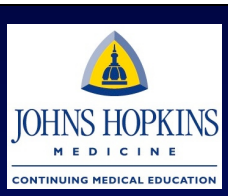
Genome-wide association studies

Karen Mohlke, PhD
Department of Genetics
University of North Carolina



Current Topics in Genome Analysis 2012

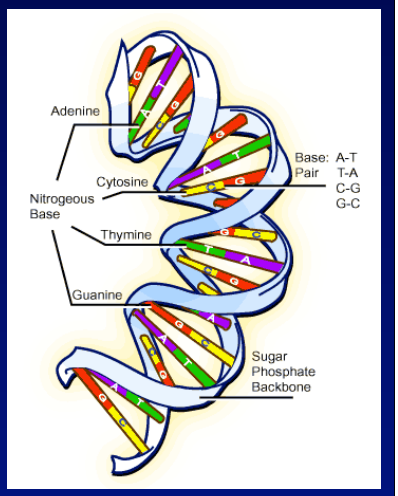
Karen Mohlke, PhD

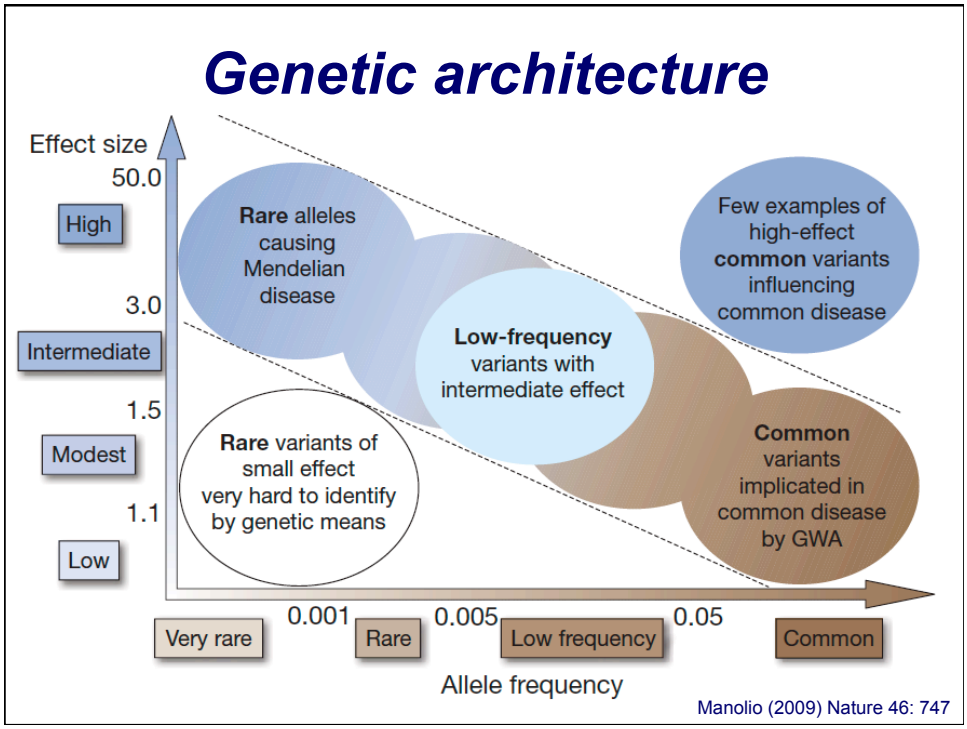
No Relevant Financial Relationships with Commercial Interests

Complex traits

Common and rare variants

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





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

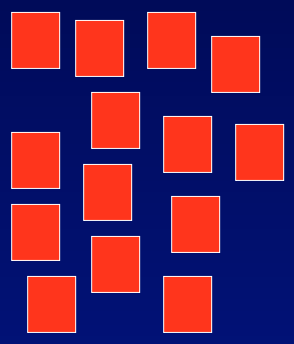
- Samples
- Genotyping
- Quality control
- Statistical analysis
- Replication

Phenotype

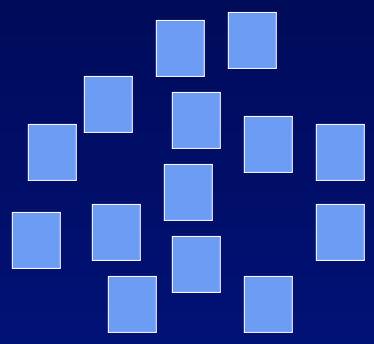
- Disease (case/control)
  - Rare
  - Common
- Quantitative trait
  - Easy to measure: Weight, height
  - Requires testing: Coronary artery thickness
  - Requires experiment: Gene expression

Selection of cases and controls

Cases



Controls

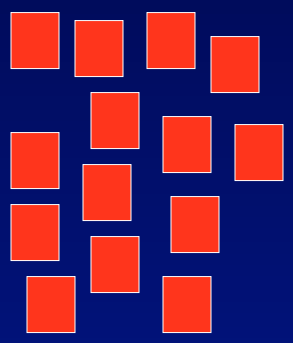


Definition of case?

Definition of control?

Selection of cases

Cases

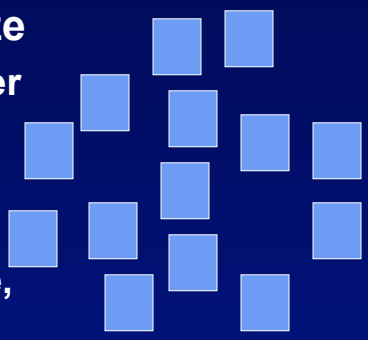


- Potential criteria to enrich genetic effect size
  - More severely affected individuals
  - Require other family member to have disease
  - Younger age-of-disease onset

Selection of controls

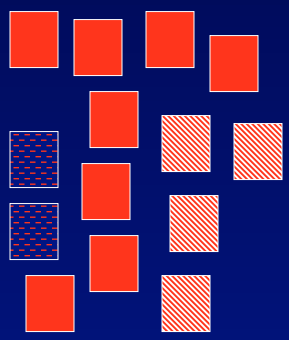
- Potential criteria to enrich genetic effect size
  - Low risk of disease rather than population-based samples
  - Same ancestry as cases
  - Matched to cases on age, sex, demographics

Controls

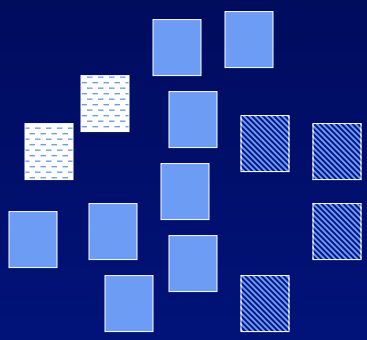


Matched ancestry

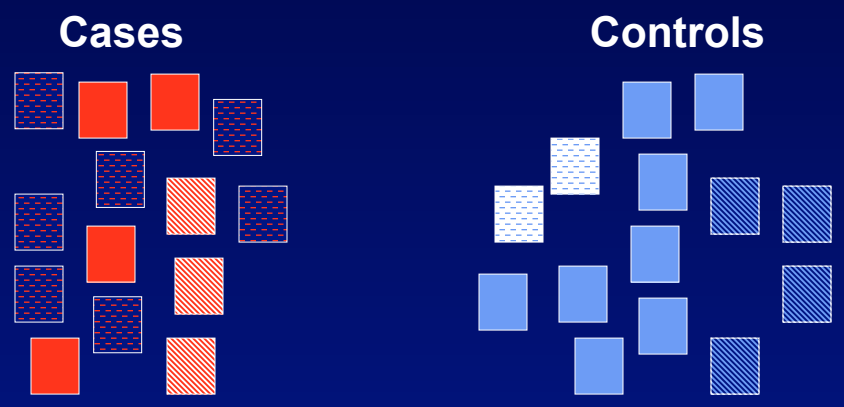
Cases



Controls



Unmatched ancestry



May have inadequate ancestry information prior to genotyping

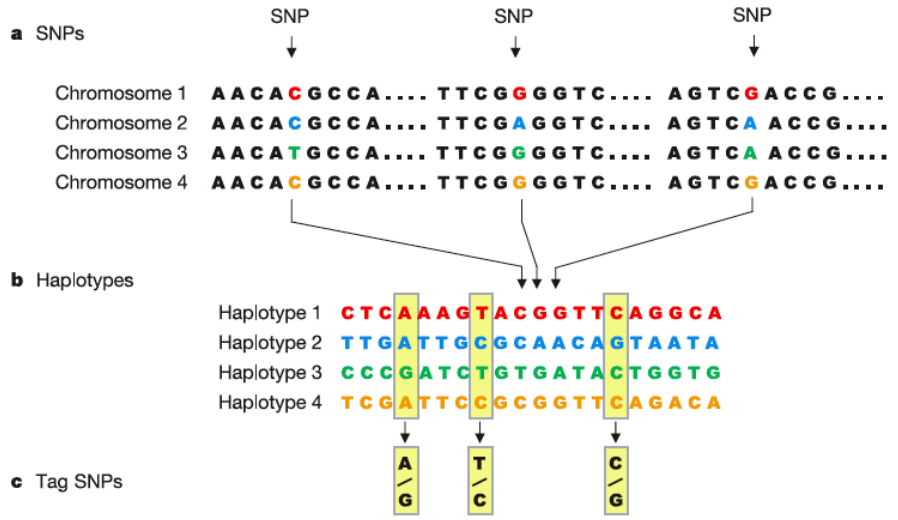
Population stratification and cryptic relatedness

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

Genome-wide SNP panels

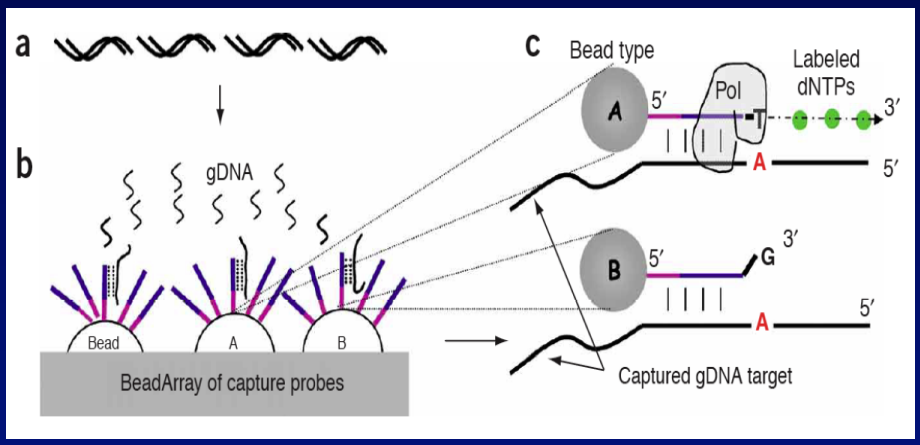
- 10,000 - 5 million SNPs
- Affymetrix, Illumina
  - Random SNPs
  - Selected haplotype tag SNPs
  - Copy number probes
  - Some arrays allow SNPs to be added

Selecting 'haplotype tag' SNPs



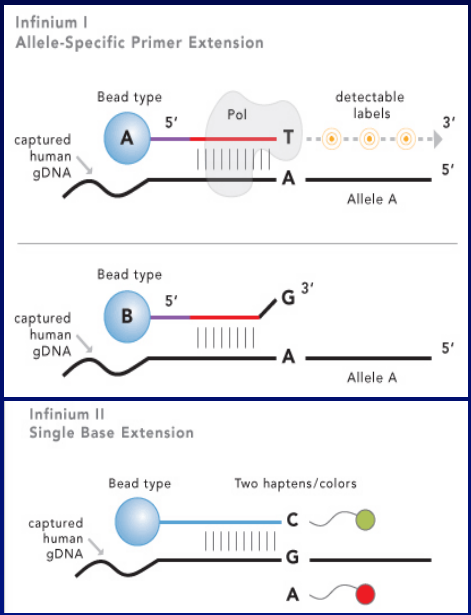
International HapMap Consortium (2003) Nature 426:789

Illumina Infinium Assay



Gunderson et al. (2005) NatGen 37:549

Illumina Infinium Assays



Illumina.com

Affymetrix GeneChip Array

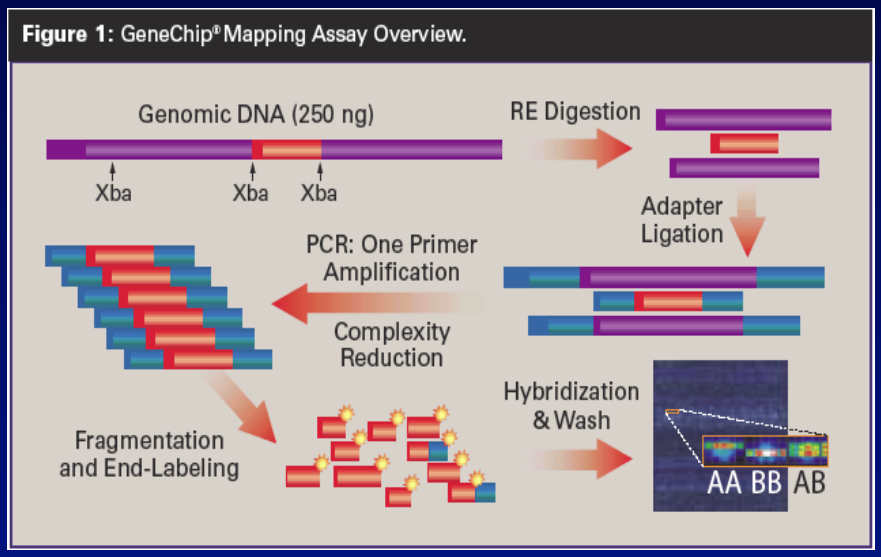
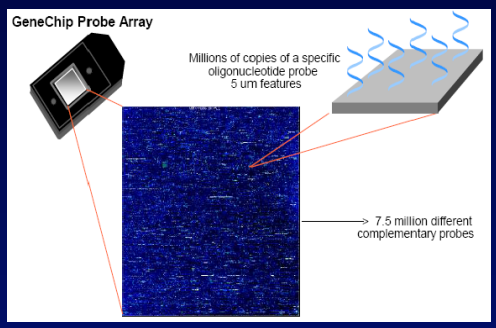
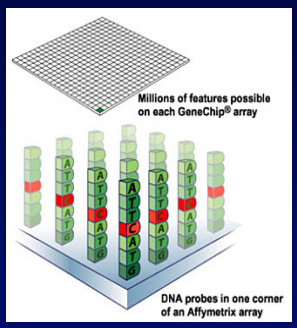
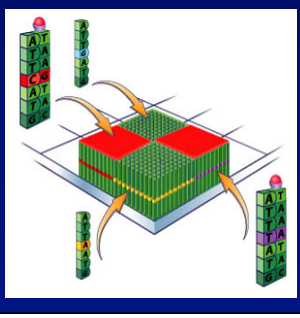


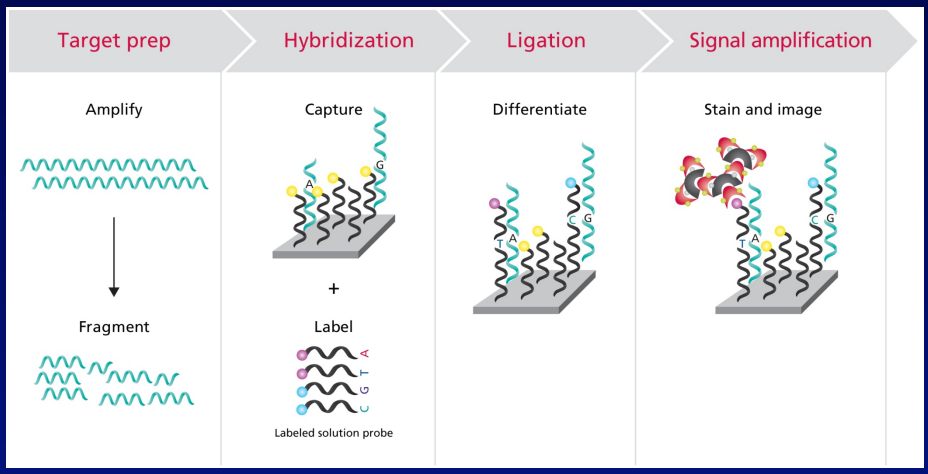
image from affymetrix.com







Affymetrix Axiom Array



Affymetrix.com

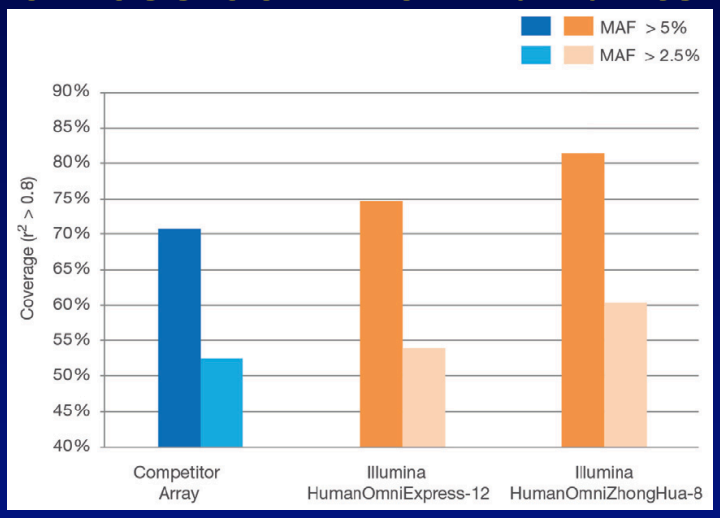
Global genomic coverage

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

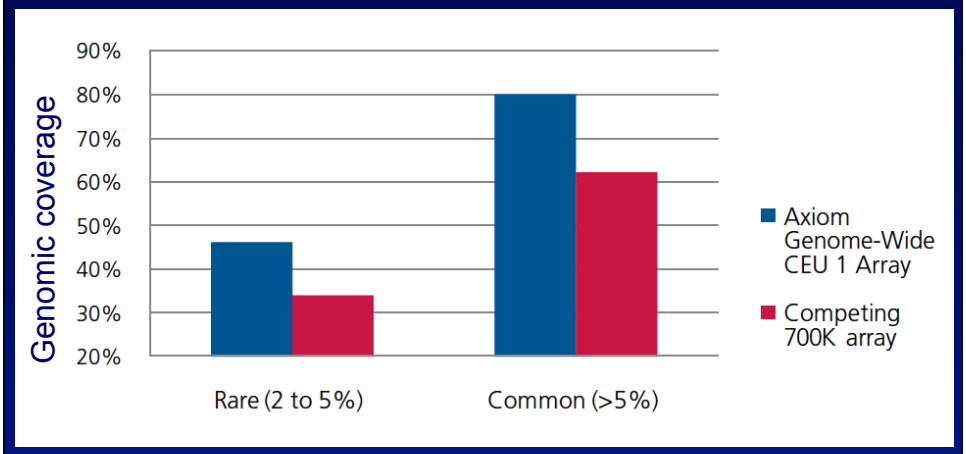
Newer arrays improve coverage of less common variants



Coverage calculations based on known common and rare Chinese population variants identified in the HapMap and 1000 Genomes projects

Illumina.com

Newer arrays improve coverage of less common variants

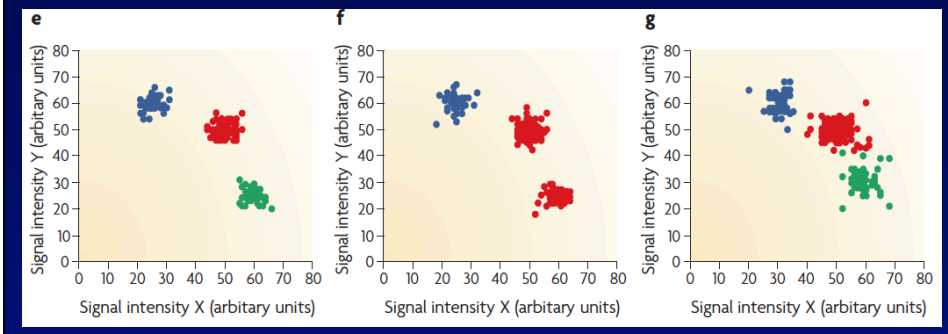


Affymetrix.com

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

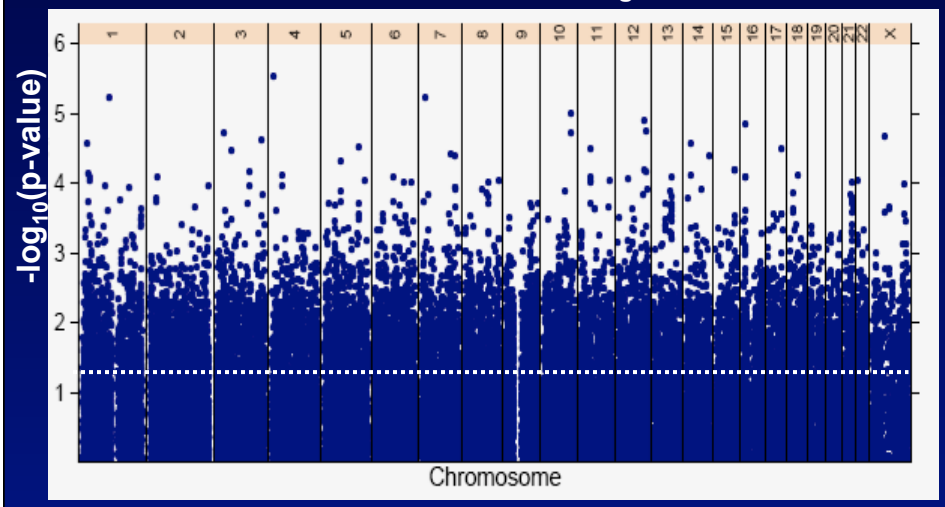
- Genotype and test > 300K – 5M SNPs
- Correct for the multiple tests

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

- Need large effect or large sample size

Type 2 diabetes association results

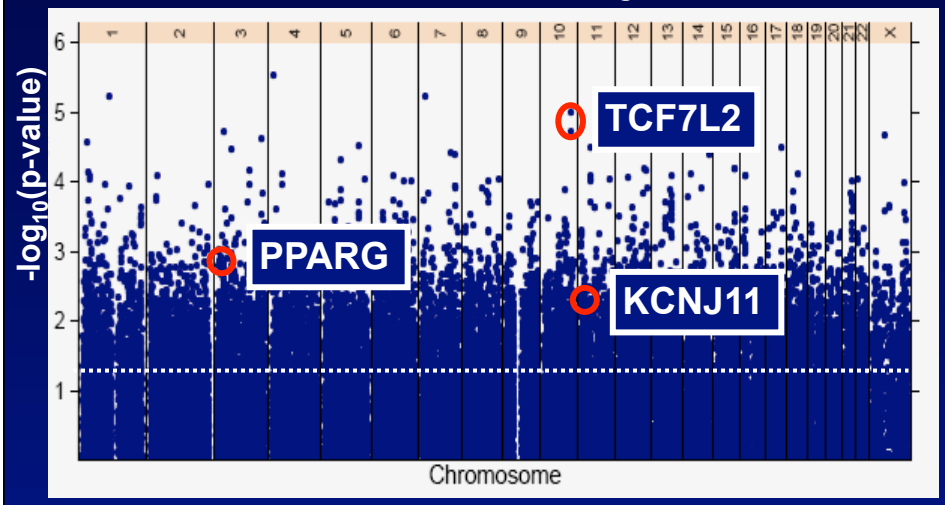
1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls



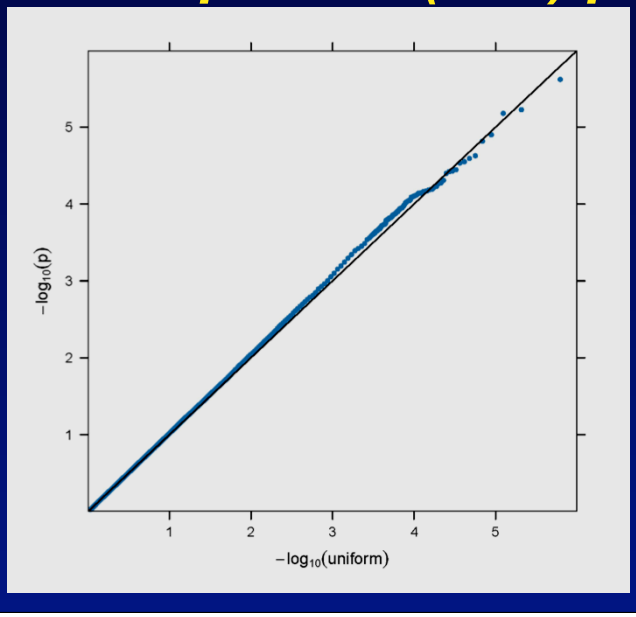
Logistic regression using additive model adjusted for age, gender, birth province

Which results are true positives?

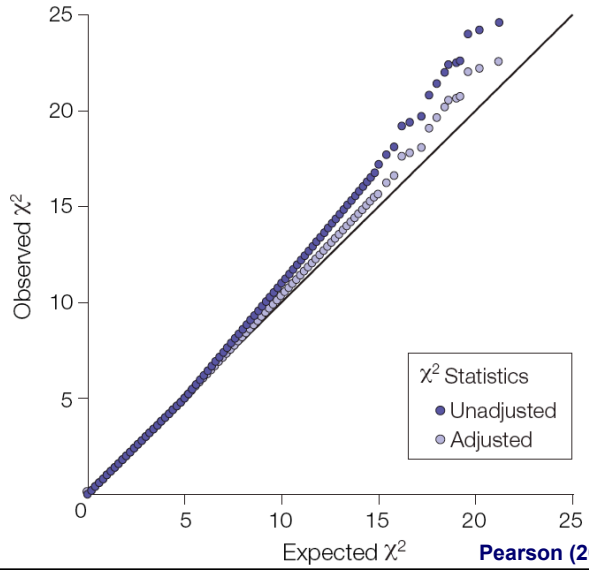
1161 Finnish T2D cases + 1174 Finnish normal glucose tolerant controls



Quantile-quantile (Q-Q) plot



Before and after adjustment of population stratification



Pearson (2008) JAMA 299:1335

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 Gen 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 Gen 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
```

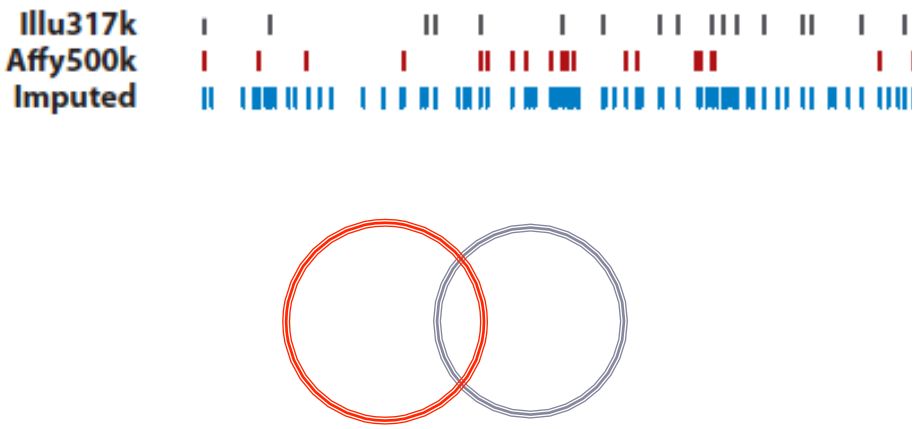
Reference Haplotypes

```
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 Gen Hum Genet 10:387

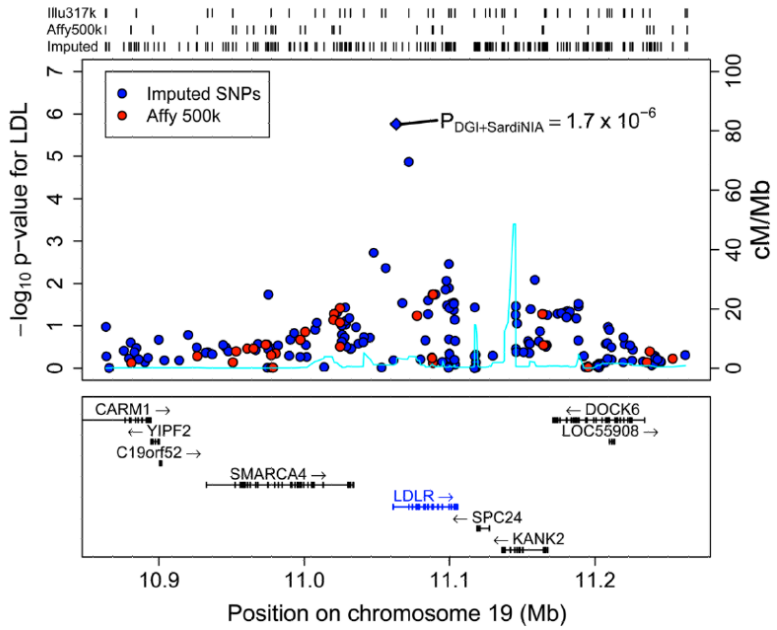
Gonçalo Abecasis

Imputation facilitates meta-analysis

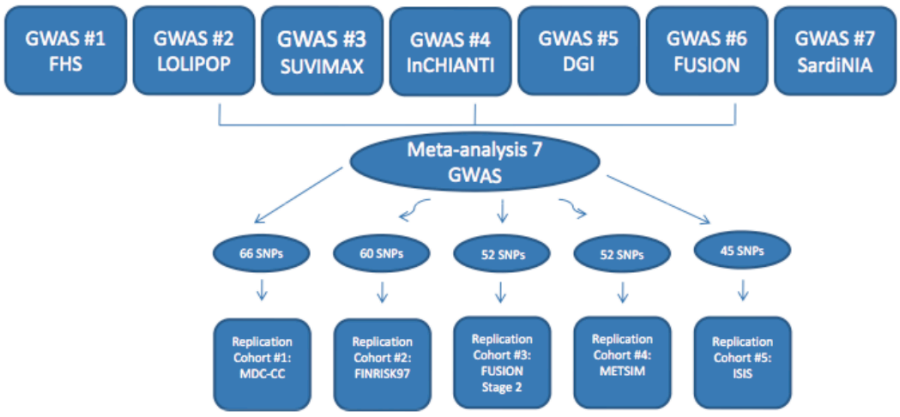


Li (2009) Ann Rev Genomics Hum Genet 10:387

LDLR locus and LDL cholesterol

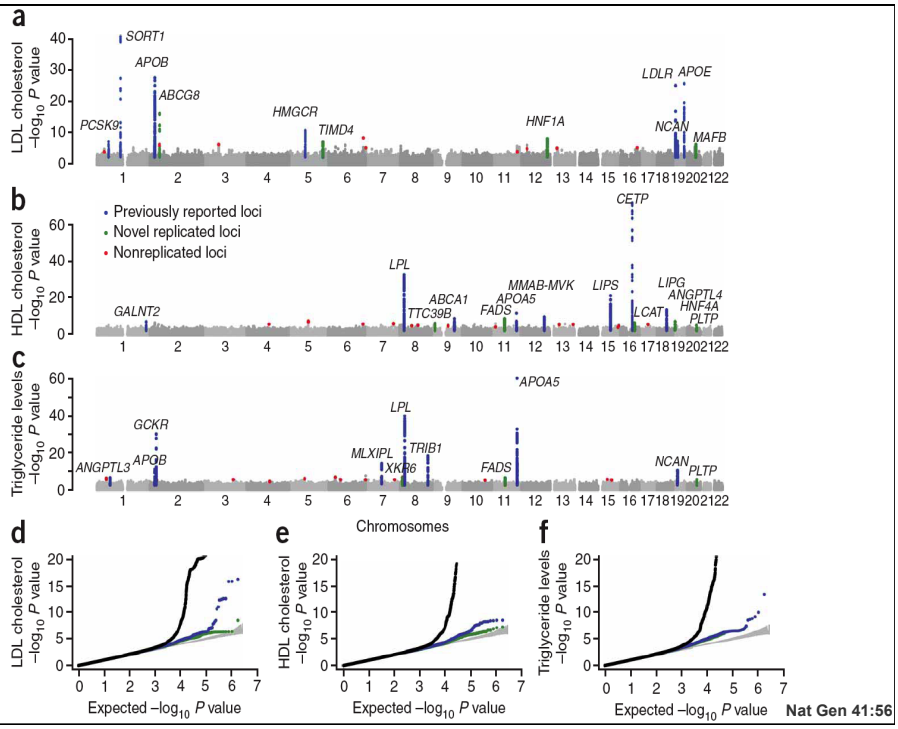


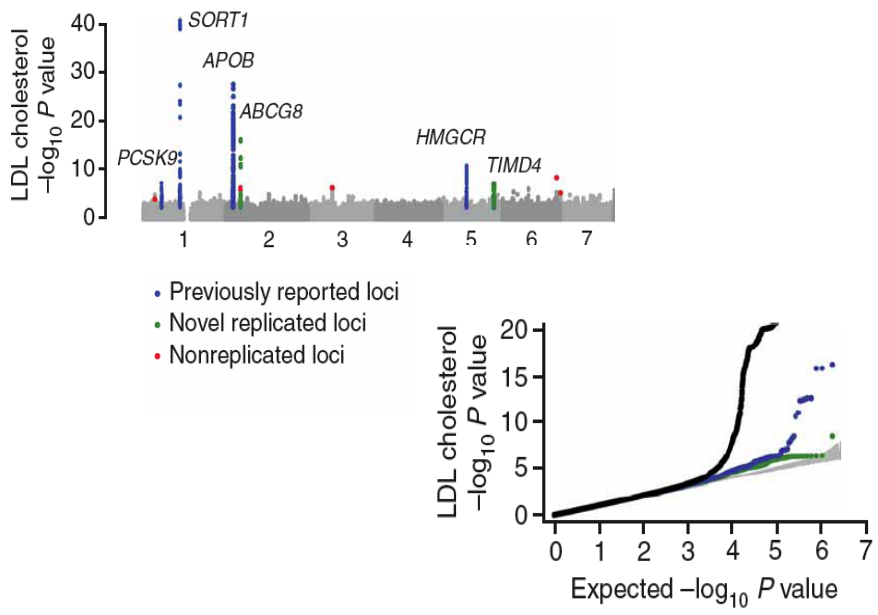
Li (2009) Ann Rev Genomics Hum Genet 10:387



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

Kathiresan (2009) Nat Gen 41:56





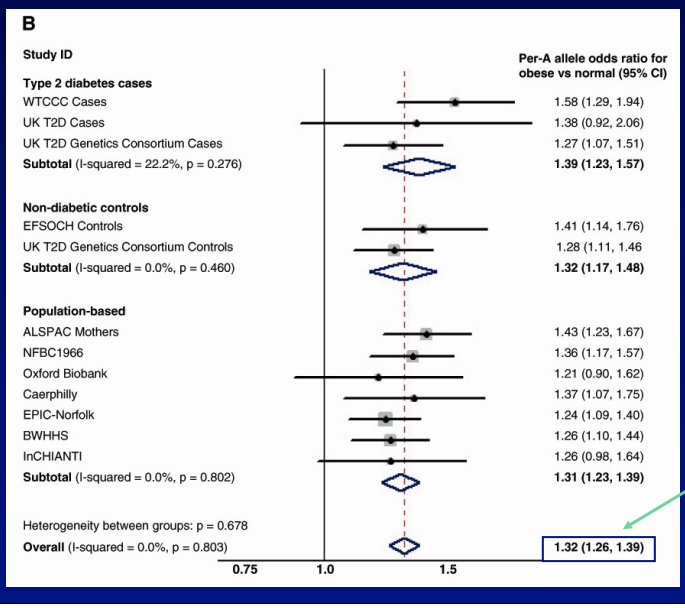
Nat Gen 41:56

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

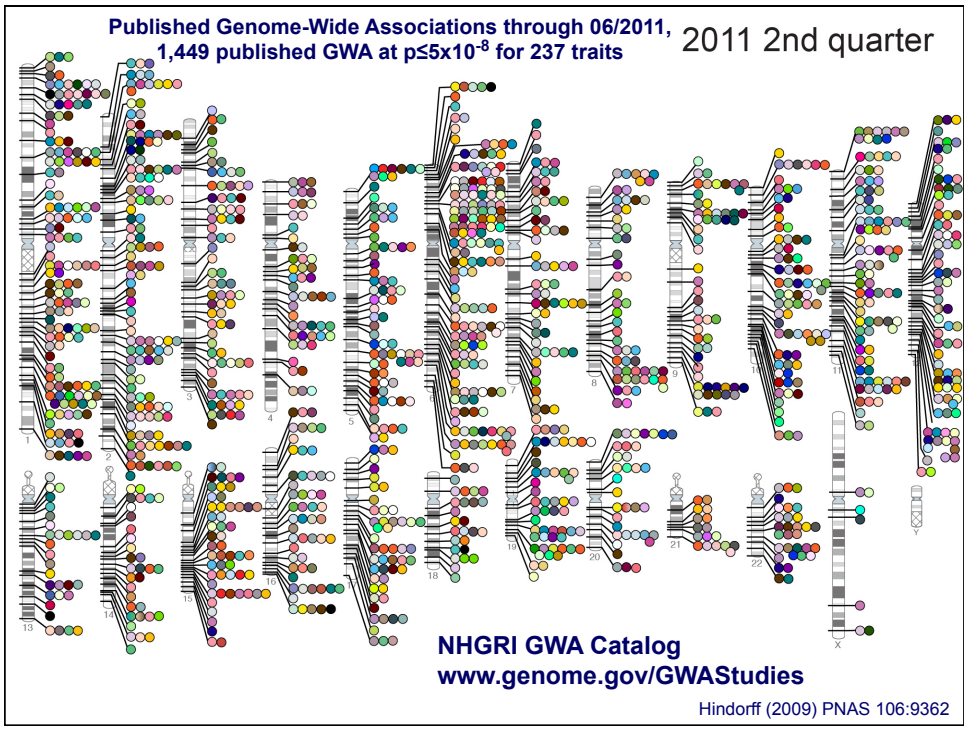
Replication of FTO with obesity



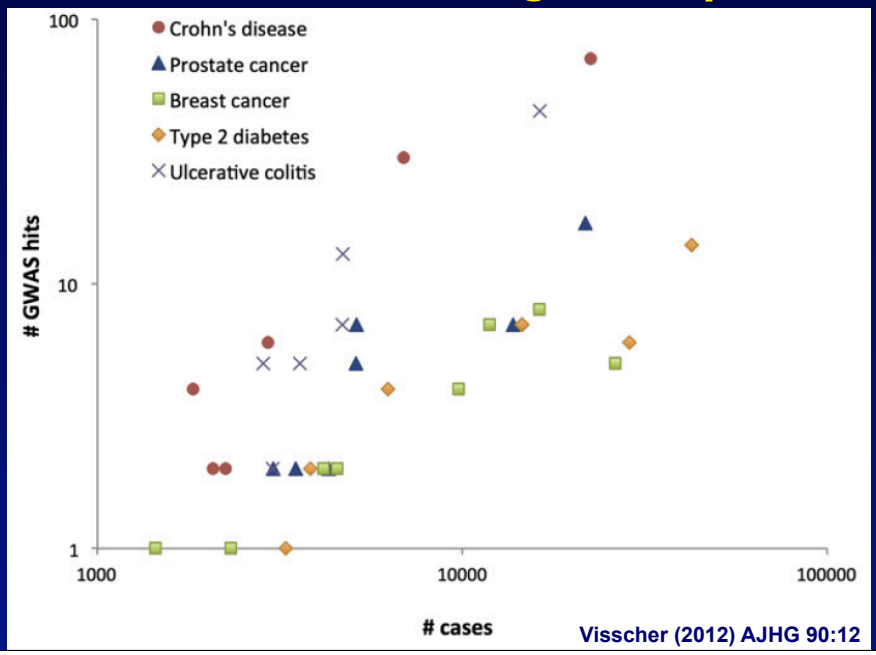
Odds of obesity for each copy of the A allele of rs9939609

Meta-analytic OR $P=3 \times 10^{-26}$

Frayling 2007 Science 316:889



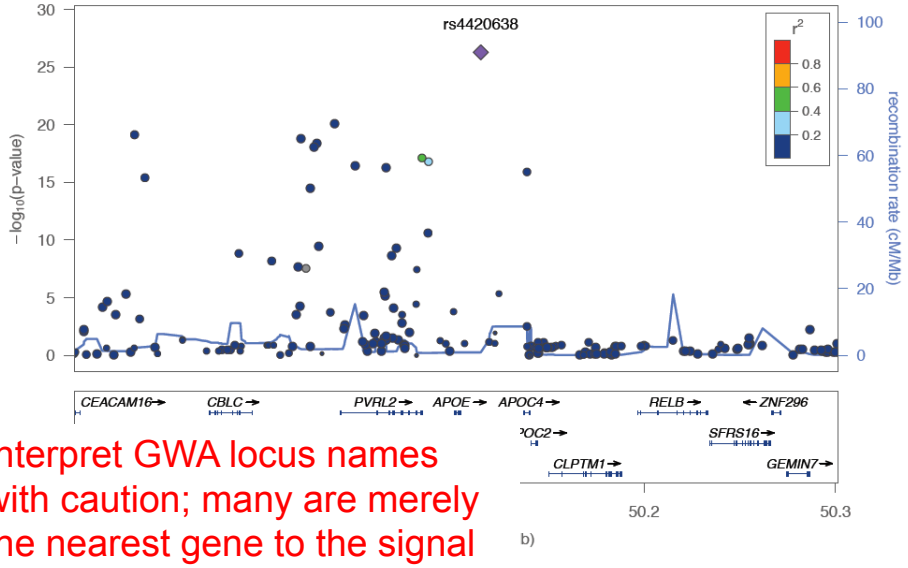
More loci found with larger sample size



Visscher (2012) AJHG 90:12

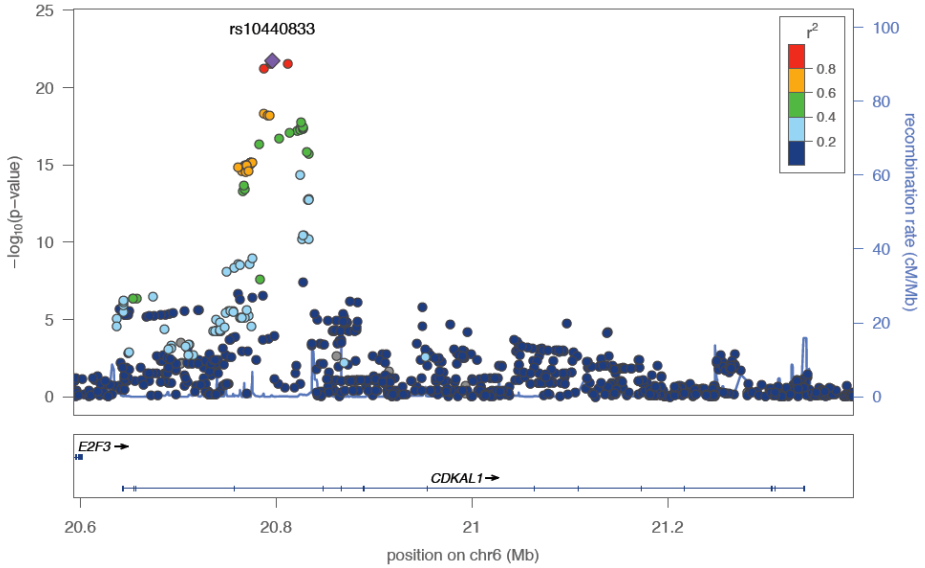
Replicate known association signal

APOE and LDL-cholesterol

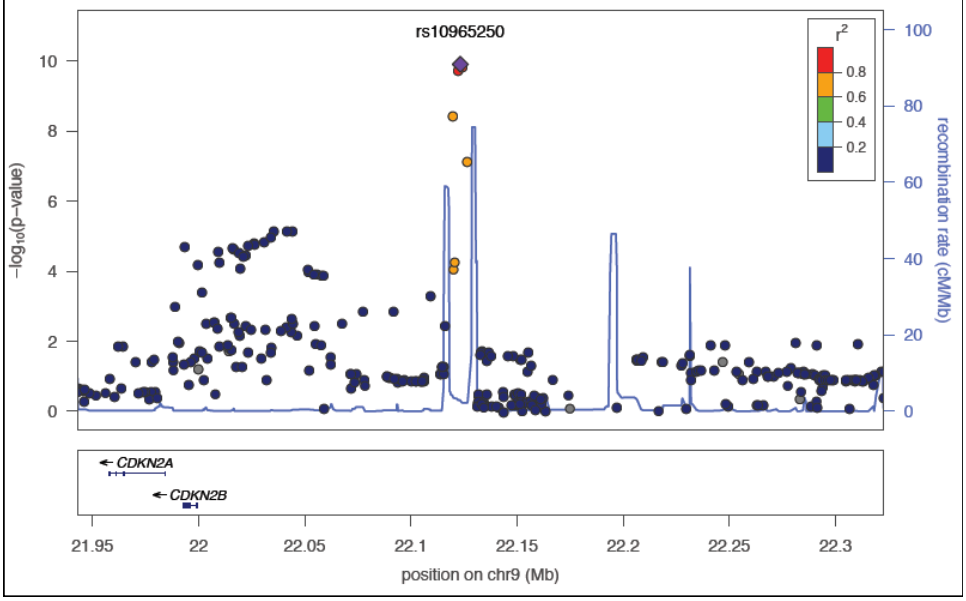


Interpret GWA locus names with caution; many are merely the nearest gene to the signal

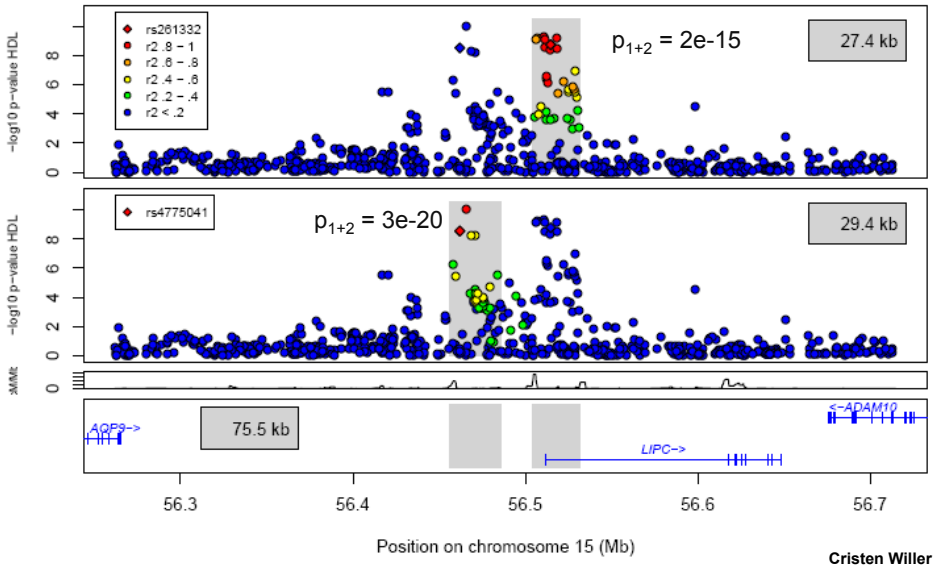
Novel signal: strongest within intron



Novel signal: outside of gene



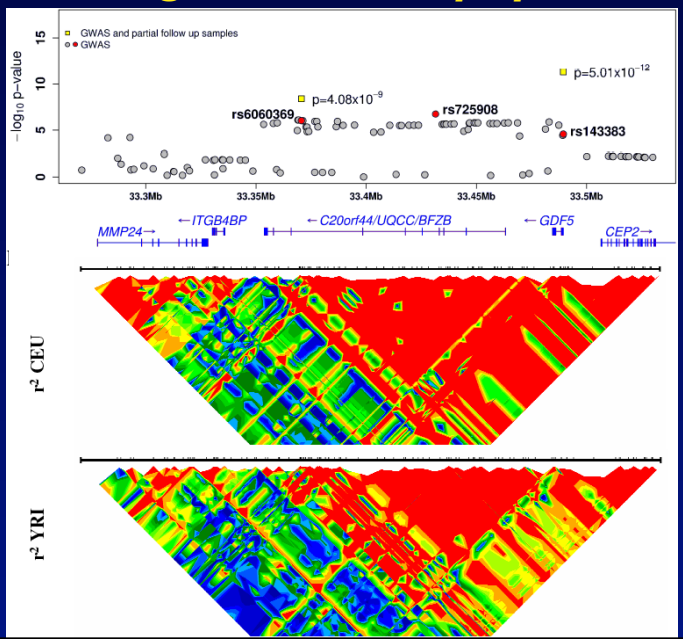
Nearby independent signals



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

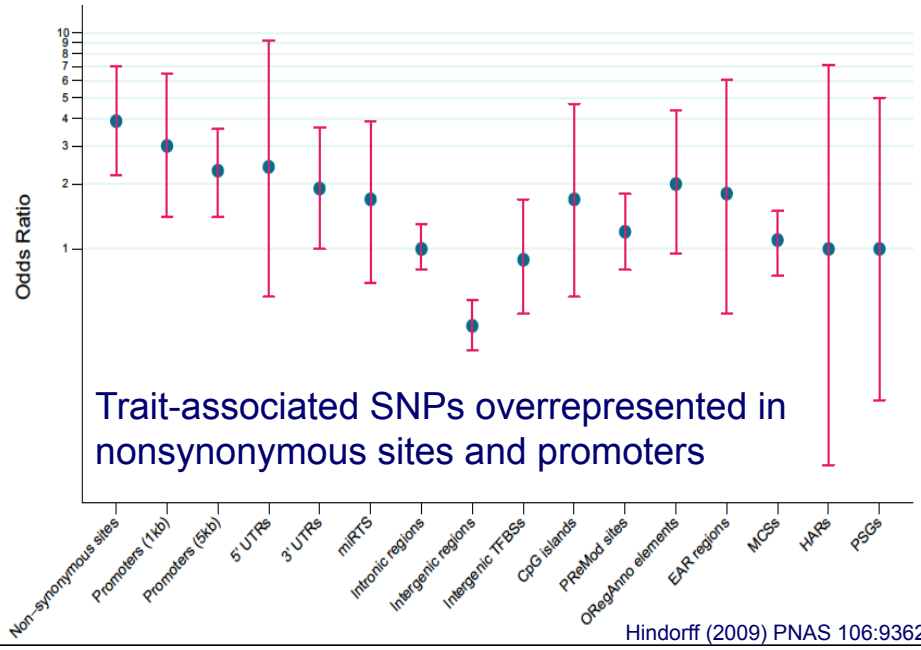
Sanna (2008) Nat Gen 40:198

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

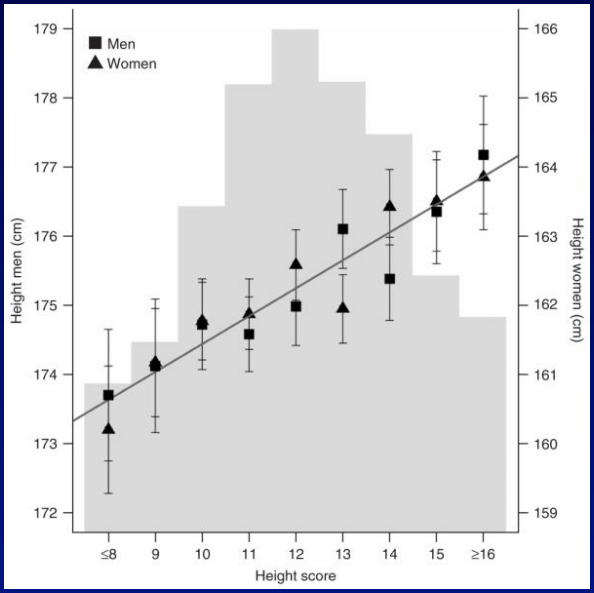
Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]	h^2 All GWAS SNPs[b]
Type 1 diabetes	0.9[98]	0.6[99,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	

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

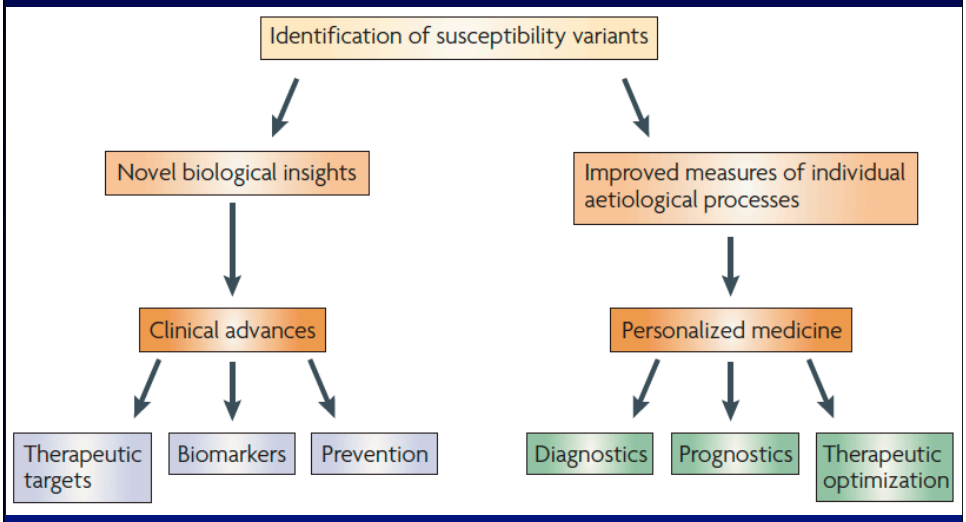
Partial table from Visscher (2012) AJHG 90:12

Prediction of height



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

Clinical translation



McCarthy (2008) Nat Rev Gen 9:356

Summary

- Need careful attention to design and QC
- Need large samples to find small signals
- 1,449 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
- Population-specific GWA panels
- More diverse populations
- Other sequence variants
- Multiple trait analysis
- Gene-gene and -environment interactions
- Molecular and biological mechanisms