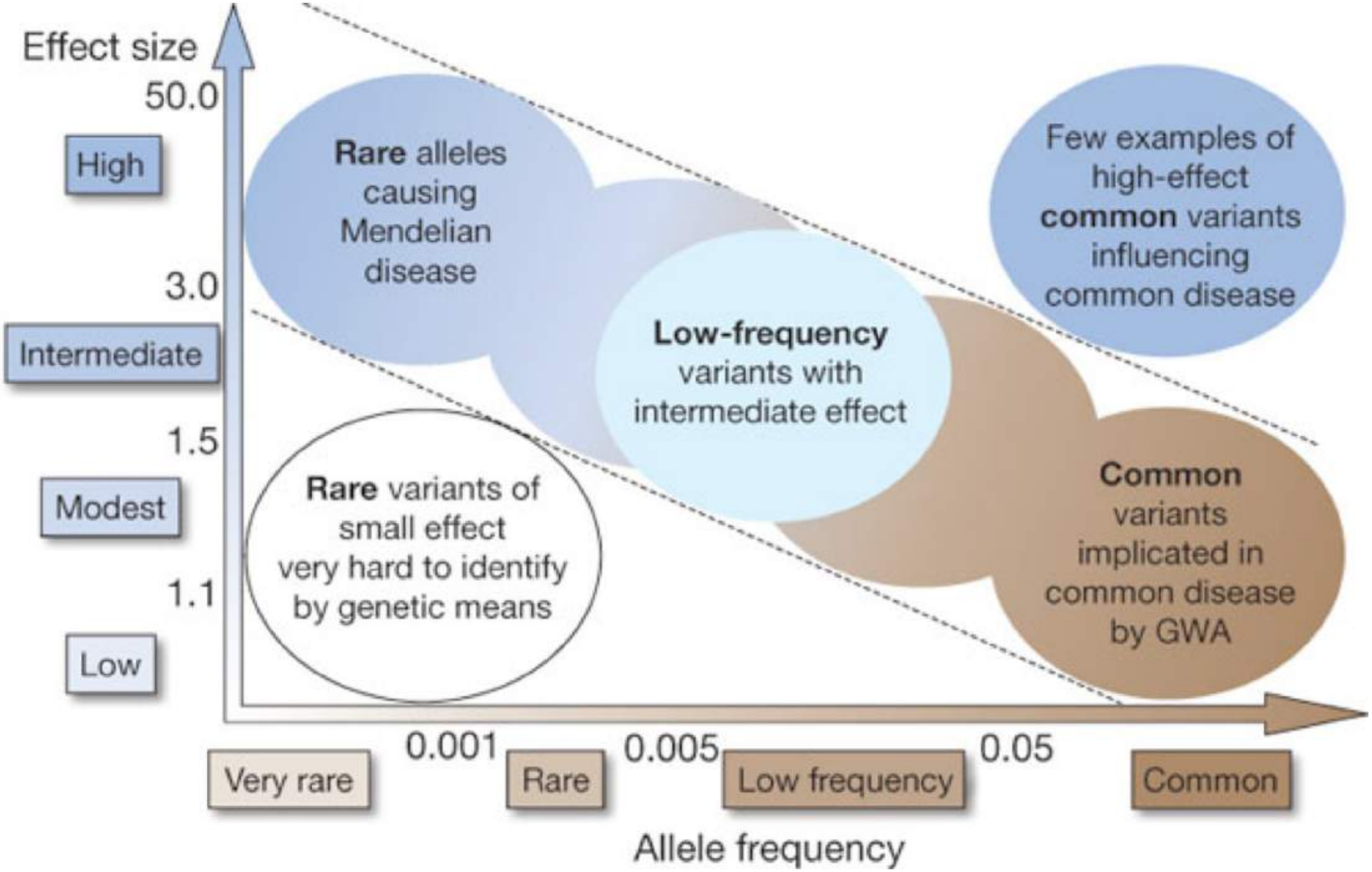
LD in Population



LD extends further in CEPH and the Han/Japanese than in the Yoruba

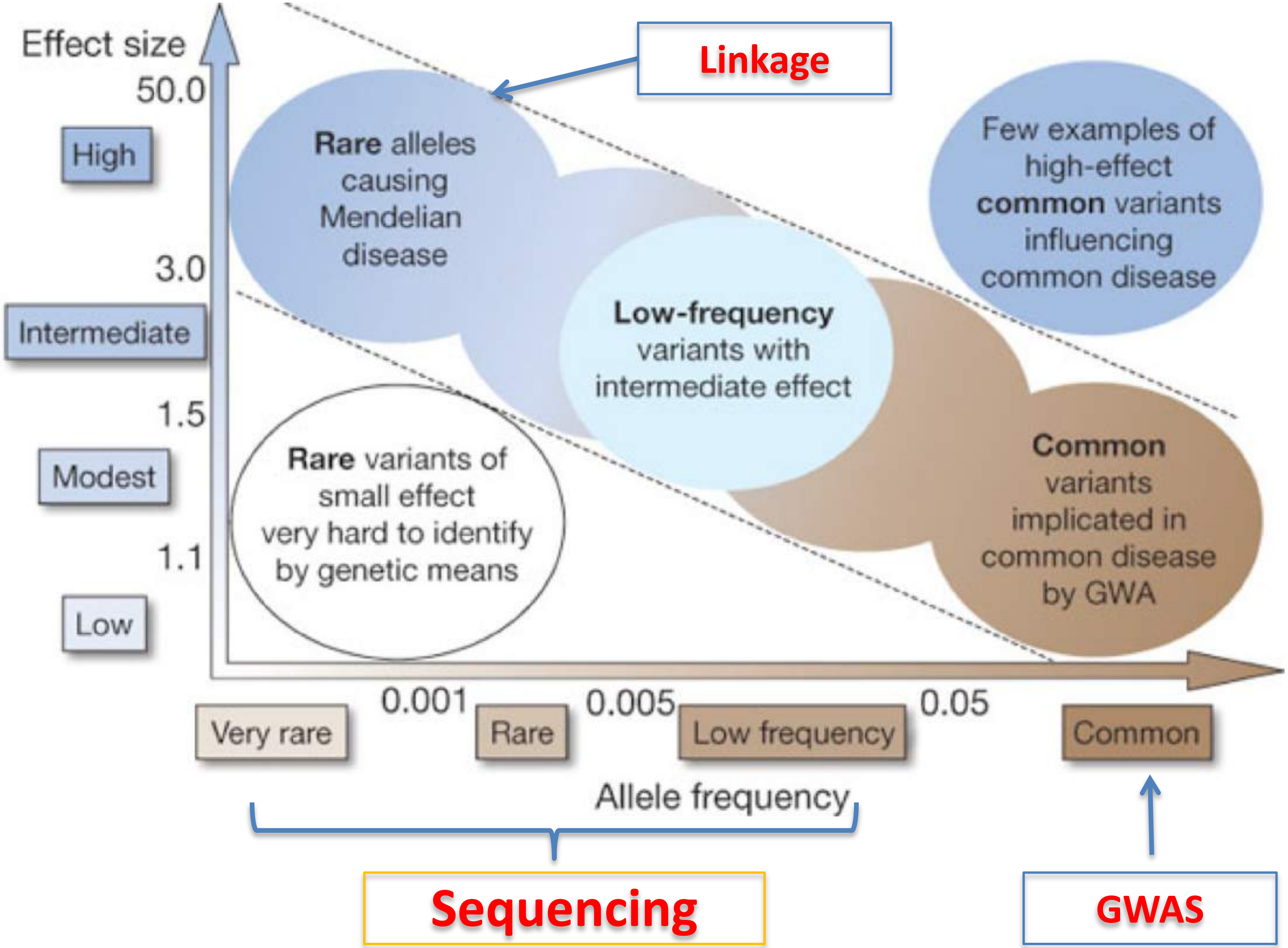
International HapMap Consortium, *Nature,* 2005

Genetic Spectrum of Complex Diseases


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Manolio et.al. 2009 Nature 461, 747-753

Genetic Spectrum of Complex Diseases



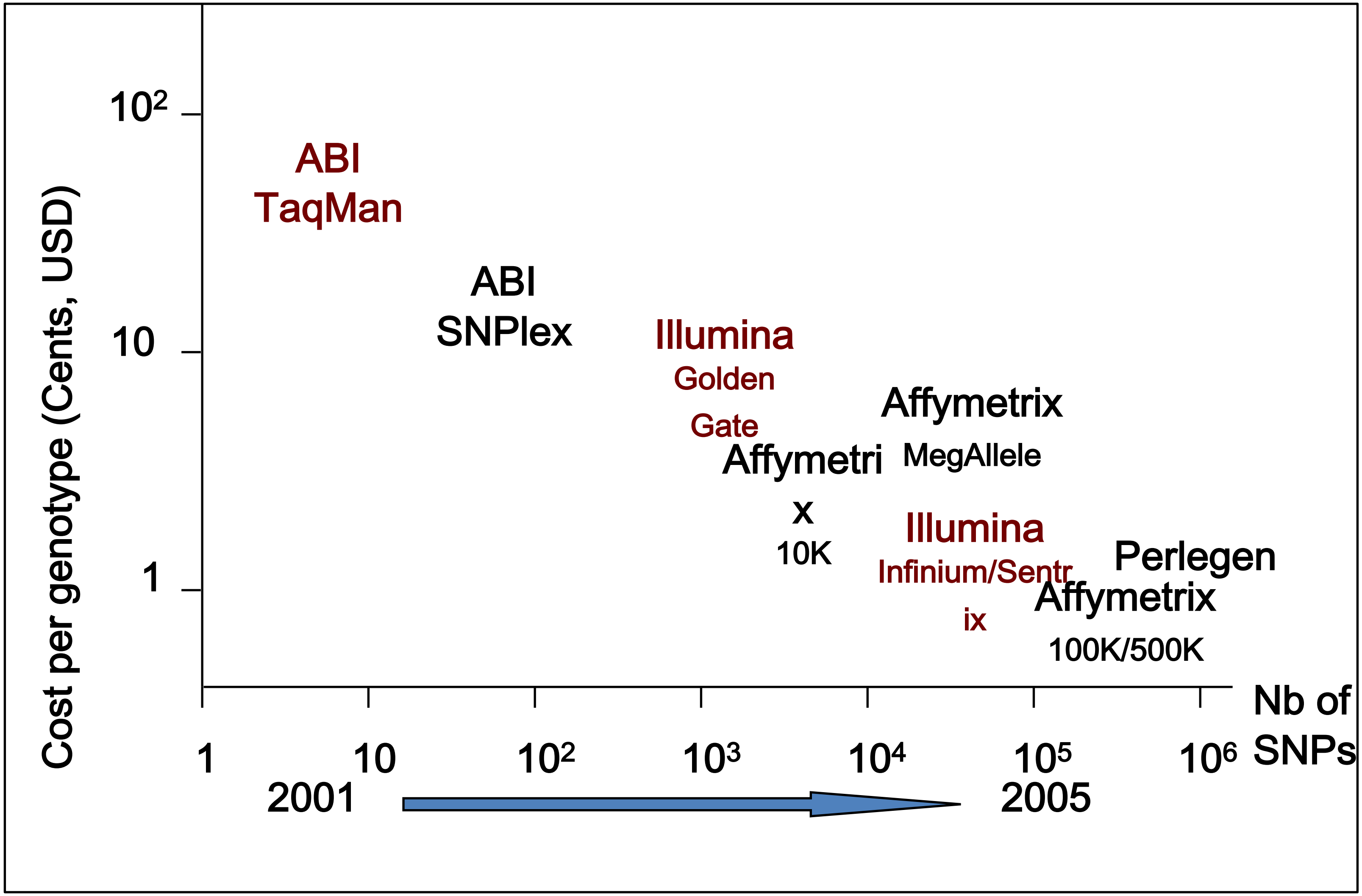
Linkage

Sequencing

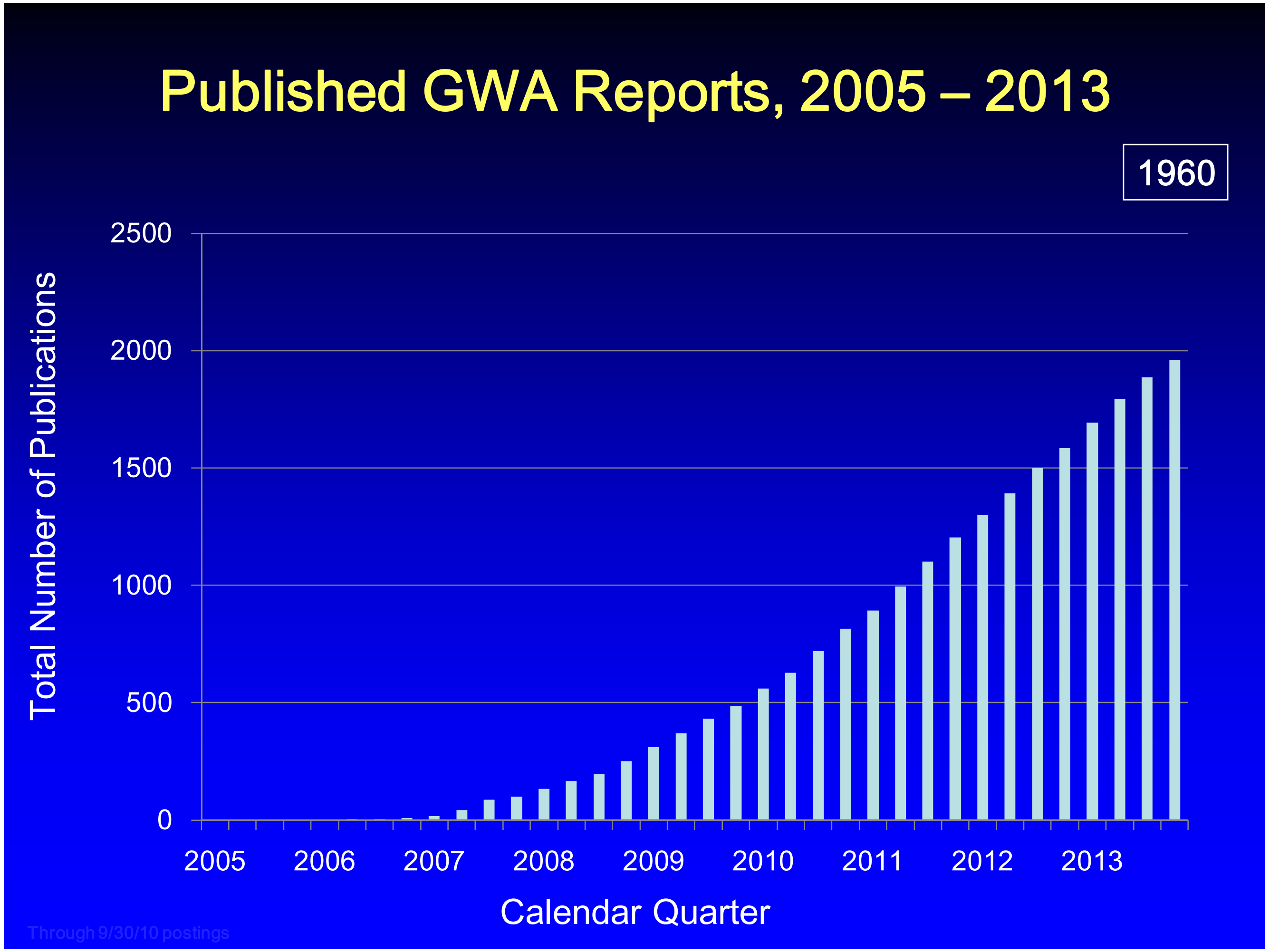
GWAS

Genome-Wide Association Study

Progress in Genotyping Technologies

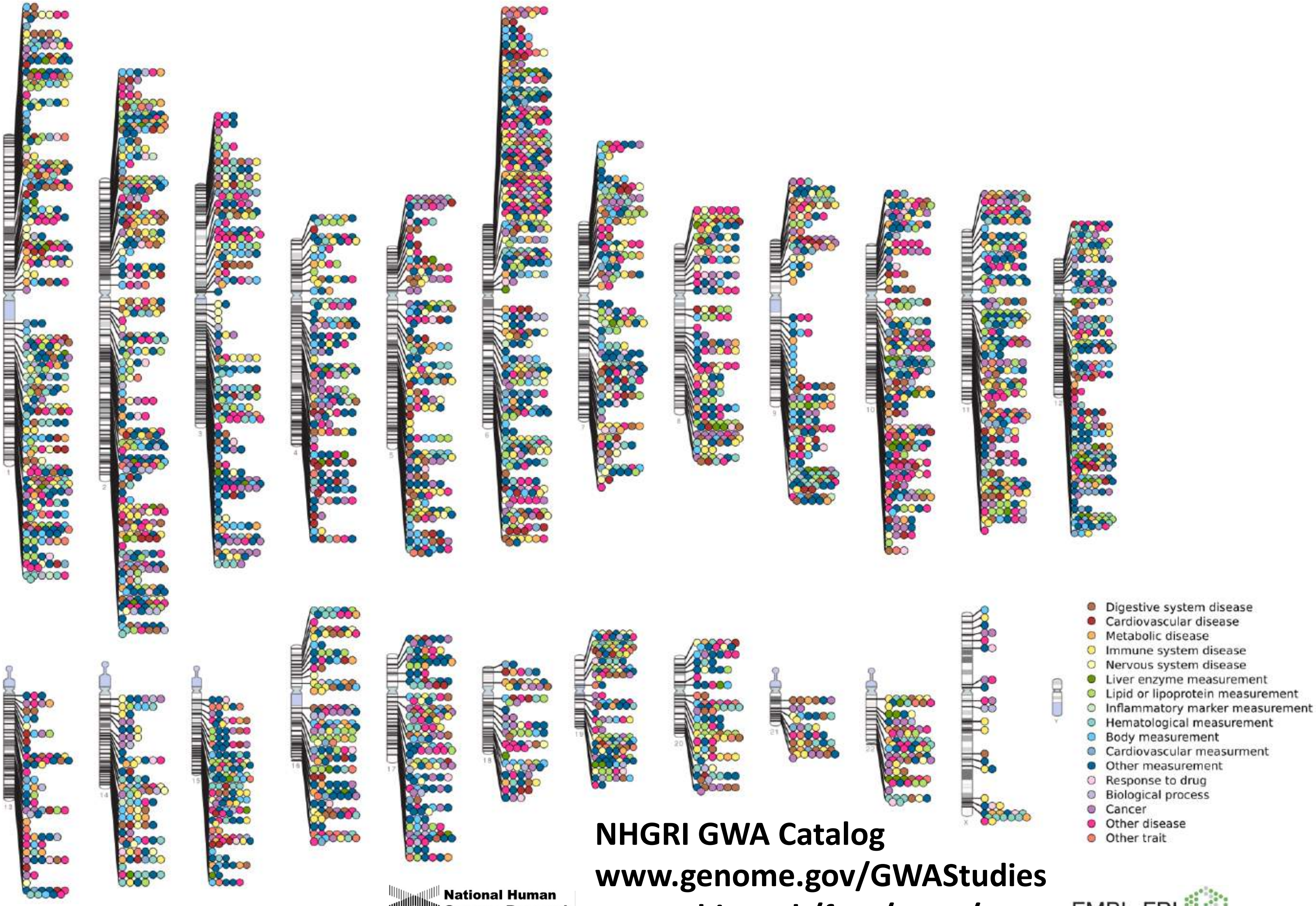


Publications



http://www.genome.gov/

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

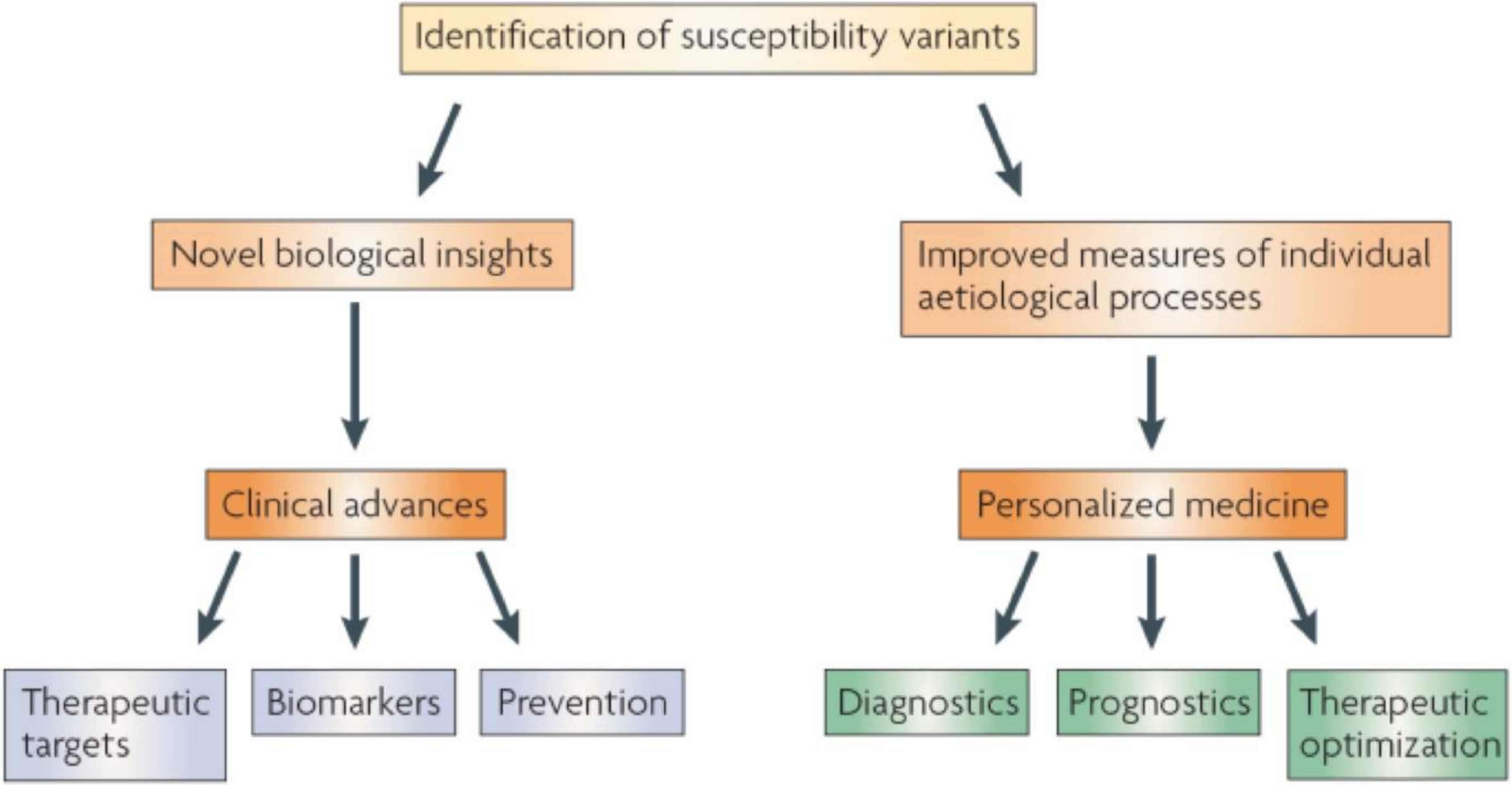


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Clinical translation of findings from GWAS



Identification of susceptibility variants

Novel biological insights

Improved measures of individual aetiological processes

Clinical advances

Personalized medicine

Therapeutic targets

Biomarkers

Prevention

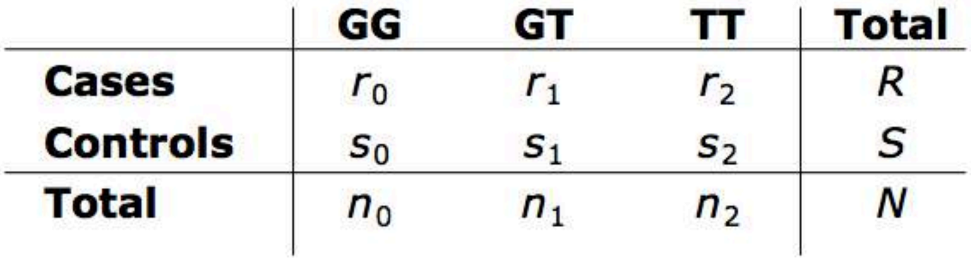
Diagnostics

Prognostics

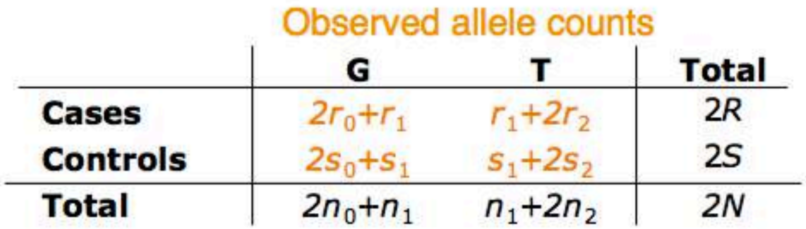
Therapeutic optimization

McCarthy, M. et al. 2008 *Nature Reviews Genetics*

Allelic Test for Association

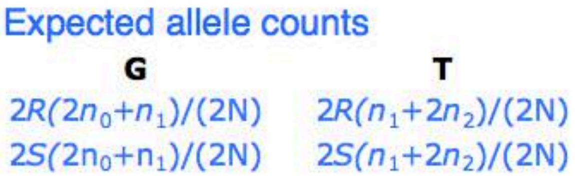


	GG	GT	TT	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N



Observed allele counts

	G	T	Total
Cases	$2r_0+r_1$	r_1+2r_2	$2R$
Controls	$2s_0+s_1$	s_1+2s_2	$2S$
Total	$2n_0+n_1$	n_1+2n_2	$2N$



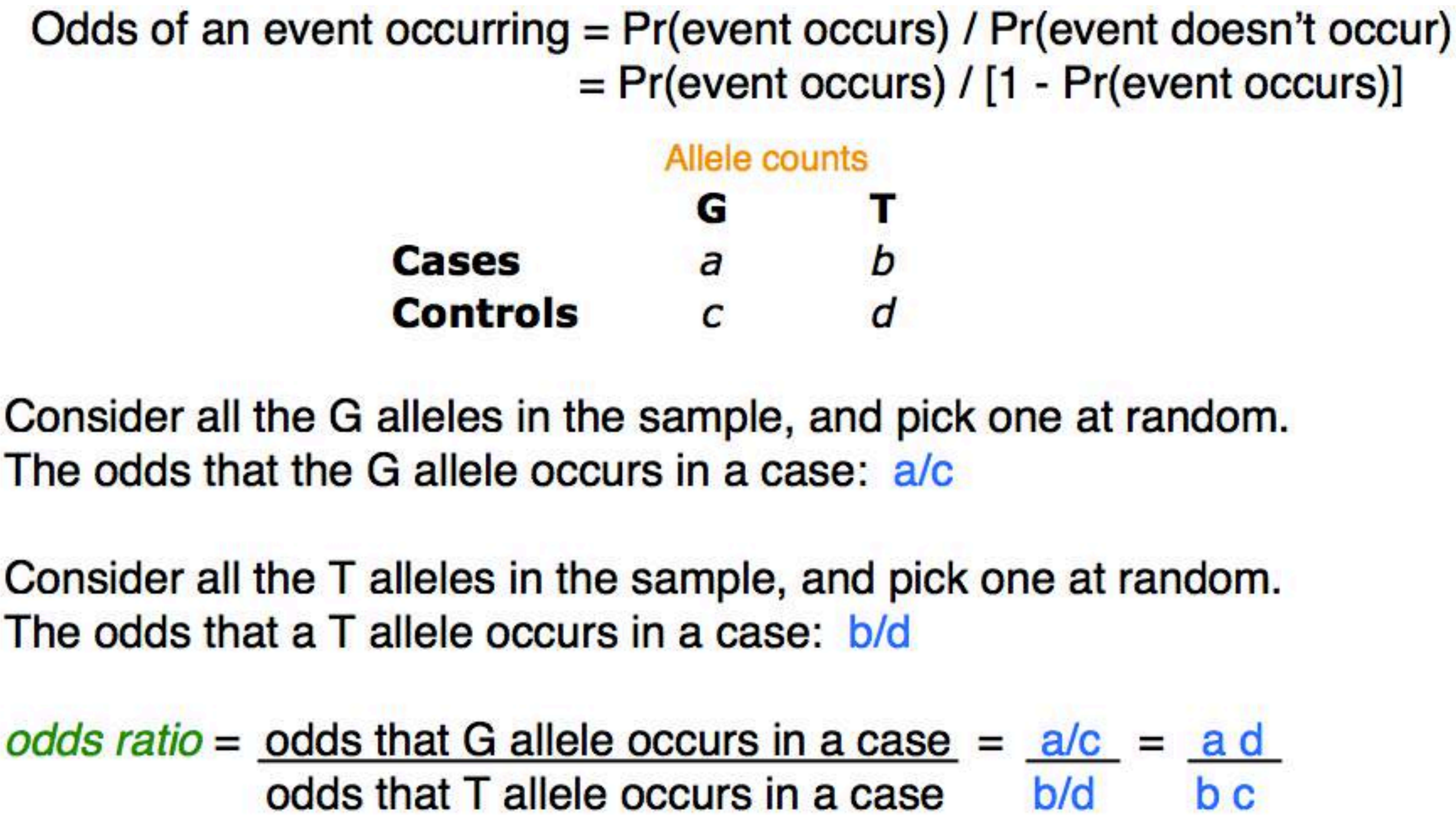
Expected allele counts

G	T
$2R(2n_0+n_1)/(2N)$	$2R(n_1+2n_2)/(2N)$
$2S(2\mathrm{n}_0+\mathrm{n}_1)/(2N)$	$2S(n_1+2n_2)/(2N)$

Chi-square test for independence of rows and columns (null hypothesis):

$$\sum \frac{(\mathrm{Obs} - \mathrm{Exp})^2}{\mathrm{Exp}} \sim \chi^2 \text{ with 1 df}$$

Odds Ratio



Odds of an event occurring = Pr(event occurs) / Pr(event doesn't occur)
= Pr(event occurs) / [1 - Pr(event occurs)]

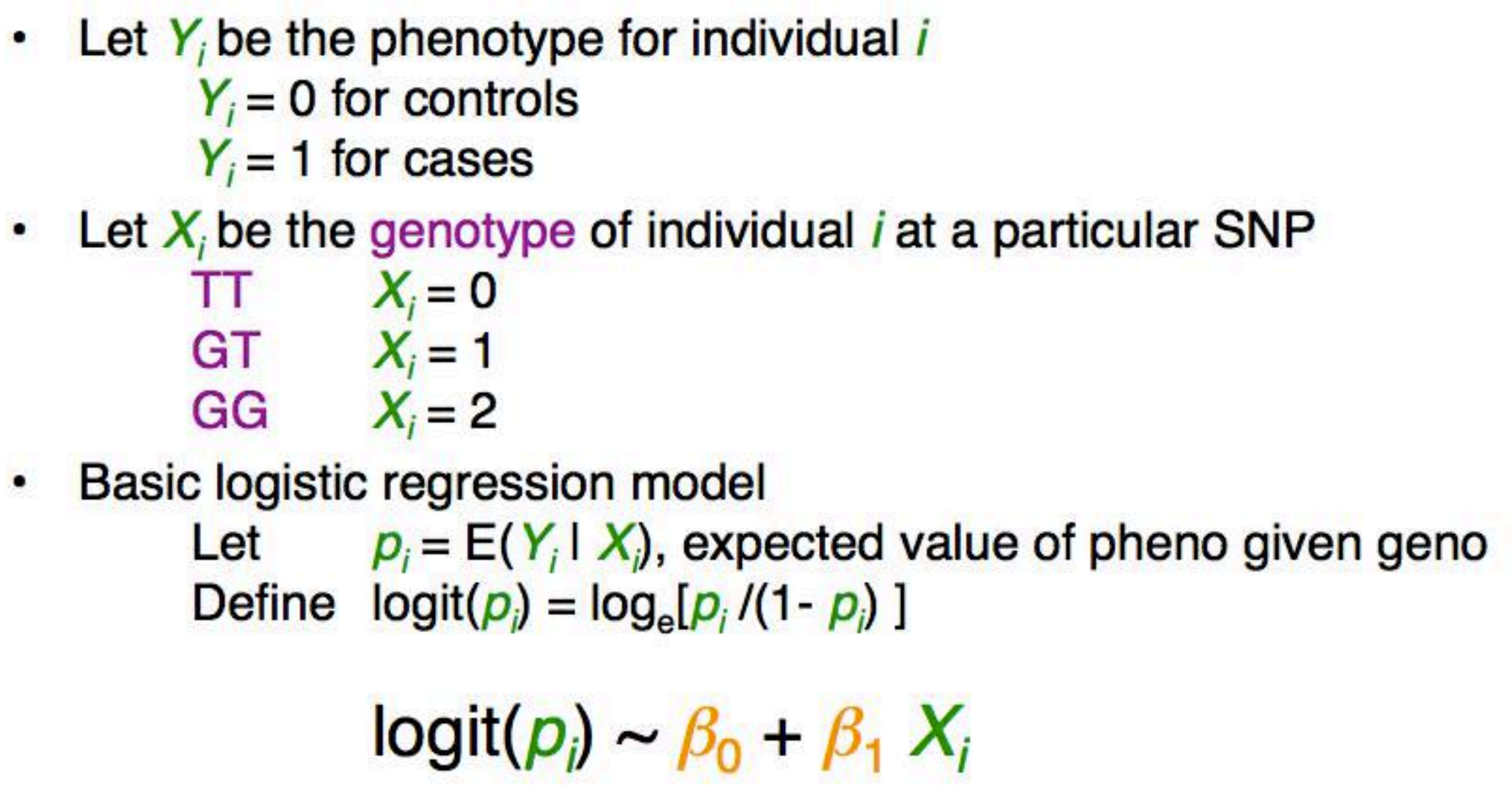
	Allele counts	
	G	**T**
Cases	*a*	*b*
Controls	*c*	*d*

Consider all the G alleles in the sample, and pick one at random.
The odds that the G allele occurs in a case: a/c

Consider all the T alleles in the sample, and pick one at random.
The odds that a T allele occurs in a case: b/d

$$\textit{odds ratio} = \frac{\text{odds that G allele occurs in a case}}{\text{odds that T allele occurs in a case}} = \frac{a/c}{b/d} = \frac{a\,d}{b\,c}$$

Logistic regression



- Let Y_i be the phenotype for individual i

  $Y_i = 0$ for controls

  $Y_i = 1$ for cases

- Let X_i be the genotype of individual i at a particular SNP

TT	$X_i = 0$
GT	$X_i = 1$
GG	$X_i = 2$

- Basic logistic regression model

  Let $p_i = E(Y_i \mid X_i)$, expected value of pheno given geno

  Define $\text{logit}(p_i) = \log_e[p_i /(1- p_i)]$

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i$$

Analogously, linear regression for continuous phenotype:

$$Y_i = \beta_0 + \beta_1 X_i$$

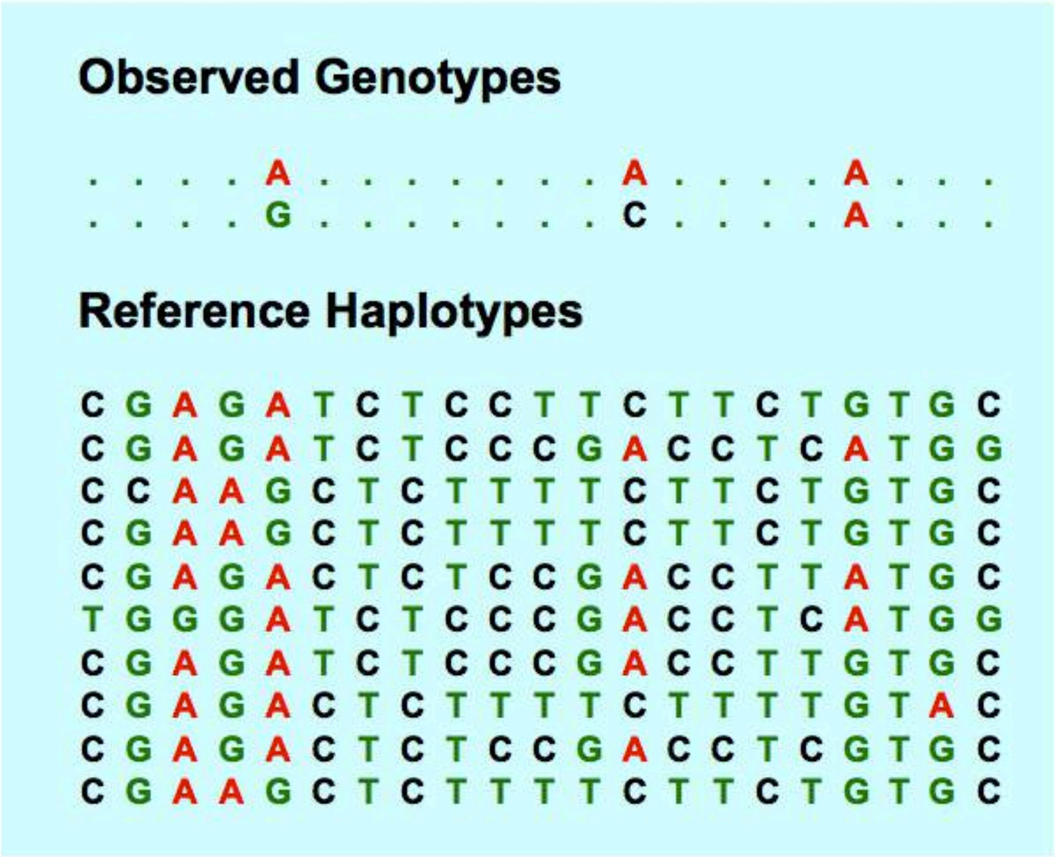
Genotype Imputation

- Use genotypes at a few markers to infer genotypes at other unobserved markers

- Closely related individuals
  - Long segments of identify by descent

- Distantly related individuals
  - Shorter segments of identify by descent

Genotype Imputation for unrelated Individuals



Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

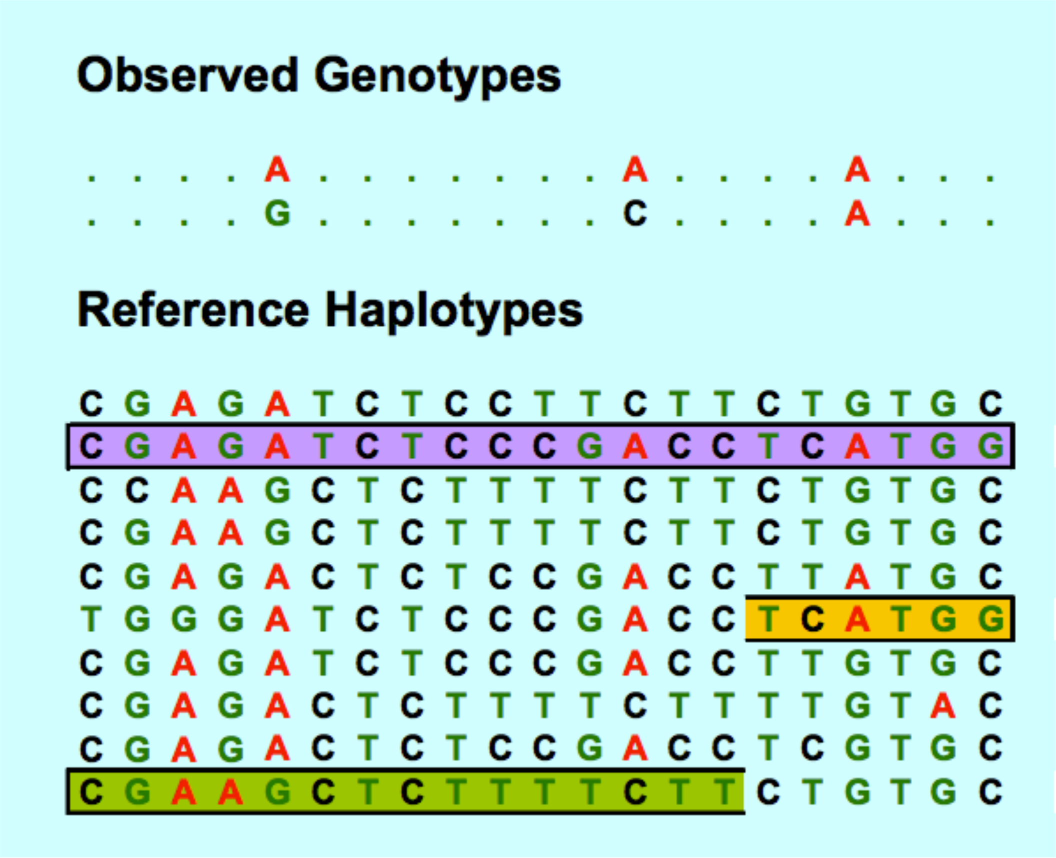
Study Sample

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

HapMap

Identify Match Among Reference



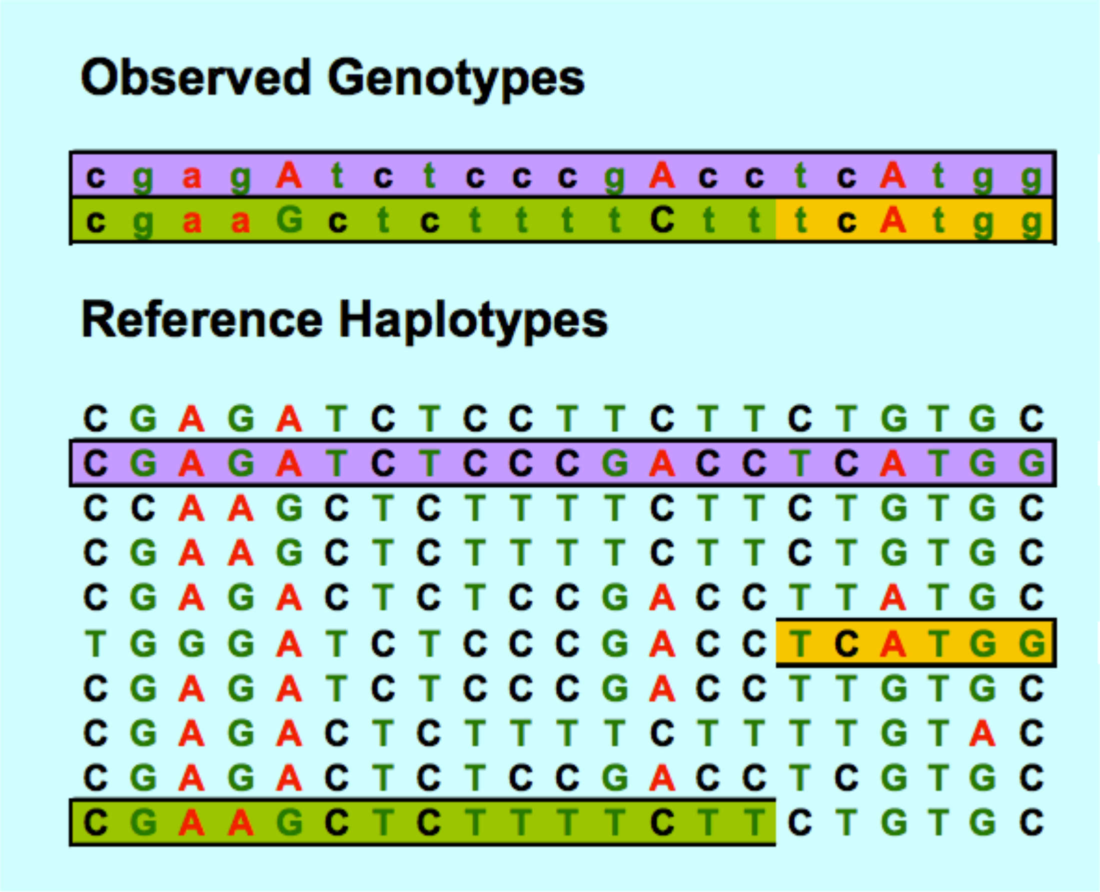
Observed Genotypes

. . . . A . . . . A . . . . . A . . .
. . . . G . . . . C . . . . . A . . .

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

Impute Missing Genotypes and Phase Chromosome



Observed Genotypes

c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

Implementation

- Markov model is used to model each haplotype, conditional on all others

- At each position, we assume the haplotype being modeled copies as a template haplotype

- Each individual has two haplotypes, and therefore copies two template haplotypes

Does This Really Work?

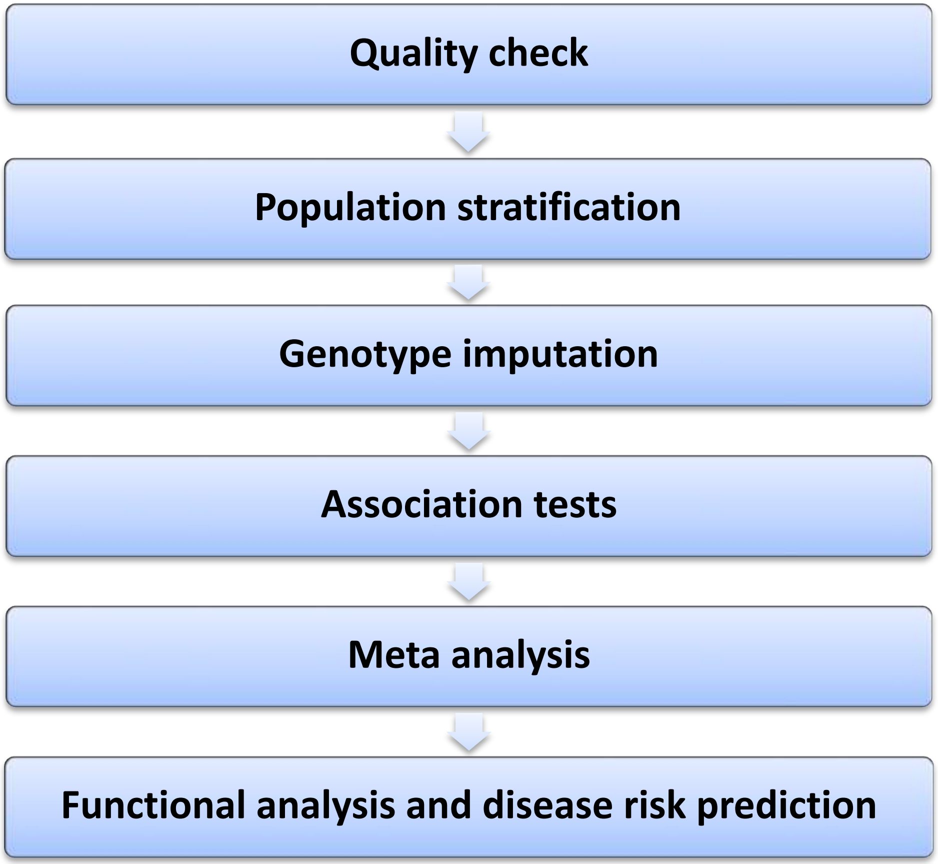
- Used about ~300,000 SNPs from Illumina HumanHap300 to impute 2.1M HapMap SNPs in 2500 individuals from a study of type II diabetes

- Compared imputed genotypes with actual experimental genotypes in a candidate region on chromosome 14
  - 1190 individuals, 521 markers not on Illumina chip

- Results of comparison
  - Average r^2 with true genotypes 0.92 (median 0.97)
  - 1.4% of imputed alleles mismatch original
  - 2.8% of imputed genotypes mismatch
  - Most errors concentrated on worst 3% of SNPs

Scott et al, *Science*, 2007

GWAS Workflow



Quality check

Population stratification

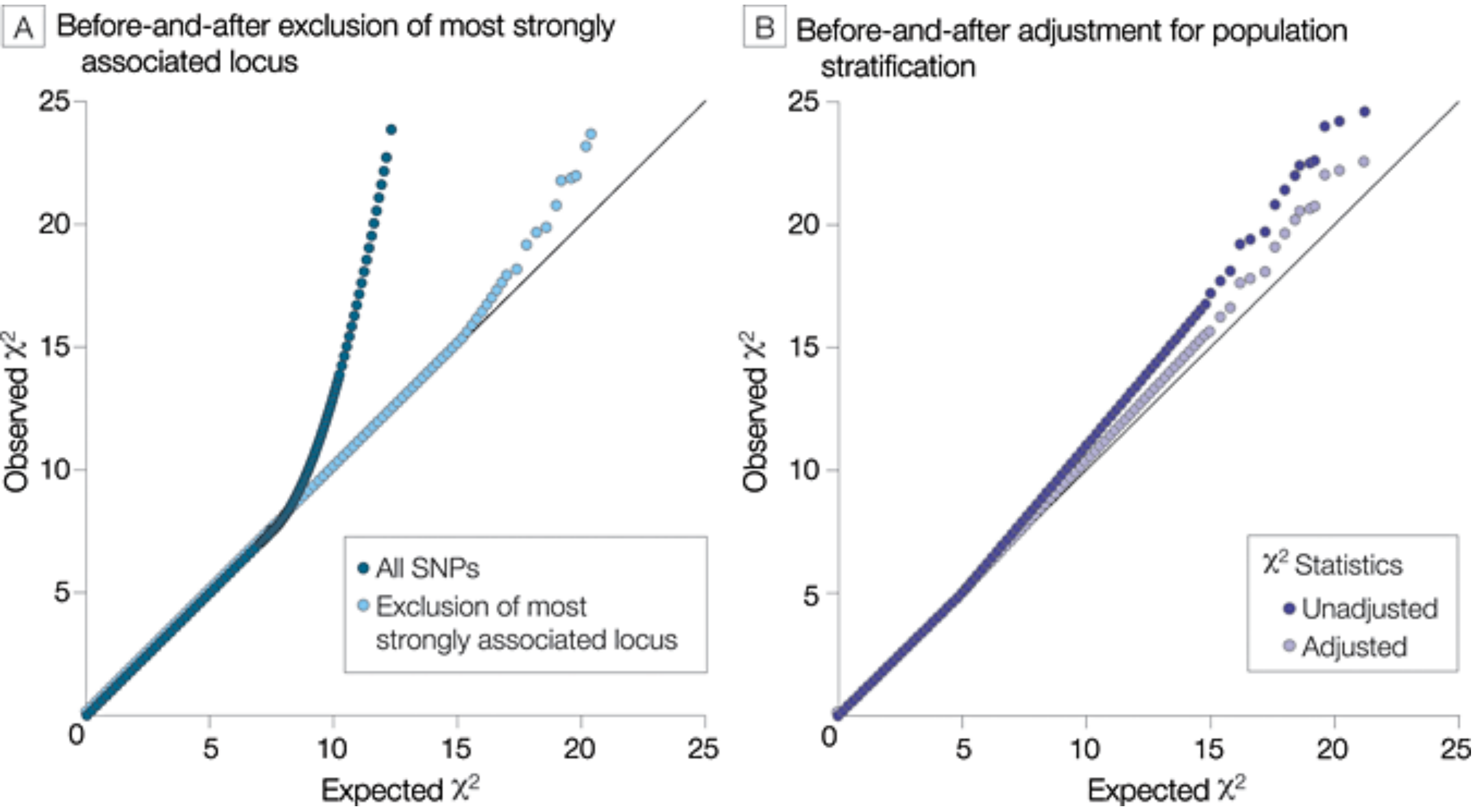
Genotype imputation

Association tests

Meta analysis

Functional analysis and disease risk prediction

Hypothetical Quantile-Quantile Plots in Genome-wide Association Studies



Pearson, T. A. et al. JAMA 2008;299:1335-1344

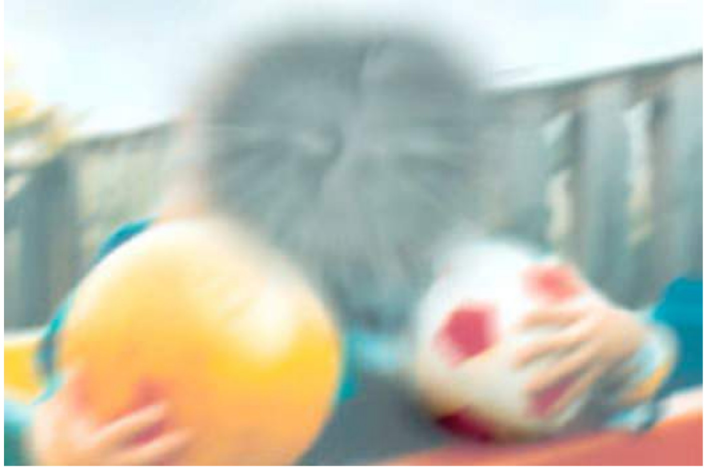
A Successful Example

Age Related Macular Degeneration (AMD)

- Progressive neurodegenerative disorder which leads to a loss of vision through the death of photoreceptors and/or retinal pigment epithelium (RPE) in the macula
- Late stage of the disease is associated with a debilitating loss of central vision and/or blindness

Images from National Eye Institute (http://www.nei.nih.gov)

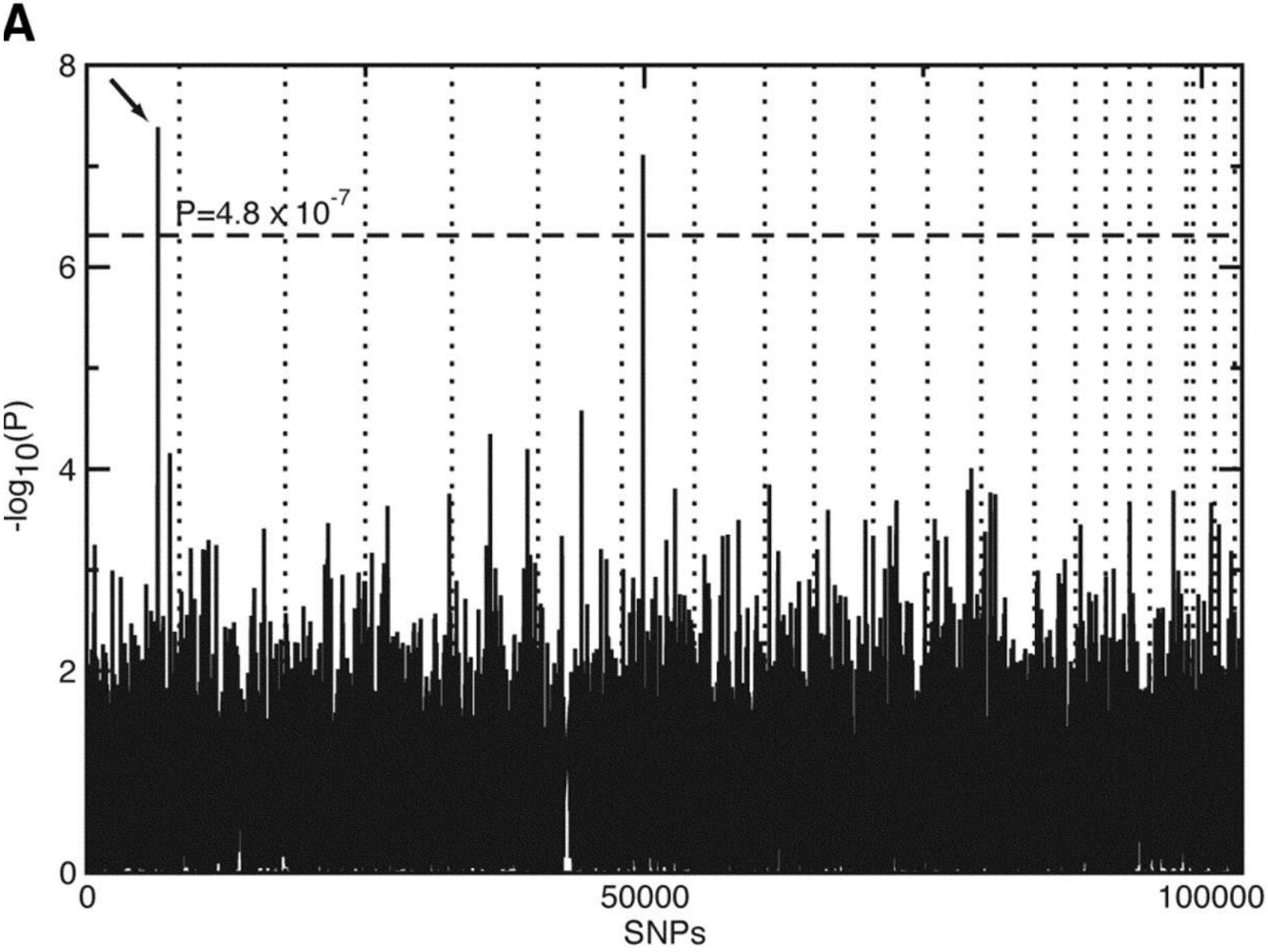
Normal Vision



Advanced AMD impairment

First GWAS of Age-related Macular Degeneration (AMD)

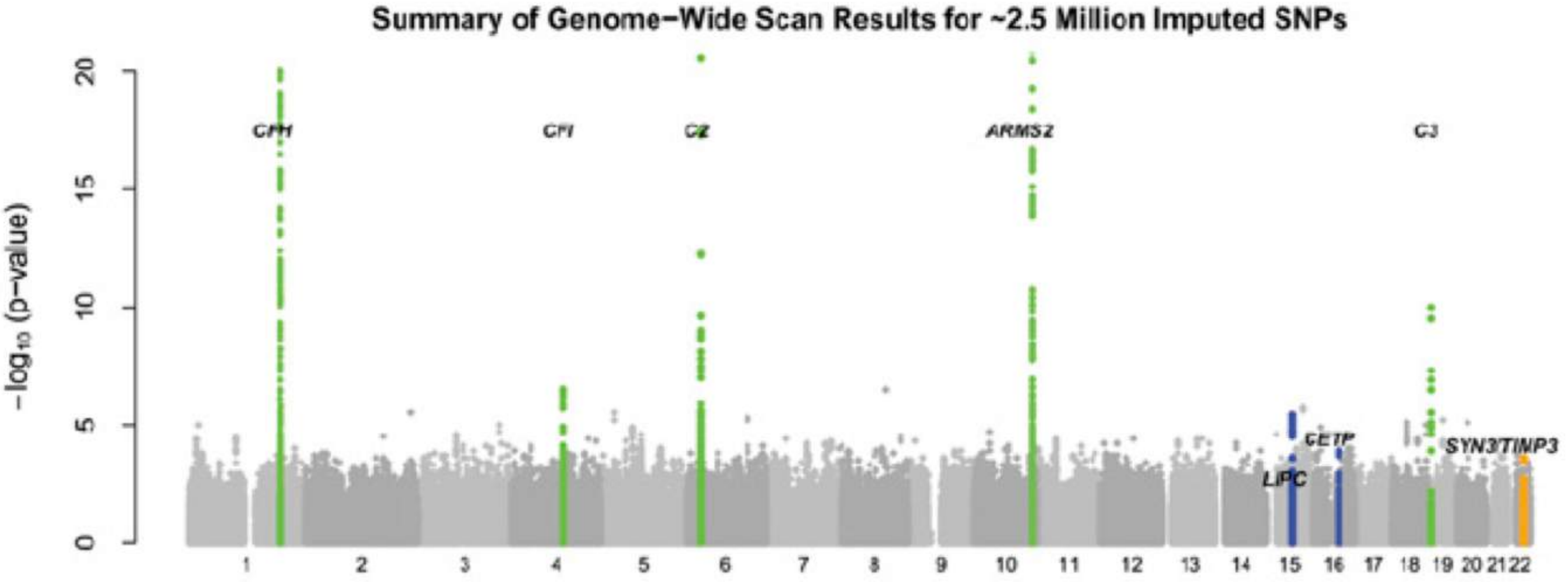
96 cases and 50 controls , 100K SNPs



Klein et al, *Science* 2005; 308:385-389.

Later GWAS of AMD

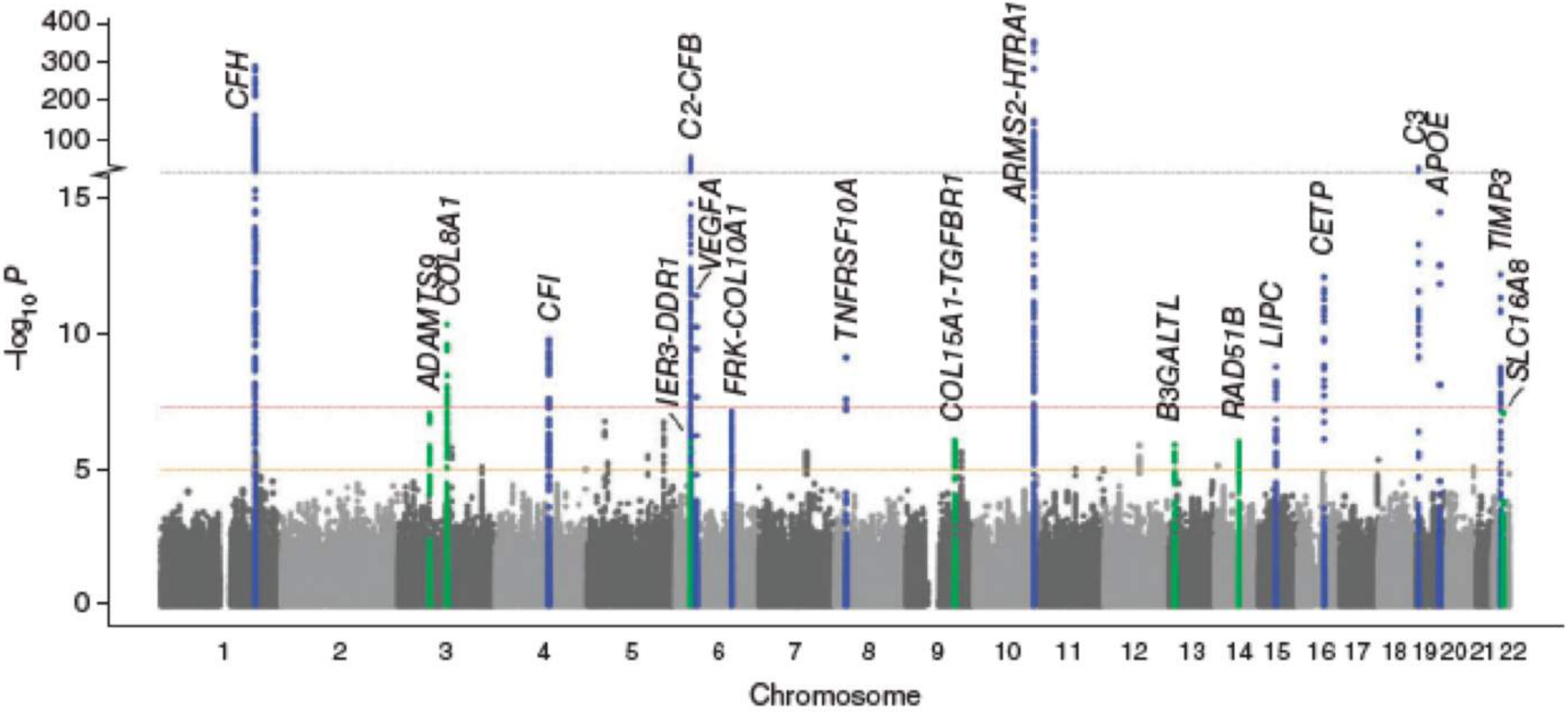
2150 cases and 1157controls , 370K SNPs



Chen et al, *PNAS* 2010

Largest Meta-analysis of AMD

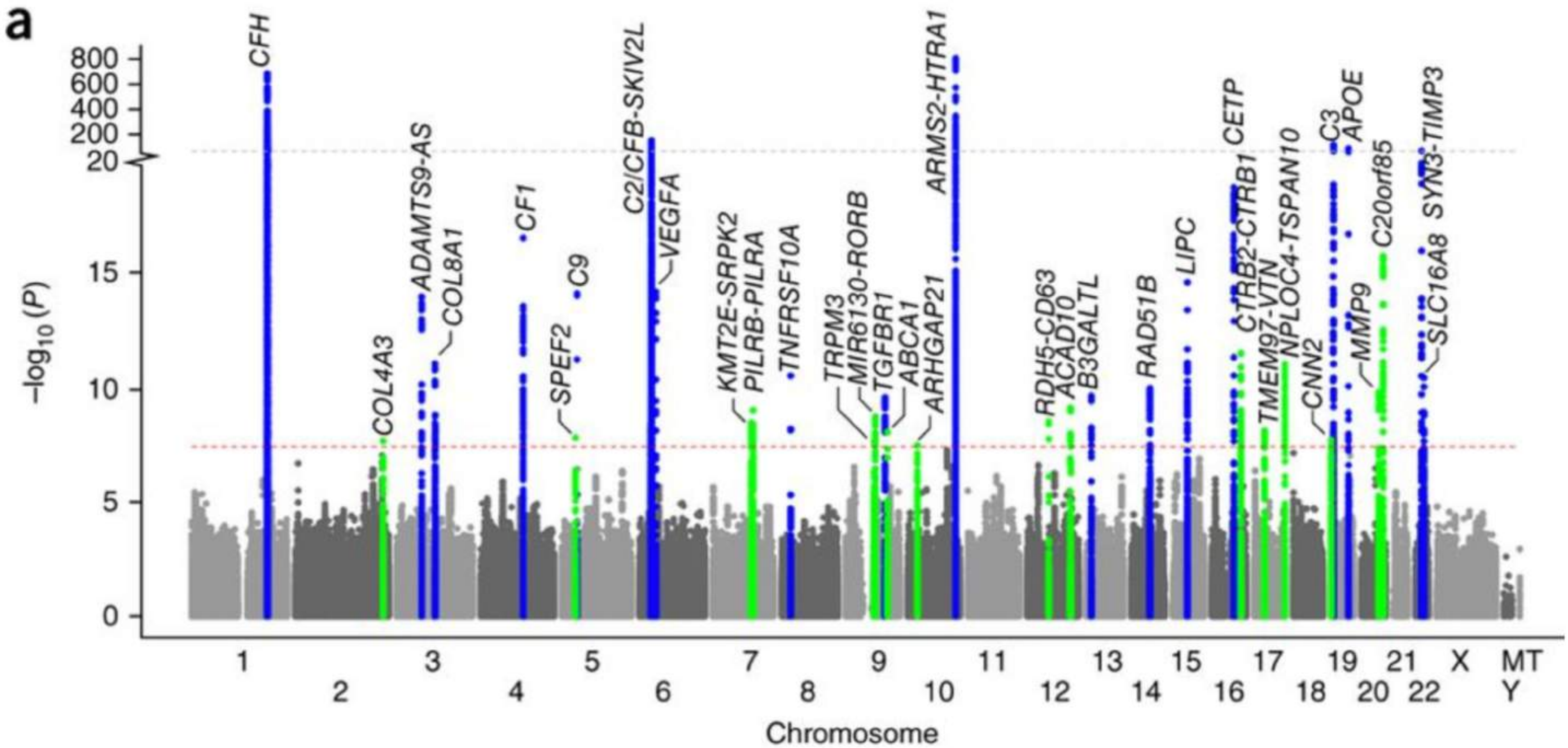
> 17,000 cases, > 60,000 controls, 2 M imputed HapMap SNPs



The AMD Gene Consortium, *Nat Genet* 2013

Latest Meta-analysis of AMD

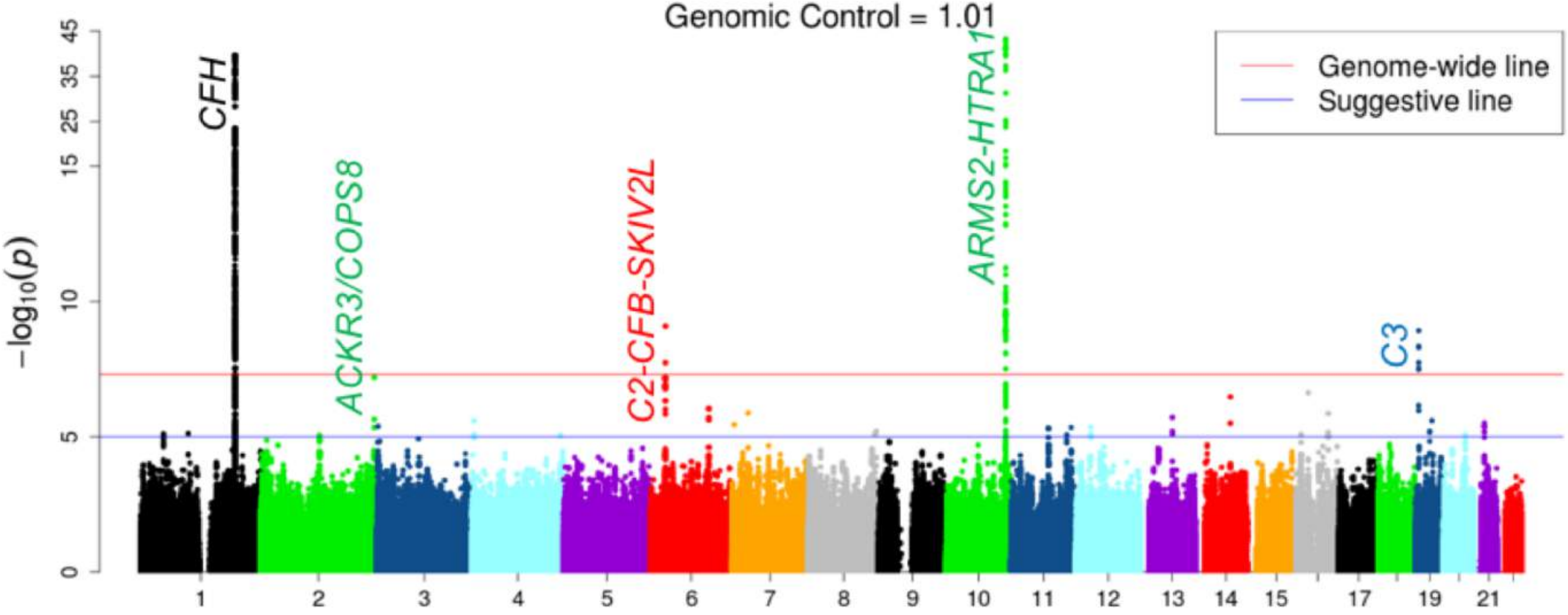
16,144 cases, 17,832 controls, 12 M imputed HapMap SNPs



the International AMD Genomics Consortium (IAMDGC), *Nat Genet* 2016

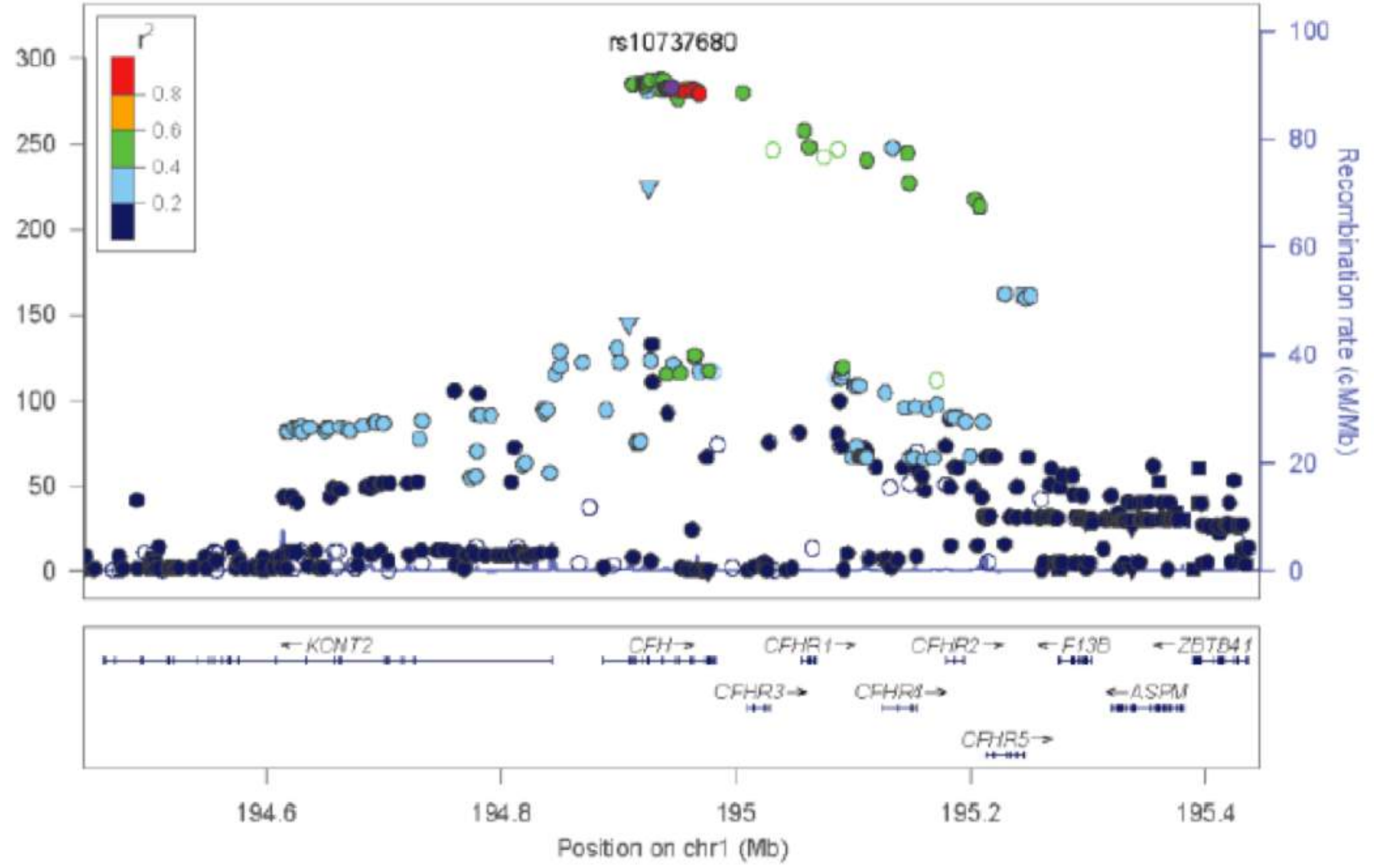
GWAS of AMD Progression

2,721 subjects, 9 M imputed 1000G phase3 SNPs

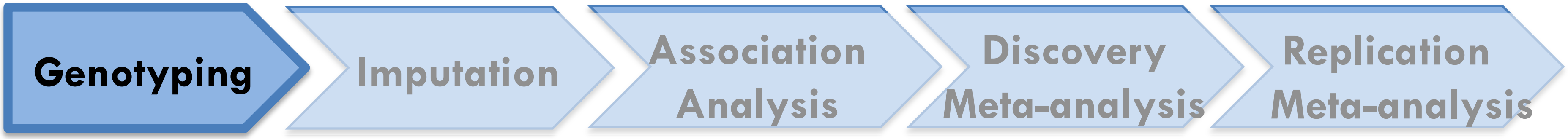


Yan et al, *Human Molecular Genetics 2017*

Regional Plots



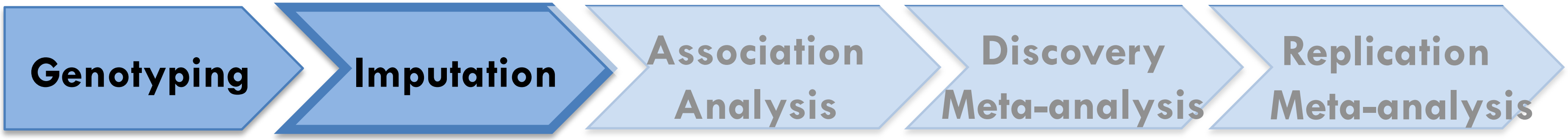
GWA Study Design



- Sample Collection
  - Genotyping of single nucleotide polymorphisms (SNPs) was performed using a variety of platforms
  - Array densities ranged from roughly 200k to 1M SNPs/chip
  - Most samples were population based case-control studies, though some data came from family based (sib-pair) studies
- Quality Control (**PLINK**)
  - Samples screened unknown for population stratification
  - Rare SNPs (MAF < 1%-5%) and SNPs with high missing rate were excluded from the analysis
  - Hardy Weinberg Equilibrium for genotype frequency
  - Potential familial samples

Nat Genet 2013

GWA Study Design

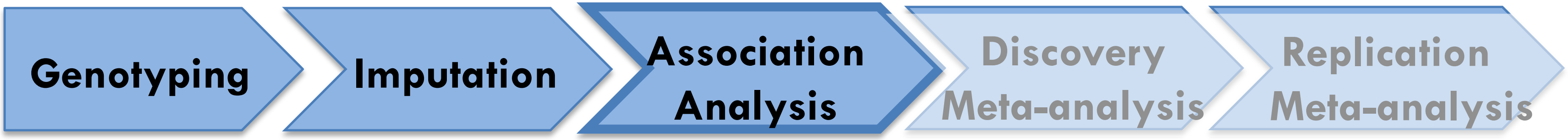


Genotying → Imputation → Association Analysis → Discovery Meta-analysis → Replication Meta-analysis

- Imputation
  - Each group participating in the discovery analysis calculated the allelic dosages using **IMPUTE2**
  - All imputation was performed using the **1000 Genome Project phase 3** reference panels
- Quality Control (**PLINK**)
  - SNPs of low imputation quality and/or extreme effect size which tend to indicate spurious associations were removed
  - After imputation and quality control measures, most data sets contain dosages for over 2 million SNPs per sample

Nat Genet 2013

GWA Study Design



Genotyping

Imputation

Association Analysis

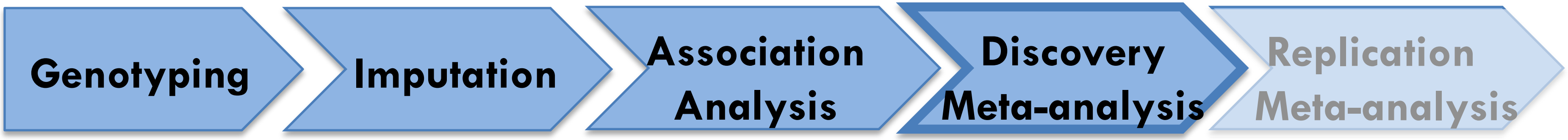
Discovery Meta-analysis

Replication Meta-analysis

- Statistical Methods
  - A logistic regression model, or equivalent analysis, was used to test for association between allelic frequency and AMD risk
  - Contributing studies adjusted for population substructure as needed
  - The primary analysis model was unadjusted for age, though subsequent analysis did included age as a covariate
  - Primary model compared allelic frequencies between all advanced stages of AMD (neovascular AMD and GA) vs controls
  - **PLINK** for logistic and linear regression

Nat Genet 2013

GWA Study Design



Genotyping

Imputation

Association Analysis

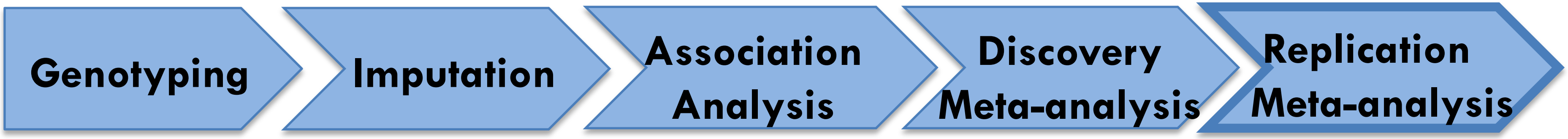
Discovery Meta-analysis

Replication Meta-analysis

- Meta-analysis details
  - Meta-analysis of all the discovery GWAS was performed via **METAL** using the inverse fixed affects model
  - Total number of samples in the discovery analysis was approximately 7,600 cases and 50,000 controls
- Discovery Results
  - From this analysis, 32 loci show promising evidence for association an were further considered for the subsequent stage of replication analysis

Nat Genet 2013

GWA Study Design



- Follow-up Analysis
  - 32 candidate SNPs from discovery analysis were sent for genotyping in an additional set of non-overlapping case-control samples (N_{case} > 9,500; $N_{control}$ > 8,200)
- Replication Results
  - After meta-analyzing these results with our discovery data, 19 loci attain genome-wide significance (p-values $< 5.0 \times 10^{-8}$)
  - Final tally of samples analyzed for SNPs in the replication data set comes to over 17,000 cases and over 60,000 controls

Nat Genet 2013

12 Loci previously observed to have genome-wide association with AMD risk

SNP/					**Discovery**		**Follow-up**		**Meta**	
Risk Allele	**Chr**	**Pos(Mb)**	**Nearby Genes**	**EAF**	***P***	**OR**	***P***	**OR**	***P***	**OR**
rs10490924/T	10	124.2	*ARMS2*	0.3	4×10^{-353}	2.7	2.8×10^{-190}	2.9	4×10^{-540}	2.8
rs10737680/A	1	195.0	*CFH*	0.64	1×10^{-283}	2.4	2.7×10^{-152}	2.5	1×10^{-434}	2.4
rs429608/G	6	32.0	*C2/CFB*	0.86	2×10^{-54}	1.6	2.4×10^{-37}	1.9	4×10^{-89}	1.7
rs2230199/C	19	6.7	*C3*	0.2	2×10^{-26}	1.4	3.4×10^{-17}	1.4	1×10^{-41}	1.4
rs5749482/G	22	31.4	*SYN3/TIMP3*	0.74	6×10^{-13}	1.3	9.7×10^{-17}	1.4	2×10^{-26}	1.3
rs4420638/A	19	50.1	*APOE*	0.83	3×10^{-15}	1.3	4.2×10^{-7}	1.3	2×10^{-20}	1.3
rs1864163/G	16	55.6	*CETP*	0.76	8×10^{-13}	1.2	8.7×10^{-5}	1.2	7×10^{-16}	1.2
rs943080/T	6	43.9	*VEGFA*	0.51	4×10^{-12}	1.2	1.6×10^{-5}	1.1	9×10^{-16}	1.2
rs13278062/T	8	23.1	*TNFRSF10A*	0.48	7×10^{-10}	1.2	6.4×10^{-7}	1.1	3×10^{-15}	1.2
rs920915/C	15	56.5	*LIPC*	0.48	2×10^{-9}	1.1	0.004	1.1	3×10^{-11}	1.1
rs4698775/G	4	110.8	*CFI*	0.31	2×10^{-10}	1.2	0.025	1.1	7×10^{-11}	1.1
rs3812111/T	6	116.6	*FRK/COL10A1*	0.64	7×10^{-8}	1.1	0.022	1.1	2×10^{-8}	1.1

Nat Genet 2013

7 loci showing genome-wide significant association with AMD risk for the first time

					Discovery		Follow-up		Meta	
SNP/Risk Allele	**Chr**	**Pos**	**Nearby Genes**	**EAF**	***P***	**OR**	***P***	**OR**	***P***	**OR**
rs13081855/T	3	101.0 Mb	*COL8A1*	0.1	4×10^{-11}	1.3	6.0×10^{-4}	1.2	4×10^{-13}	1.2
rs3130783/A	6	30.9 Mb	*IER3/DDR1*	0.79	1×10^{-6}	1.2	3.5×10^{-6}	1.2	2×10^{-11}	1.2
rs8135665/T	22	36.8 Mb	*SLC16A8*	0.21	8×10^{-8}	1.2	5.6×10^{-5}	1.1	2×10^{-11}	1.2
rs334353/T	9	100.9 Mb	*COL15A1/TGFBR1*	0.73	9×10^{-7}	1.1	6.7×10^{-6}	1.1	3×10^{-11}	1.1
rs8017304/A	14	67.9 Mb	*RAD51B*	0.61	9×10^{-7}	1.1	2.1×10^{-5}	1.1	9×10^{-11}	1.1
rs6795735/T	3	64.7 Mb	*ADAMTS9*	0.46	9×10^{-8}	1.1	0.0066	1.1	5×10^{-9}	1.1
rs9542236/C	13	30.7 Mb	*B3GALTL*	0.44	2×10^{-6}	1.1	0.0018	1.1	2×10^{-8}	1.1

Nat Genet 2013

Functional Analysis

- Gene set enrichment of all implicated results was run using Ingenuity Pathway Analysis (IPA) software.

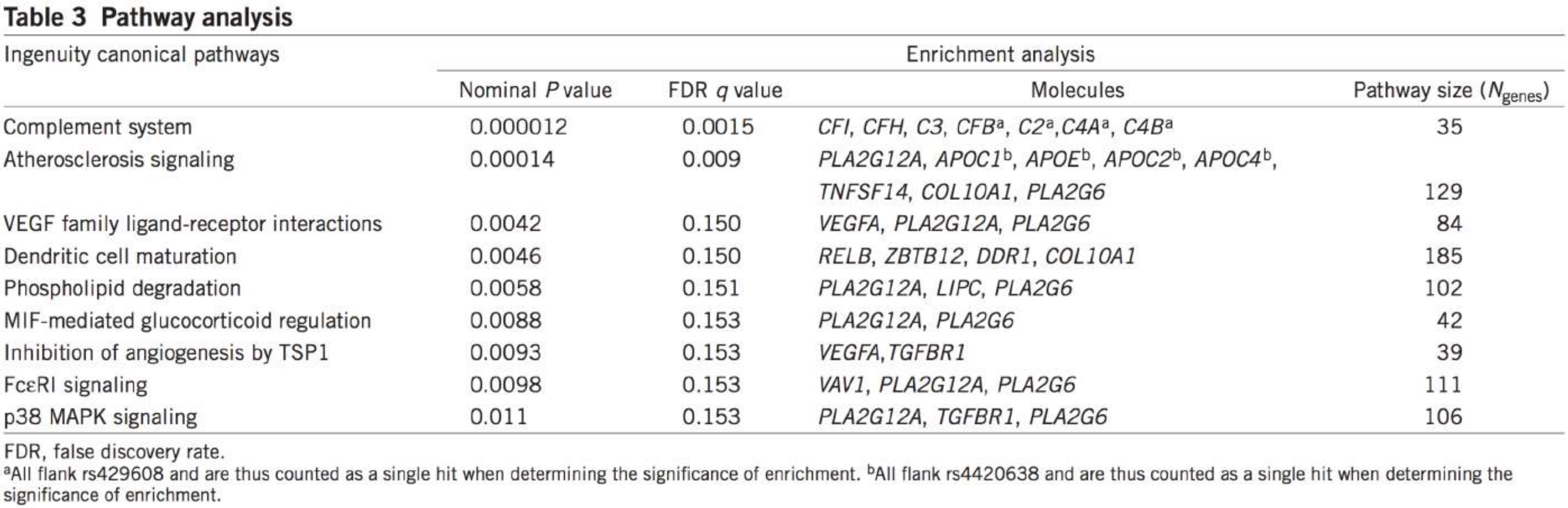


Table 3 Pathway analysis

Ingenuity canonical pathways	Enrichment analysis			
	Nominal *P* value	FDR *q* value	Molecules	Pathway size (N_{genes})
Complement system	0.000012	0.0015	*CFI, CFH, C3, CFB*[a], *C2*[a],*C4A*[a], *C4B*[a]	35
Atherosclerosis signaling	0.00014	0.009	*PLA2G12A, APOC1*[b], *APOE*[b], *APOC2*[b], *APOC4*[b], *TNFSF14, COL10A1, PLA2G6*	129
VEGF family ligand-receptor interactions	0.0042	0.150	*VEGFA, PLA2G12A, PLA2G6*	84
Dendritic cell maturation	0.0046	0.150	*RELB, ZBTB12, DDR1, COL10A1*	185
Phospholipid degradation	0.0058	0.151	*PLA2G12A, LIPC, PLA2G6*	102
MIF-mediated glucocorticoid regulation	0.0088	0.153	*PLA2G12A, PLA2G6*	42
Inhibition of angiogenesis by TSP1	0.0093	0.153	*VEGFA,TGFBR1*	39
FcεRI signaling	0.0098	0.153	*VAV1, PLA2G12A, PLA2G6*	111
p38 MAPK signaling	0.011	0.153	*PLA2G12A, TGFBR1, PLA2G6*	106

FDR, false discovery rate.

[a] All flank rs429608 and are thus counted as a single hit when determining the significance of enrichment. [b] All flank rs4420638 and are thus counted as a single hit when determining the significance of enrichment.

Nat Genet 2013

Summary

- GWAS have been successful in identifying genetic variants associated with common diseases and traits.

- A large proportion of heritability remains unexplained by GWAS and very limited functional knowledge is known at most identified loci.

- Next generation sequencing will be the next step to dissect the genetic basis beyond GWAS.

References

- http://www.genome.gov/gwastudies/
- http://pngu.mgh.harvard.edu/~purcell/plink/

- Mark I. McCarthy et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Review Genetics. 2008
- The AMD Gene Consortium. Seven New Loci Associated with Age-Related Macular Degeneration. Nature Genetics. 2013
- The International AMD Genomics Consortium (IAMDGC). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature Genetics. 2016