Genomic Approaches to the Study of Complex Genetic Diseases

Karen Mohlke, PhD
Department of Genetics
University of North Carolina
April 23, 2014

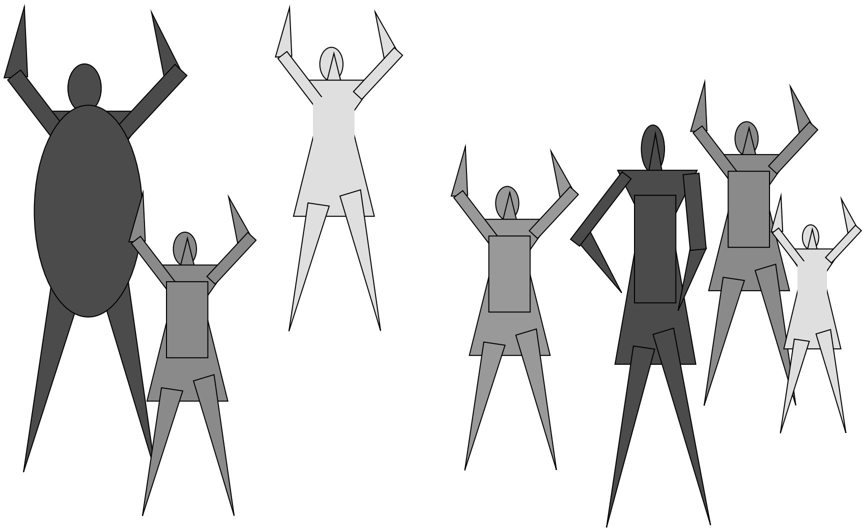
JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2014

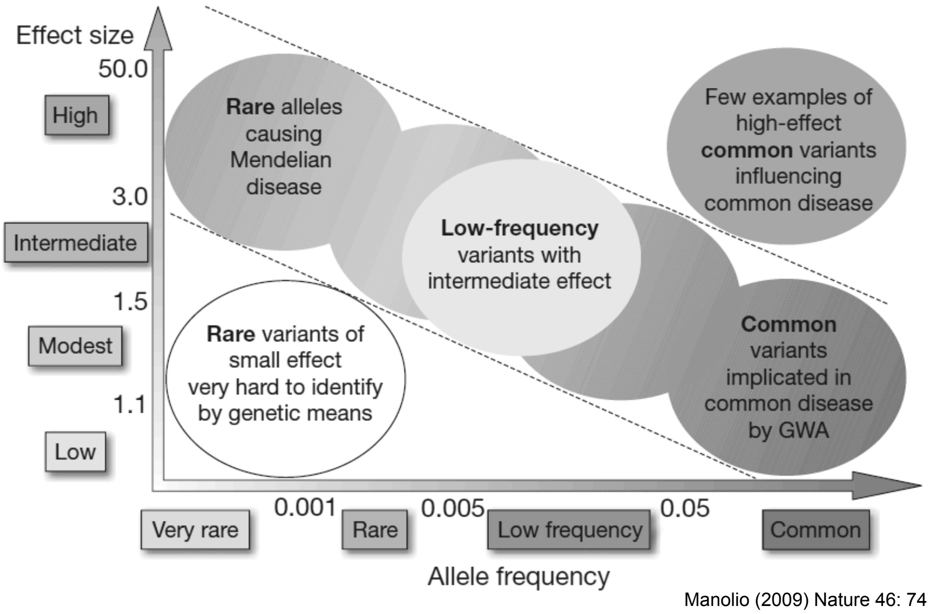
Karen Mohlke

No Relevant Financial Relationships with Commercial Interests

Complex diseases & traits

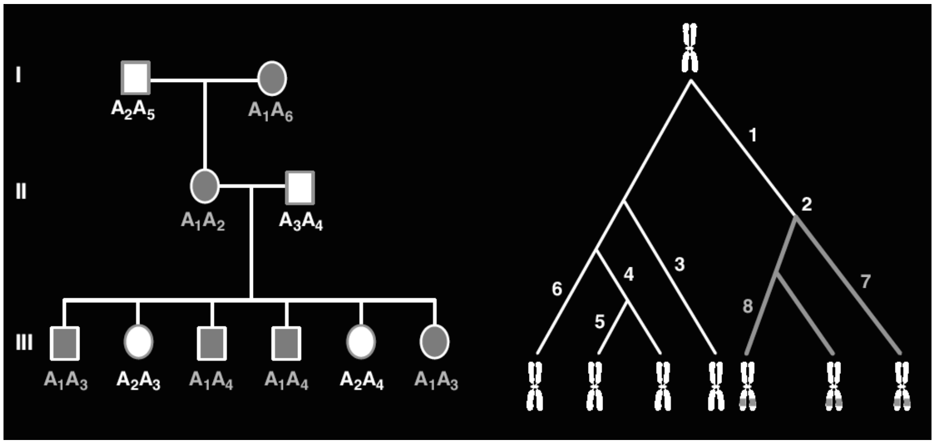


Genetic architecture


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

Manolio (2009) Nature 46: 747

Gene mapping in populations

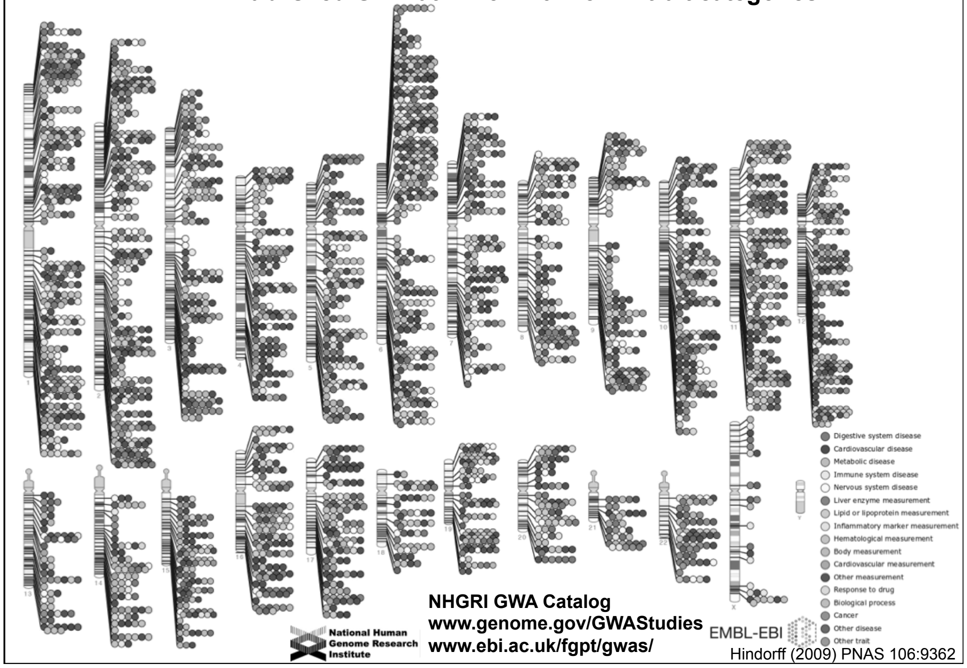


Altshuler and Clark (2005) Science 307:1052

Genome-wide association study goals

- 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

Published Genome-Wide Associations through 12/2012
Published GWA at *P* ≤ 5 x 10^{-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

Hindorff (2009) PNAS 106:9362

Outline

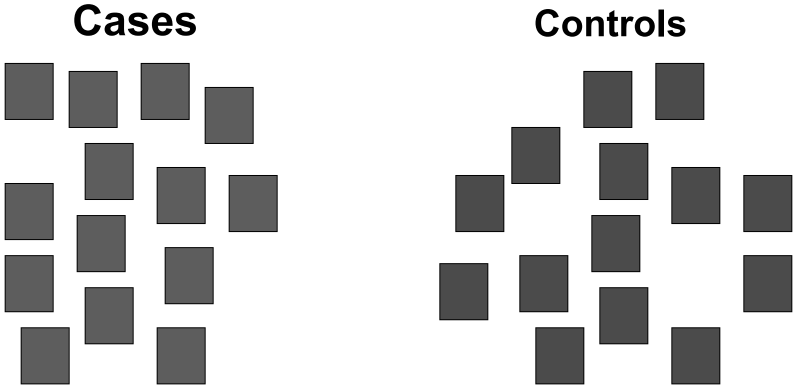
- **Genome-wide association study design**
  - **Samples/study participants**
  - **Genotyping**
  - **Tests of association**
  - **Imputation and meta-analysis**
- **Interpretation of results**
  - **Effect size and significance**
  - **Example locus characteristics**
- **Sequencing/rare variant studies**

Study design depends on disease or trait

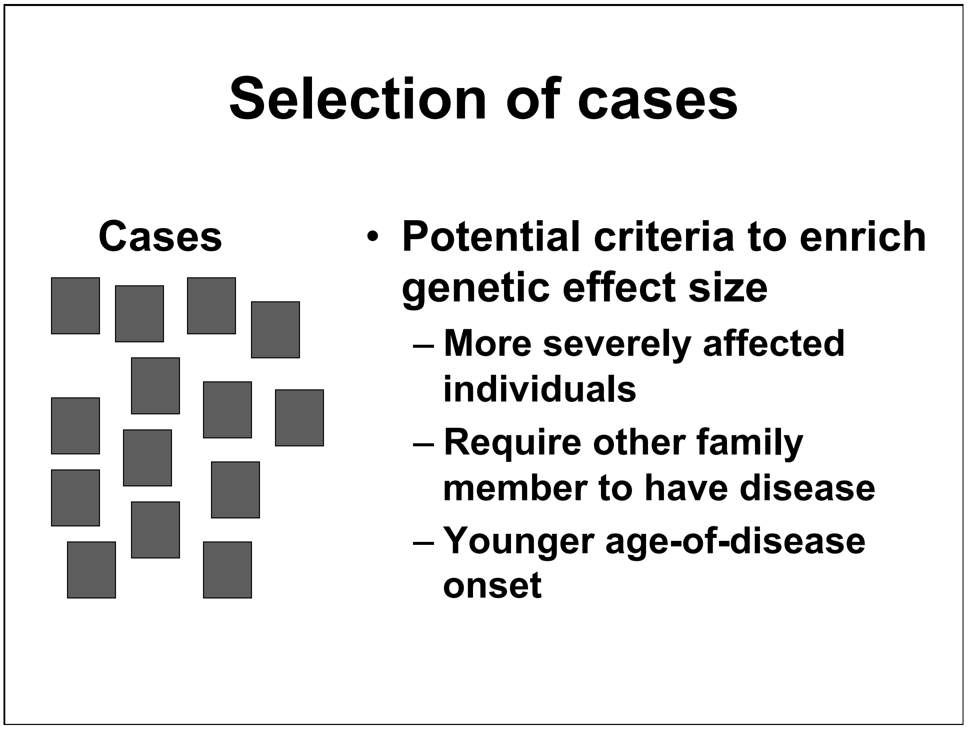
- **Disease (case/control)**
  - **Rare**
  - **Common**

- **Quantitative traits**
  - **Easy to measure: Weight, height**
  - **Requires testing: Coronary artery thickness**
  - **Requires experiment: Gene expression**

Selection of cases and controls



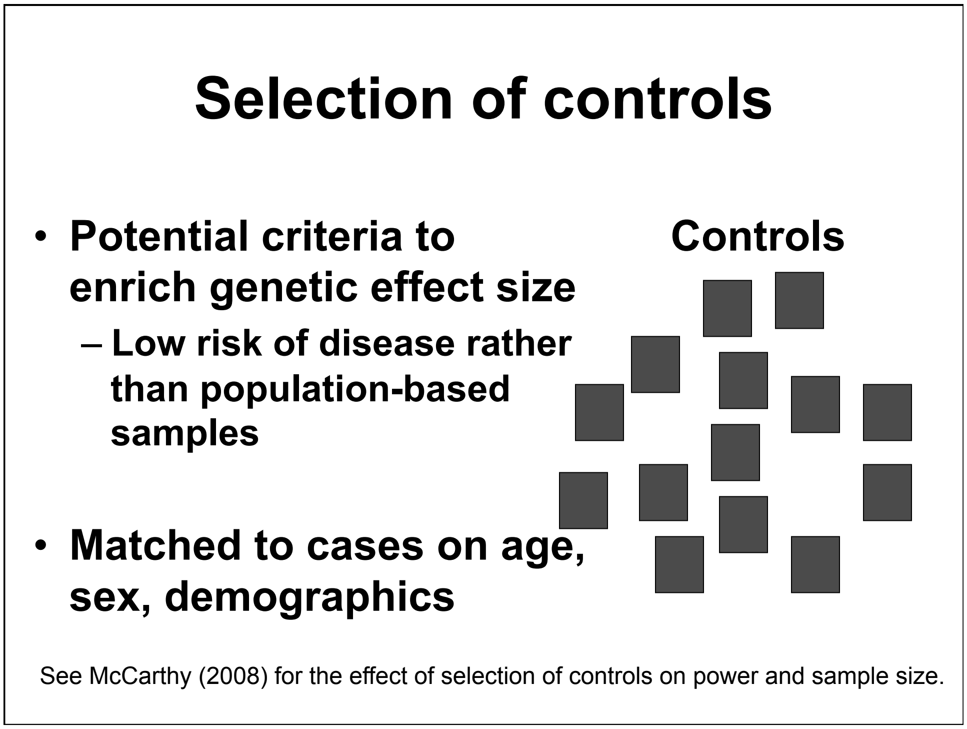
Cases and controls should be comparable in other respects except disease status.



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**



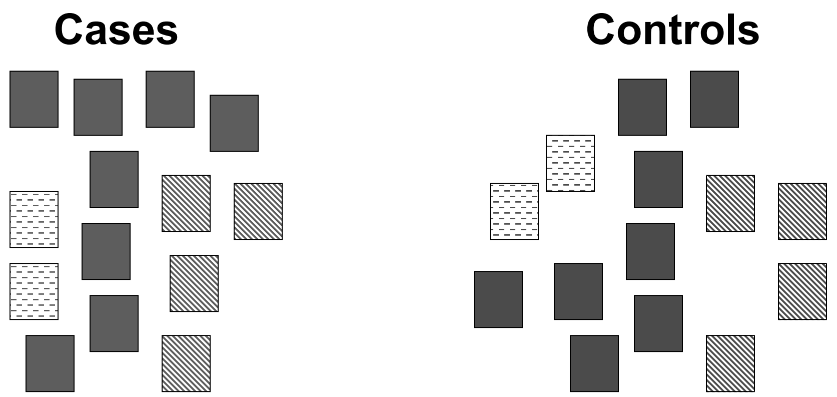
Selection of controls

Controls

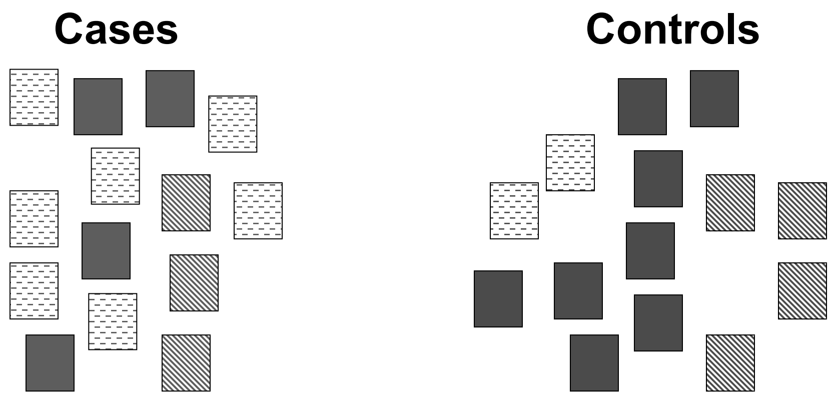
- **Potential criteria to enrich genetic effect size**
  - **Low risk of disease rather than population-based samples**
- **Matched to cases on age, sex, demographics**

See McCarthy (2008) for the effect of selection of controls on power and sample size.

Comparable ancestry



Ancestry differences



May have inadequate ancestry information prior to genotyping

Population stratification

- **Systematic differences in allele frequencies between subpopulations that may be due to different ancestry**
- **Can produce spurious associations in case-control studies**

Population stratification

Example: IgG haplotype ‘Gm’ association with type 2 diabetes in Gila River Indian Community

Presence of Gm marker associated with lower prevalence of diabetes

Apparent association with diabetes is due to an association between Gm marker and amount of Indian heritage

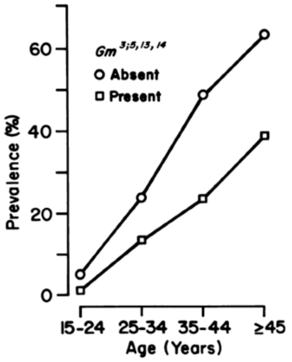


Figure 1 Prevalence of diabetes by age and the presence of the haplotype $Gm^{3;5,13,14}$ among residents of the Gila River Indian Community.

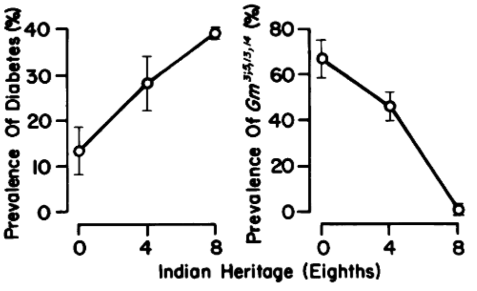


Figure 3 Age-adjusted prevalence (±1 standard error) of diabetes (left) and of $Gm^{3;5,13,14}$ (right), according to Indian heritage, among residents of the Gila River Indian Community.

Knowler (1998) AJHG 43:520

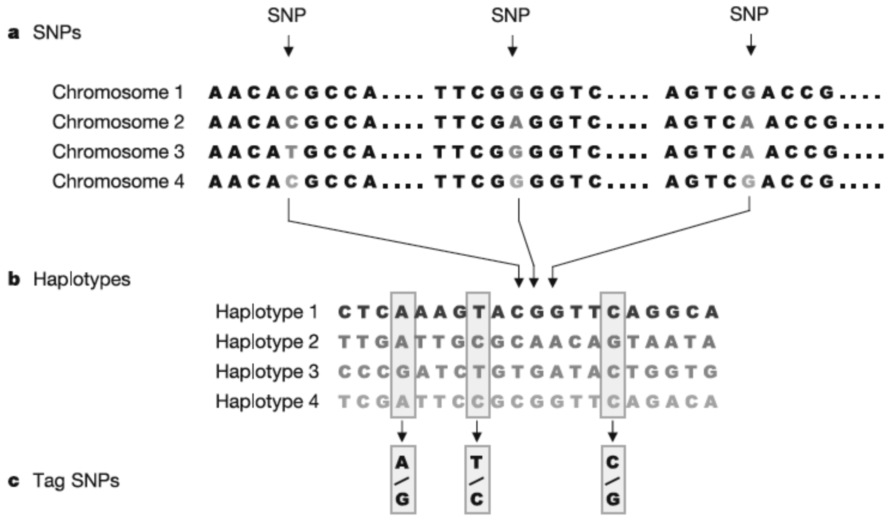
Account for or avoid population stratification

- Match cases with controls
- Restrict to one subgroup
- Adjust for genetic background
  - Use principle components (PCs) to infer ancestry from genotype data and adjust for PCs in association analysis
- Family-based study design – genotype relatives and analyze transmission of alleles from heterozygous parents to offspring
  - Transmission disequilibrium test (TDT), family-based association test (FBAT)

Genome-wide SNP panels

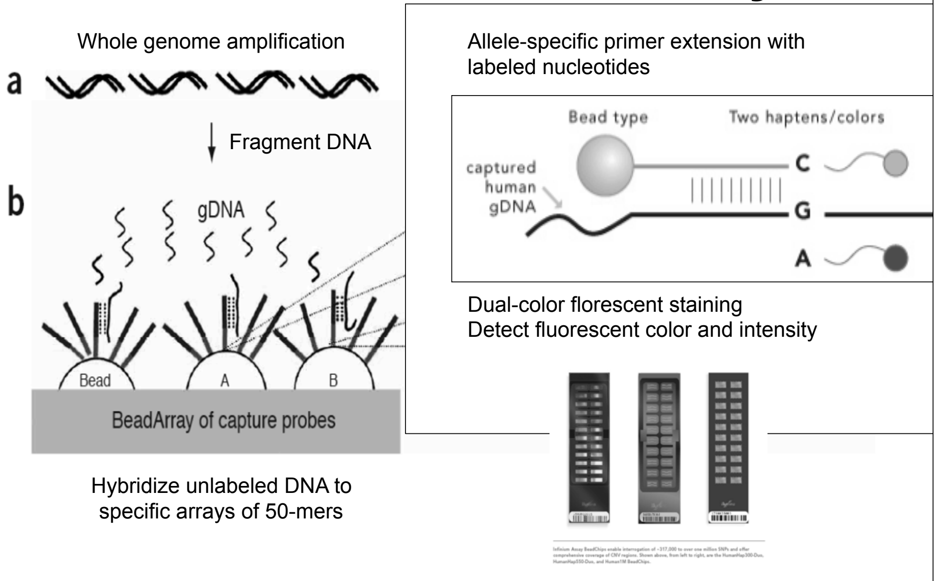
- 10,000 - 5 million SNPs
- Affymetrix, Illumina
  - Random SNPs
  - Selected haplotype tag SNPs
  - Copy number probes
  - More lower frequency variants
  - Exome variants
  - Some arrays allow SNPs to be added

Selecting ‘haplotype tag’ SNPs


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
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



Whole genome amplification

Fragment DNA

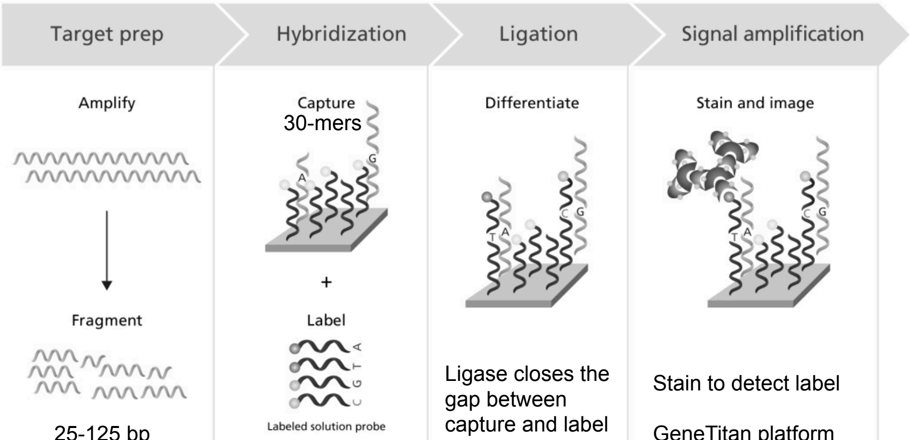
Hybridize unlabeled DNA to specific arrays of 50-mers

Allele-specific primer extension with labeled nucleotides

Dual-color florescent staining
Detect fluorescent color and intensity

Illumina.com and adapted from Gunderson (2005) NatGen 37:549

Affymetrix Axiom Array



30-mers

25-125 bp

Cocktail of labeled 9-mer oligos

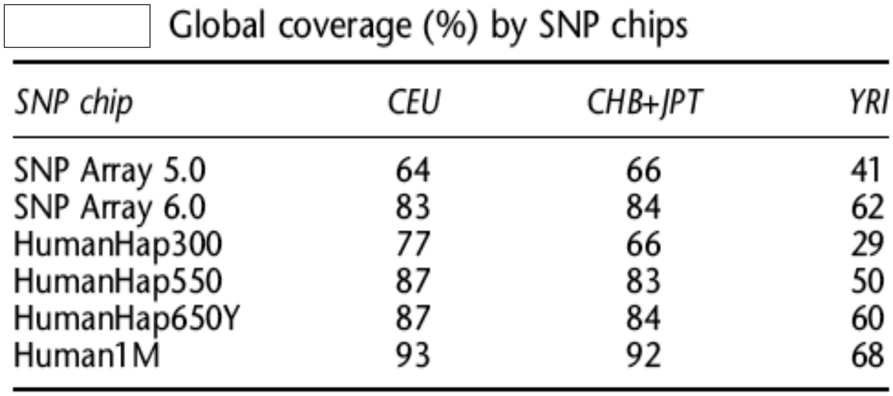
Ligase closes the gap between capture and label probe if complete complementarity; wash off others

Stain to detect label

GeneTitan platform

Affymetrix.com

Global genomic coverage

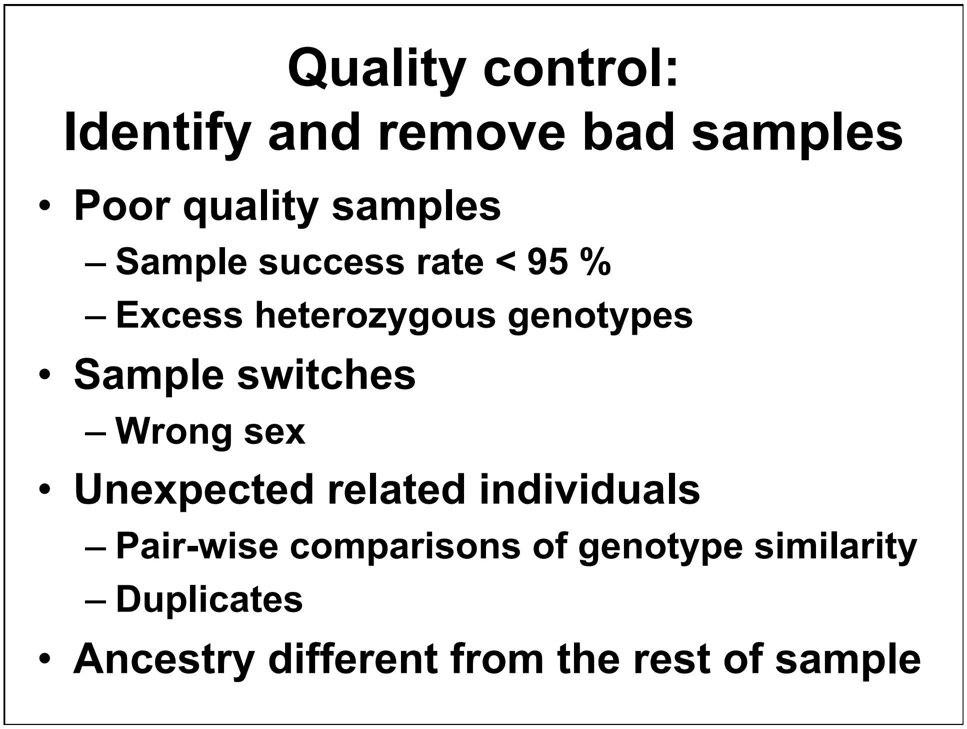


Global coverage (%) by SNP chips

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

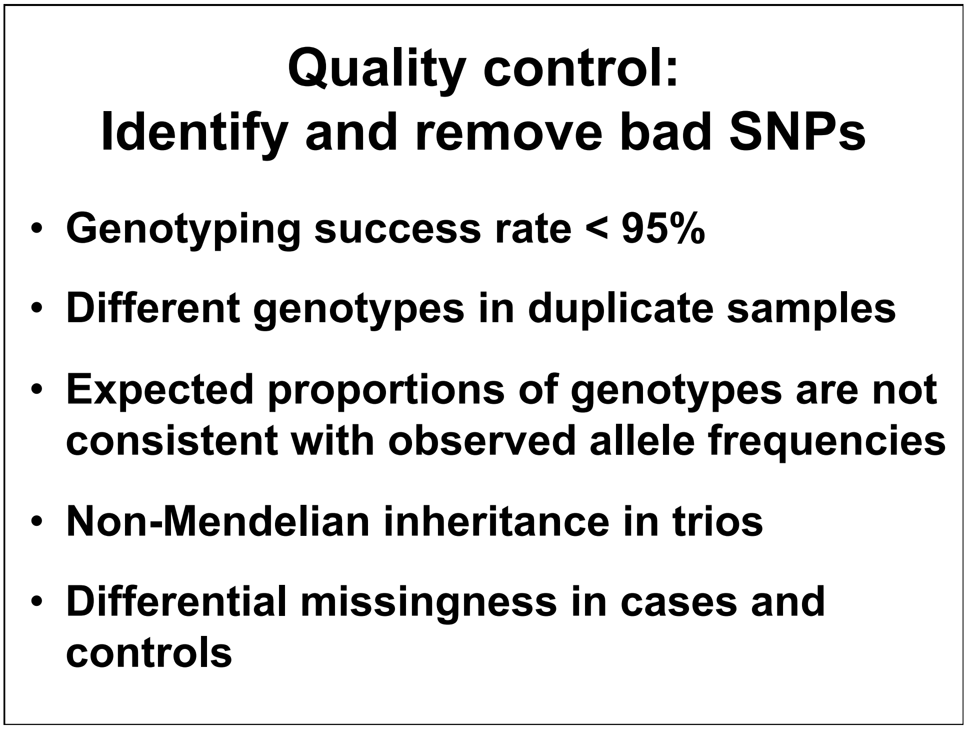
Percent of SNPs present on the chip or tagged at r^2>0.8 by at least one SNP in the chip within 250 kb

Li (2008) EJHG 16:625



Quality control: Identify and remove bad samples

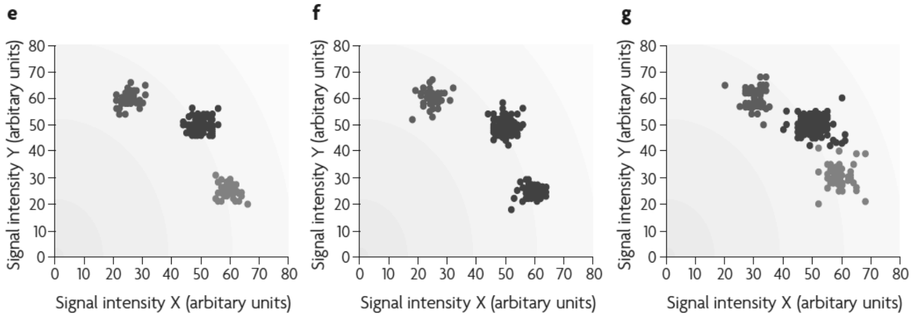
- **Poor quality samples**
  - **Sample success rate < 95 %**
  - **Excess heterozygous genotypes**
- **Sample switches**
  - **Wrong sex**
- **Unexpected related individuals**
  - **Pair-wise comparisons of genotype similarity**
  - **Duplicates**
- **Ancestry different from the rest of sample**



Quality control: Identify and remove bad SNPs

- **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**

Quality control: Identify and remove bad SNPs



Ideal genotyping plot Clusters mis-called Clusters overlap

McCarthy (2008) Nat Rev Gen 9:356

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

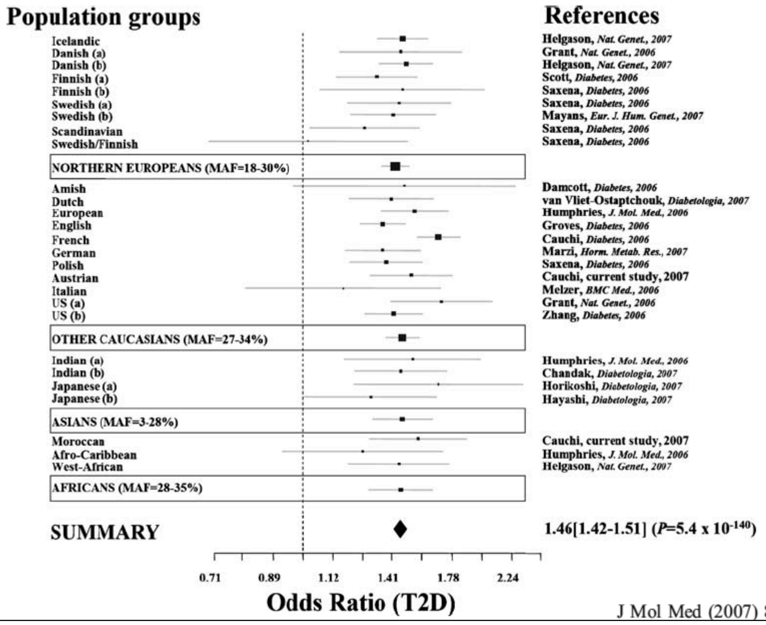
- **Surrogate measure of effect of allele on risk of developing disease**

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

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

Association study odds ratio plot



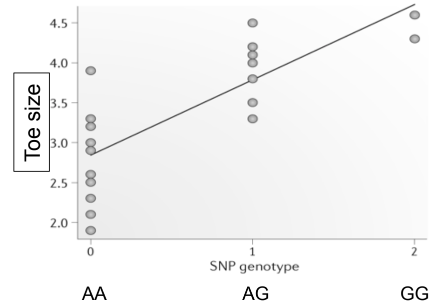
J Mol Med (2007) 85:777–782

Linear regression

$y = \beta_0 + \beta_1 x$

$\text{Trait} = \beta_0 + \beta_1 \text{SNP}_1$

$\text{Toe size} = \beta_0 + \beta_1 \text{rs123456}$



Linear regression

$y = \beta_0 + \beta_1 x$

$\text{Trait} = \beta_0 + \beta_1 \text{SNP}_1$

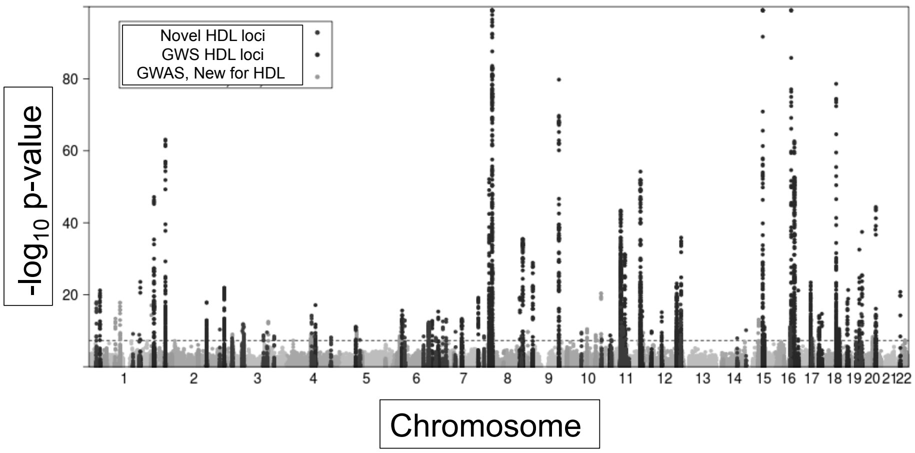
$\text{Toe size} = \beta_0 + \beta_1 \text{rs123456}$

$\text{Toe size} = \beta_0 + \beta_1 \text{rs123456} + \beta_2 \text{sex} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{BMI}$

covariates

- **Assumptions**
  - Trait is normally distributed for each genotype, with a common variance
  - Subjects independent (e.g. unrelated)

‘Manhattan plot’ for HDL-cholesterol



Global Lipids Genetics Consortium
188,577 individuals from 60 studies, GWAS + metabochip variants

GLGC (2013) Nat Gen 45:1274

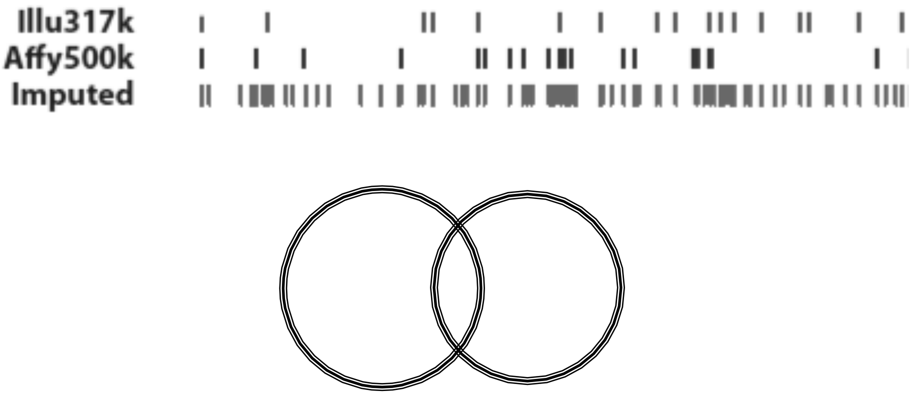
Multiple testing

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

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

- **Need large effect or large sample size**

Imputation of ungenotyped variants



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

Imputation: Observed genotypes

Observed Genotypes

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

Study Sample

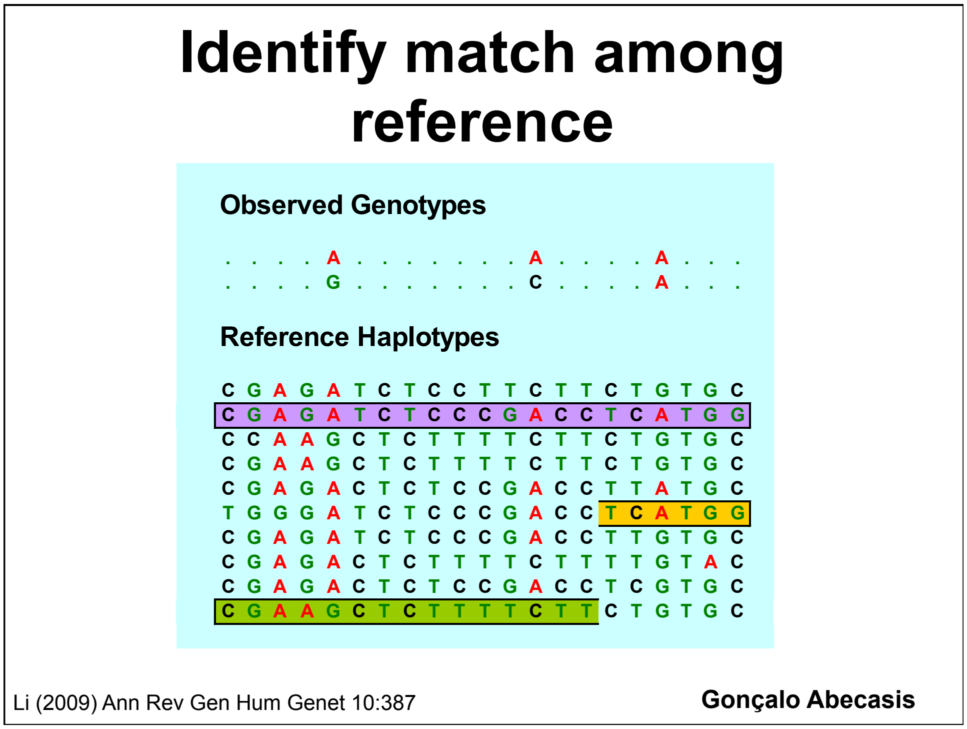
Reference Haplotypes

```
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 or 1000 Genomes

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

Gonçalo Abecasis



Identify match among reference

Observed Genotypes

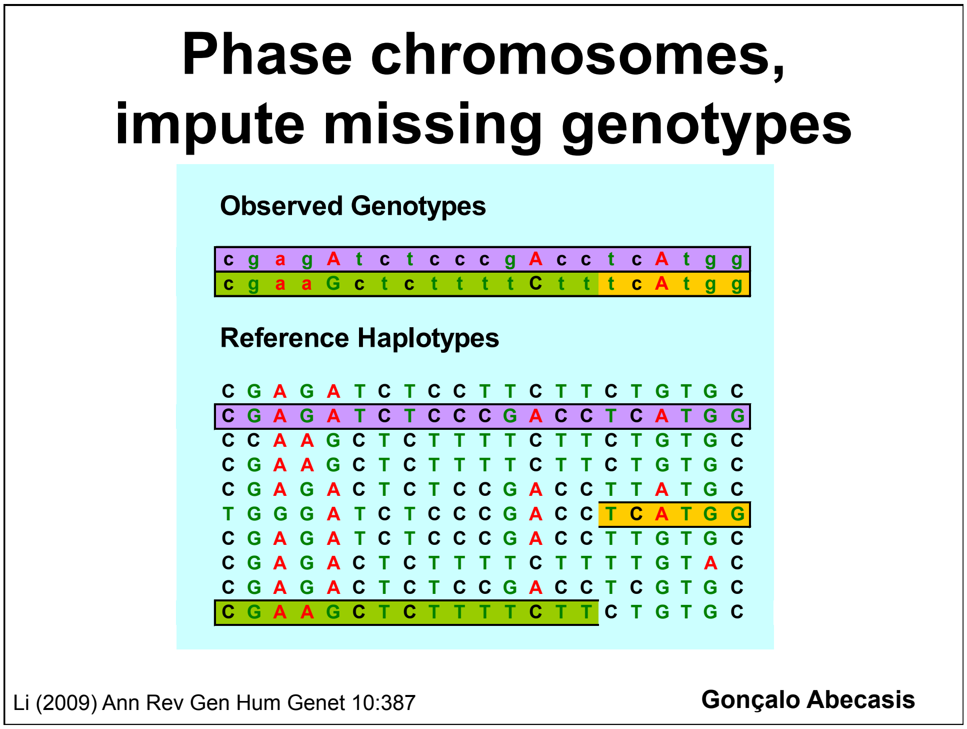
```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

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
```

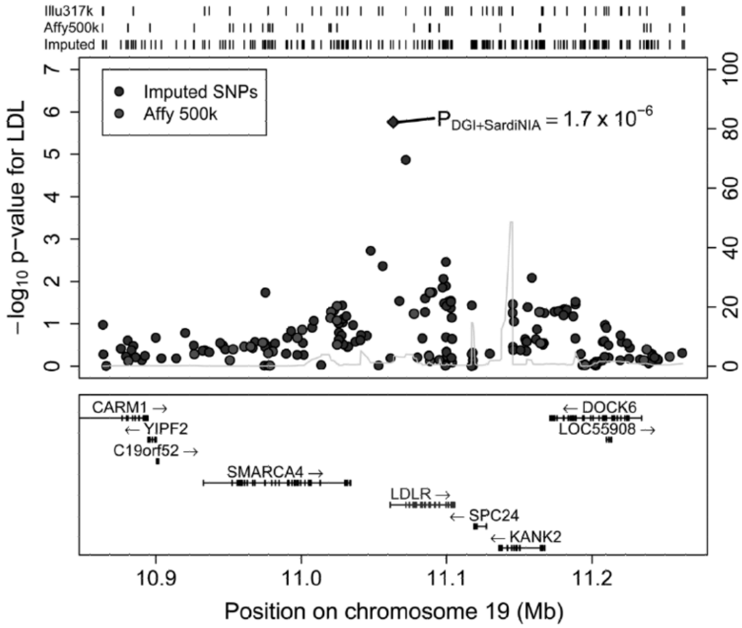
Reference Haplotypes

```
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

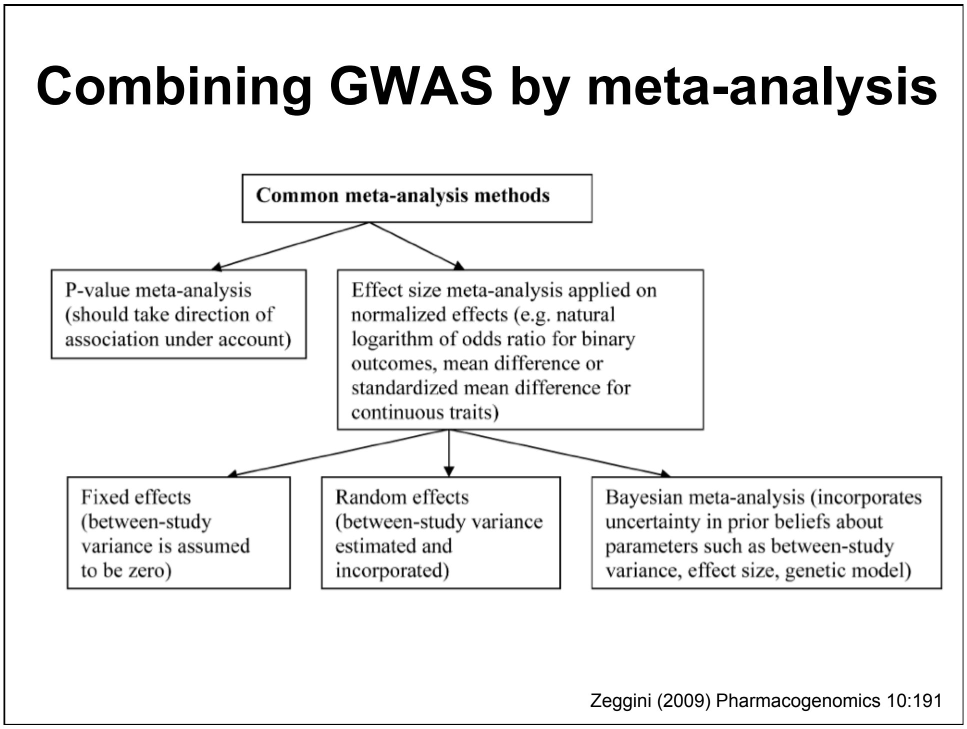
LDLR locus and LDL cholesterol



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

Combining GWAS by meta-analysis

- **Combine studies giving more weight to studies with greater precision**
- **Increase power vs individual studies**
- **Can investigate consistency of effects across studies**
- **Potential sources of heterogeneity:**
  - **Phenotype definitions are different**
  - **Different genotyping and analysis strategies**
  - **Environmental effects may differ**



Combining GWAS by meta-analysis

Common meta-analysis methods

P-value meta-analysis (should take direction of association under account)

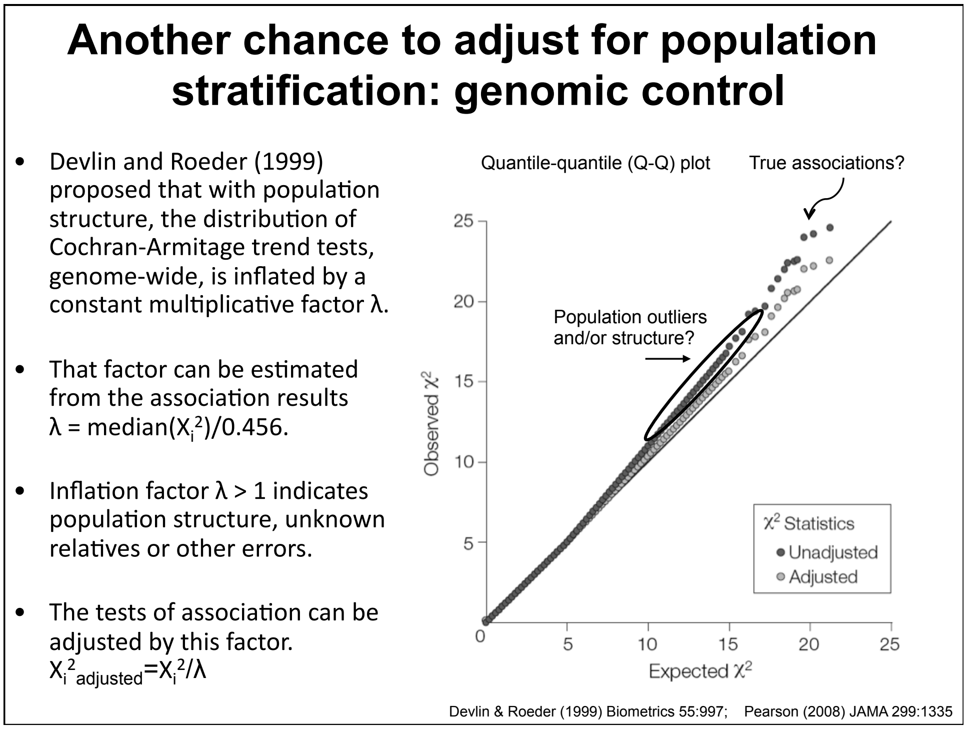
Effect size meta-analysis applied on normalized effects (e.g. natural logarithm of odds ratio for binary outcomes, mean difference or standardized mean difference for continuous traits)

Fixed effects (between-study variance is assumed to be zero)

Random effects (between-study variance estimated and incorporated)

Bayesian meta-analysis (incorporates uncertainty in prior beliefs about parameters such as between-study variance, effect size, genetic model)

Zeggini (2009) Pharmacogenomics 10:191



Another chance to adjust for population stratification: genomic control

- Devlin and Roeder (1999) proposed that with population structure, the distribution of Cochran-Armitage trend tests, genome-wide, is inflated by a constant multiplicative factor λ.
- That factor can be estimated from the association results λ = median(X_i^2)/0.456.
- Inflation factor λ > 1 indicates population structure, unknown relatives or other errors.
- The tests of association can be adjusted by this factor. $X_{i\ adjusted}^2 = X_i^2/\lambda$

Quantile-quantile (Q-Q) plot

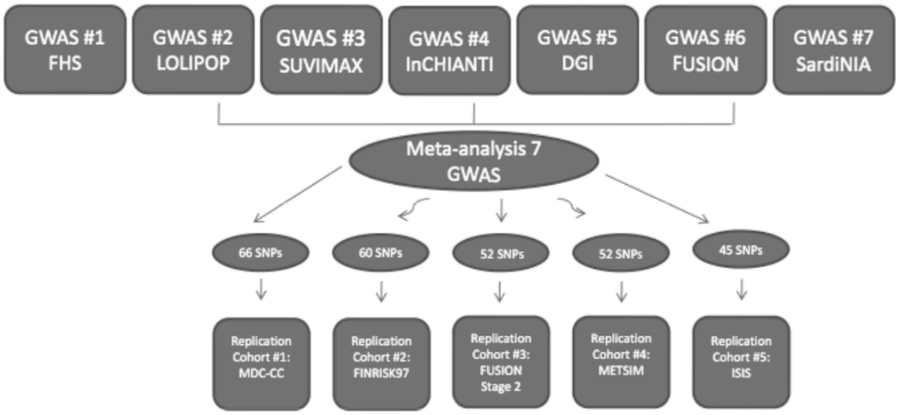
True associations?

Population outliers and/or structure?

Devlin & Roeder (1999) Biometrics 55:997; Pearson (2008) JAMA 299:1335

Outline

- **Genome-wide association study design**
  - **Samples/study participants**
  - **Genotyping**
  - **Tests of association**
  - **Imputation and meta-analysis**
- **Interpretation of results**
  - **Effect size and significance**
  - **Example locus characteristics**
- **Sequencing/rare variant studies**



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

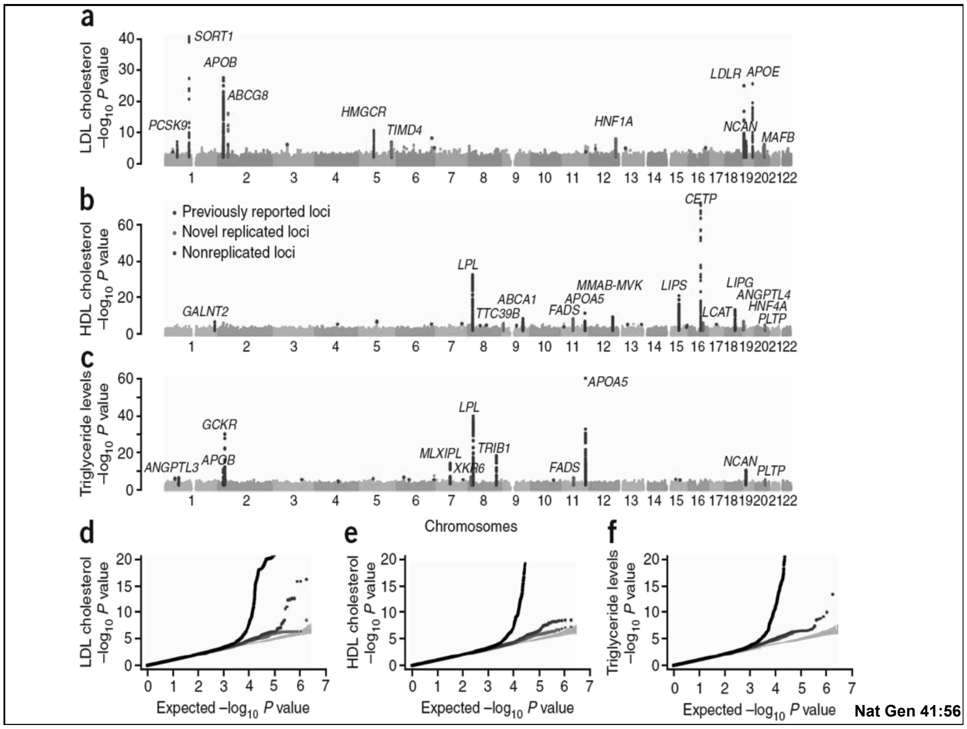
52 SNPs → Replication Cohort #3: FUSION Stage 2

52 SNPs → Replication Cohort #4: METSIM

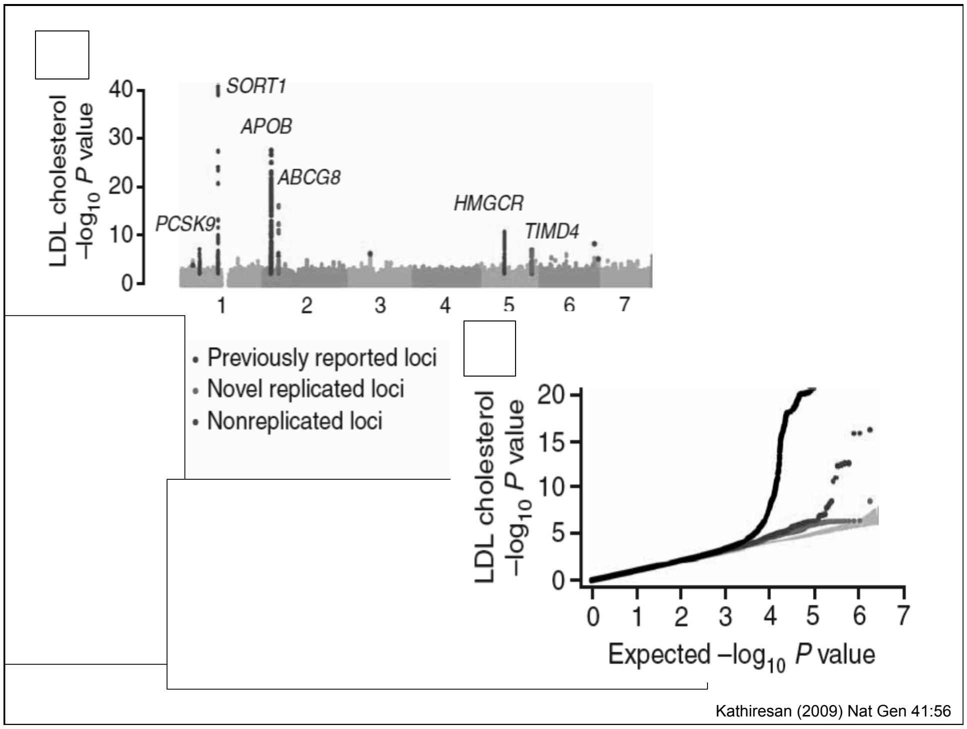
45 SNPs → Replication Cohort #5: ISIS

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

Kathiresan (2009) Nat Gen 41:56

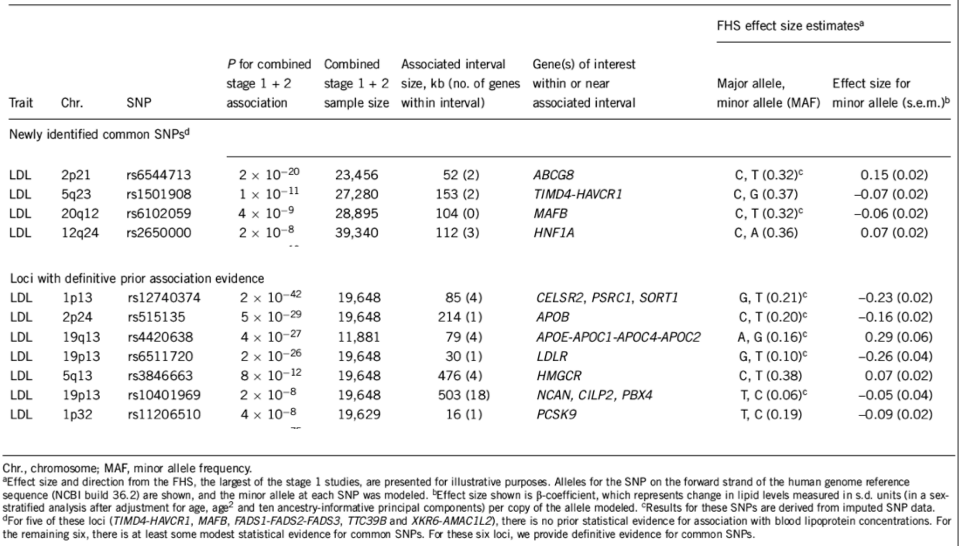


Nat Gen 41:56



Kathiresan (2009) Nat Gen 41:56

Effect size and significance



							FHS effect size estimates[a]	
Trait	Chr.	SNP	*P* for combined stage 1 + 2 association	Combined stage 1 + 2 sample size	Associated interval size, kb (no. of genes within interval)	Gene(s) of interest within or near associated interval	Major allele, minor allele (MAF)	Effect size for minor allele (s.e.m.)[b]
Newly identified common SNPs[d]								
LDL	2p21	rs6544713	2×10^{-20}	23,456	52 (2)	*ABCG8*	C, T (0.32)[c]	0.15 (0.02)
LDL	5q23	rs1501908	1×10^{-11}	27,280	153 (2)	*TIMD4-HAVCR1*	C, G (0.37)	–0.07 (0.02)
LDL	20q12	rs6102059	4×10^{-9}	28,895	104 (0)	*MAFB*	C, T (0.32)[c]	–0.06 (0.02)
LDL	12q24	rs2650000	2×10^{-8}	39,340	112 (3)	*HNF1A*	C, A (0.36)	0.07 (0.02)
Loci with definitive prior association evidence								
LDL	1p13	rs12740374	2×10^{-42}	19,648	85 (4)	*CELSR2, PSRC1, SORT1*	G, T (0.21)[c]	–0.23 (0.02)
LDL	2p24	rs515135	5×10^{-29}	19,648	214 (1)	*APOB*	C, T (0.20)[c]	–0.16 (0.02)
LDL	19q13	rs4420638	4×10^{-27}	11,881	79 (4)	*APOE-APOC1-APOC4-APOC2*	A, G (0.16)[c]	0.29 (0.06)
LDL	19p13	rs6511720	2×10^{-26}	19,648	30 (1)	*LDLR*	G, T (0.10)[c]	–0.26 (0.04)
LDL	5q13	rs3846663	8×10^{-12}	19,648	476 (4)	*HMGCR*	C, T (0.38)	0.07 (0.02)
LDL	19p13	rs10401969	2×10^{-8}	19,648	503 (18)	*NCAN, CILP2, PBX4*	T, C (0.06)[c]	–0.05 (0.04)
LDL	1p32	rs11206510	4×10^{-8}	19,629	16 (1)	*PCSK9*	T, C (0.19)	–0.09 (0.02)

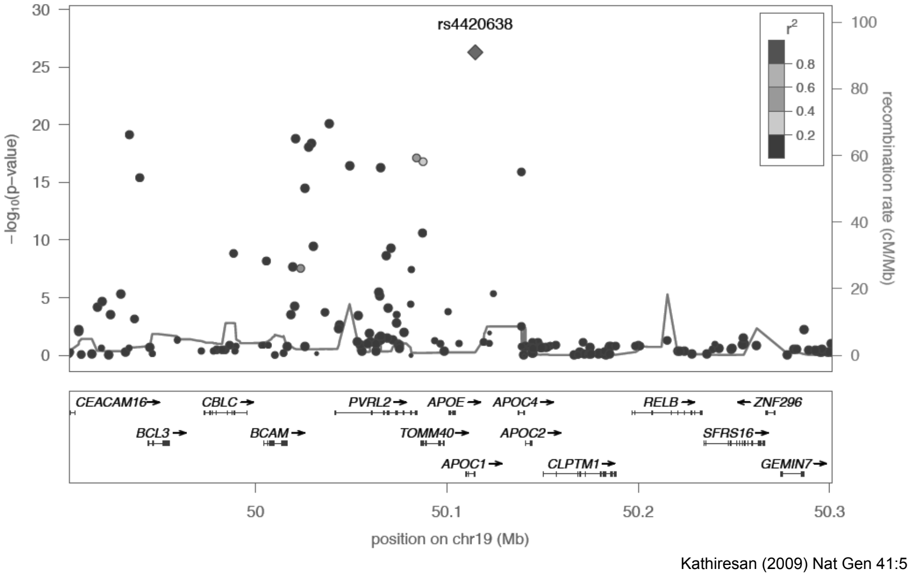
Chr., chromosome; MAF, minor allele frequency.

[a] Effect size and direction from the FHS, the largest of the stage 1 studies, are presented for illustrative purposes. Alleles for the SNP on the forward strand of the human genome reference sequence (NCBI build 36.2) are shown, and the minor allele at each SNP was modeled. [b] Effect size shown is β-coefficient, which represents change in lipid levels measured in s.d. units (in a sex-stratified analysis after adjustment for age, age^2 and ten ancestry-informative principal components) per copy of the allele modeled. [c] Results for these SNPs are derived from imputed SNP data. [d] For five of these loci (*TIMD4-HAVCR1*, *MAFB*, *FADS1-FADS2-FADS3*, *TTC39B* and *XKR6-AMAC1L2*), there is no prior statistical evidence for association with blood lipoprotein concentrations. For the remaining six, there is at least some modest statistical evidence for common SNPs. For these six loci, we provide definitive evidence for common SNPs.

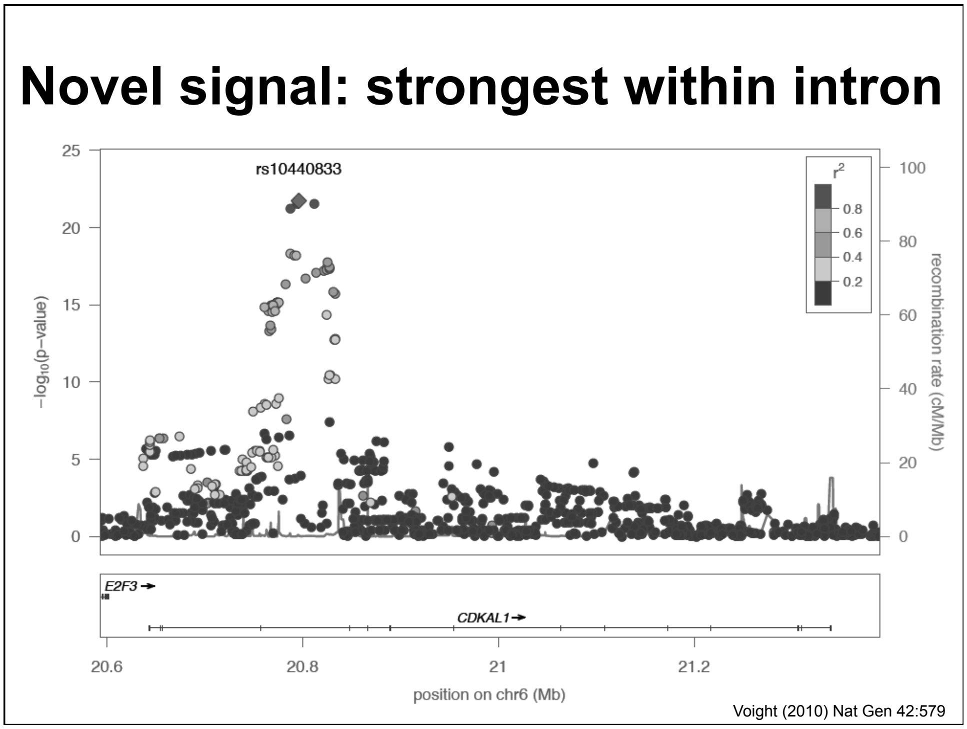
Kathiresan (2009) Nat Gen 41:56

Replicate known association signal

APOE and LDL-cholesterol

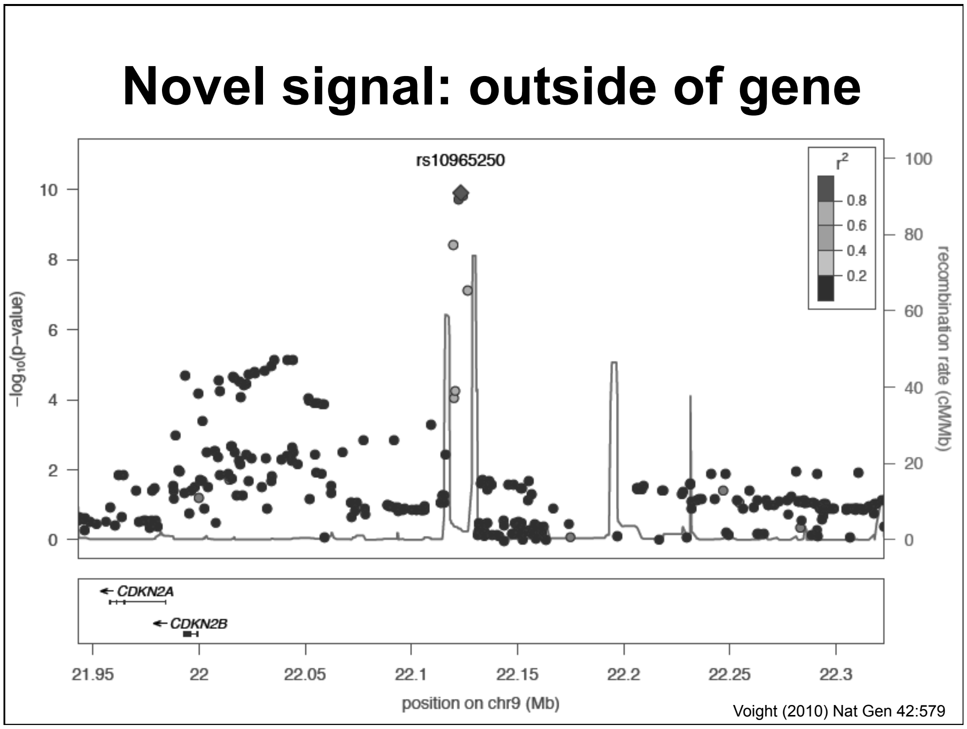


Kathiresan (2009) Nat Gen 41:56



Novel signal: strongest within intron

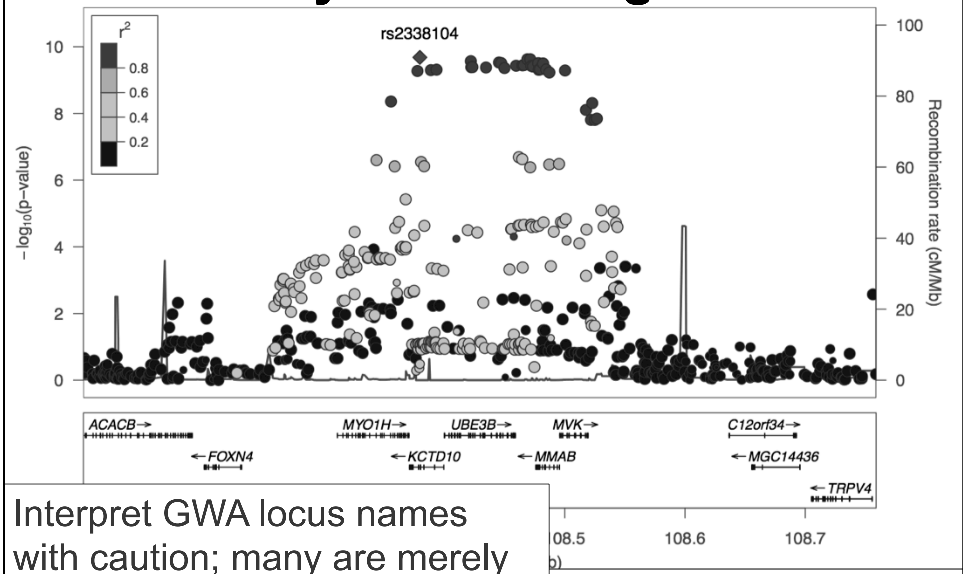
Voight (2010) Nat Gen 42:579



Novel signal: outside of gene

Voight (2010) Nat Gen 42:579

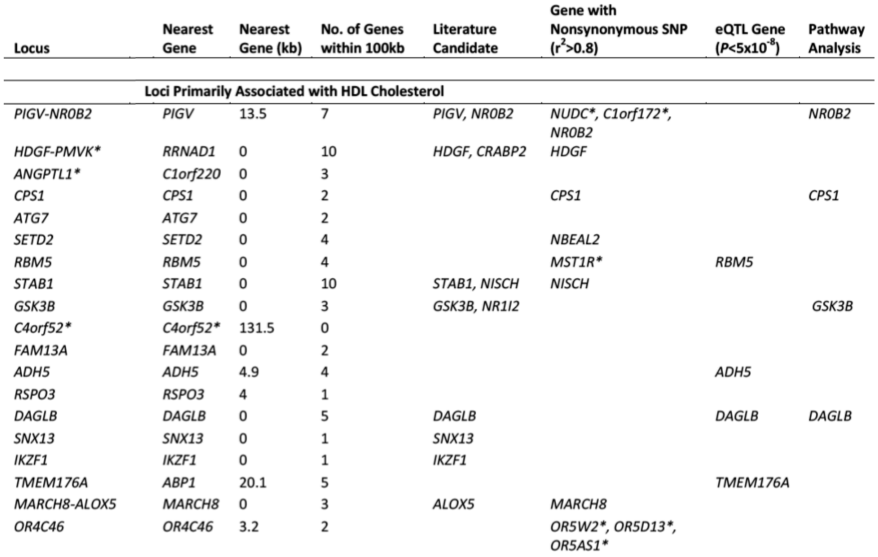
Many candidate genes



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

Kathiresan (2009) Nat Gen 41:56

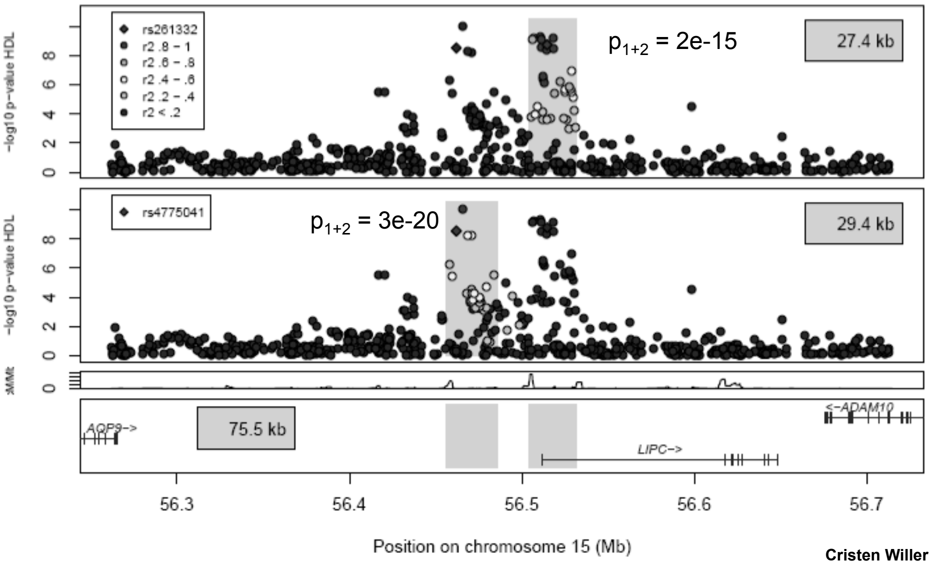
Interpret plausible candidate genes



Locus	Nearest Gene	Nearest Gene (kb)	No. of Genes within 100kb	Literature Candidate	Gene with Nonsynonymous SNP $(r^2>0.8)$	eQTL Gene $(P<5x10^{-8})$	Pathway Analysis
	Loci Primarily Associated with HDL Cholesterol						
PIGV-NR0B2	*PIGV*	13.5	7	*PIGV, NR0B2*	*NUDC*, C1orf172*, NR0B2*		*NR0B2*
*HDGF-PMVK**	*RRNAD1*	0	10	*HDGF, CRABP2*	*HDGF*		
*ANGPTL1**	*C1orf220*	0	3				
CPS1	*CPS1*	0	2		*CPS1*		*CPS1*
ATG7	*ATG7*	0	2				
SETD2	*SETD2*	0	4		*NBEAL2*		
RBM5	*RBM5*	0	4		*MST1R**	*RBM5*	
STAB1	*STAB1*	0	10	*STAB1, NISCH*	*NISCH*		
GSK3B	*GSK3B*	0	3	*GSK3B, NR1I2*			*GSK3B*
*C4orf52**	*C4orf52**	131.5	0				
FAM13A	*FAM13A*	0	2				
ADH5	*ADH5*	4.9	4			*ADH5*	
RSPO3	*RSPO3*	4	1				
DAGLB	*DAGLB*	0	5	*DAGLB*		*DAGLB*	*DAGLB*
SNX13	*SNX13*	0	1	*SNX13*			
IKZF1	*IKZF1*	0	1	*IKZF1*			
TMEM176A	*ABP1*	20.1	5			*TMEM176A*	
MARCH8-ALOX5	*MARCH8*	0	3	*ALOX5*	*MARCH8*		
OR4C46	*OR4C46*	3.2	2		*OR5W2*, OR5D13*, OR5AS1**		

GLGC (2013) Nat Gen 45:1274

Nearby independent signals



p_{1+2} = 2e-15

p_{1+2} = 3e-20

Cristen Willer

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

Conditional analysis

$$y = \beta_0 + \beta_1 x$$

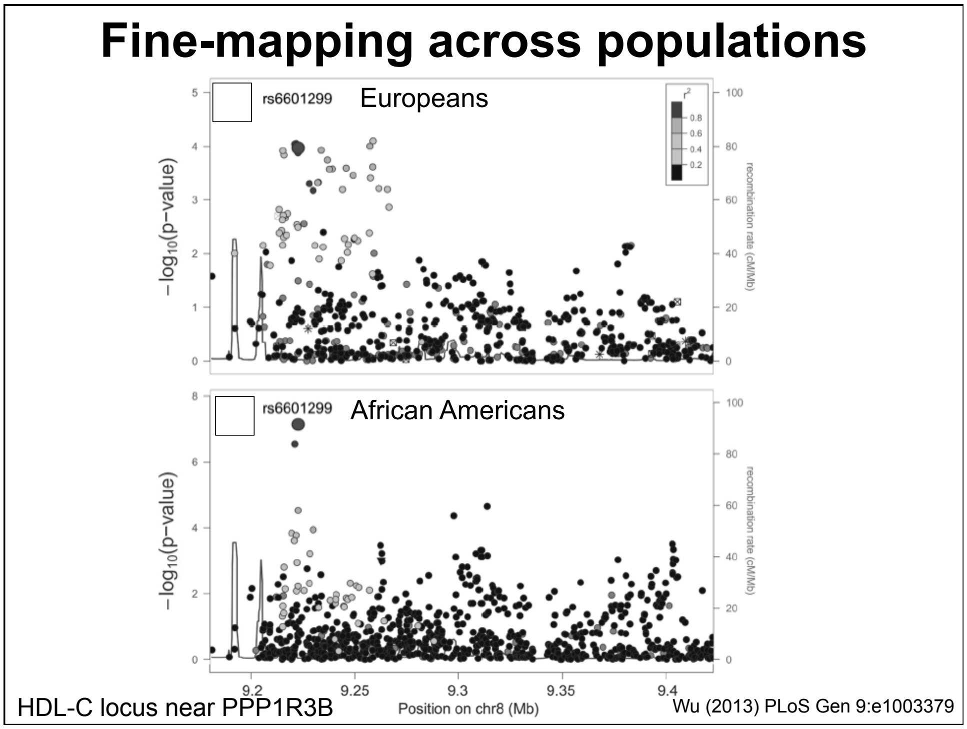
$$\text{Trait} = \beta_0 + \beta_1\text{SNP}_1 + \beta_2\text{SNP}_2$$

$$[\text{HDL}] = \beta_0 + \beta_1\text{rs261332} + \beta_2\text{rs4775041}$$

$$[\text{HDL}] = \beta_0 + \beta_1\text{rs261332} + \beta_2\text{rs4775041} + \beta_3\text{sex} + \beta_4\text{age} + \beta_5\text{age}^2$$

Tests independence of SNP effects

- If β_1 changes when β_2 is included in the model, then SNP_1 is sometimes inherited with SNP_2
- If neither β changes in reciprocal tests, then the two SNPs independently affect the trait



Fine-mapping across populations

HDL-C locus near PPP1R3B

Wu (2013) PLoS Gen 9:e1003379

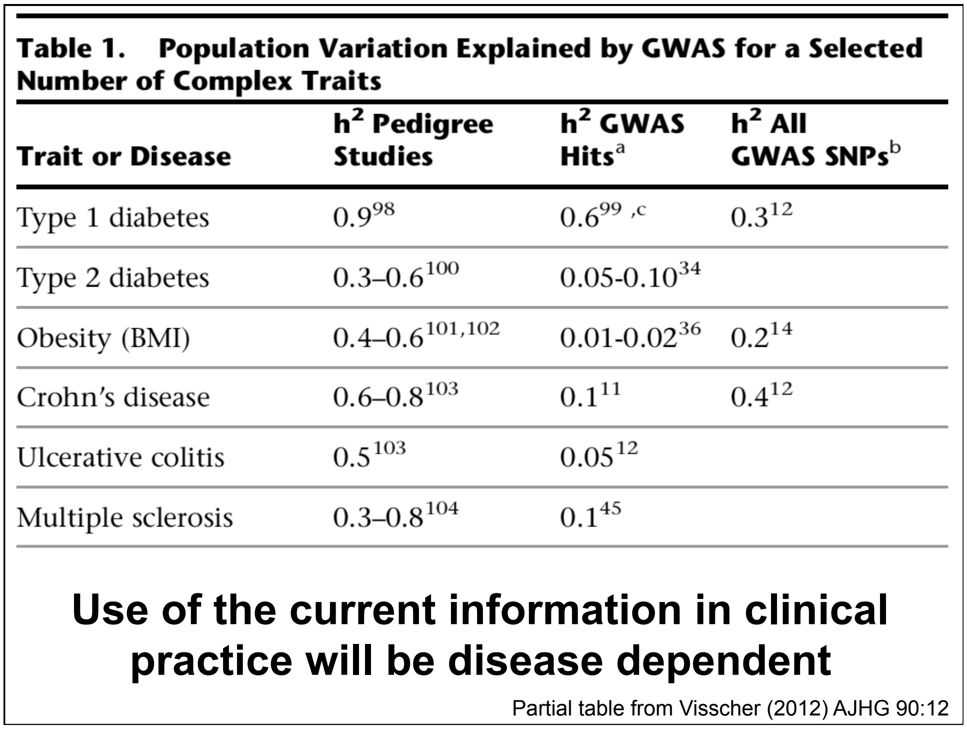


Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

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

Outline

- **Genome-wide association study design**
  - **Samples/study participants**
  - **Genotyping**
  - **Tests of association**
  - **Imputation and meta-analysis**
- **Interpretation of results**
  - **Effect size and significance**
  - **Example locus characteristics**
- **Sequencing/rare variant studies**

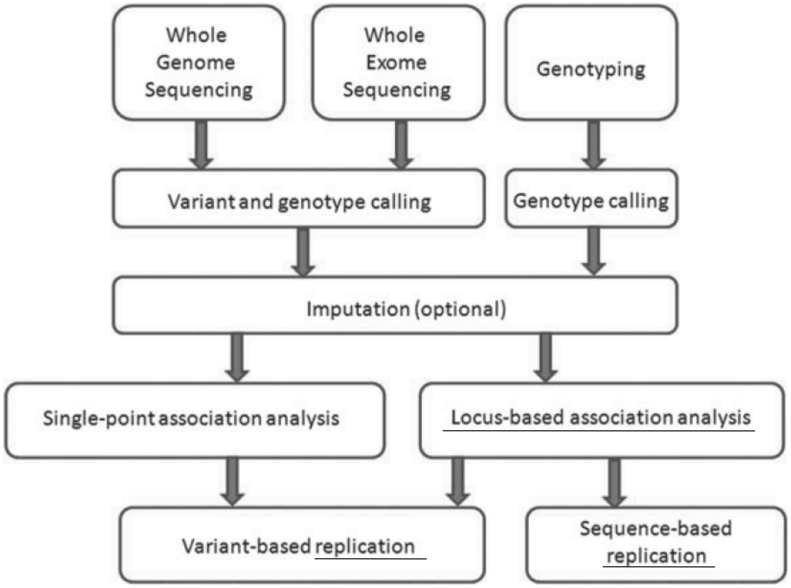
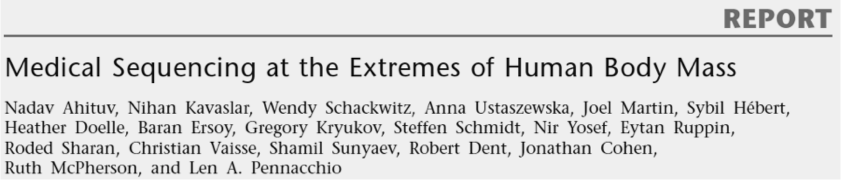

Whole Genome Sequencing
Whole Exome Sequencing
Genotyping
Variant and genotype calling
Genotype calling
Imputation (optional)
Single-point association analysis
Locus-based association analysis
Variant-based replication
Sequence-based replication

Figure 1. An overview of steps taken in the search for low-frequency and rare variants affecting complex traits.

Panoutsopoulou (2013) Hum Mol Gen 22:R16

Some sequencing study designs for complex traits

- Sequence selected individuals
  - extreme trait values (>95% vs <5% level)
  - cases and controls
- Increase the number of individuals
  - by decreasing sequencing coverage ($)
  - by collecting rare variants onto a less expensive genotyping array
- Sequence population isolates, where rare variants may have drifted to higher frequencies and LD may be longer



REPORT

Medical Sequencing at the Extremes of Human Body Mass

Nadav Ahituv, Nihan Kavaslar, Wendy Schackwitz, Anna Ustaszewska, Joel Martin, Sybil Hébert, Heather Doelle, Baran Ersoy, Gregory Kryukov, Steffen Schmidt, Nir Yosef, Eytan Ruppin, Roded Sharan, Christian Vaisse, Shamil Sunyaev, Robert Dent, Jonathan Cohen, Ruth McPherson, and Len A. Pennacchio

Sequenced coding regions and splice junctions of 58 genes in
379 obese individuals with mean BMI 49 and 378 lean individuals with mean BMI 19

Found >1000 variants, including 8 in *MC4R* that were subsequently tested for function

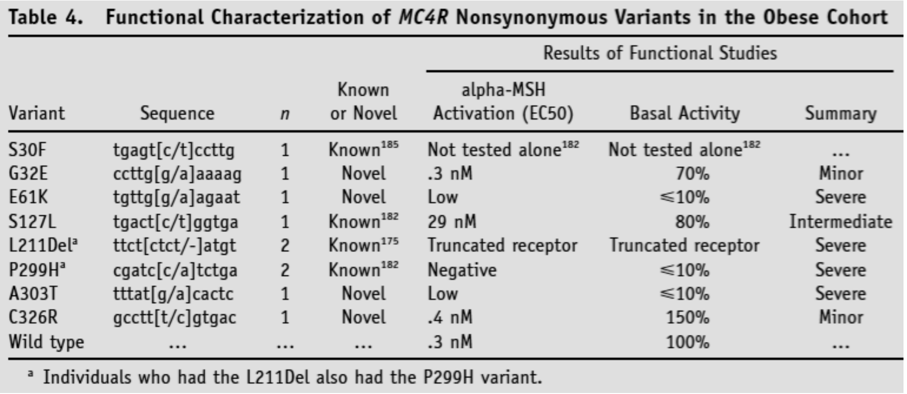


Table 4. Functional Characterization of *MC4R* Nonsynonymous Variants in the Obese Cohort

				Results of Functional Studies		
Variant	Sequence	*n*	Known or Novel	alpha-MSH Activation (EC50)	Basal Activity	Summary
S30F	tgagt[c/t]ccttg	1	Known[185]	Not tested alone[182]	Not tested alone[182]	...
G32E	ccttg[g/a]aaaag	1	Novel	.3 nM	70%	Minor
E61K	tgttg[g/a]agaat	1	Novel	Low	≤10%	Severe
S127L	tgact[c/t]ggtga	1	Known[182]	29 nM	80%	Intermediate
L211Del[a]	ttct[ctct/-]atgt	2	Known[175]	Truncated receptor	Truncated receptor	Severe
P299H[a]	cgatc[c/a]tctga	2	Known[182]	Negative	≤10%	Severe
A303T	tttat[g/a]cactc	1	Novel	Low	≤10%	Severe
C326R	gcctt[t/c]gtgac	1	Novel	.4 nM	150%	Minor
Wild type	...	...	...	.3 nM	100%	...

[a] Individuals who had the L211Del also had the P299H variant.

Am. J. Hum. Genet. 2007;80:779–791.

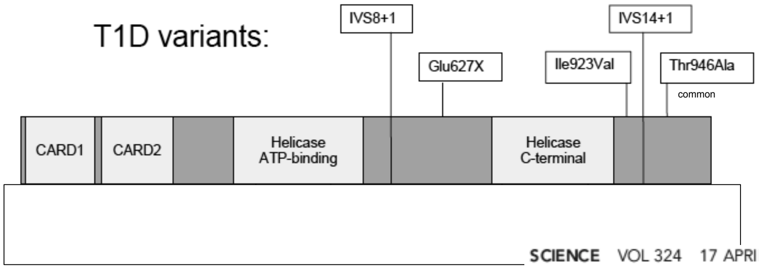
Sequencing at a GWAS locus

- **Sequence 'positional candidate' genes in cases & controls or individuals with extreme trait values**
- **Identify variants in cases (one extreme) that are absent from controls (other extreme)**
- **Hypothesize that occasional 'smoking gun' variants with strong effect will be identified**
- **Use evidence that variants affect gene function and lead to the same disease/trait to implicate that gene at the association signal**
- **Does not require finding the variant(s) responsible for association signal that may have a weaker effect**

Rare Variants of *IFIH1*, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes

Sergey Nejentsev,[1,2]* Neil Walker,[1] David Riches,[3] Michael Egholm,[3] John A. Todd[1]

Resequenced exons and splice sites of 10 candidate genes
in pools of DNA from 480 pts & 480 controls
Tested variants for association in >30,000 subjects



SCIENCE VOL 324 17 APRIL 2009

Rare variants confirmed to be associated with T1D in more samples

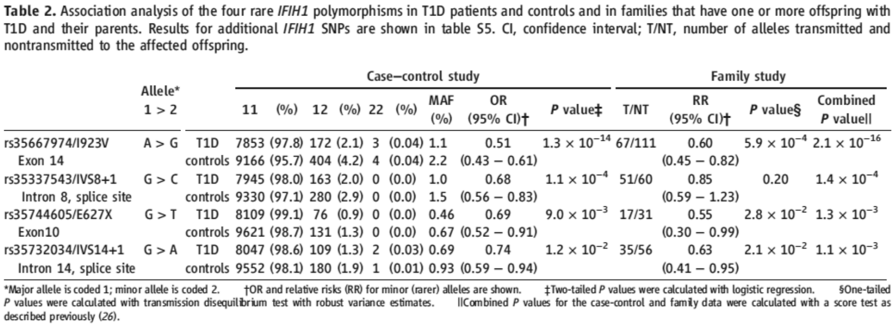


Table 2. Association analysis of the four rare *IFIH1* polymorphisms in T1D patients and controls and in families that have one or more offspring with T1D and their parents. Results for additional *IFIH1* SNPs are shown in table S5. CI, confidence interval; T/NT, number of alleles transmitted and nontransmitted to the affected offspring.

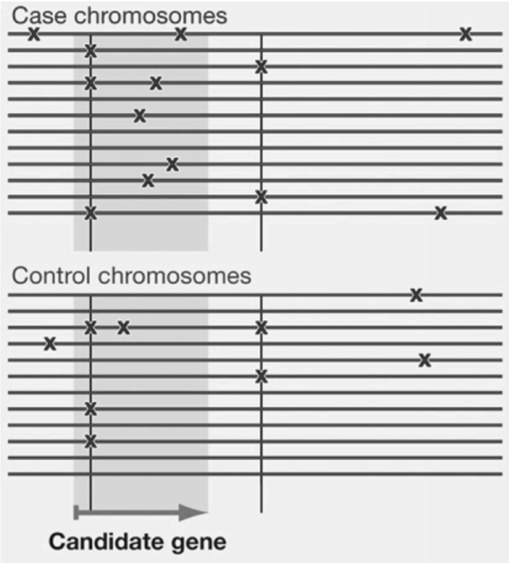
			Case–control study									Family study			
	Allele* 1 > 2		11	(%)	12	(%)	22	(%)	MAF (%)	OR (95% CI)†	*P* value‡	T/NT	RR (95% CI)†	*P* value§	Combined *P* value‖
rs35667974/I923V	A > G	T1D	7853	(97.8)	172	(2.1)	3	(0.04)	1.1	0.51	1.3×10^{-14}	67/111	0.60	5.9×10^{-4}	2.1×10^{-16}
Exon 14		controls	9166	(95.7)	404	(4.2)	4	(0.04)	2.2	(0.43 – 0.61)			(0.45 – 0.82)		
rs35337543/IVS8+1	G > C	T1D	7945	(98.0)	163	(2.0)	0	(0.0)	1.0	0.68	1.1×10^{-4}	51/60	0.85	0.20	1.4×10^{-4}
Intron 8, splice site		controls	9330	(97.1)	280	(2.9)	0	(0.0)	1.5	(0.56 – 0.83)			(0.59 – 1.23)		
rs35744605/E627X	G > T	T1D	8109	(99.1)	76	(0.9)	0	(0.0)	0.46	0.69	9.0×10^{-3}	17/31	0.55	2.8×10^{-2}	1.3×10^{-3}
Exon10		controls	9621	(98.7)	131	(1.3)	0	(0.0)	0.67	(0.52 – 0.91)			(0.30 – 0.99)		
rs35732034/IVS14+1	G > A	T1D	8047	(98.6)	109	(1.3)	2	(0.03)	0.69	0.74	1.2×10^{-2}	35/56	0.63	2.1×10^{-2}	1.1×10^{-3}
Intron 14, splice site		controls	9552	(98.1)	180	(1.9)	1	(0.01)	0.93	(0.59 – 0.94)			(0.41 – 0.95)		

*Major allele is coded 1; minor allele is coded 2. †OR and relative risks (RR) for minor (rarer) alleles are shown. ‡Two-tailed *P* values were calculated with logistic regression. §One-tailed *P* values were calculated with transmission disequilibrium test with robust variance estimates. ‖Combined *P* values for the case-control and family data were calculated with a score test as described previously (*26*).

Establishes the role of IFIH1 in T1D and demonstrates that resequencing studies can pinpoint disease-causing genes in regions initially identified by GWASs.

SCIENCE VOL 324 17 APRIL 2009

Identify an increased ‘burden’ of variants in a single gene or locus



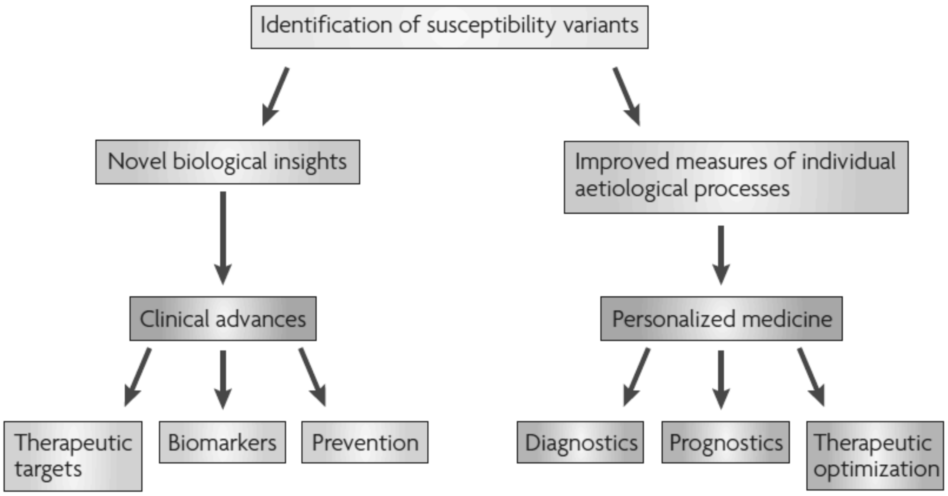
- Many individually important variants will be too rare to detect the association with the trait; however, there will often be more than one important variant in a gene
- Gene-based tests combine information from multiple variants into a single test statistic to be used as predictor in genetic association tests
- What information about the variants should we use?

Raychaudhuri (2011) Cell 147:57

Rare variant burden tests

- Many alternative forms – an active area of research
- Collapse information from multiple variants into single test
- Some tests allow the direction of effect of each variant to be different
- The choice of variants included in tests has a large impact on the test
- Including too many null variants can kill statistical power but so can not including the right ones
- Filter missense variants on minor allele frequency and predictive function?
- Restrict tests to obvious functional variants?

Clinical translation



Identification of susceptibility variants

Novel biological insights

Improved measures of individual aetiological processes

Clinical advances

Personalized medicine

Therapeutic targets

Biomarkers

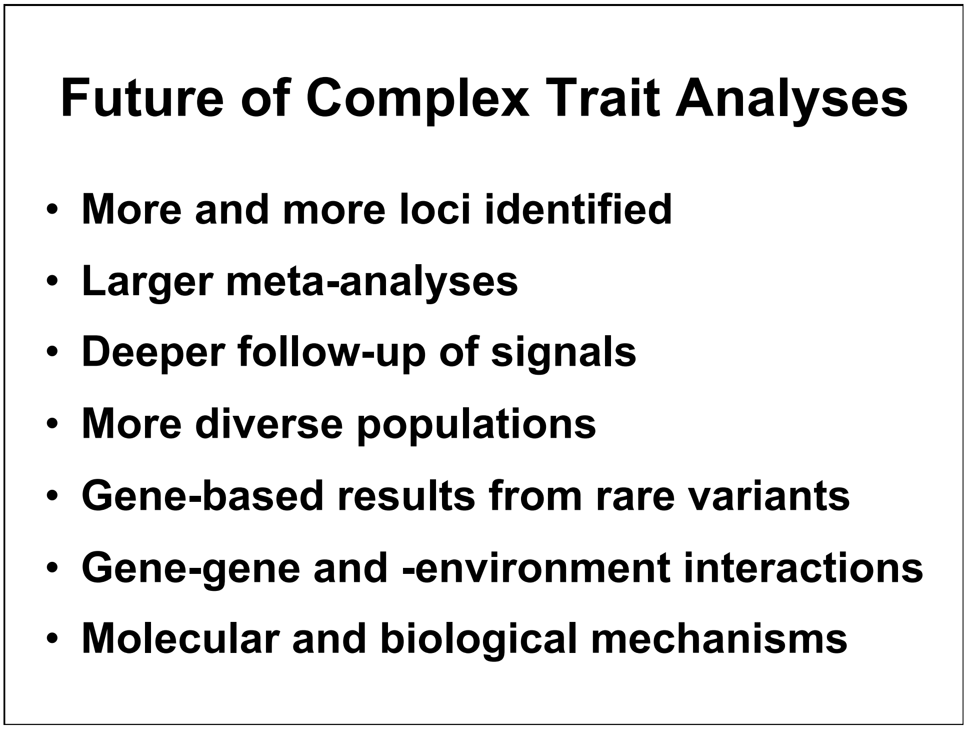
Prevention

Diagnostics

Prognostics

Therapeutic optimization

McCarthy (2008) Nat Rev Gen 9:356



Future of Complex Trait Analyses

- More and more loci identified
- Larger meta-analyses
- Deeper follow-up of signals
- More diverse populations
- Gene-based results from rare variants
- Gene-gene and -environment interactions
- Molecular and biological mechanisms