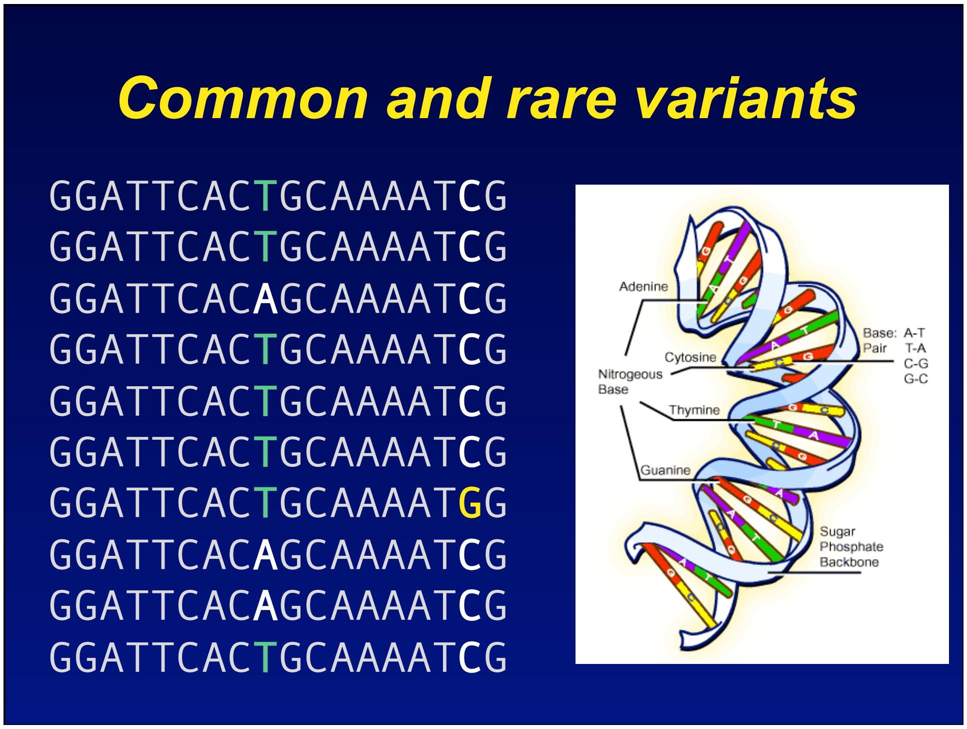
Genome-wide association studies

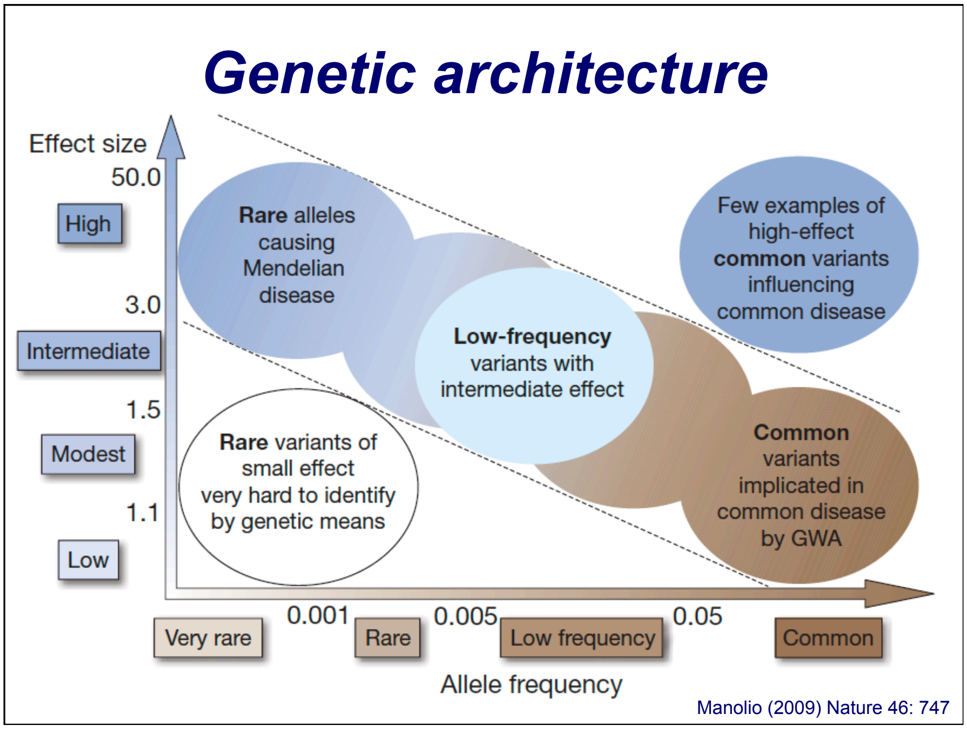
Karen Mohlke, PhD
Department of Genetics
University of North Carolina

Complex traits



Common and rare variants

```
GGATTCACTGCAAAATCG
GGATTCACTGCAAAATCG
GGATTCACAGCAAAATCG
GGATTCACTGCAAAATCG
GGATTCACTGCAAAATCG
GGATTCACTGCAAAATCG
GGATTCACTGCAAAATGG
GGATTCACAGCAAAATCG
GGATTCACAGCAAAATCG
GGATTCACTGCAAAATCG
```



Genetic architecture

Rare alleles causing Mendelian disease

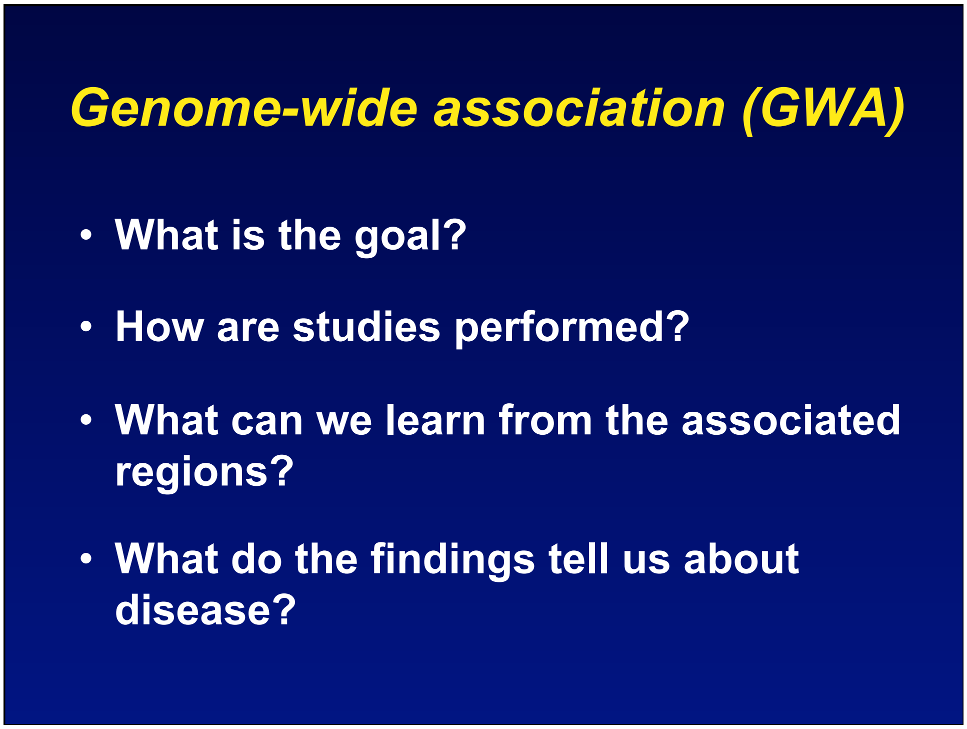
Few examples of high-effect **common** variants influencing common disease

Low-frequency variants with intermediate effect

Common variants implicated in common disease by GWA

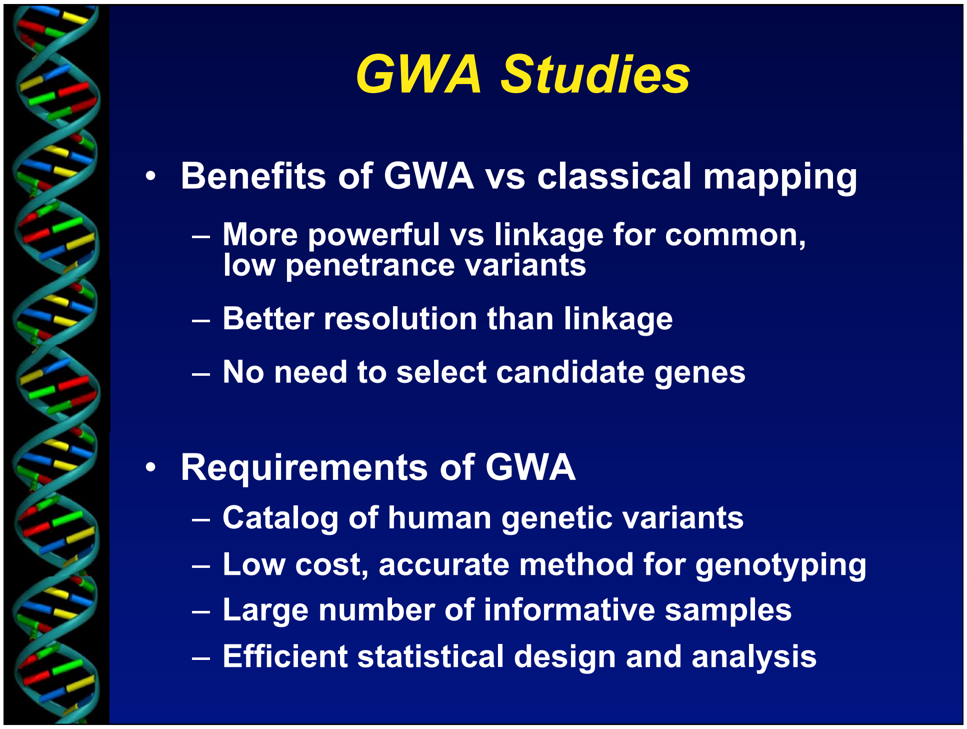
Rare variants of small effect very hard to identify by genetic means

Manolio (2009) Nature 46: 747



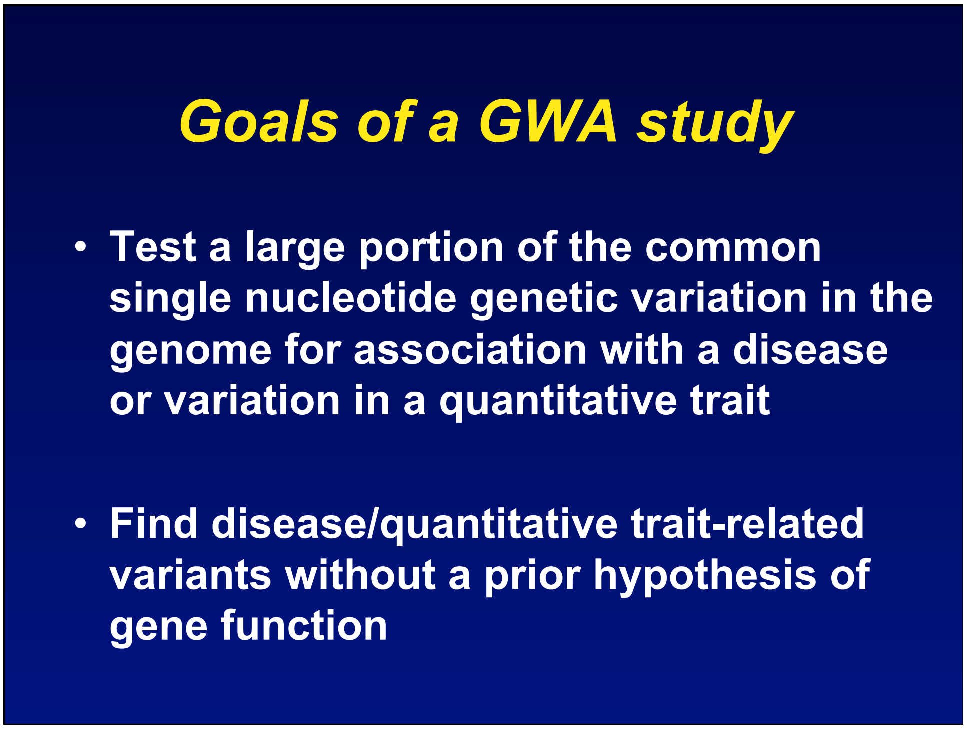
Genome-wide association (GWA)

- What is the goal?
- How are studies performed?
- What can we learn from the associated regions?
- What do the findings tell us about disease?



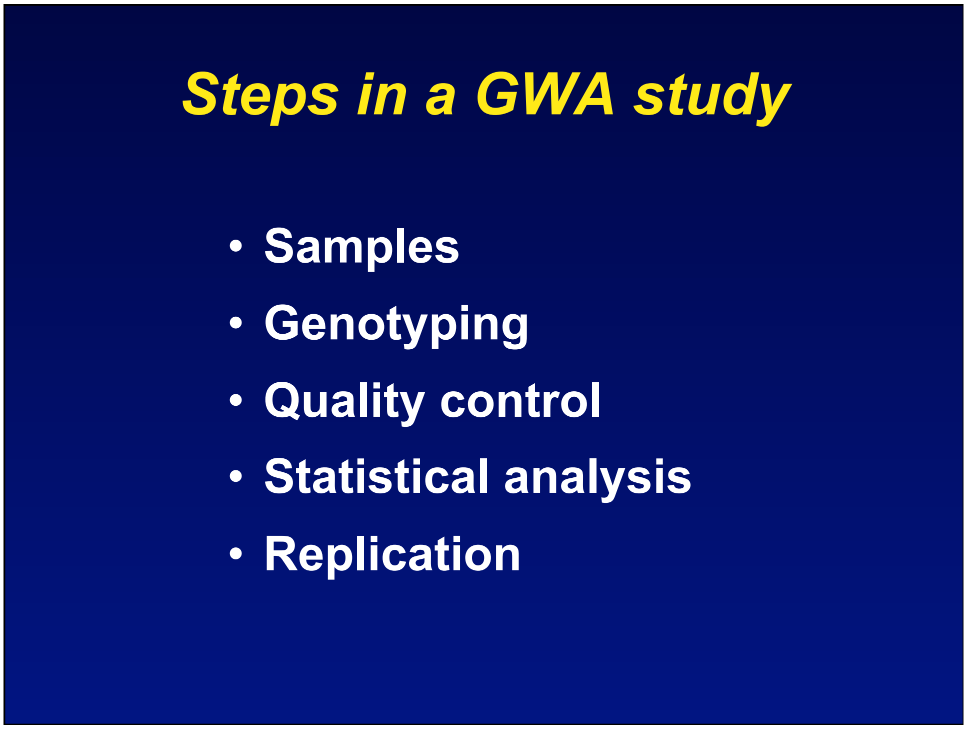
GWA Studies

- Benefits of GWA vs classical mapping
  - More powerful vs linkage for common, low penetrance variants
  - Better resolution than linkage
  - No need to select candidate genes
- Requirements of GWA
  - Catalog of human genetic variants
  - Low cost, accurate method for genotyping
  - Large number of informative samples
  - Efficient statistical design and analysis



Goals of a GWA study

- Test a large portion of the common single nucleotide genetic variation in the genome for association with a disease or variation in a quantitative trait
- Find disease/quantitative trait-related variants without a prior hypothesis of gene function



Steps in a GWA study

- Samples
- Genotyping
- Quality control
- Statistical analysis
- Replication

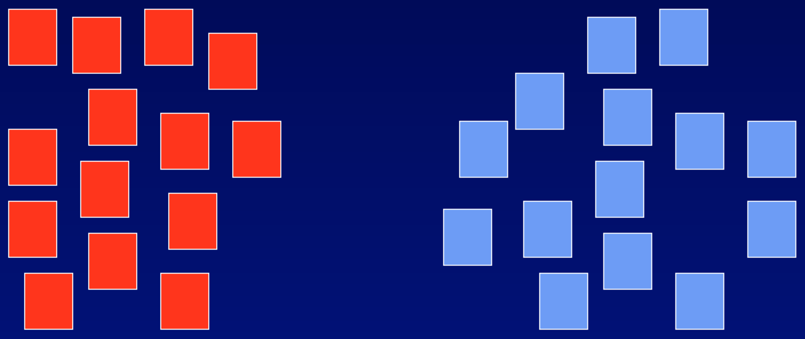
Phenotype

- Disease (case/control)
  - Rare
  - Common
- Quantitative trait
  - Easy to measure: Weight, height
  - Requires testing: Coronary artery thickness
  - Requires experiment: Gene expression

Selection of cases and controls

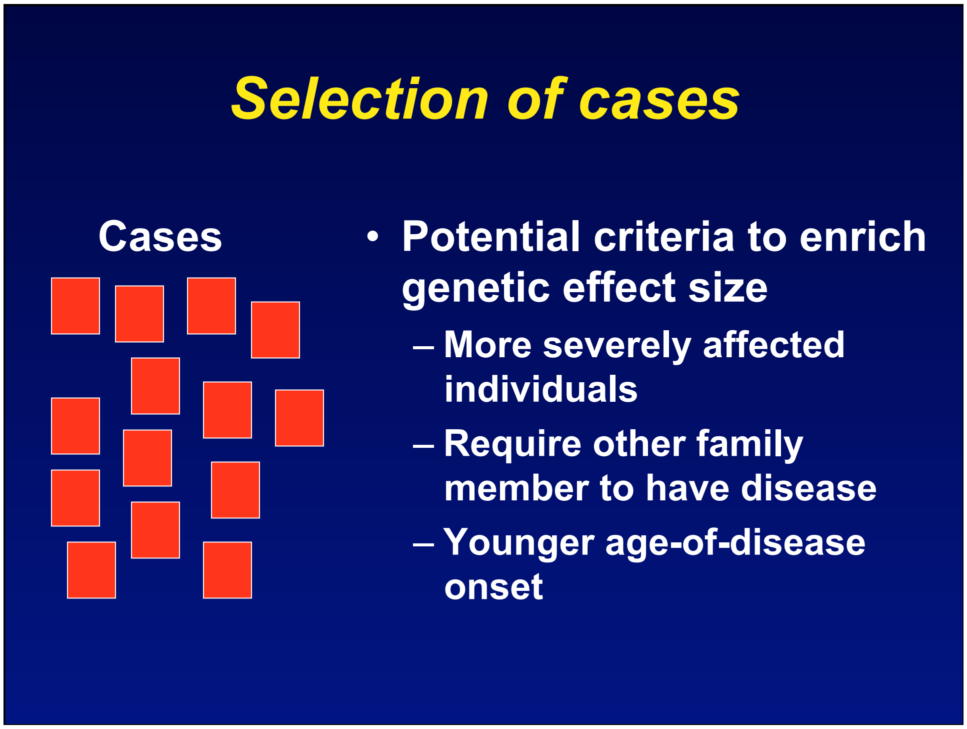
Cases

Controls



Definition of case?

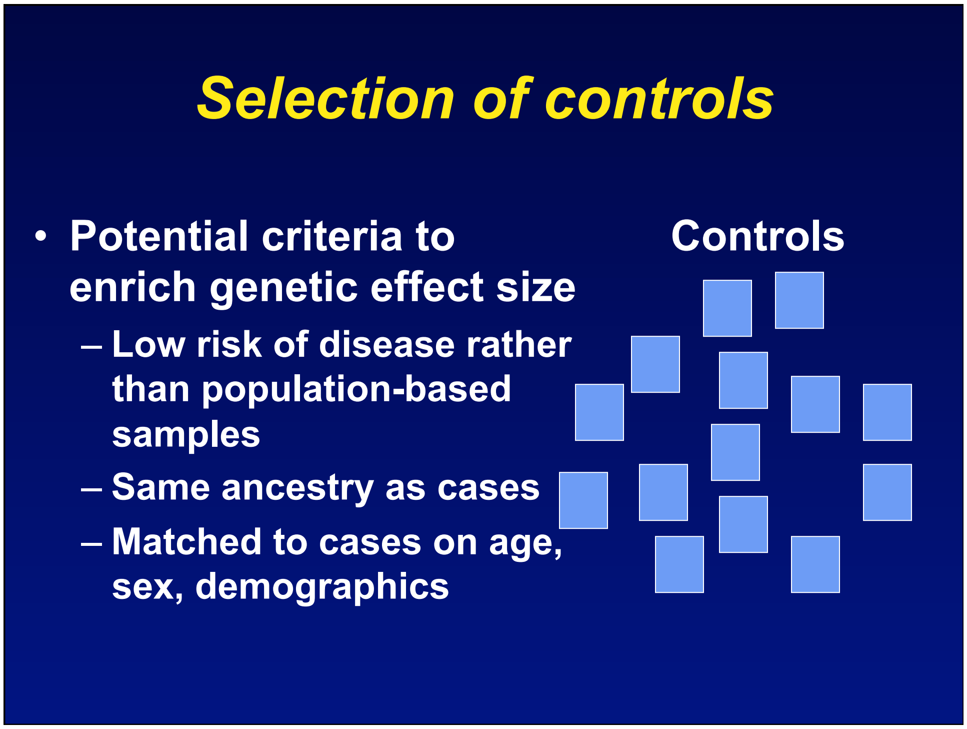
Definition of control?



Selection of cases

Cases

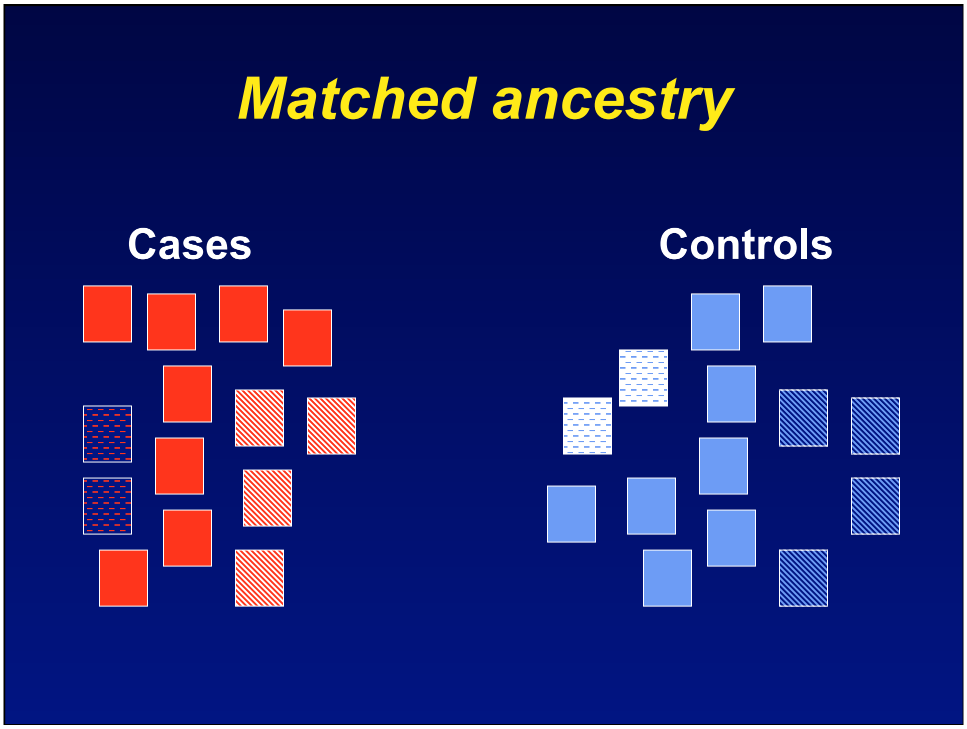
- Potential criteria to enrich genetic effect size
  - More severely affected individuals
  - Require other family member to have disease
  - Younger age-of-disease onset



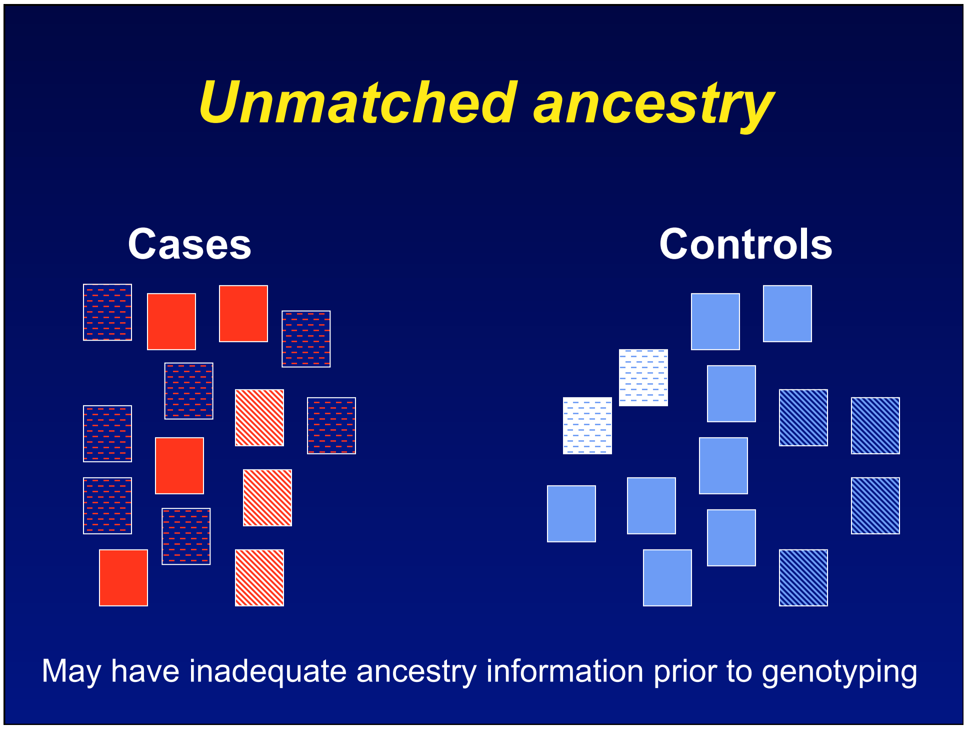
Selection of controls

- Potential criteria to enrich genetic effect size
  - Low risk of disease rather than population-based samples
  - Same ancestry as cases
  - Matched to cases on age, sex, demographics

Controls

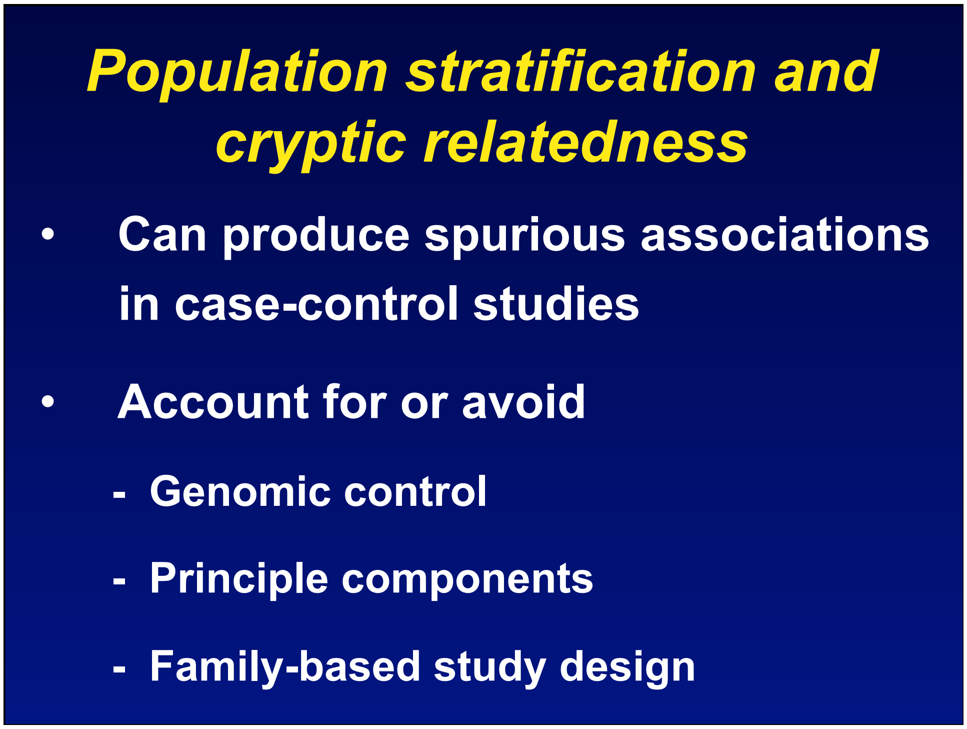


Matched ancestry



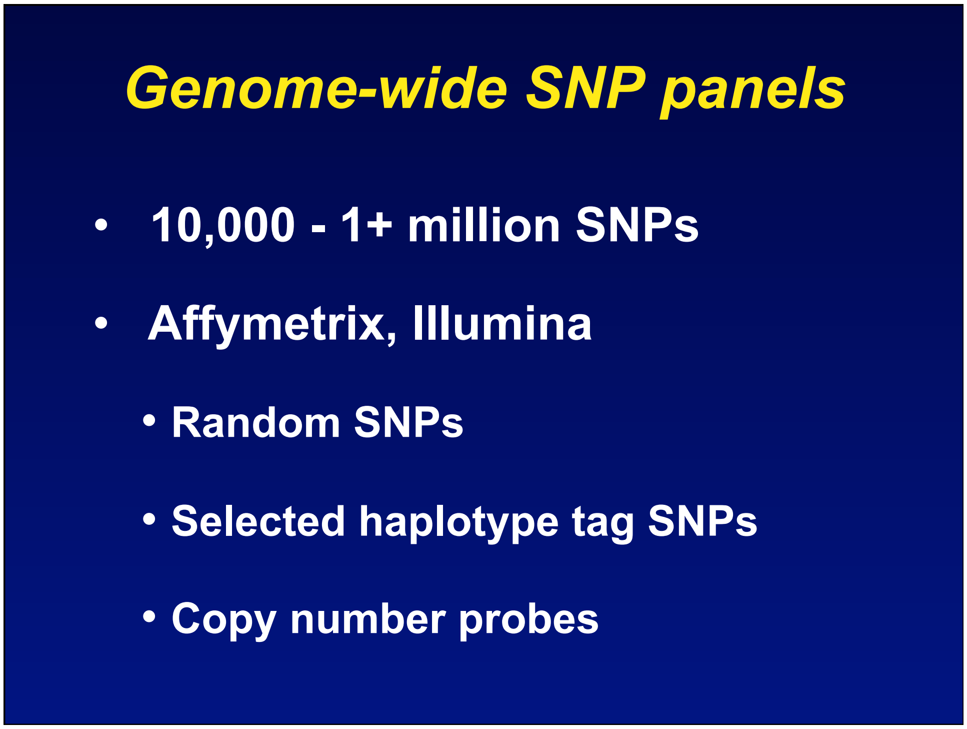
Unmatched ancestry

May have inadequate ancestry information prior to genotyping



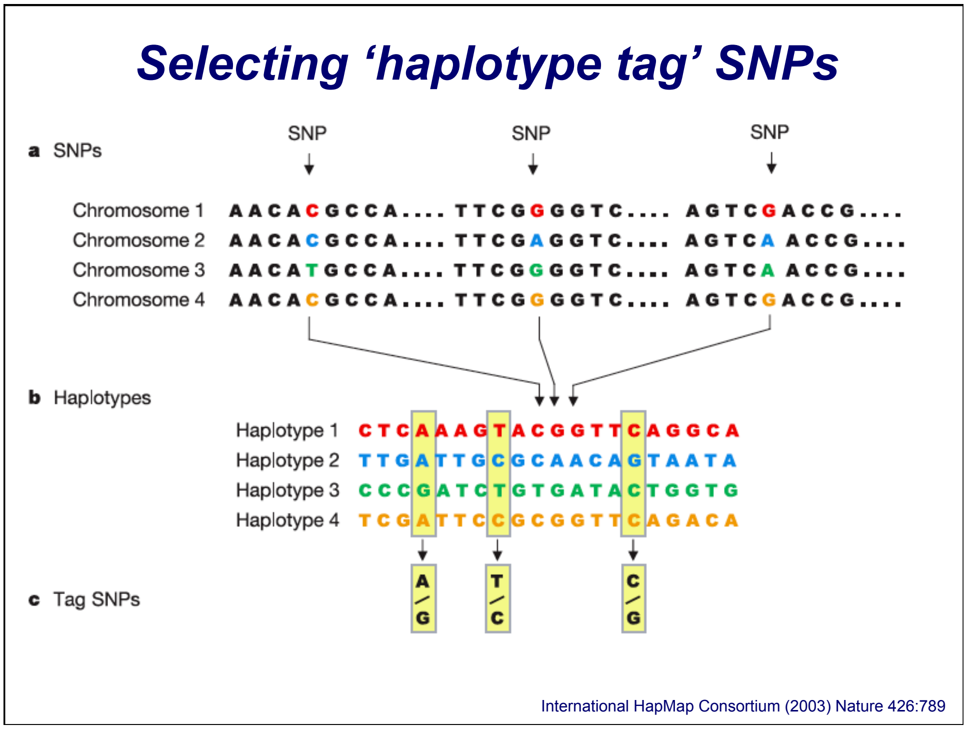
Population stratification and cryptic relatedness

- Can produce spurious associations in case-control studies
- Account for or avoid
  - Genomic control
  - Principle components
  - Family-based study design



Genome-wide SNP panels

- 10,000 - 1+ million SNPs
- Affymetrix, Illumina
  - Random SNPs
  - Selected haplotype tag SNPs
  - Copy number probes



Selecting 'haplotype tag' SNPs

a SNPs

SNP SNP SNP

Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....

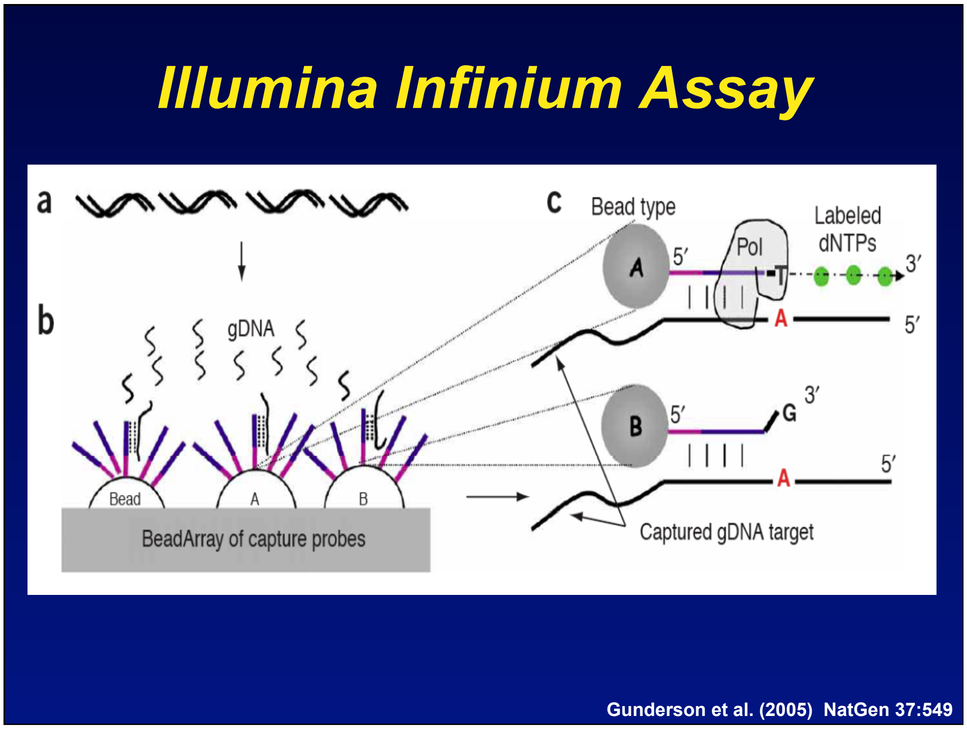
b Haplotypes

Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA

c Tag SNPs

A/G T/C C/G

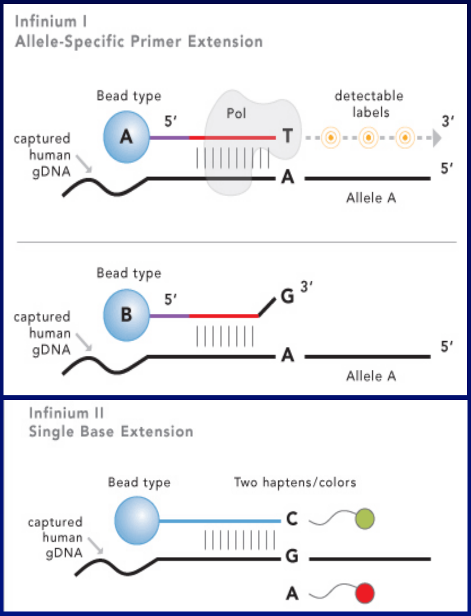
International HapMap Consortium (2003) Nature 426:789



Illumina Infinium Assay

Gunderson et al. (2005) NatGen 37:549

Illumina Infinium Assays



Infinium Assay BeadChips enable interrogation of ~317,000 to over one million SNPs and offer comprehensive coverage of CNV regions. Shown above, from left to right, are the HumanHap300-Duo, HumanHap550-Duo, and Human1M BeadChips.

Illumina.com

Affymetrix GeneChip Array

Figure 1: GeneChip® Mapping Assay Overview.

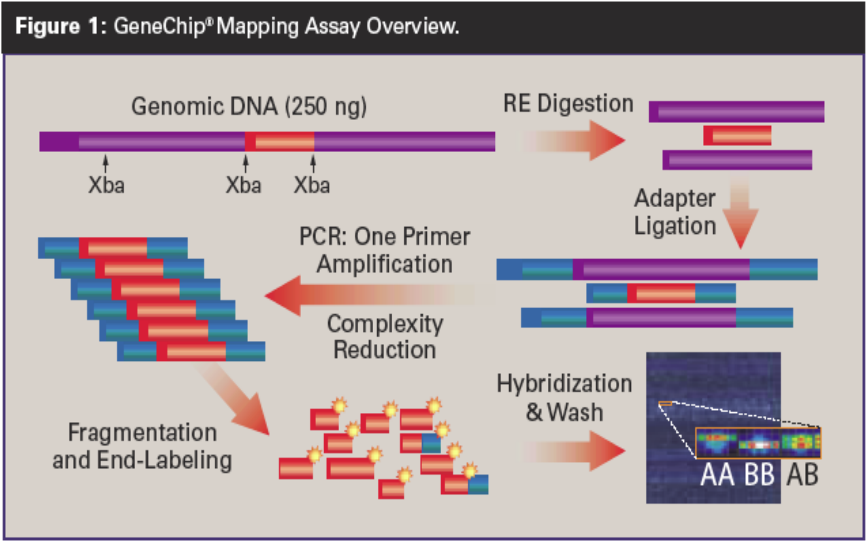
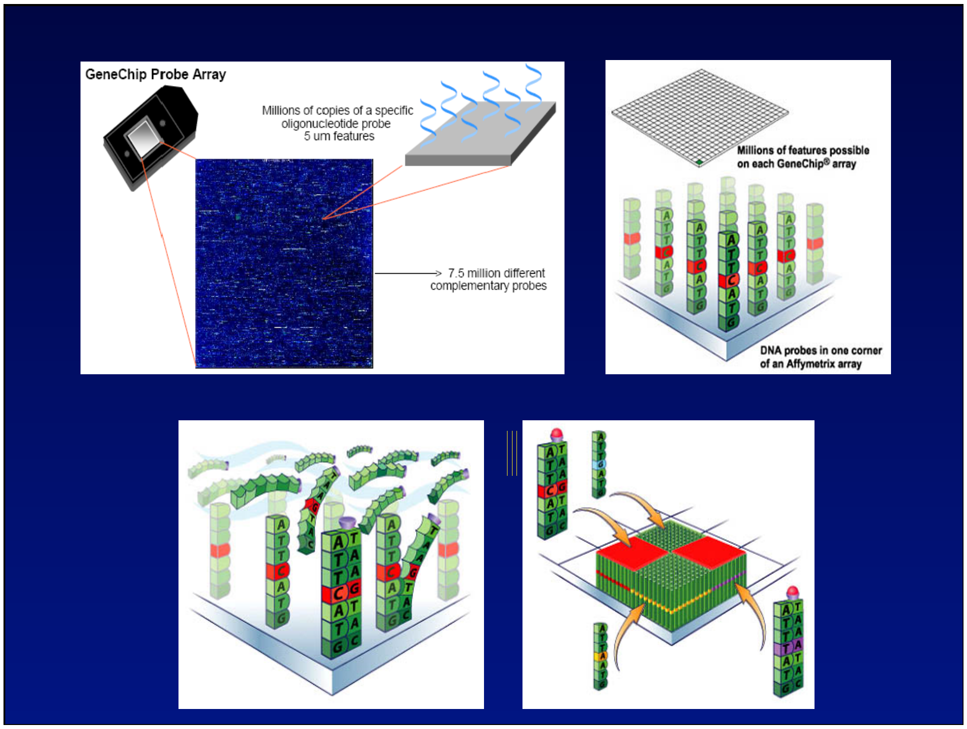
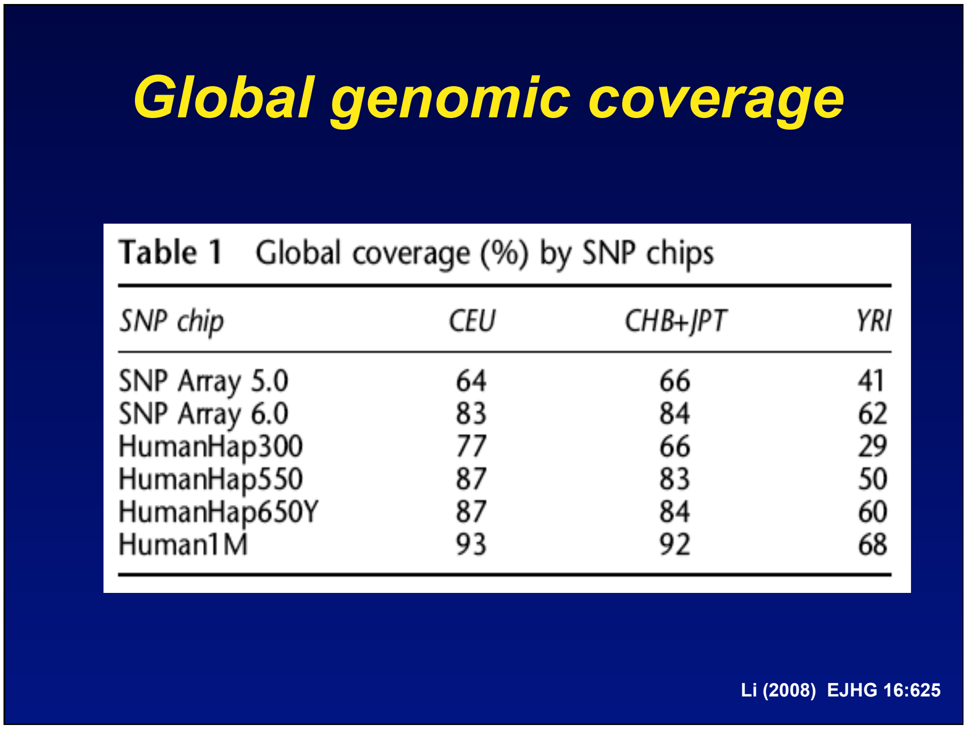


image from affymetrix.com


GeneChip Probe Array
Millions of copies of a specific oligonucleotide probe 5 um features
7.5 million different complementary probes
Millions of features possible on each GeneChip® array
DNA probes in one corner of an Affymetrix array



Global genomic coverage

Table 1 Global coverage (%) by SNP chips

SNP chip	CEU	CHB+JPT	YRI
SNP Array 5.0	64	66	41
SNP Array 6.0	83	84	62
HumanHap300	77	66	29
HumanHap550	87	83	50
HumanHap650Y	87	84	60
Human1M	93	92	68

Li (2008) EJHG 16:625

Local genomic coverage

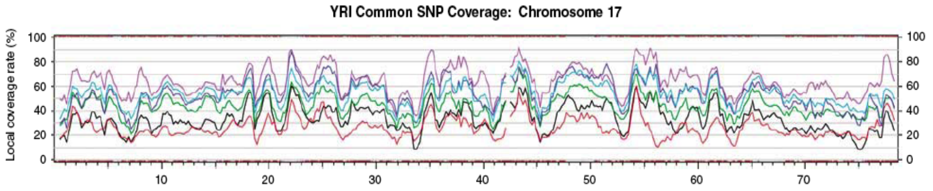
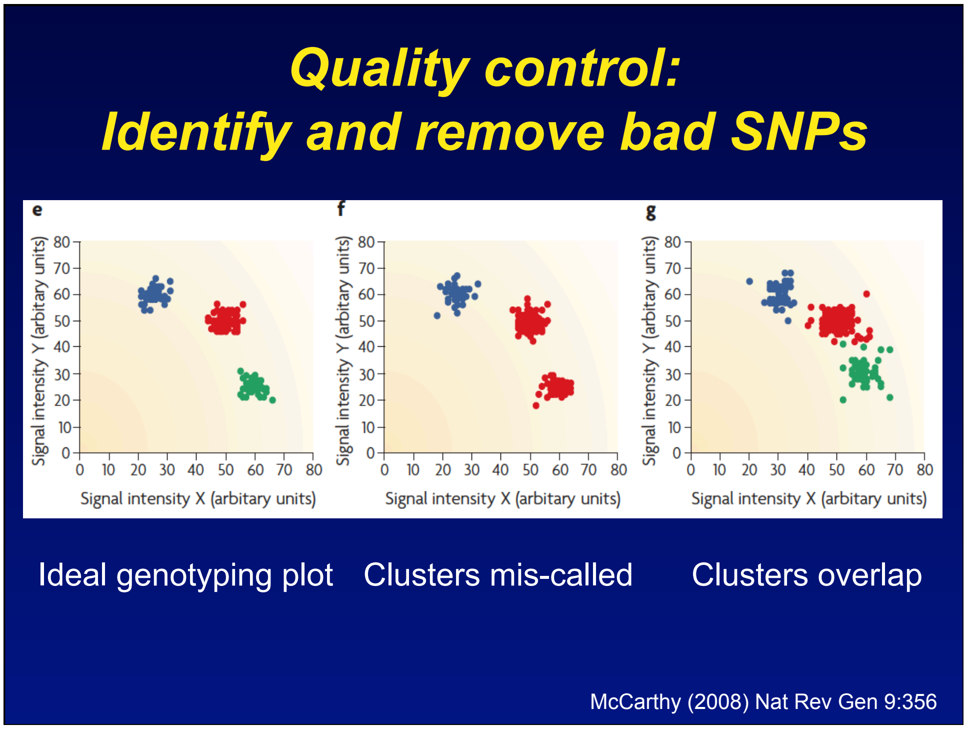


Figure 1 Local coverage map for each HapMap population for chromosome 17. The six SNP chips that were evaluated are SNP Array 5.0 (black), SNP Array 6.0 (blue), HumanHap300 (red), HumanHap550 (green), HumanHap650Y (cyan), and Human1M (purple). The red bars at the top and bottom indicate the transcription regions of known protein coding genes.

Li (2008) EJHG 16:625

Quality control: Identify and remove bad samples

- Poor quality samples
  - Sample success rate < 95 %
  - Excess heterozygous genotypes
- Sample switches
  - Wrong sex
- Unexpected related individuals
  - Pair-wise comparisons of genotype similarity
  - Duplicates
- Ancestry different from the rest of sample



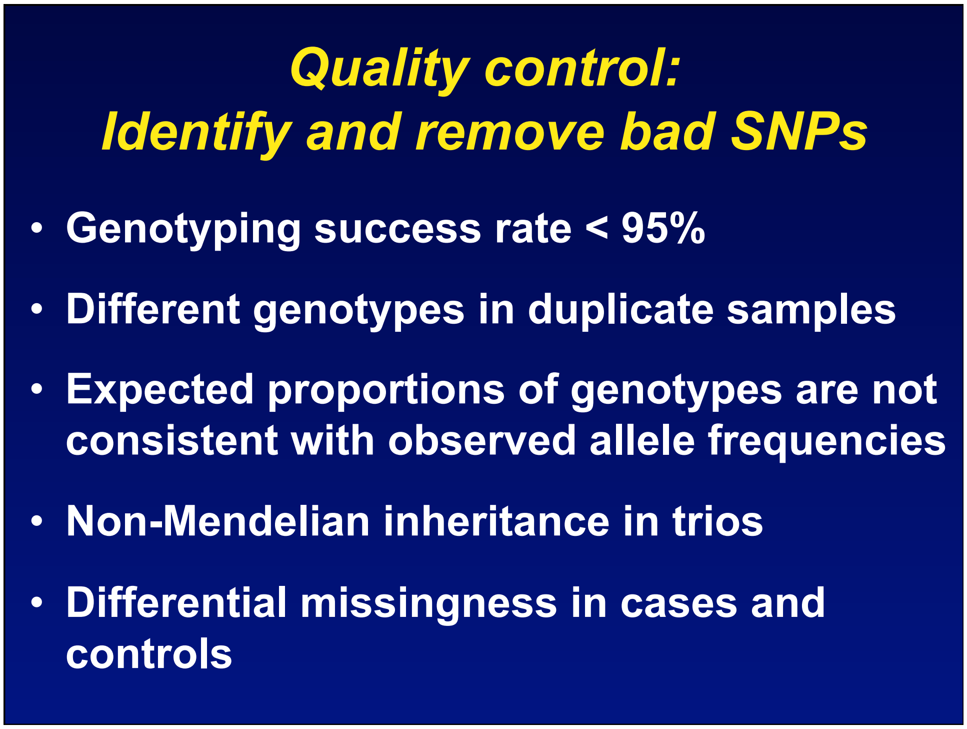
Quality control:
Identify and remove bad SNPs

Ideal genotyping plot

Clusters mis-called

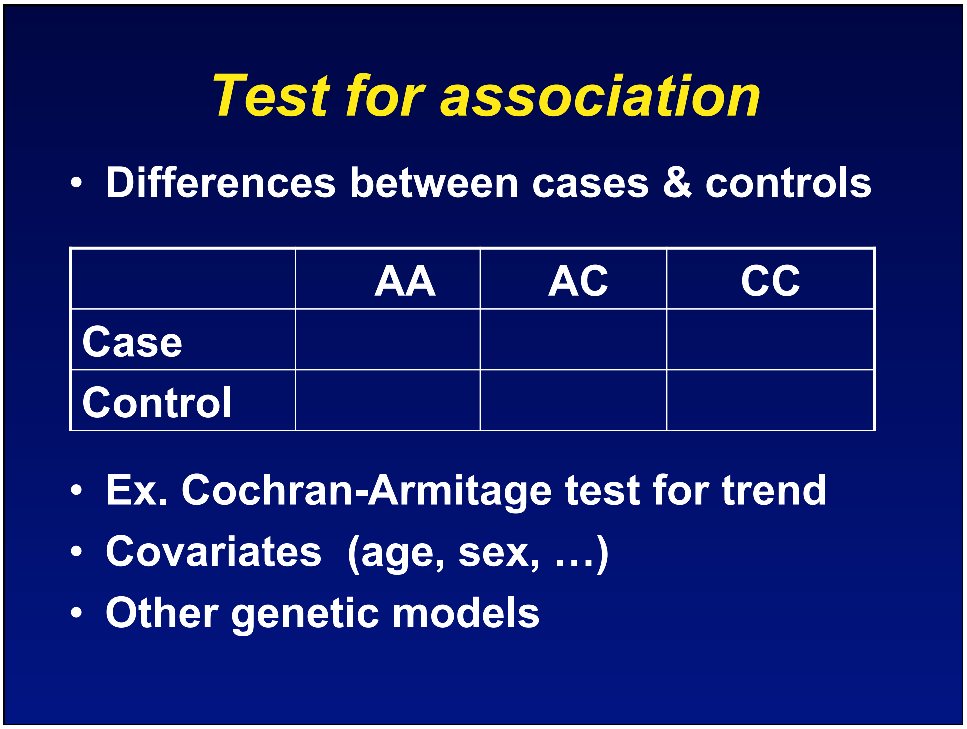
Clusters overlap

McCarthy (2008) Nat Rev Gen 9:356



Quality control:
Identify and remove bad SNPs

- Genotyping success rate < 95%
- Different genotypes in duplicate samples
- Expected proportions of genotypes are not consistent with observed allele frequencies
- Non-Mendelian inheritance in trios
- Differential missingness in cases and controls

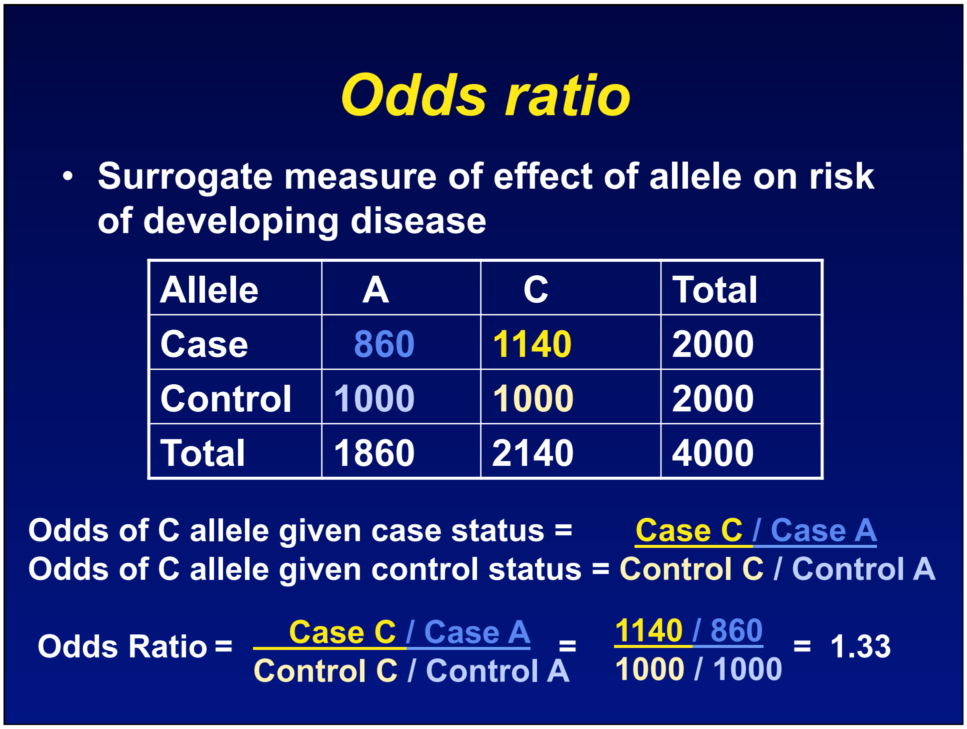


Test for association

- Differences between cases & controls

	AA	AC	CC
Case			
Control			

- Ex. Cochran-Armitage test for trend
- Covariates (age, sex, …)
- Other genetic models



Odds ratio

- Surrogate measure of effect of allele on risk of developing disease

Allele	A	C	Total
Case	860	1140	2000
Control	1000	1000	2000
Total	1860	2140	4000

Odds of C allele given case status = Case C / Case A
Odds of C allele given control status = Control C / Control A

$$\text{Odds Ratio} = \frac{\text{Case C / Case A}}{\text{Control C / Control A}} = \frac{1140 / 860}{1000 / 1000} = 1.33$$

Multiple testing

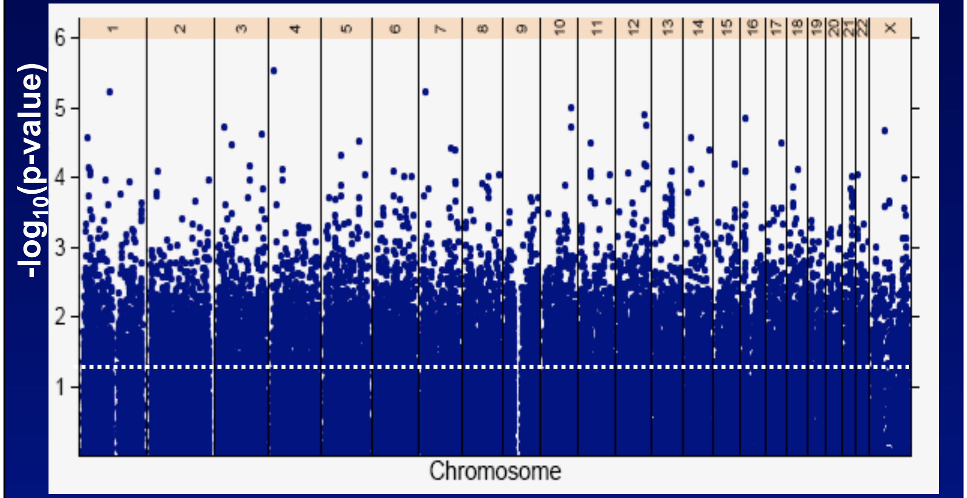
- Genotype and test > 300K – 1M SNPs
- Correct for the multiple tests

$$\frac{.05\ P\text{-value}}{1\text{ million SNPs}} = 5 \times 10^{-8}$$

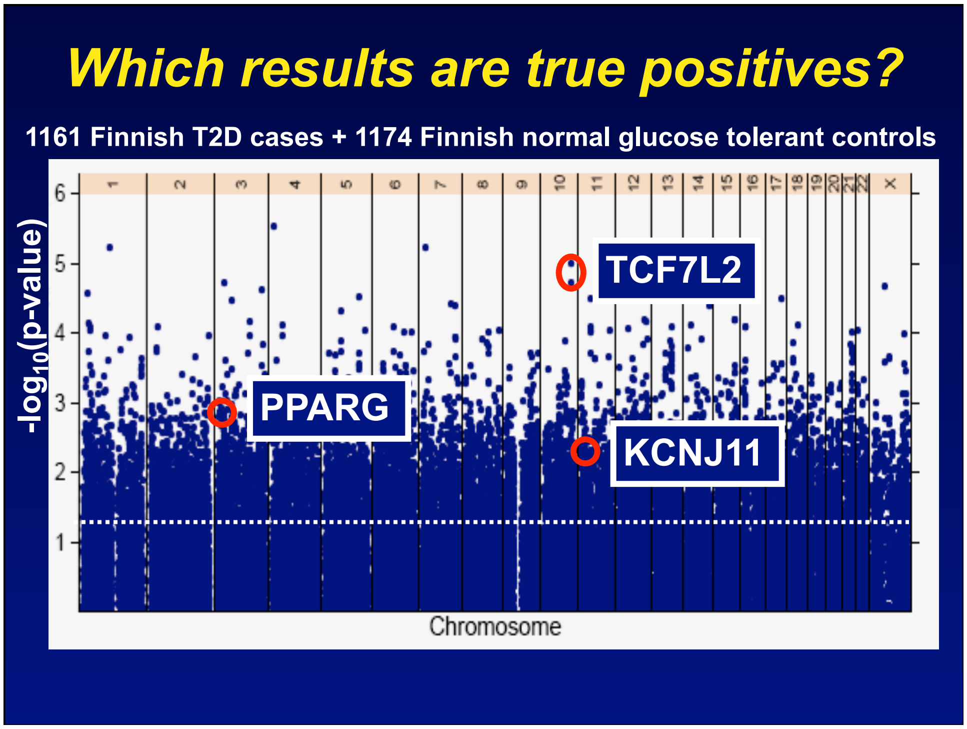
- Need large effect or large sample size

Type 2 diabetes association results

1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls

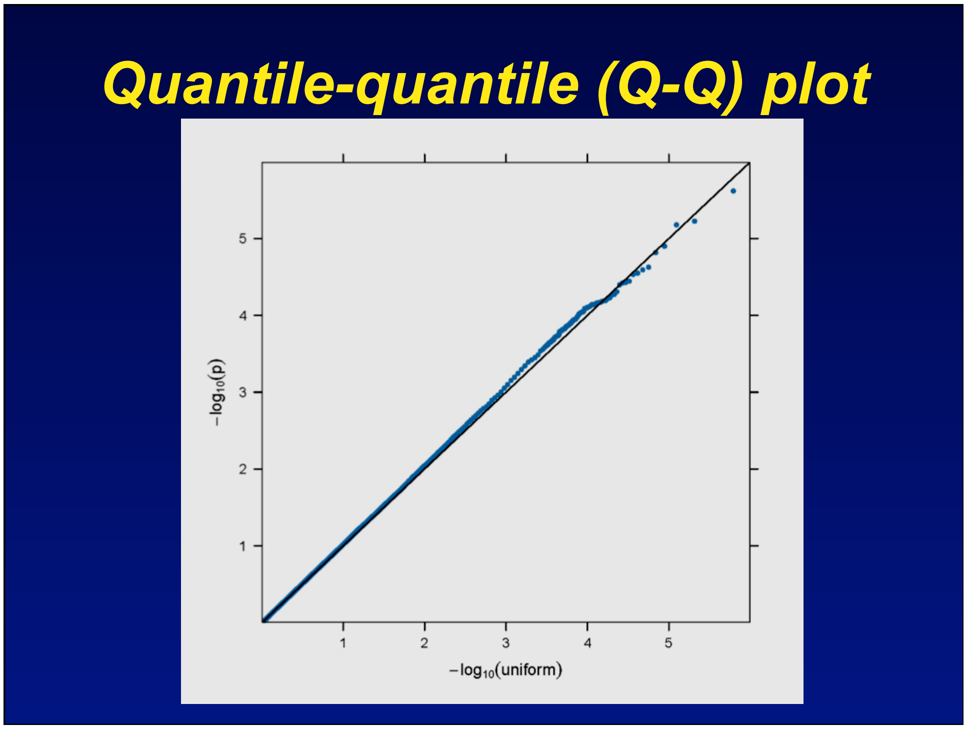


Logistic regression using additive model adjusted for age, gender, birth province



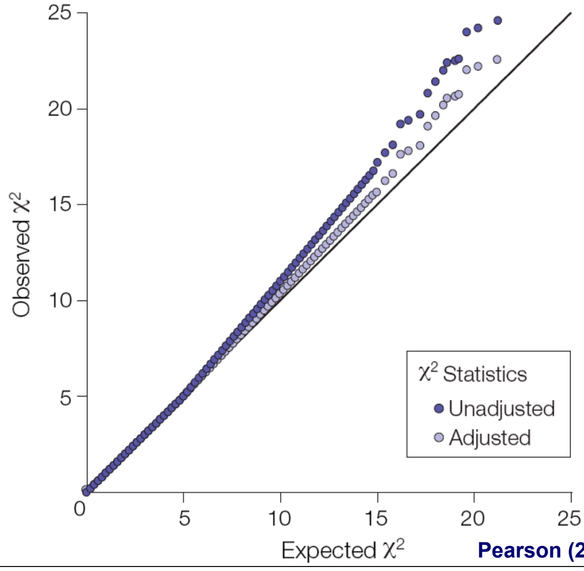
Which results are true positives?

1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls



Quantile-quantile (Q-Q) plot

Before and after adjustment of population stratification



Pearson (2008) JAMA 299:1335

Power to detect association

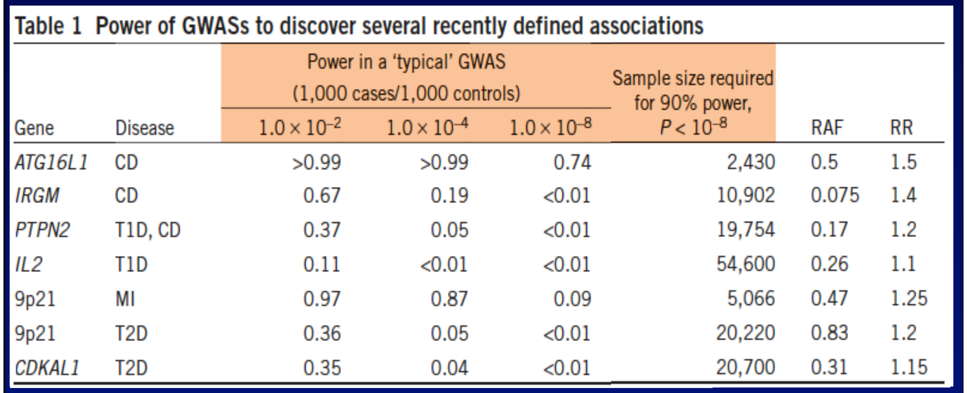
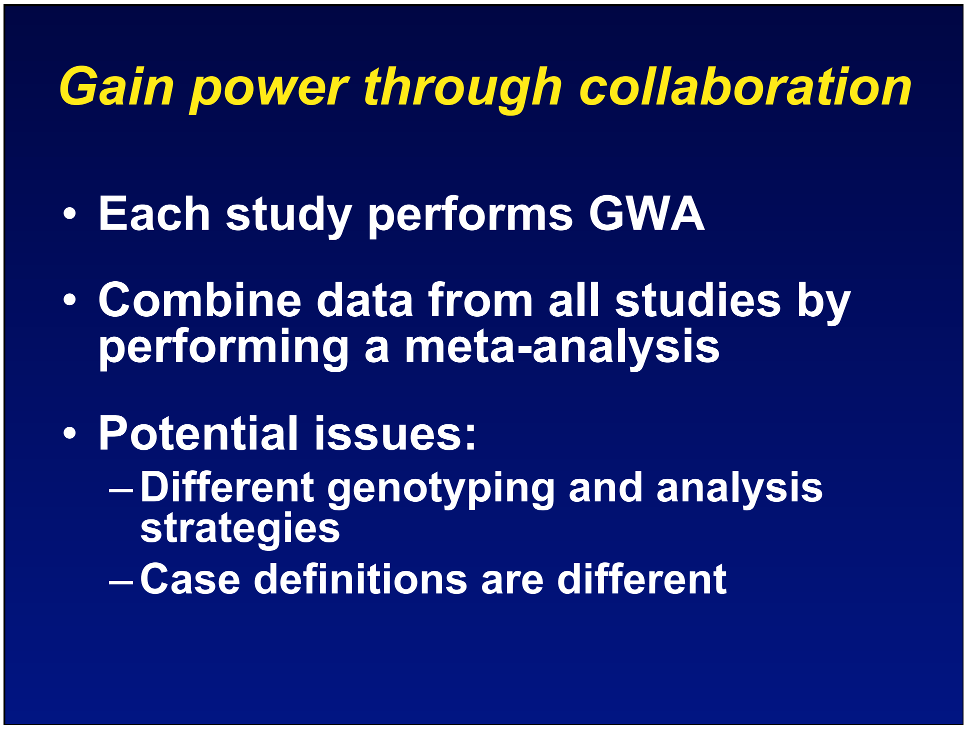


Table 1 Power of GWASs to discover several recently defined associations

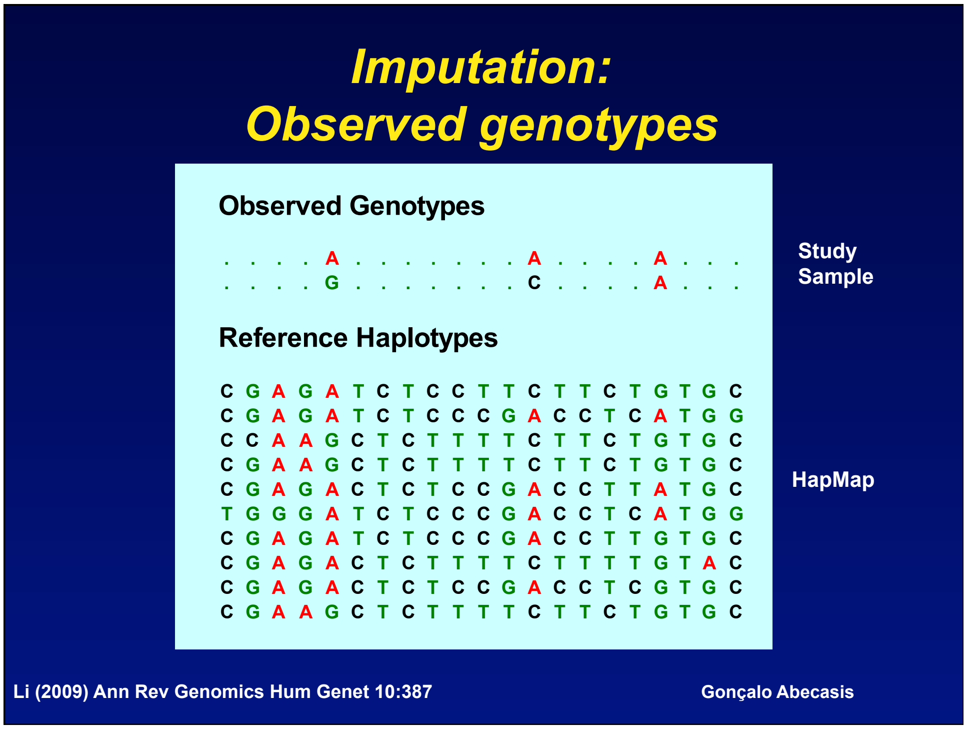
		Power in a 'typical' GWAS (1,000 cases/1,000 controls)			Sample size required for 90% power,		
Gene	Disease	1.0×10^{-2}	1.0×10^{-4}	1.0×10^{-8}	$P < 10^{-8}$	RAF	RR
ATG16L1	CD	>0.99	>0.99	0.74	2,430	0.5	1.5
IRGM	CD	0.67	0.19	<0.01	10,902	0.075	1.4
PTPN2	T1D, CD	0.37	0.05	<0.01	19,754	0.17	1.2
IL2	T1D	0.11	<0.01	<0.01	54,600	0.26	1.1
9p21	MI	0.97	0.87	0.09	5,066	0.47	1.25
9p21	T2D	0.36	0.05	<0.01	20,220	0.83	1.2
CDKAL1	T2D	0.35	0.04	<0.01	20,700	0.31	1.15

Altshuler (2007) Nat Gen 7:813



Gain power through collaboration

- Each study performs GWA
- Combine data from all studies by performing a meta-analysis
- Potential issues:
  - Different genotyping and analysis strategies
  - Case definitions are different



Imputation: Observed genotypes

Observed Genotypes

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

Study Sample

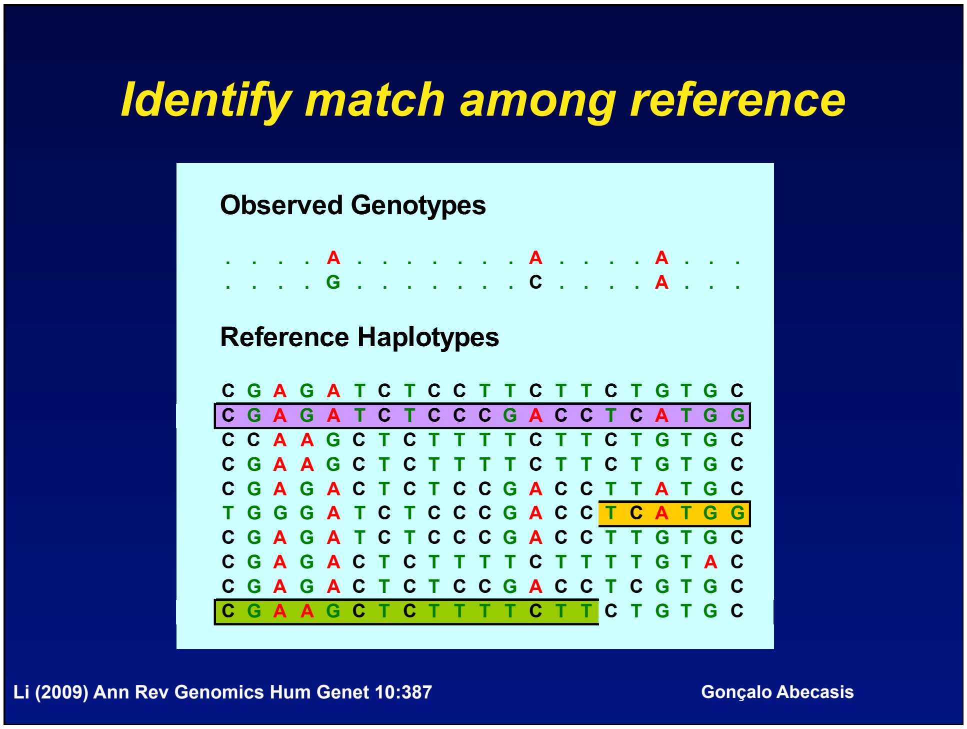
Reference Haplotypes

```
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

Li (2009) Ann Rev Genomics Hum Genet 10:387

Gonçalo Abecasis



Identify match among reference

Observed Genotypes

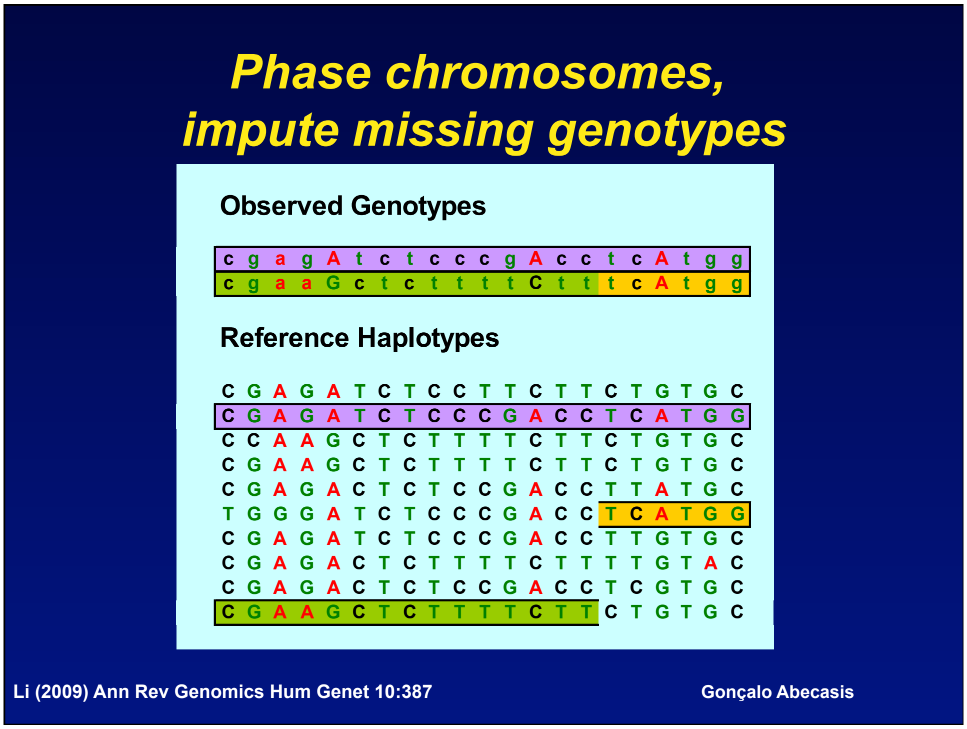
```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

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
```

Li (2009) Ann Rev Genomics Hum Genet 10:387

Gonçalo Abecasis



Phase chromosomes, impute missing genotypes

Observed Genotypes

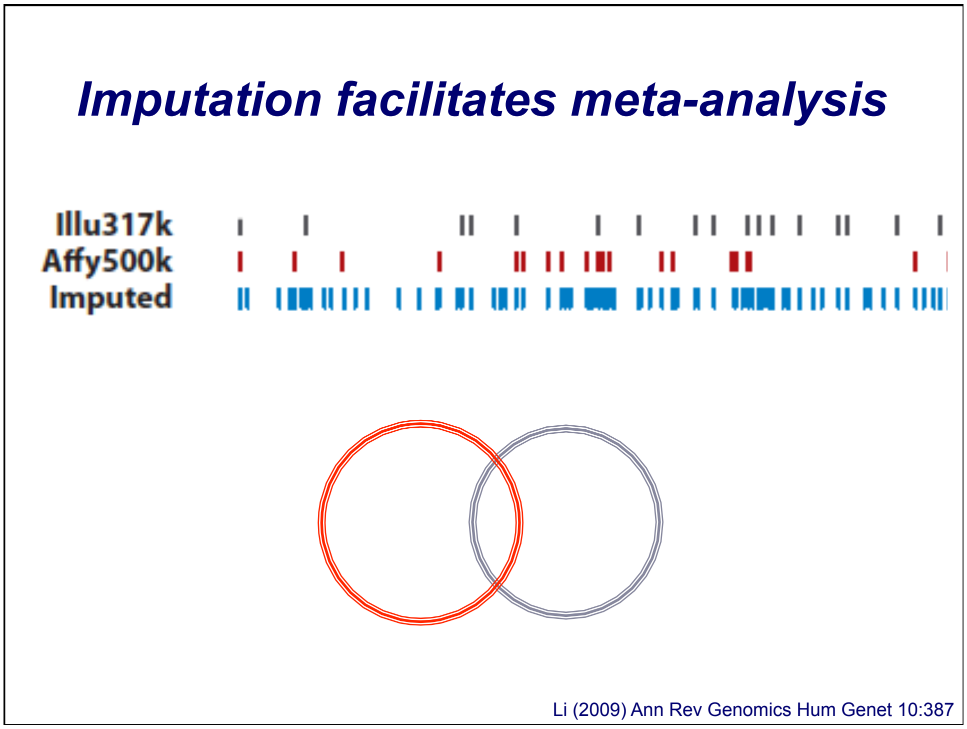
```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

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
```

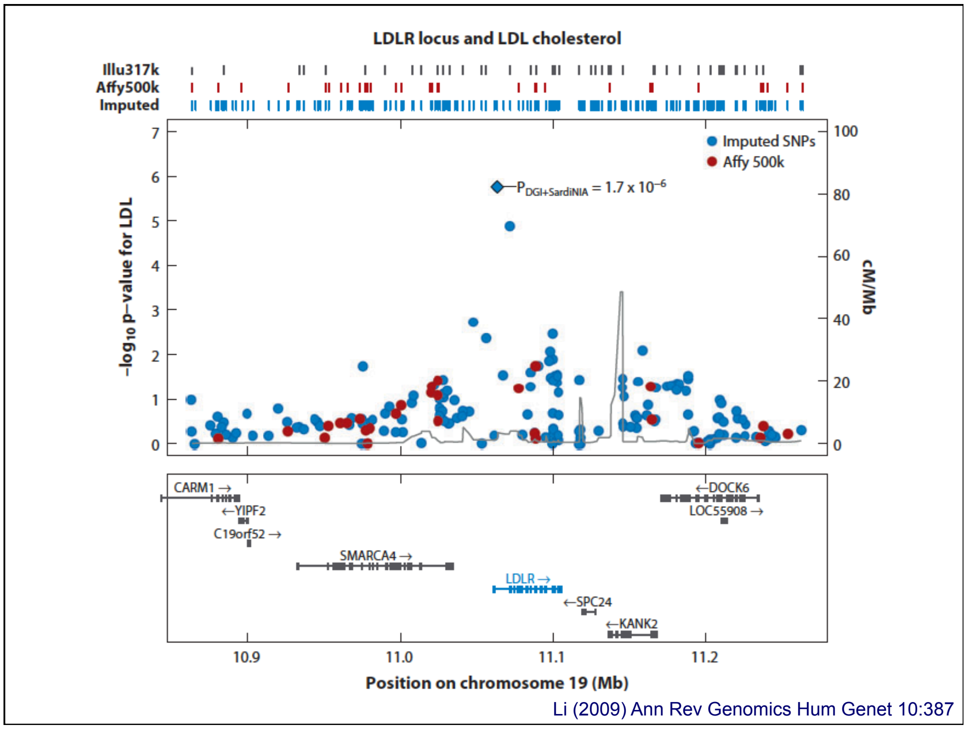
Li (2009) Ann Rev Genomics Hum Genet 10:387

Gonçalo Abecasis



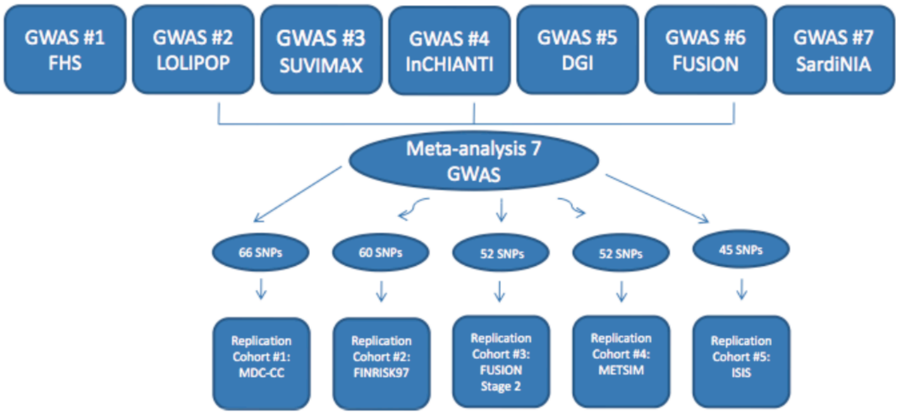
Imputation facilitates meta-analysis

Li (2009) Ann Rev Genomics Hum Genet 10:387



LDLR locus and LDL cholesterol

Li (2009) Ann Rev Genomics Hum Genet 10:387



GWAS #1 FHS

GWAS #2 LOLIPOP

GWAS #3 SUVIMAX

GWAS #4 InCHIANTI

GWAS #5 DGI

GWAS #6 FUSION

GWAS #7 SardiNIA

Meta-analysis 7 GWAS

66 SNPs → Replication Cohort #1: MDC-CC

60 SNPs → Replication Cohort #2: FINRISK97

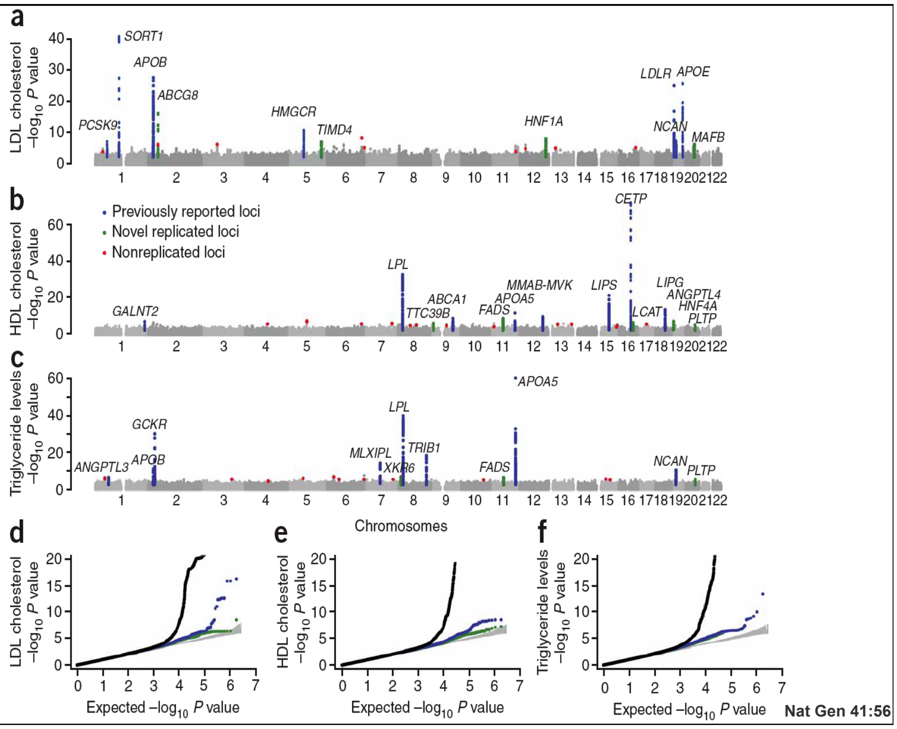
52 SNPs → Replication Cohort #3: FUSION Stage 2

52 SNPs → Replication Cohort #4: METSIM

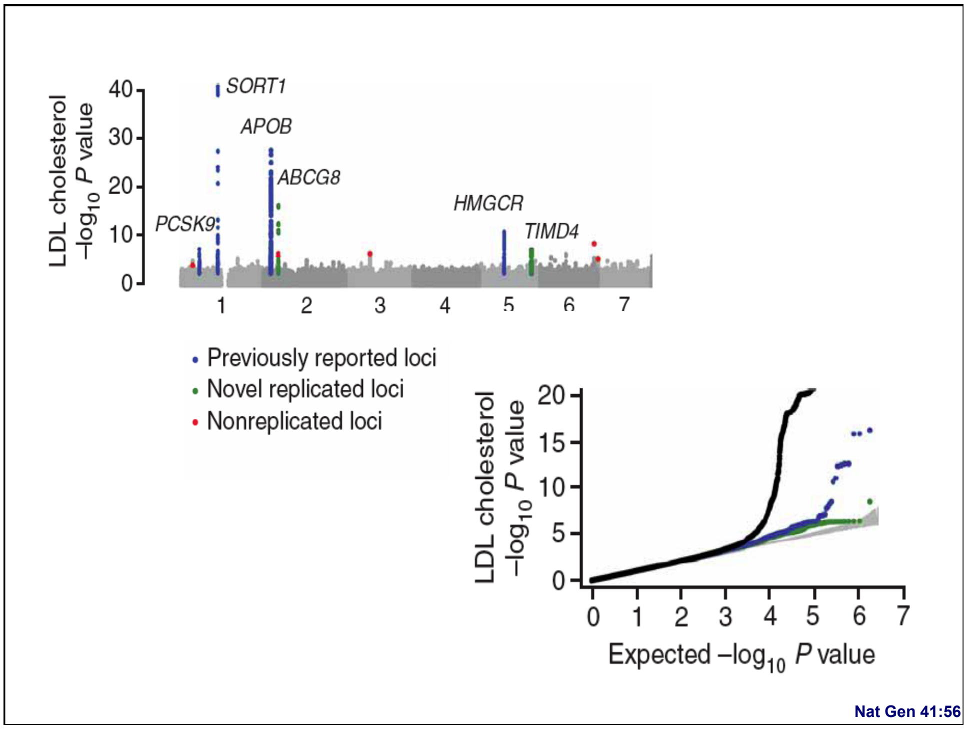
45 SNPs → Replication Cohort #5: ISIS

GWA in ~19,840 individuals
Follow-up in ~20,623 individuals

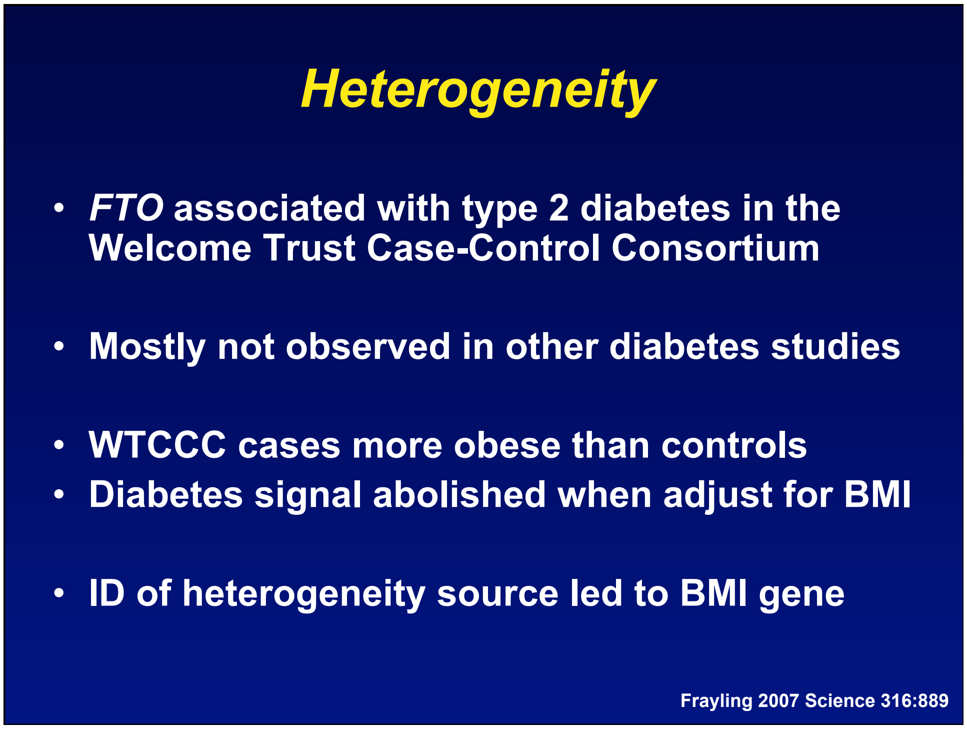
Kathiresan (2009) Nat Gen 41:56



Nat Gen 41:56



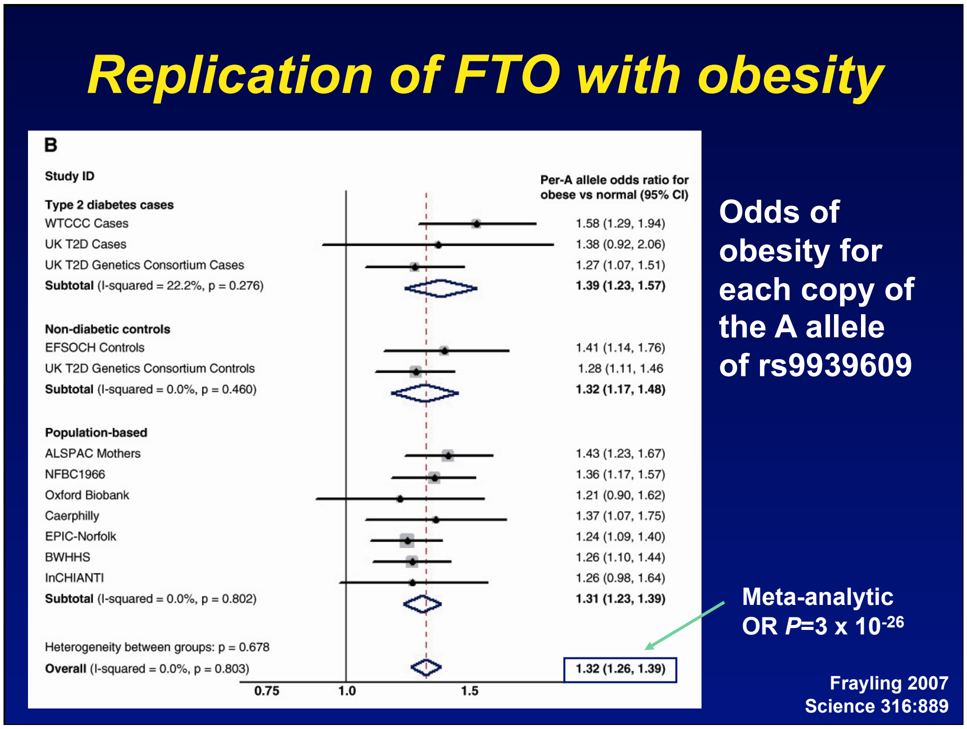
Nat Gen 41:56



Heterogeneity

- *FTO* associated with type 2 diabetes in the Welcome Trust Case-Control Consortium
- Mostly not observed in other diabetes studies
- WTCCC cases more obese than controls
- Diabetes signal abolished when adjust for BMI
- ID of heterogeneity source led to BMI gene

Frayling 2007 Science 316:889



Replication of FTO with obesity

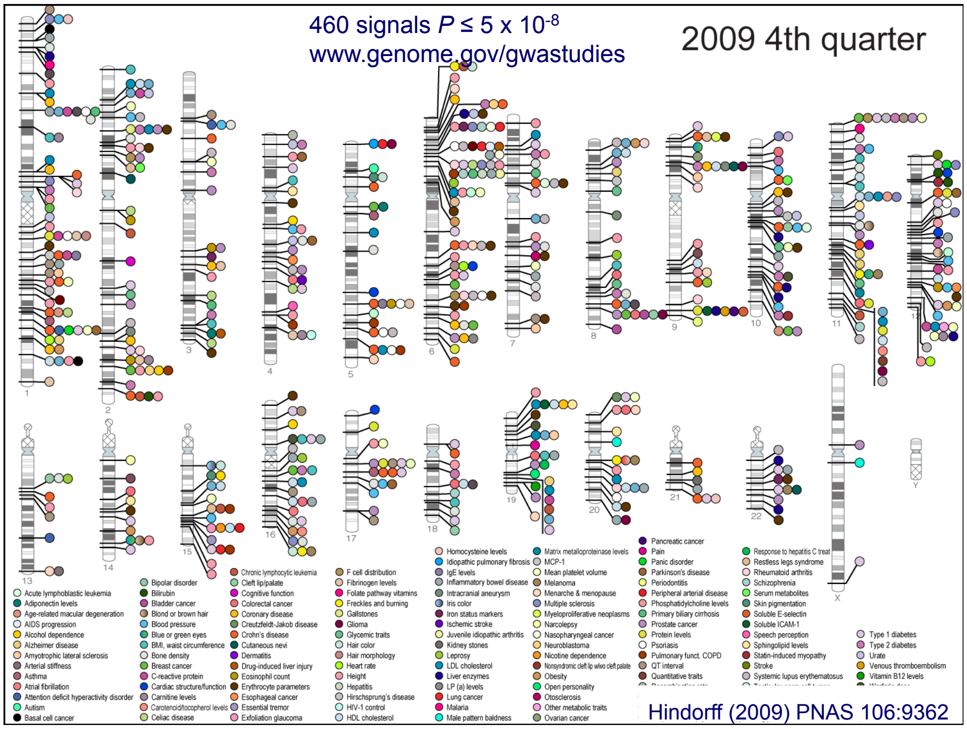
B

Study ID	Per-A allele odds ratio for obese vs normal (95% CI)
Type 2 diabetes cases	
WTCCC Cases	1.58 (1.29, 1.94)
UK T2D Cases	1.38 (0.92, 2.06)
UK T2D Genetics Consortium Cases	1.27 (1.07, 1.51)
Subtotal (I-squared = 22.2%, p = 0.276)	**1.39 (1.23, 1.57)**
Non-diabetic controls	
EFSOCH Controls	1.41 (1.14, 1.76)
UK T2D Genetics Consortium Controls	1.28 (1.11, 1.46
Subtotal (I-squared = 0.0%, p = 0.460)	**1.32 (1.17, 1.48)**
Population-based	
ALSPAC Mothers	1.43 (1.23, 1.67)
NFBC1966	1.36 (1.17, 1.57)
Oxford Biobank	1.21 (0.90, 1.62)
Caerphilly	1.37 (1.07, 1.75)
EPIC-Norfolk	1.24 (1.09, 1.40)
BWHHS	1.26 (1.10, 1.44)
InCHIANTI	1.26 (0.98, 1.64)
Subtotal (I-squared = 0.0%, p = 0.802)	**1.31 (1.23, 1.39)**
Heterogeneity between groups: p = 0.678	
Overall (I-squared = 0.0%, p = 0.803)	**1.32 (1.26, 1.39)**

Odds of obesity for each copy of the A allele of rs9939609

Meta-analytic OR *P*=3 x 10^{-26}

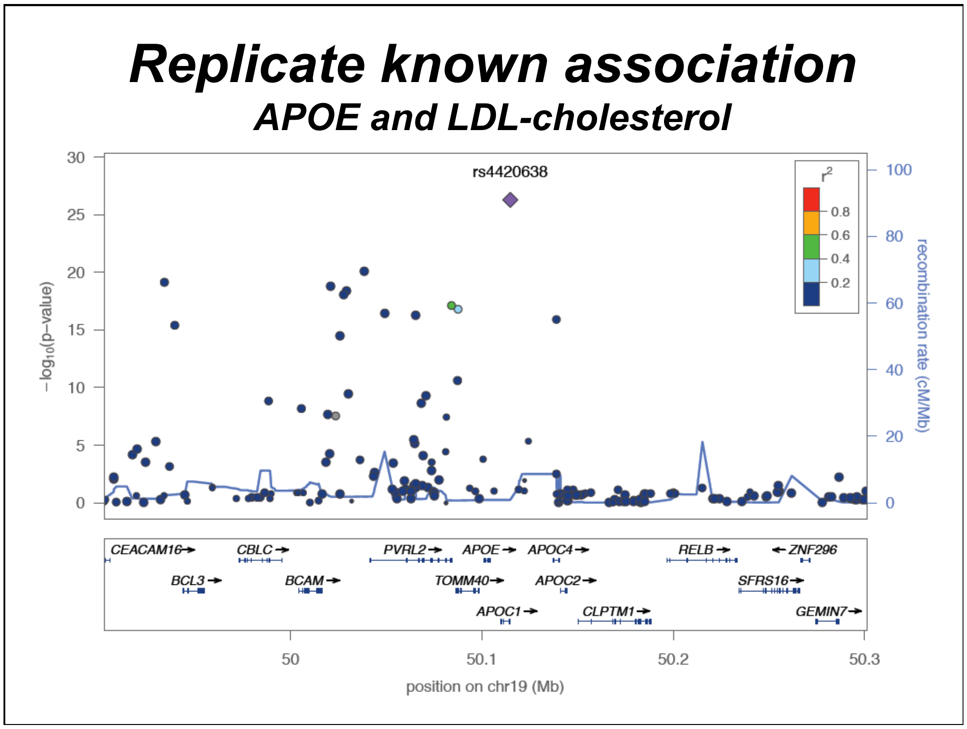
Frayling 2007
Science 316:889



460 signals $P \leq 5 \times 10^{-8}$
www.genome.gov/gwastudies

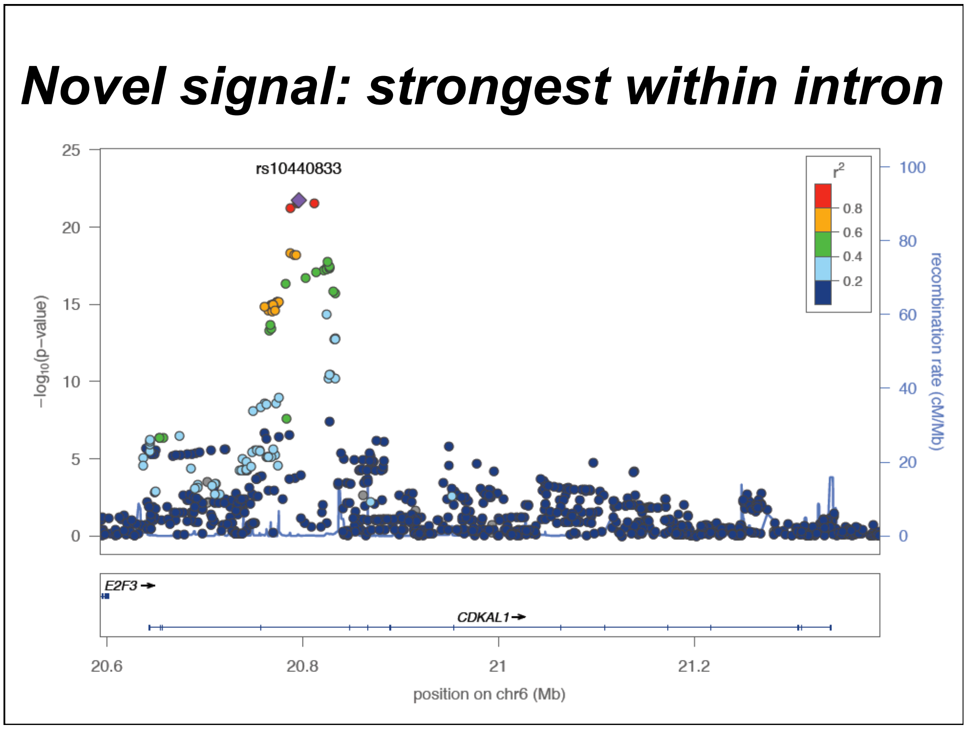
2009 4th quarter

Hindorff (2009) PNAS 106:9362



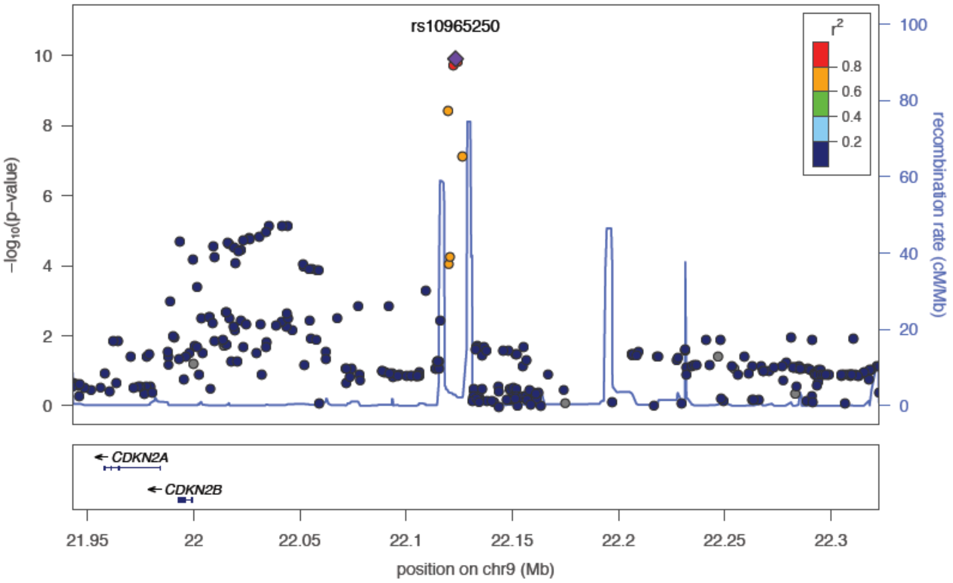
Replicate known association

APOE and LDL-cholesterol

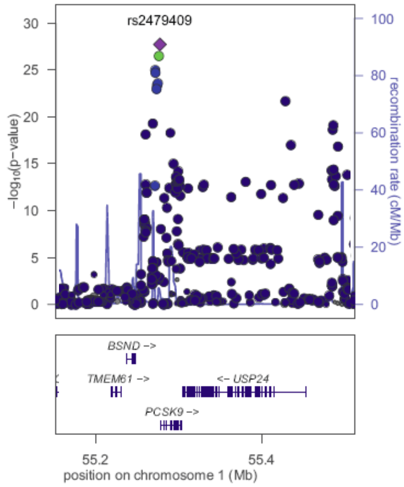


Novel signal: strongest within intron

Novel signal: outside of gene



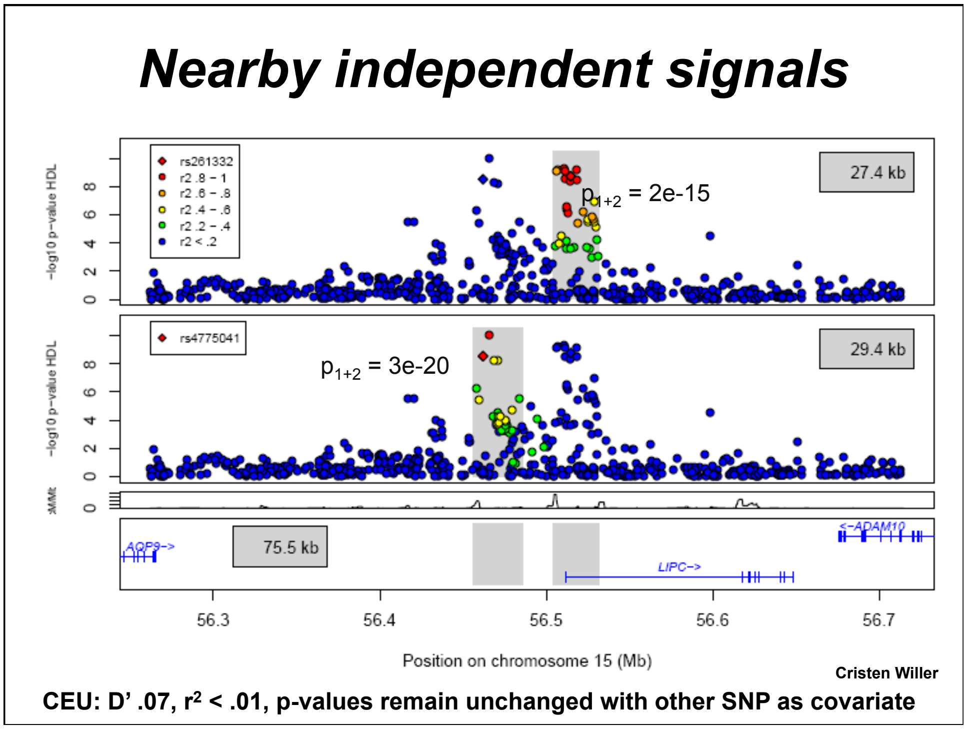
New common variant near gene with known rare variants



Private mutations in *PCSK9* change LDL by >100 mg/dl
Abifadel et al, 2003

Rare variants (MAF 1%) in *PCSK9* can change LDL by ~16 mg/dl
Cohen et al, 2005

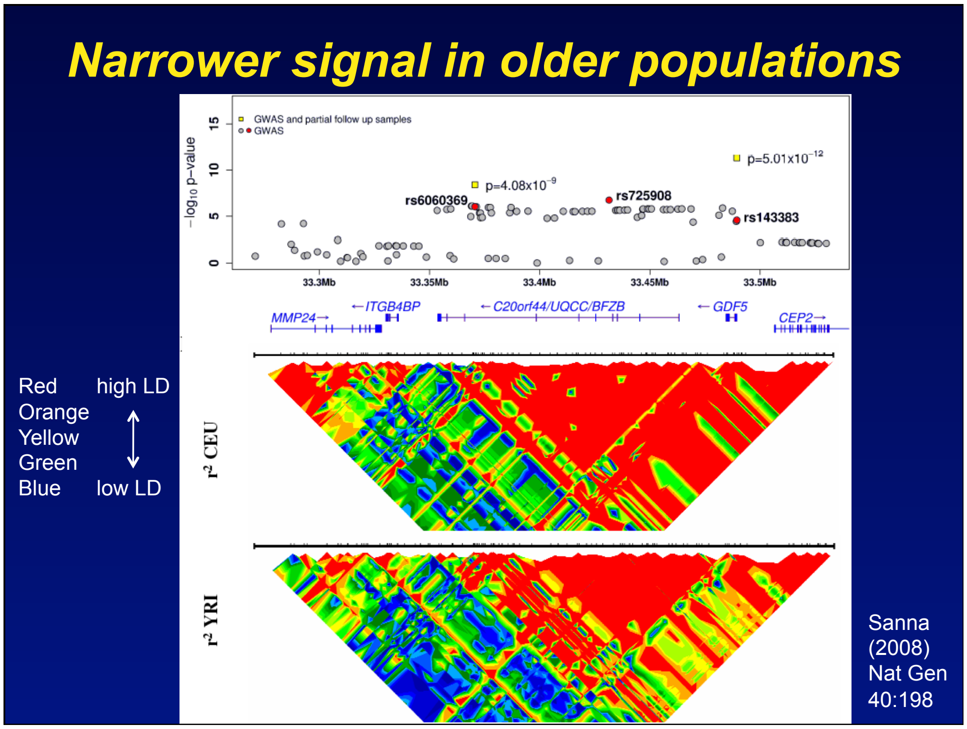
Common variants (MAF 20%) in *PCSK9* change LDL by ~3 mg/dl
Willer et al, 2008



Nearby independent signals

Cristen Willer

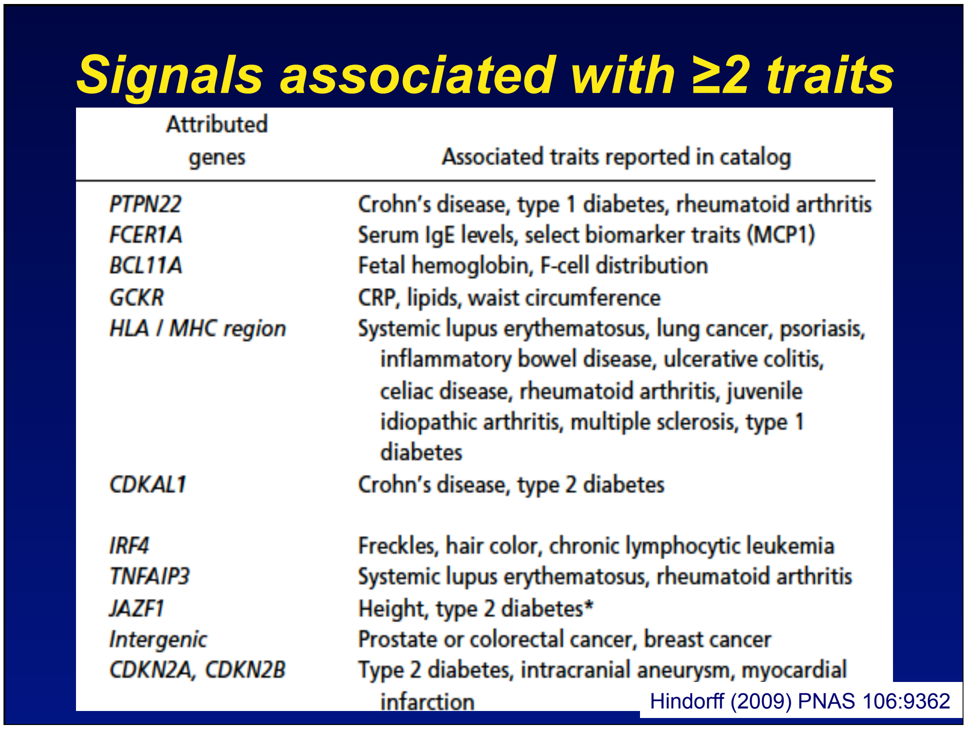
CEU: D' .07, r^2 < .01, p-values remain unchanged with other SNP as covariate



Narrower signal in older populations

Red high LD
Orange
Yellow
Green
Blue low LD

Sanna (2008) Nat Gen 40:198

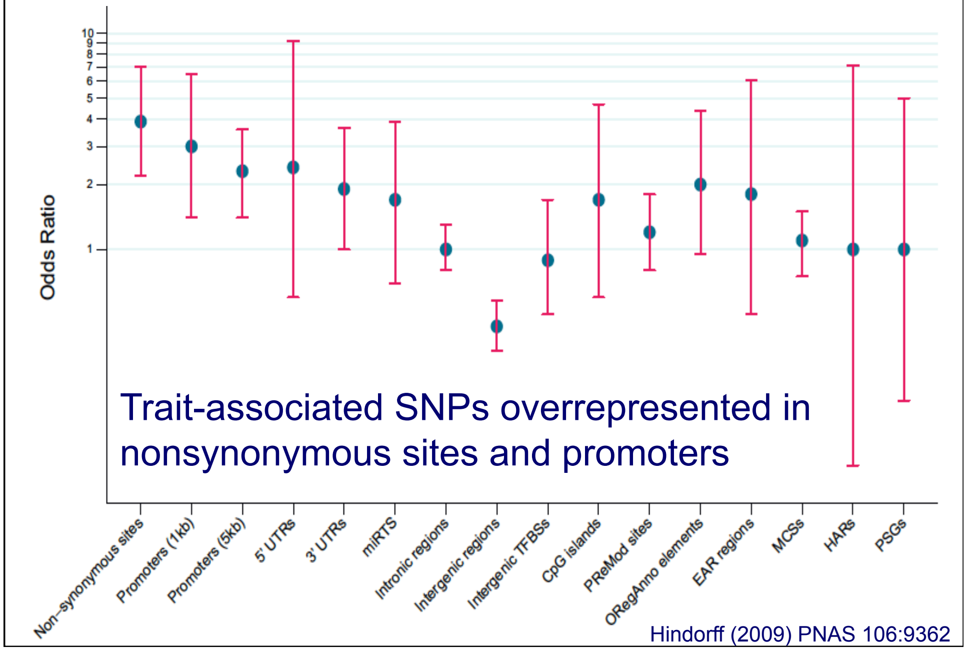


Signals associated with ≥2 traits

Attributed genes	Associated traits reported in catalog
PTPN22	Crohn's disease, type 1 diabetes, rheumatoid arthritis
FCER1A	Serum IgE levels, select biomarker traits (MCP1)
BCL11A	Fetal hemoglobin, F-cell distribution
GCKR	CRP, lipids, waist circumference
HLA / MHC region	Systemic lupus erythematosus, lung cancer, psoriasis, inflammatory bowel disease, ulcerative colitis, celiac disease, rheumatoid arthritis, juvenile idiopathic arthritis, multiple sclerosis, type 1 diabetes
CDKAL1	Crohn's disease, type 2 diabetes
IRF4	Freckles, hair color, chronic lymphocytic leukemia
TNFAIP3	Systemic lupus erythematosus, rheumatoid arthritis
JAZF1	Height, type 2 diabetes*
Intergenic	Prostate or colorectal cancer, breast cancer
CDKN2A, CDKN2B	Type 2 diabetes, intracranial aneurysm, myocardial infarction

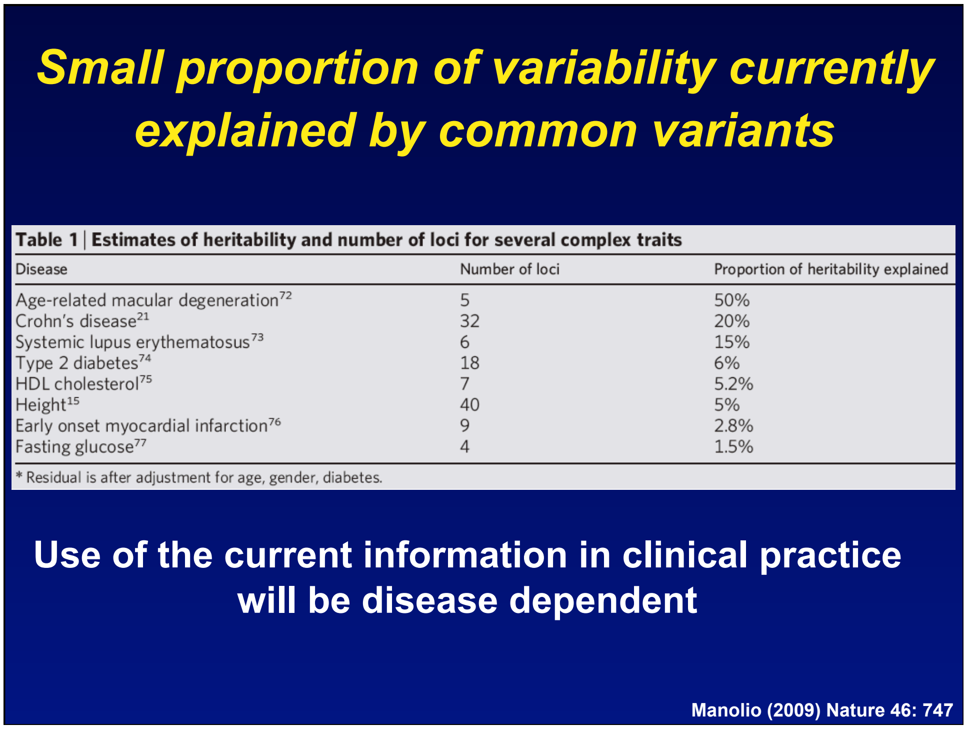
Hindorff (2009) PNAS 106:9362

Types of associated variants



Trait-associated SNPs overrepresented in nonsynonymous sites and promoters

Hindorff (2009) PNAS 106:9362



Small proportion of variability currently explained by common variants

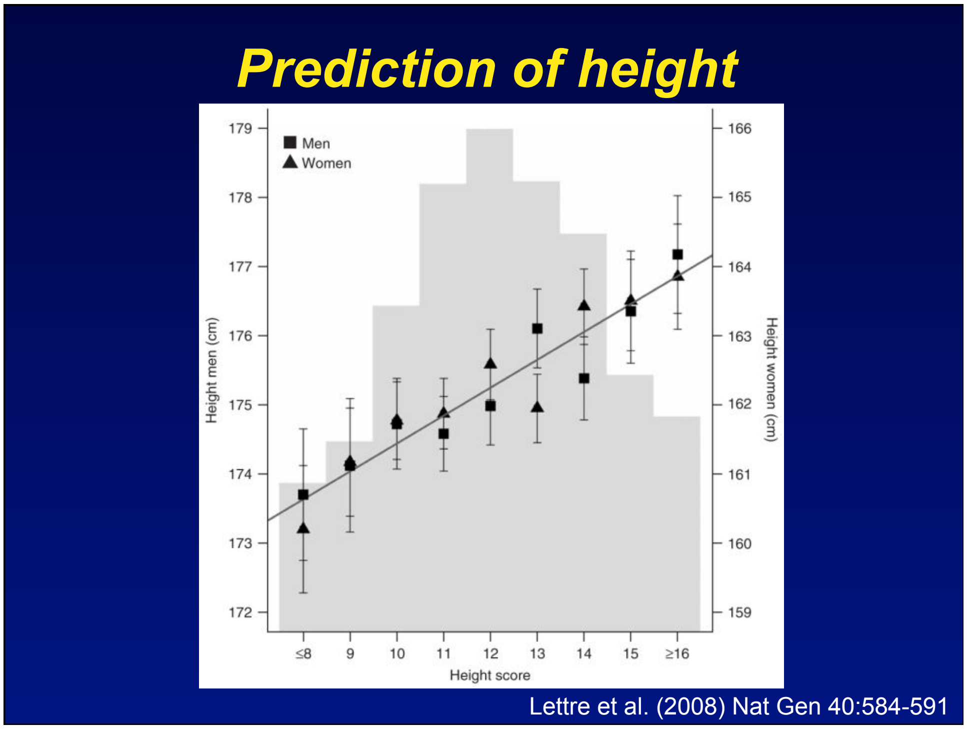
Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

* Residual is after adjustment for age, gender, diabetes.

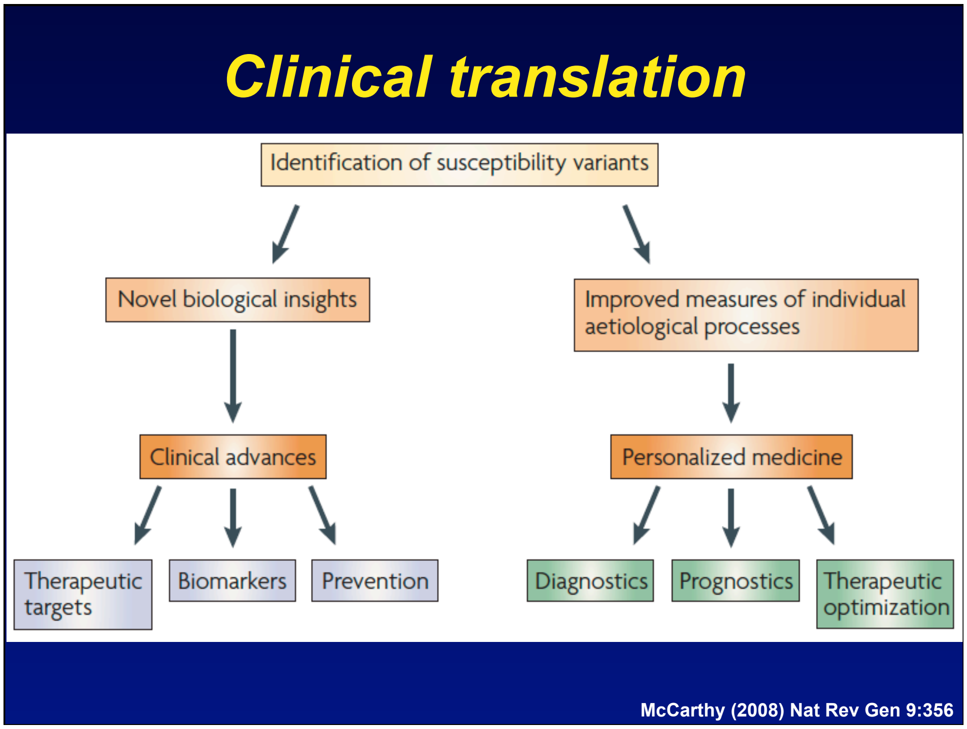
Use of the current information in clinical practice will be disease dependent

Manolio (2009) Nature 46: 747



Prediction of height

Lettre et al. (2008) Nat Gen 40:584-591



Clinical translation

Identification of susceptibility variants

Novel biological insights

Improved measures of individual aetiological processes

Clinical advances

Personalized medicine

Therapeutic targets

Biomarkers

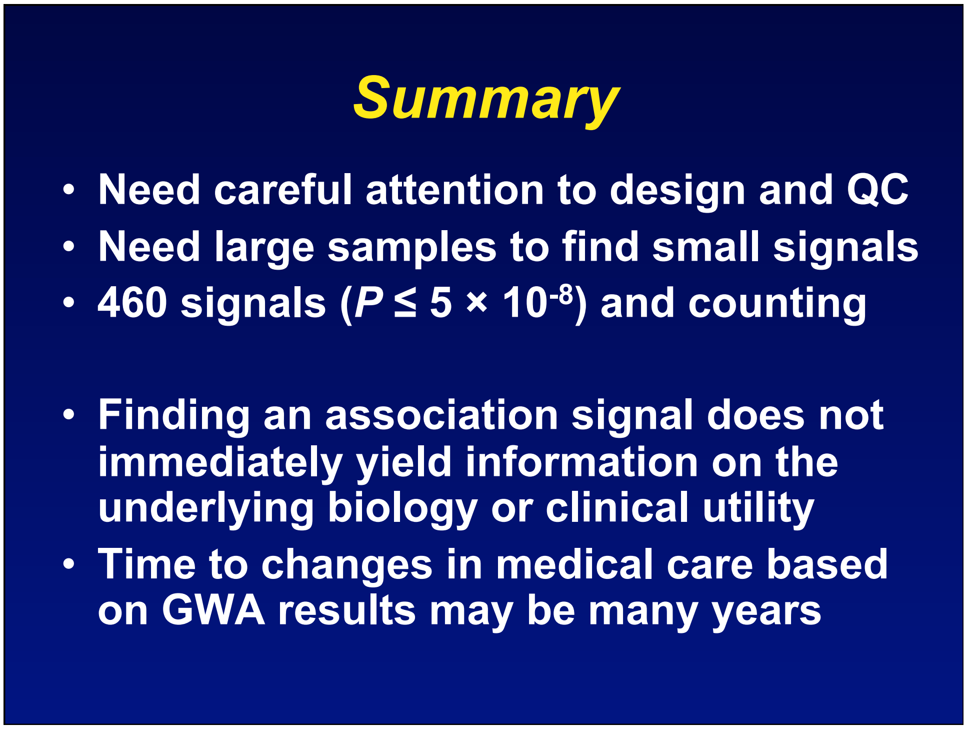
Prevention

Diagnostics

Prognostics

Therapeutic optimization

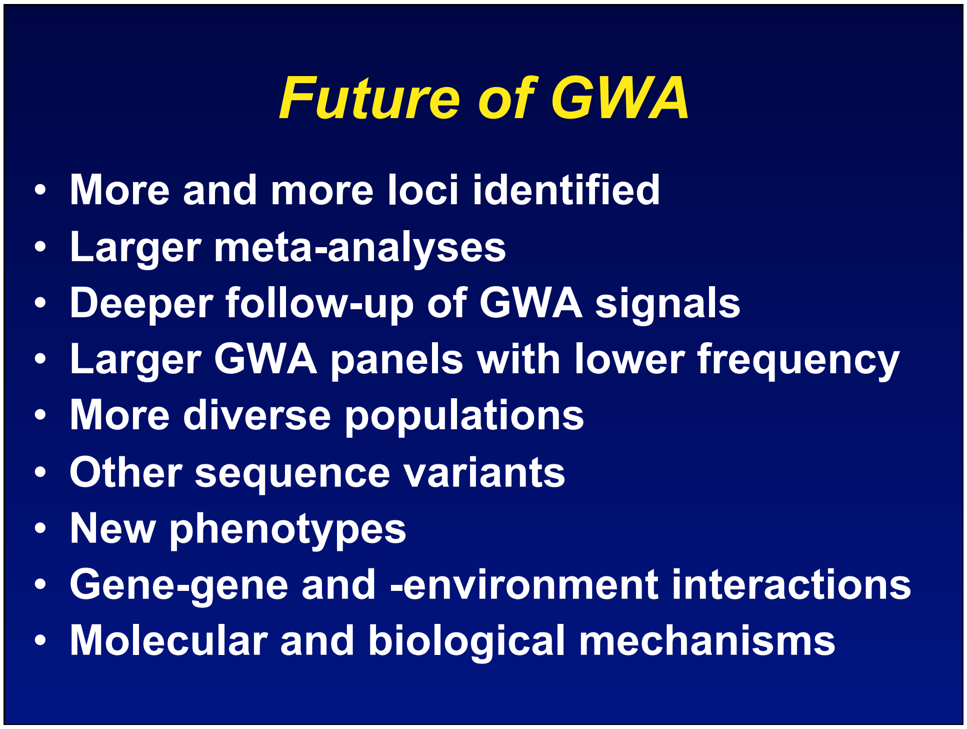
McCarthy (2008) Nat Rev Gen 9:356



Summary

- Need careful attention to design and QC
- Need large samples to find small signals
- 460 signals ($P \leq 5 \times 10^{-8}$) and counting

- Finding an association signal does not immediately yield information on the underlying biology or clinical utility
- Time to changes in medical care based on GWA results may be many years



Future of GWA

- More and more loci identified
- Larger meta-analyses
- Deeper follow-up of GWA signals
- Larger GWA panels with lower frequency
- More diverse populations
- Other sequence variants
- New phenotypes
- Gene-gene and -environment interactions
- Molecular and biological mechanisms