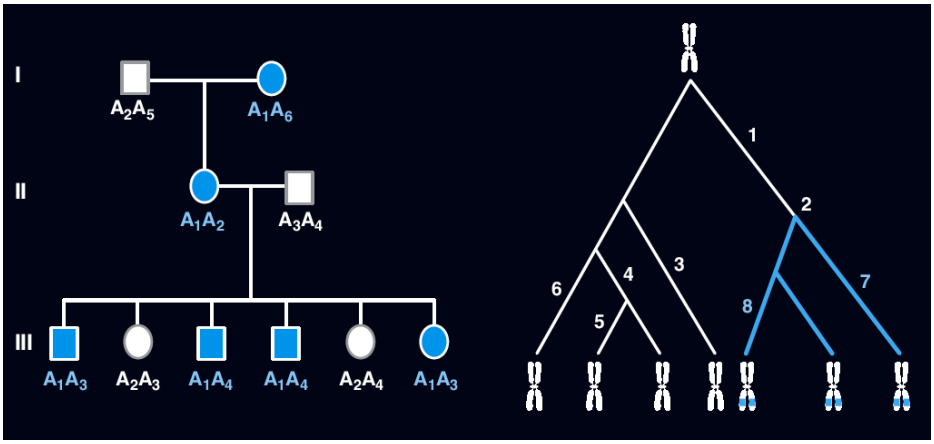
Genomic Approaches to the Study of Complex Genetic Diseases

Karen Mohlke, PhD
Department of Genetics
University of North Carolina
April 20, 2016

Complex diseases & traits

Gene mapping in populations

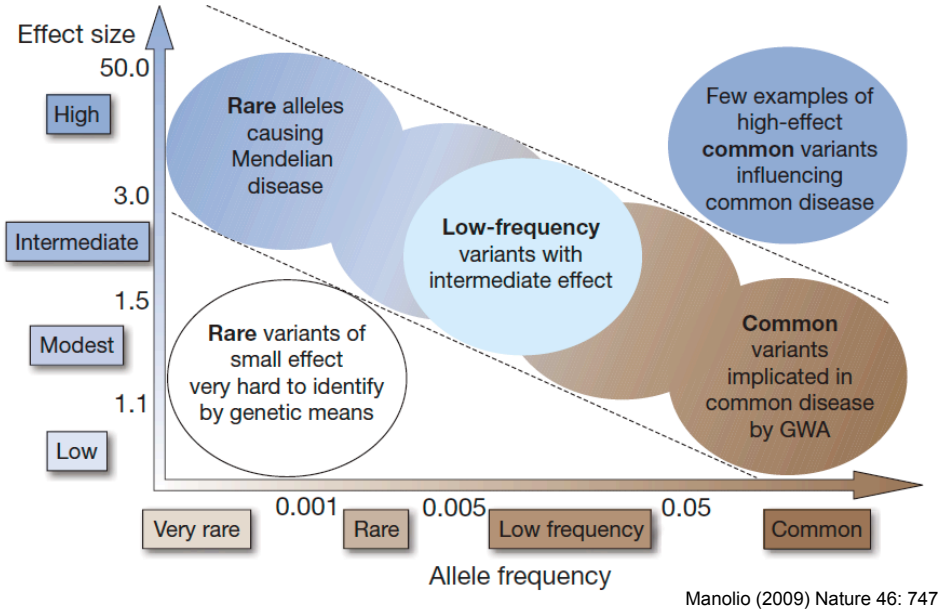


Altshuler and Clark (2005) Science 307:1052

Genome-wide association studies

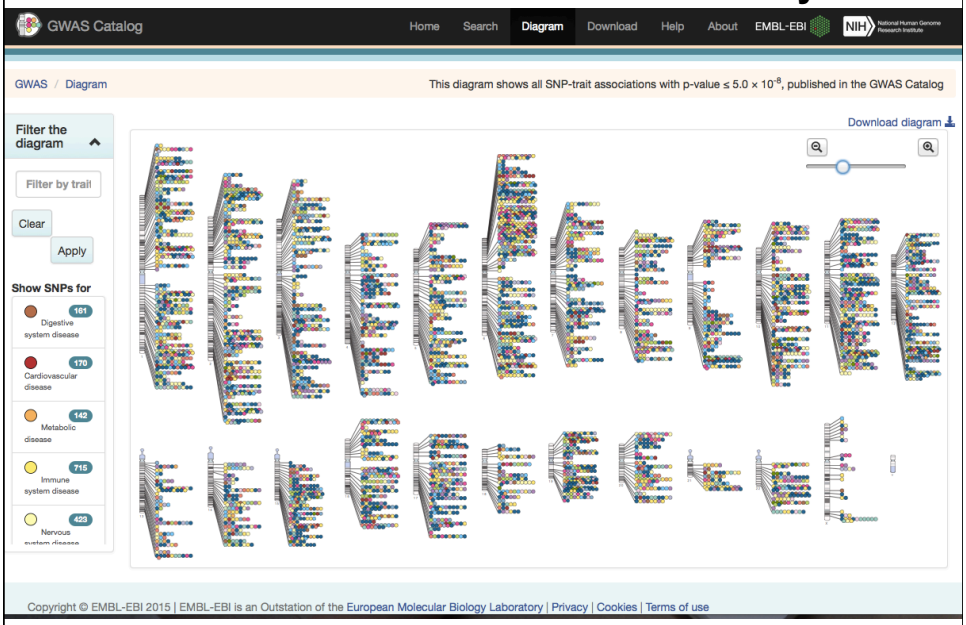
- **Test a large portion of the common single nucleotide genetic variation in the genome for association with a disease or variation in a quantitative trait**
- **Find disease/quantitative trait-related variants without a prior hypothesis of gene function**

Genetic architecture



Manolio (2009) Nature 46: 747

Genome-wide association studies identify loci



https://www.ebi.ac.uk/gwas/diagram

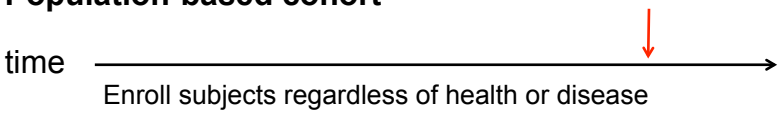
Welter (2014) Nuc Acids Res 42: D1001

Outline

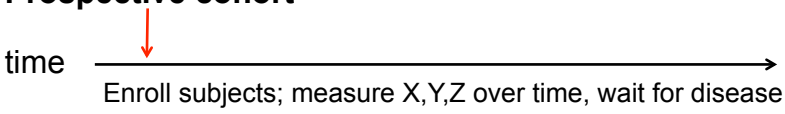
- **Genome-wide association study design**
  - **Samples/study participants**
  - **Genotyping**
  - **Tests of association**
  - **Imputation and meta-analysis**
- **Interpretation of results**
  - **Effect size and significance**
  - **Example locus characteristics**
- **Sequencing/rare variant studies**

Study designs

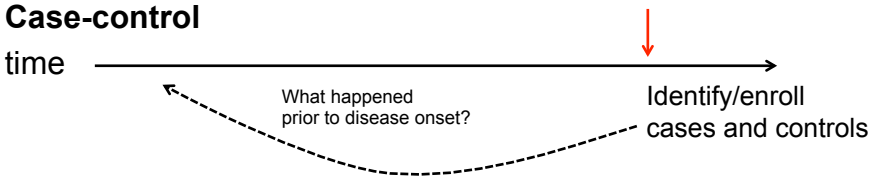
Population-based cohort



Prospective cohort

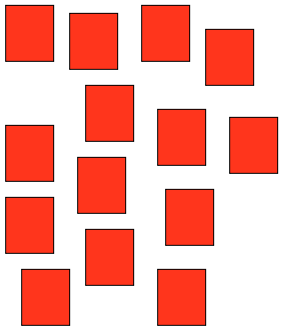


Case-control

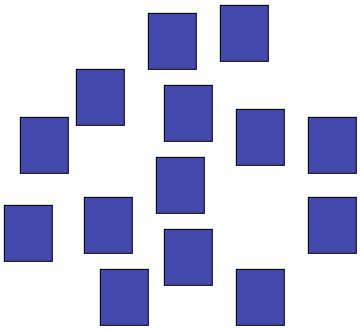


Matching of cases and controls

Cases



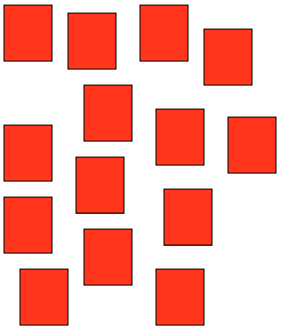
Controls



Cases and controls should be comparable in all respects except disease status (e.g. age, sex, demographics)

Selection of cases

Cases

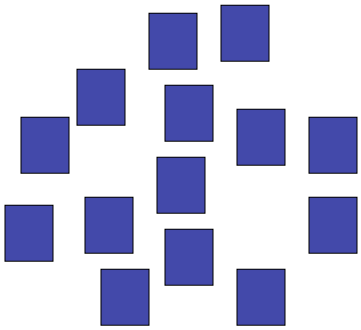


- Potential criteria to enrich genetic effect size
  - More severely affected individuals
  - Require other family member to have disease
  - Younger age-of-disease onset

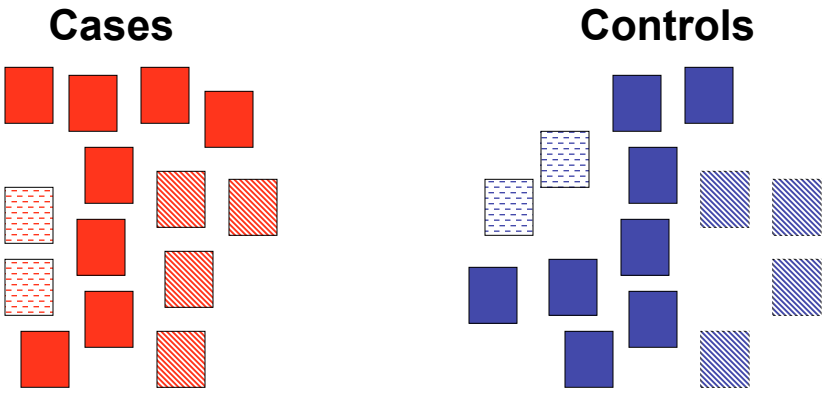
Selection of controls

- Potential criterion to enrich genetic effect size
  - Low risk of disease rather than population-based samples

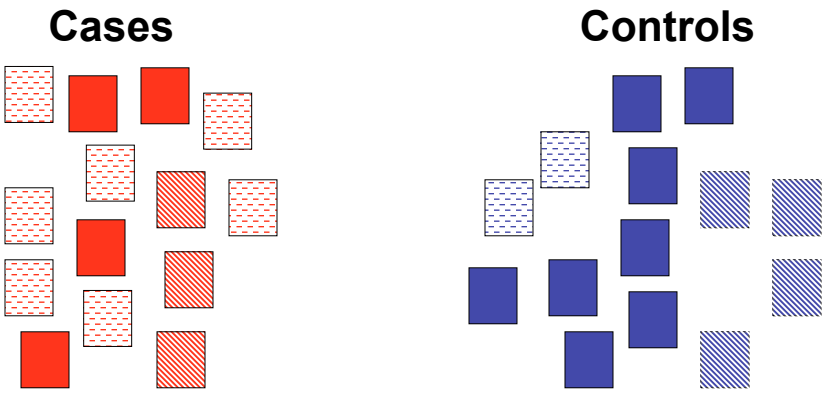
Controls



Comparable ancestry

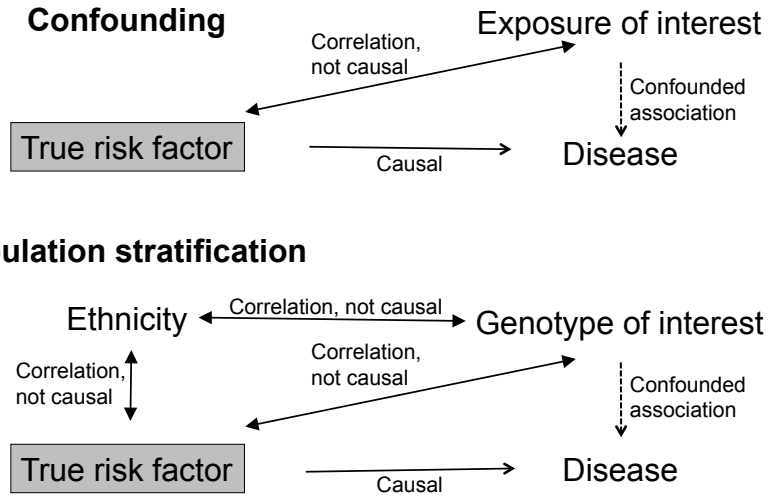


Ancestry differences



May have inadequate ancestry information prior to genotyping

Confounding and population stratification



Cancer Epidemiol Biomarkers Prev 11: 513

Population stratification

- **Systematic differences in allele frequencies between subpopulations that may be due to different ancestry**
- **Oversampled individuals from one sub-population for cases in a case-control genetic association study can produce spurious associations**

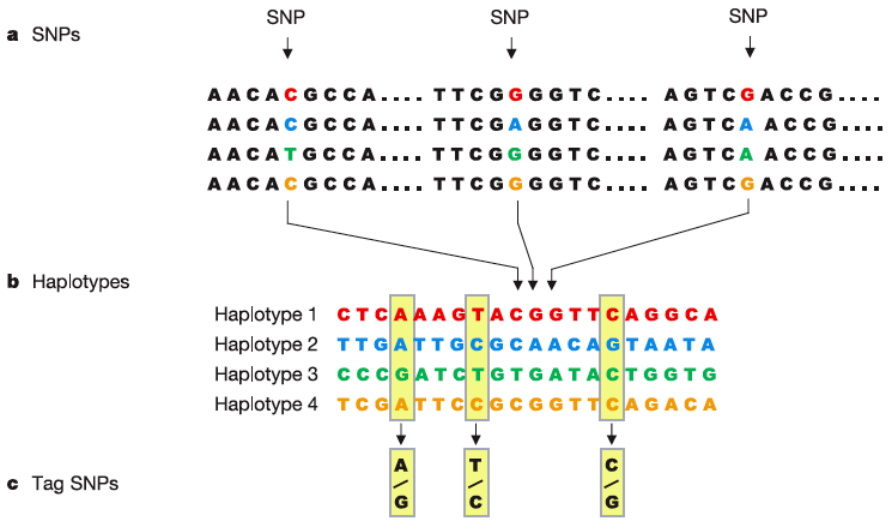
Account for or avoid population stratification

- **Match cases with controls**
- **Restrict to one subgroup**
- **Adjust for genetic background**
  - **E.g. 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 genotyping panels

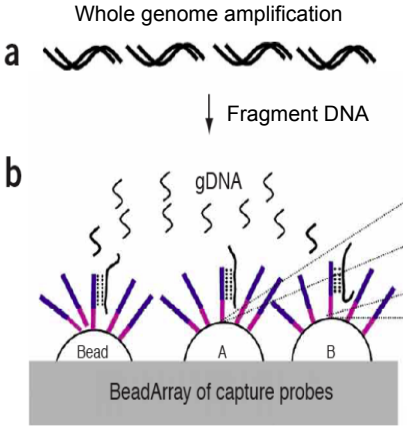
- **10,000 - 5 million variants**
- **Affymetrix, Illumina**
  - **Random SNPs**
  - **Selected haplotype tag variants**
  - **Copy number probes**
  - **More lower frequency variants**
  - **Exome variants**
  - **Some arrays allow variants to be added**

Selecting 'haplotype tag' SNPs



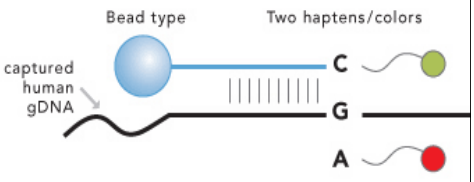
International HapMap Consortium (2003) Nature 426:789

Illumina Infinium Assay



Hybridize unlabeled DNA to specific arrays of 50-mers

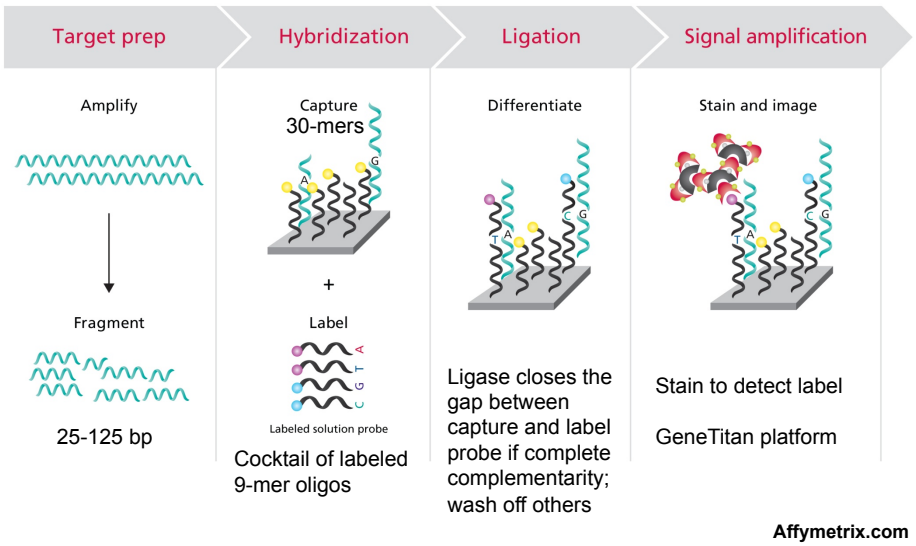
Allele-specific primer extension with labeled nucleotides



Dual-color florescent staining
Detect fluorescent color and intensity

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

Affymetrix Axiom Array



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

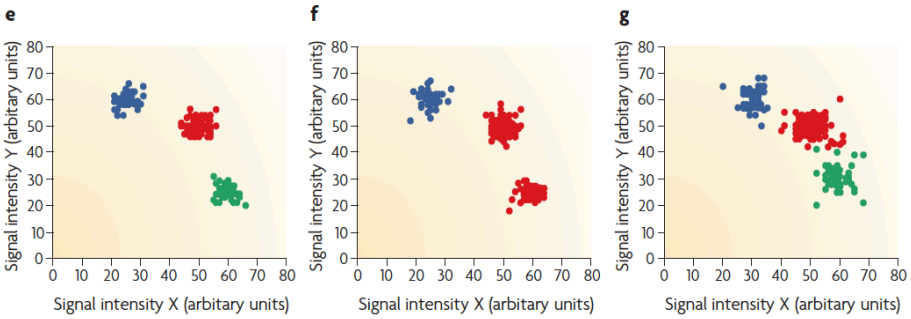
Quality control: Identify and remove bad samples

- 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

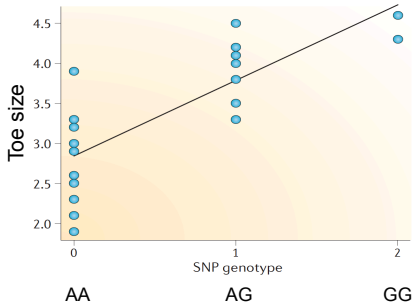
Statistical analysis: linear regression

Two main parameters: p-value and effect size

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

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

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



Statistical analysis: linear regression

Two main parameters: p-value and effect size

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

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

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

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

covariates

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

Odds ratio

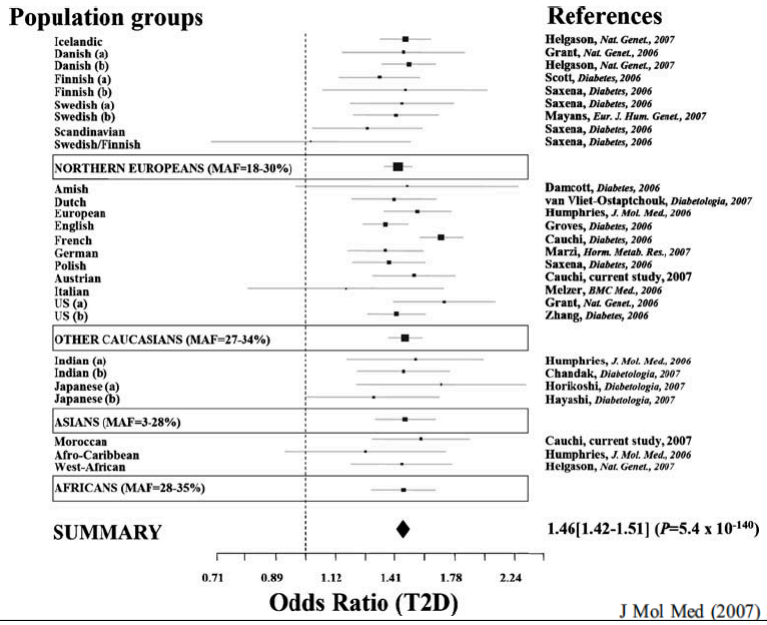
- **Surrogate measure of effect of allele on risk of developing disease**

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

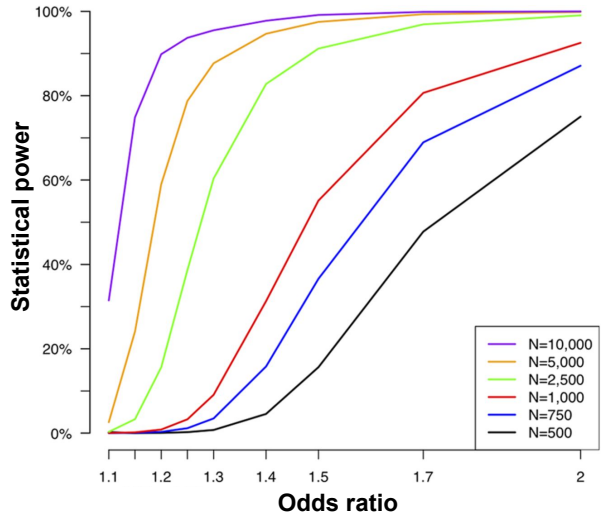
$$\text{Odds Ratio} = \frac{\text{Case C / Case A}}{\text{Control C / Control A}} = \frac{1140 / 860}{1000 / 1000} = 1.33$$

Association study odds ratio plot



J Mol Med (2007) 85:777–782

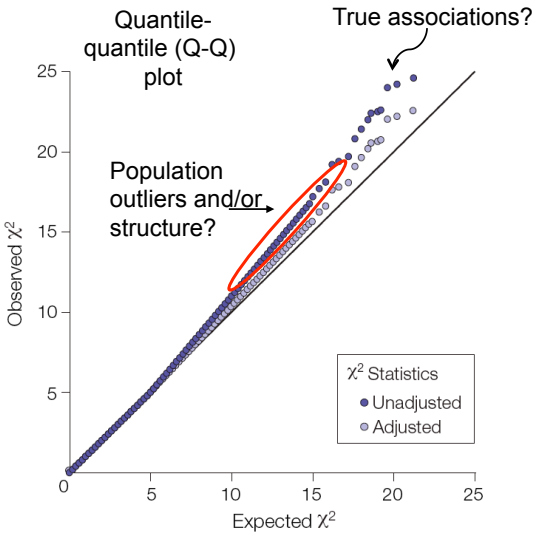
Relationship between GWAS sample size and power



Paria (2014) J Bone Joint Surg Am 96:e38

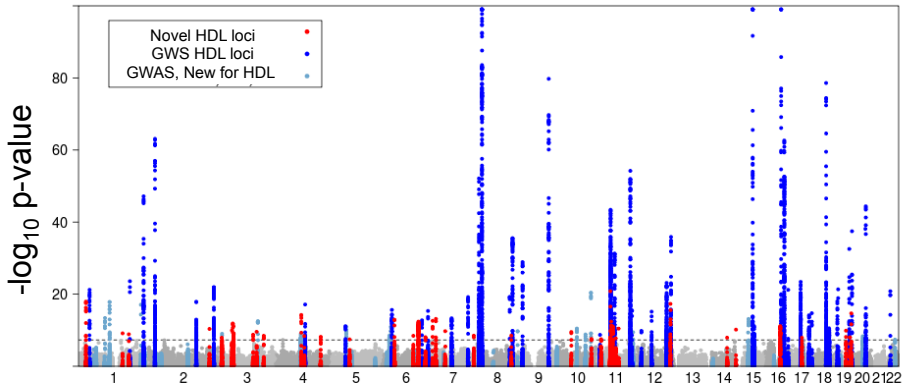
Adjust for population structure: genomic control

- 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 $\lambda = \text{median}(X_i^2)/0.456$.
- Inflation factor $\lambda > 1$ indicates population structure, unknown relatives or other errors.
- The tests of association can be adjusted by this factor. $X_i^2{}_{adjusted} = X_i^2/\lambda$



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

'Manhattan plot' for HDL-cholesterol



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

GLGC (2013) Nat Gen 45:1274

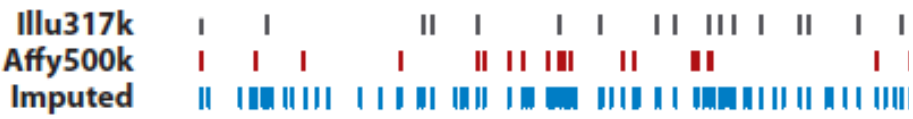
Multiple testing

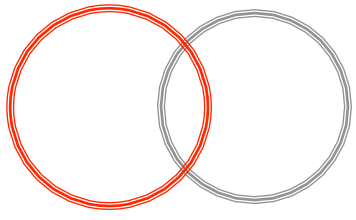
- **Genotype and test > 300K – 5M SNPs**
- **Correct for the multiple tests**

$$\frac{.05}{\text{~1 million common SNPs}} = \text{5 x } 10^{-8}$$

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

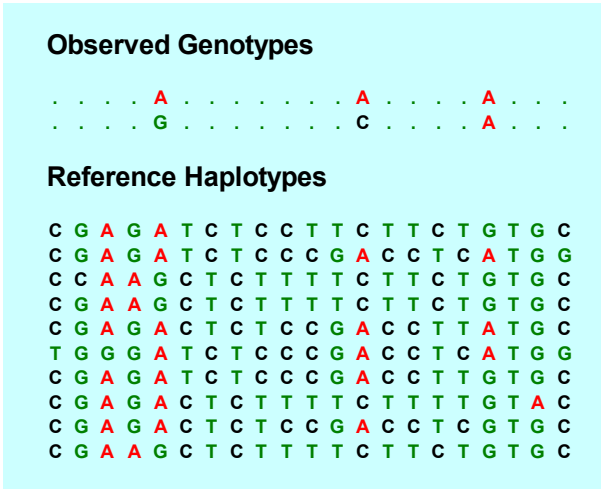
Imputation of ungenotyped variants





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

Imputation: Observed genotypes



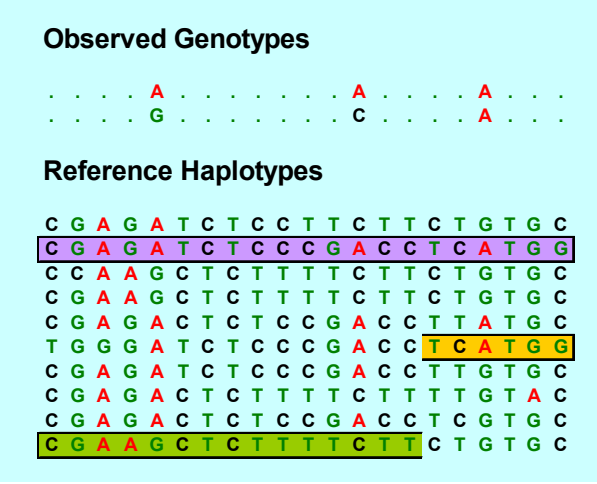
Study Sample

HapMap or 1000 Genomes or …

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

Gonçalo Abecasis

Identify match among reference



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

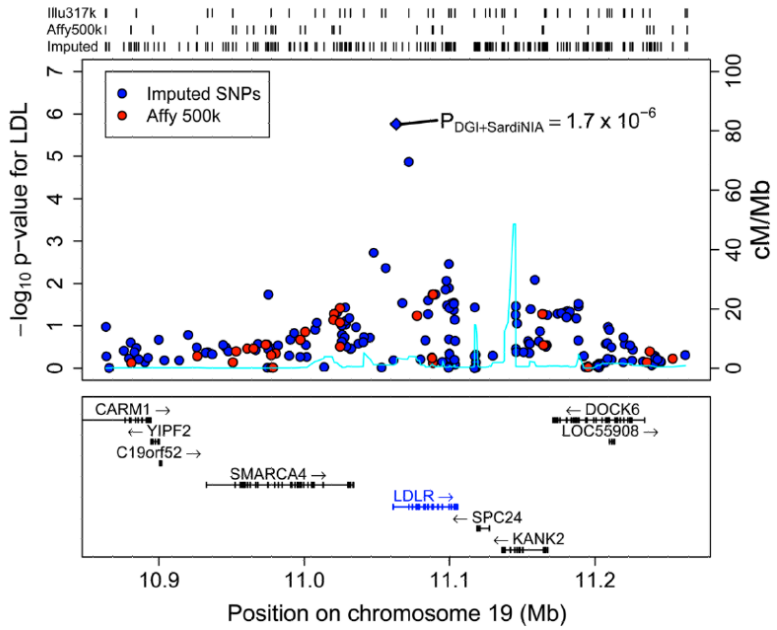
Reference Haplotypes

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

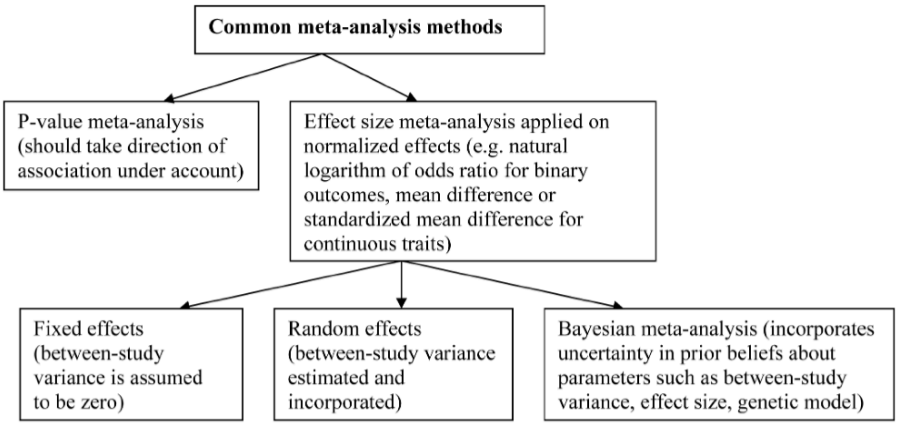


Li (2009) Ann Rev Genomics Hum Genet 10:387

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

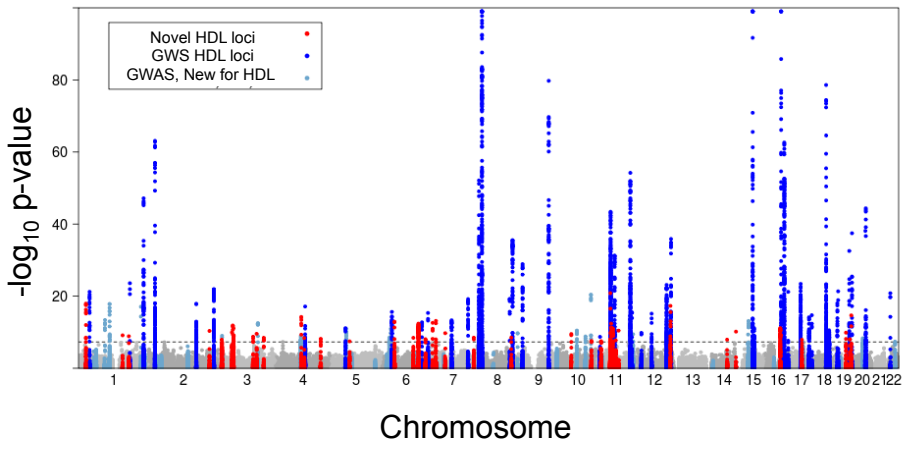


Zeggini (2009) Pharmacogenomics 10:191

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**

'Manhattan plot' for HDL-cholesterol



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

GLGC (2013) Nat Gen 45:1274

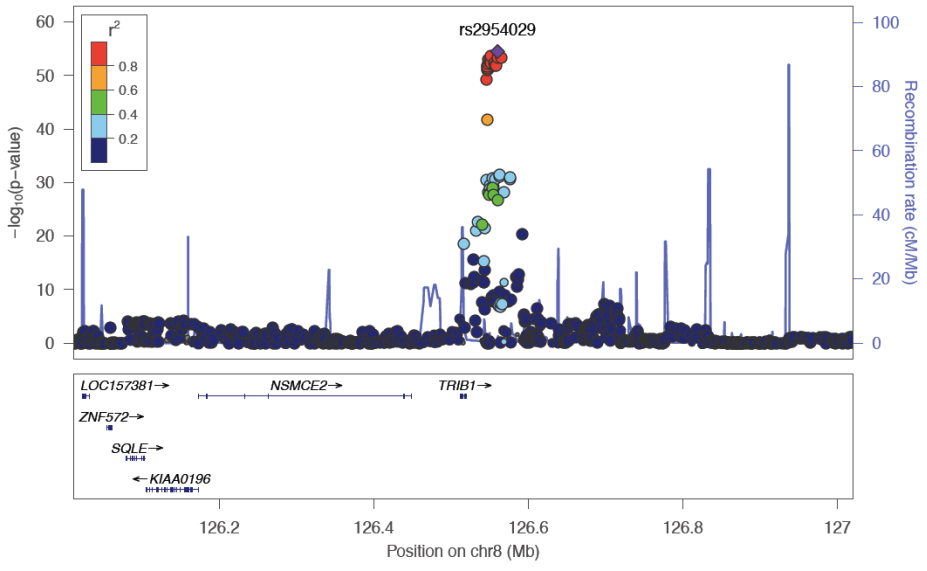
Table 1 New loci primarily associated with HDL cholesterol levels obtained from joint GWAS and Metabochip meta-analysis

Locus	Marker name	Chr.	hg19 position (Mb)	Associated trait(s)	MAF	Minor/major allele	Effect of A1	Joint *n* (×1,000)	Joint *P* value
PIGV-NR0B2	rs12748152	1	27.14	HDL, LDL, TG	0.09	T/C	−0.051, 0.050, 0.037	187, 173, 178	1×10^{-15}, 3×10^{-12}, 1×10^{-9}
HDGF-PMVK	rs12145743	1	156.70	HDL	0.34	G/T	0.020	181	2×10^{-8}
ANGPTL1	rs4650994	1	178.52	HDL	0.49	G/A	0.021	187	7×10^{-9}
CPS1	rs1047891	2	211.54	HDL	0.33	A/C	−0.027	182	9×10^{-10}
ATG7	rs2606736	3	11.40	HDL	0.39	C/T	0.025	129	5×10^{-8}
SETD2	rs2290547	3	47.06	HDL	0.20	A/G	−0.030	187	4×10^{-9}
RBM5	rs2013208	3	50.13	HDL	0.50	T/C	0.025	170	9×10^{-12}
STAB1	rs13326165	3	52.53	HDL	0.21	A/G	0.029	187	9×10^{-11}
GSK3B	rs6805251	3	119.56	HDL	0.39	T/C	0.020	186	1×10^{-8}
C4orf52	rs10019888	4	26.06	HDL	0.18	G/A	−0.027	187	5×10^{-8}
FAM13A	rs3822072	4	89.74	HDL	0.46	A/G	−0.025	187	4×10^{-12}
ADH5	rs2602836	4	100.01	HDL	0.44	A/G	0.019	187	5×10^{-8}
RSPO3	rs1936800	6	127.44	HDL, TG[a]	0.49	C/T	0.020, −0.020	187, 168	3×10^{-10}, 3×10^{-8}
DAGLB	rs702485	7	6.45	HDL	0.45	G/A	0.024	187	6×10^{-12}
SNX13	rs4142995	7	17.92	HDL	0.38	T/G	−0.026	165	9×10^{-12}
IKZF1	rs4917014	7	50.31	HDL	0.32	G/T	0.022	187	1×10^{-8}
TMEM176A	rs17173637	7	150.53	HDL	0.12	C/T	−0.036	184	2×10^{-8}
MARCH8-ALOX5	rs970548	10	46.01	HDL, TC	0.26	C/A	0.026, 0.025	187, 187	2×10^{-10}, 8×10^{-9}
OR4C46	rs11246602	11	51.51	HDL	0.15	C/T	0.034	176	2×10^{-10}
KAT5	rs12801636	11	65.39	HDL	0.23	A/G	0.024	187	3×10^{-8}
MOGAT2-DGAT2	rs499974	11	75.46	HDL	0.19	A/C	−0.026	187	1×10^{-8}
ZBTB42-AKT1	rs4983559	14	105.28	HDL	0.40	G/A	0.020	184	1×10^{-8}
FTO	rs1121980	16	53.81	HDL, TG[b]	0.43	A/G	−0.020, 0.021	186, 155	7×10^{-9}, 3×10^{-8}
HAS1	rs17695224	19	52.32	HDL	0.26	A/G	−0.029	185	2×10^{-13}

Chr., chromosome; A1, minor allele; A2, major allele; TG, triglycerides; TC, total cholesterol. Effect sizes are given with respect to the minor allele (A1) in s.d. For loci associated with two or more traits at genome-wide significance, the trait corresponding to the strongest *P* value is listed first.
[a]The secondary trait was most strongly associated with a different SNP: rs719726 (within 1 Mb of rs1936800, $r^2 = 0.74$). [b]The secondary trait was most strongly associated with a different SNP, rs9930333 (within 1 Mb of rs1121980, $r^2 = 0.99$).

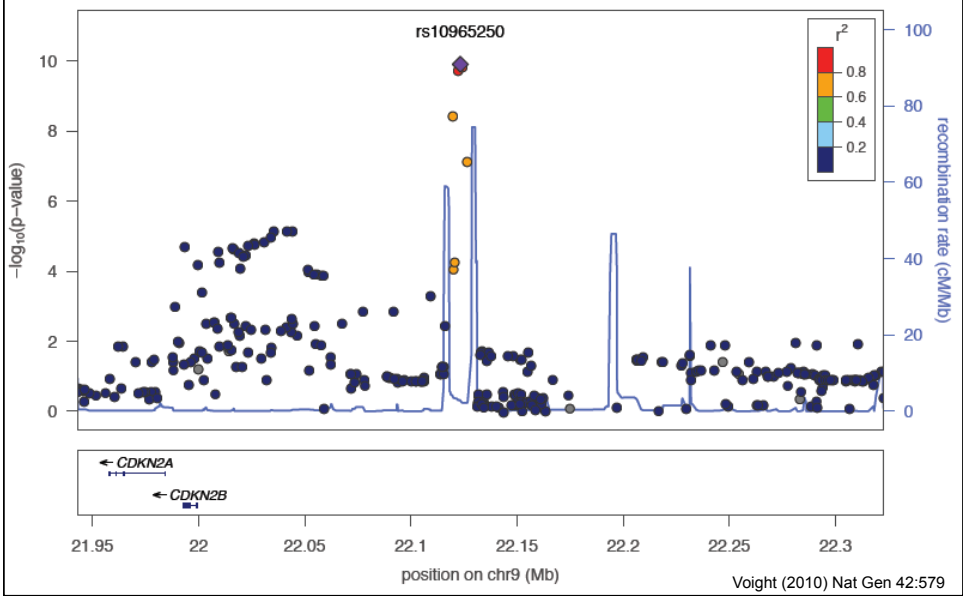
GLGC (2013) Nat Gen 45:1274

Single good candidate gene



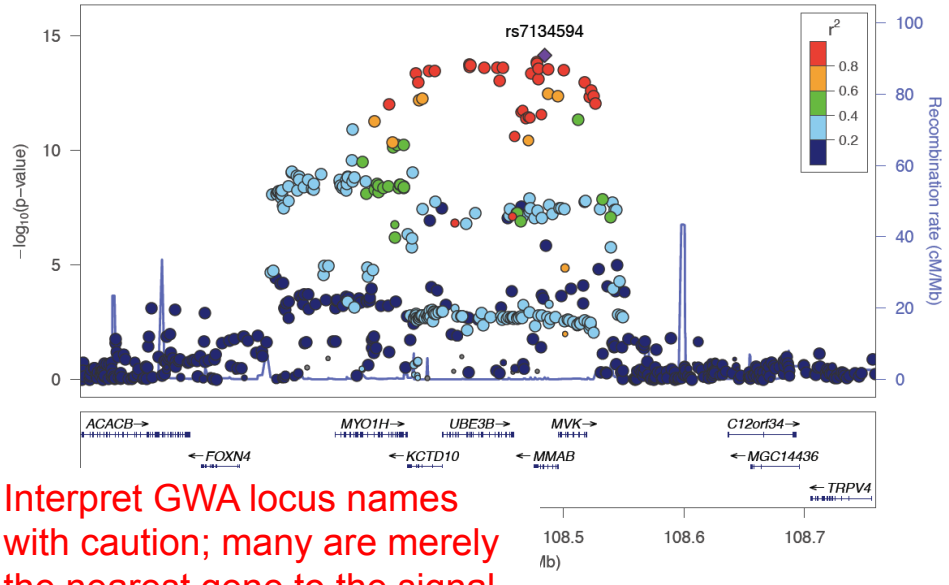
Teslovich (2010) Nature 466:707

Signal outside of genes



Voight (2010) Nat Gen 42:579

Many candidate genes



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

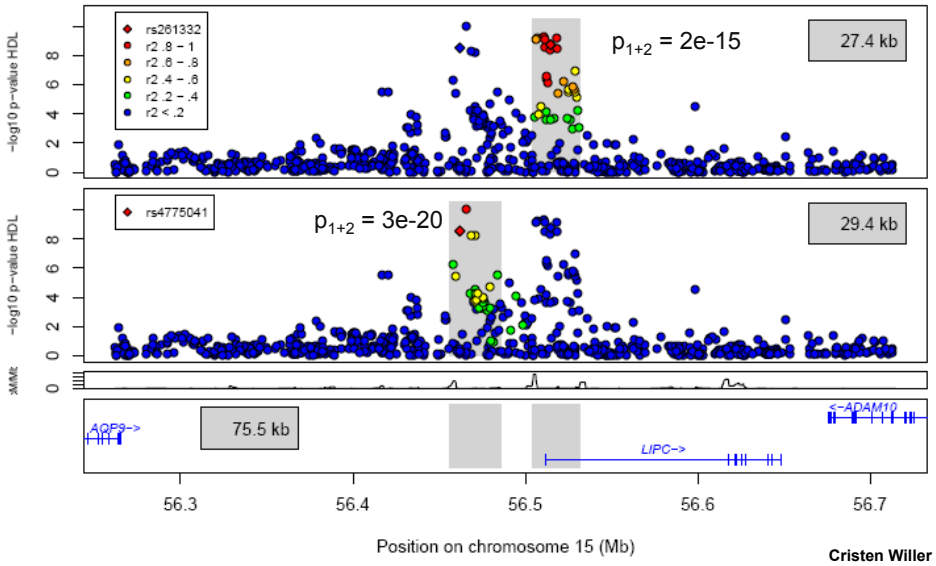
Teslovich (2010) Nature 466:707

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



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

Conditional analysis

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

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

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

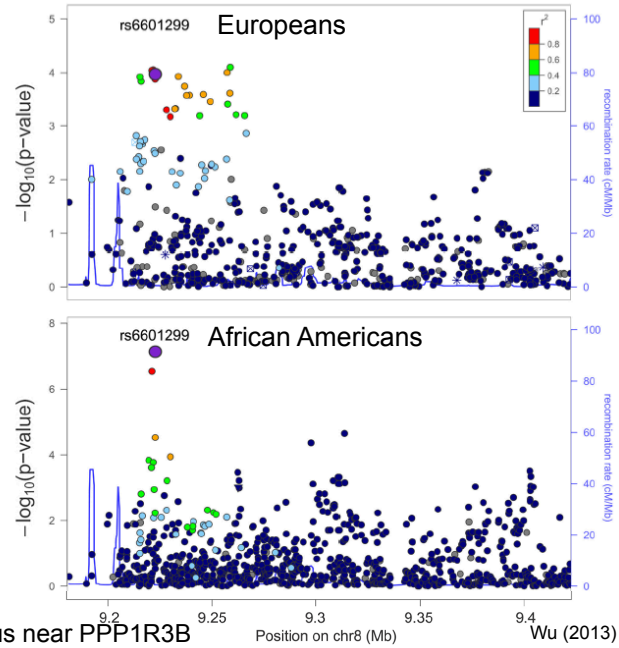
$$[\text{HDL}] = \beta_0 + \beta_1 \text{rs261332} + \beta_2 \text{rs4775041} + \beta_3 \text{sex} + \beta_4 \text{age} + \beta_5 \text{age}^2$$

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

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**

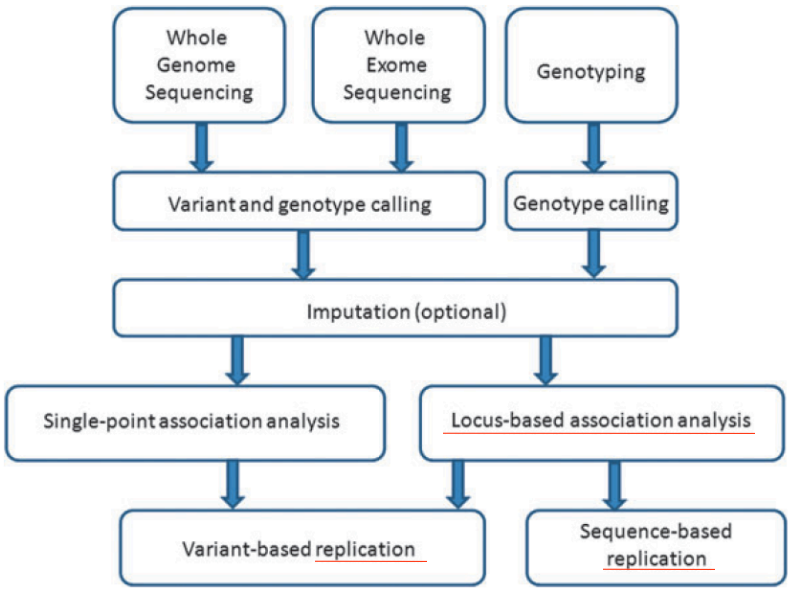


Figure 1. An overview of steps taken in the search for low-frequency and rare variants affecting complex traits.

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.

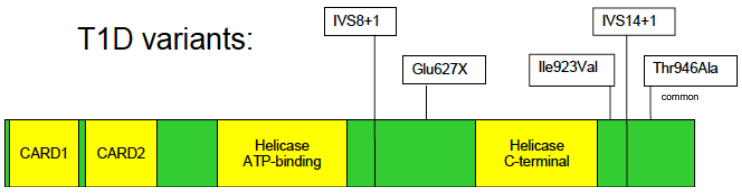
Variant discovery at GWAS locus

- **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.

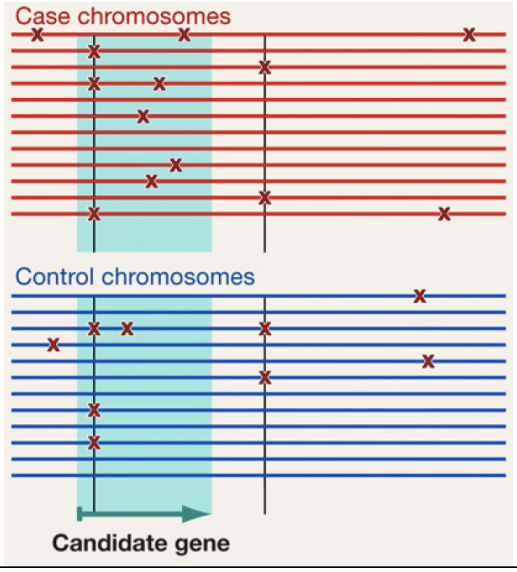
	Allele* 1 > 2		Case–control study								Family study			
			11	(%)	12	(%)	22	(%)	MAF (%)	OR (95% CI)†	*P* value‡	T/NT	RR (95% CI)†	*P* value§	Combined *P* value‖
rs35667974/I923V	A > G	T1D	7853	(97.8)	172	(2.1)	3	(0.04)	1.1	0.51	1.3×10^{-14}	67/111	0.60	5.9×10^{-4}	2.1×10^{-16}
Exon 14		controls	9166	(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

Raychaudhuri (2011) Cell 147:57

Rare variant burden (gene-based) tests

- **Collapse information from multiple variants into single test (e.g. count risk alleles across a set of variants)**
- **Some tests allow the direction of effect of each variant to be different (gain of function versus lost of function)**
- **Choice of variants to include in tests has a large impact on the test. Including too many neutral variants reduces 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 (nonsense, frameshift indels, splice errors)**

Gene-based rare variant association methods

Method name	Citation	Software	Description
Unidirectional rare variant gene-based tests			
Collapsing methods			
Combined Multivariate and Collapsing (CMC)	Liu & Leal, PLoS Comp. Bio. 2008	EPACTS	All rare variants collapsed into a single variant; individual dosage for the collapsed 'variant' is regressed against phenotype.
Weighted and un-weighted sum methods			
Variable threshold (VT)	Price et al, AJHG. 2010	PLINK-Seq	Sum of rare allele count in cases vs. controls; allele frequency threshold for inclusion is varied to maximize test statistic.
Weighted Sum Statistic (FRQWGT)	Madsen & Browning, PLoS Gen. 2009	PLINK-Seq	Permutation-based test comparing inverse-frequency-weighted rare variant counts per individual in cases vs. controls.
Weighted Sum Method (WILCOX-WSS)	Madsen & Browning, PLoS Gen. 2009	EPACTS	Wilcoxon Rank Sum test between phenotypes and inverse frequency-weighted rare variant scores.
Kernel-Based Adaptive Cluster (KBAC)	Liu & Leal, PLoS Gen. 2010	PLINK-Seq	Variant weights are determined adaptively, and are based on observed effect sizes; individuals scored by weighted sum of allele counts.
Summary case:control count methods			
BURDEN method	Purcell (PLINK-Seq)	PLINK-Seq	Permutation-based test comparing raw allele counts in cases vs. controls.
UNIQ test	Purcell (PLINK-Seq)	PLINK-Seq	Simple count of total case-unique rare alleles; permutations to assess significance.
Bi-directional variance-component gene-based tests			
C-ALPHA	Neale et al, PLoS Gen. 2011	PLINK-Seq	Detects deviation of observed case:control variant counts from expected binomial distribution.
Sequence Kernel Association Test (SKAT)	Wu et al, AJHG 2011	EPACTS	Generalized form of C-ALPHA with variants weighted by allele frequency.
Linear combination of unidirectional and variance-component tests			
SKAT-O ('Optimal' SKAT)	Lee et al, AJHG. 2012	EPACTS	Adaptive linear combination of unidirectional burden test and variance-component SKAT test.
Mixed Effects Score Test (MiST)	Sun et al, Genetic Epi. 2013	Public R package	Hierarchical regression model combining two independent test statistics which quantify variant effect sizes and 'heterogeneity'.

Moutsianas (2015) PLoS Genet 11: e1005165

An example of a gene-based test

Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes

- Initially sequenced 352 young lean T2D cases, 406 elderly obese euglycemic controls
- Then tested variants in 6,388 cases and 7,496 controls
- Found a nonsense variant in 7 cases and 21 controls, odds ratio (OR) = 0.38, *P* = 0.05
- Added this variant to the exome array and tested more individuals (N= 48,115, *P* = 0.0067).
- Difficult to increase sample size because variant mostly restricted to western Finland
- Expanded to look at more variants in the gene in other populations…

Flannick (2014) NatGen 46:357

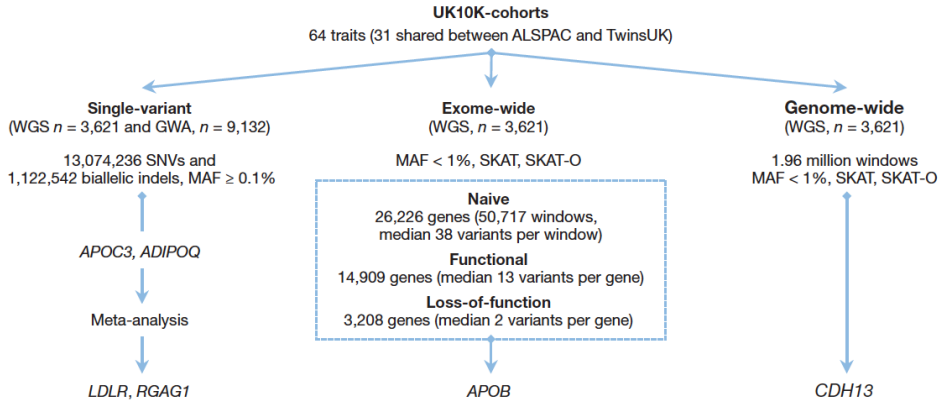
SLC30A8 variants in ~150,000 individuals

Table 1 Association of *SLC30A8* variants with T2D

				N		Carriers		Allele frequency		OR	
Variant	Ancestry	Country	Cohort	Cases	Controls	Cases	Controls	Cases (%)	Controls (%)	(95% CI)	*P*
p.Arg138*	European	Finland	Botnia	3,727	5,440	9	39	0.12	0.36	0.47 (0.27–0.81)	0.0067
	European	Sweden	Malmo	6,960	5,480	2	3	0.014	0.027		
	European	Sweden	PIVUS/ULSAM	270	1,734	1	3	0.19	0.087		
	European	Denmark	Danish	3,889	7,869	0	9	0.0	0.057		
	European	Finland	Finnish	4,050	8,696	1	2	0.012	0.011		
	South Asian	Singapore	Singapore Indians	562	585	1	1	0.089	0.085		
	European	UK	UKT2D	321	319	0	1	0.0	0.16		
p.Lys34Serfs*50	European	Iceland	deCODE	2,953	67,919	2	248	0.034	0.18	0.17 (0.05–0.52)	0.0019
	European	Norway	HUNT2	1,645	4,069	0	3	0.0	0.037		
c.71+2T>A	African American	United States	WFS	501	527	1	0	0.1	0.0	0.30 (0.14–0.64)	0.0021
	African American	United States	JHS	530	533	0	1	0.0	0.094		
p.Met50Ile	European	Germany	KORA	97	91	0	1	0.0	0.55		
c.271+G>A	East Asian	Korea	KARE	520	551	0	1	0.0	0.091		
	South Asian	Singapore	Singapore Indians	562	585	0	1	0.0	0.085		
c.419–1G>C	South Asian	UK	LOLIPOP	530	537	1	0	0.094	0.0		
p.Trp152*	European	Finland	Botnia	134	180	0	1	0.0	0.28		
p.Gln174*	South Asian	UK	LOLIPOP	530	537	1	5	0.094	0.47		
c.572+1G>A	African American	United States	JHS	530	533	0	1	0.0	0.094		
p.Tyr284*	South Asian	UK	LOLIPOP	530	537	0	2	0.0	0.19		
	South Asian	Singapore	Singapore Indians	562	585	0	1	0.0	0.085		
p.Ile291Phefs*2	African American	United States	JHS	530	533	0	1	0.0	0.094		
p.Ser327Thrs*55	African American	United States	WFS	501	527	0	2	0.0	0.19		
Combined	–	–	–	30,433	118,701	19	326	–	–	0.34 (0.21–0.53)	1.7×10^{-6}

Through sequencing and genotyping of ~150,000 individuals across 5 ancestry groups, a spectrum of 12 rare predicted protein-truncating variants was identified in *SLC30A8*. Shown for each variant are ancestry group, cohort, number of genotyped cases and controls (*N*), number of cases and controls observed to carry the variant, and observed allele frequencies in cases and controls. ORs and *P* values were computed separately for three groups of variants: p.Arg138*, p.Lys34Serfs*50 and the remaining variants. For p.Arg138* and p.Lys34Serfs*50, for which more than ten carriers were observed, statistics were computed separately for each cohort (Online Methods and **Supplementary Note**) and then combined via a fixed-effects meta-analysis. For the remaining variants, an association score was computed by comparing the aggregate frequencies of variant carriers in cases and controls. These three statistics were combined via a random-effects meta-analysis to produce combined estimates of risk and statistical significance (bottom row). Variant counts and frequencies were computed on the basis of all studied individuals, whereas ORs and *P* values were computed with correction for sample structure (population stratification and genetic relatedness; **Supplementary Note**); thus, displayed ORs differ from those computed solely from frequency estimates. CI, confidence interval.

UK10K sequencing study



UK10K (2015) Nature 526:82

UK10K association results

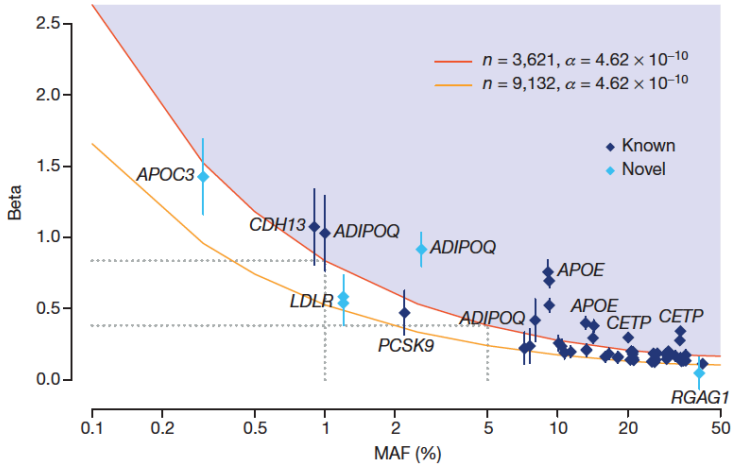
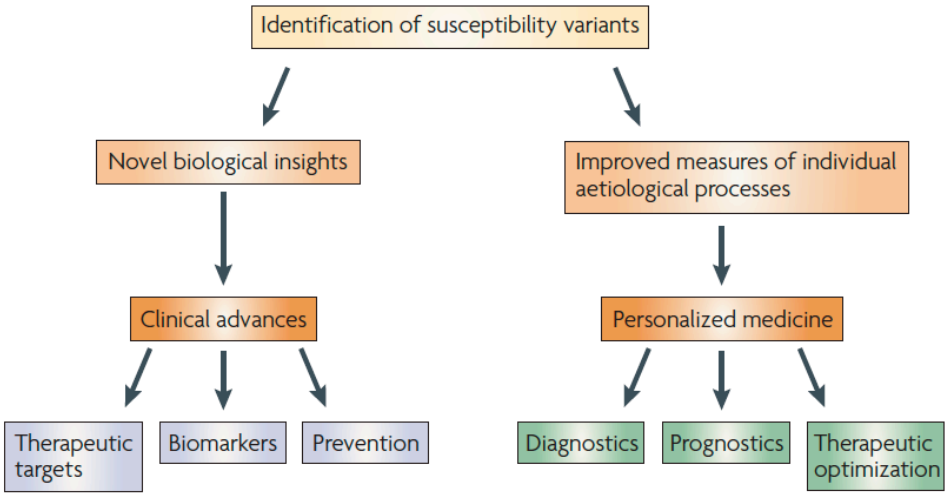


Figure 3 | Summary of association results across the UK10K-cohorts study.

UK10K (2015) Nature 526:82

Clinical translation



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