Genome-wide association studies (GWAS)

Matti Pirinen
FIMM, University of Helsinki

Manuel A. Rivas
University of Oxford and
The Broad Institute

15 January 2015
NIASC event, Helsinki

Part I: Common variants

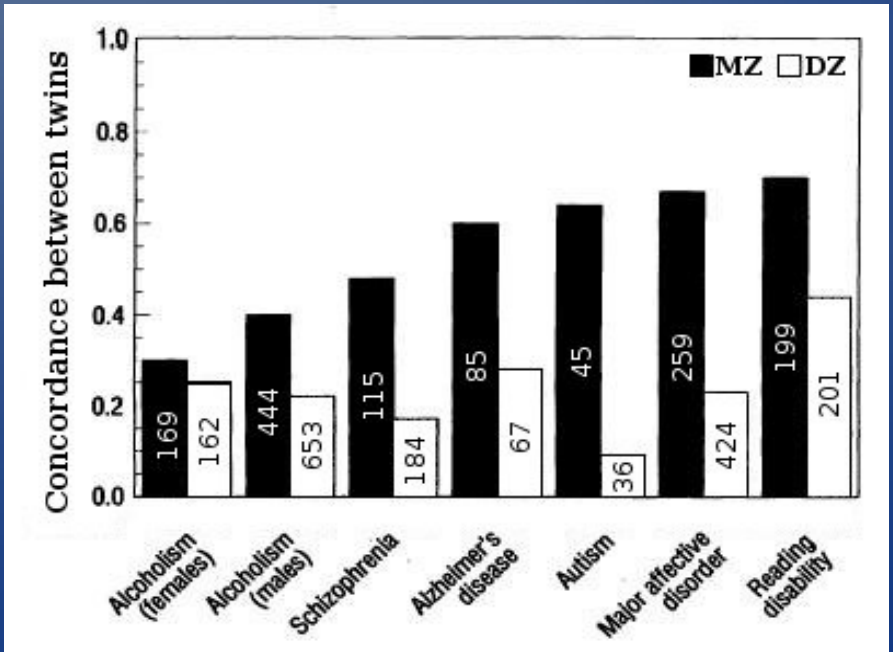
- 1. Motivation
  - What is a “genetic association” ?
  - Why is this important ?
- 2. How to read the genome (for SNPs) ?
- 3. An Example GWAS of Multiple Sclerosis
- 4. Interpreting a GWAS
- 5. Current state of GWAS
  - Criticism

Part I: Common variants

- 1. Motivation
  - Genetics contribute to traits
  - What is a “genetic association” ?
  - Why is this important ?

Does genetics affect trait? Is the trait *heritable*?

- Do more close relatives have more similar phenotypes (on average)?
  - Twin studies compare monozygotic twins with dizygotic twins (but environment may confound)



Human Genome

... G C G T T T A C G ... DNA sequence

A Human genome is 3×10^9 letters from alphabet {A, C, G, T}

Single Nucleotide Polymorphism (SNP)

- On average, 1:300 positions has (common) variation in population, called “SNP”

Single Nucleotide Polymorphism (SNP)

- On average, 1:300 positions has (common) variation in population, called “SNP”

Genomes in population

... G C G T T ... 96%

... G C T T T ... 4%

Individuals in population

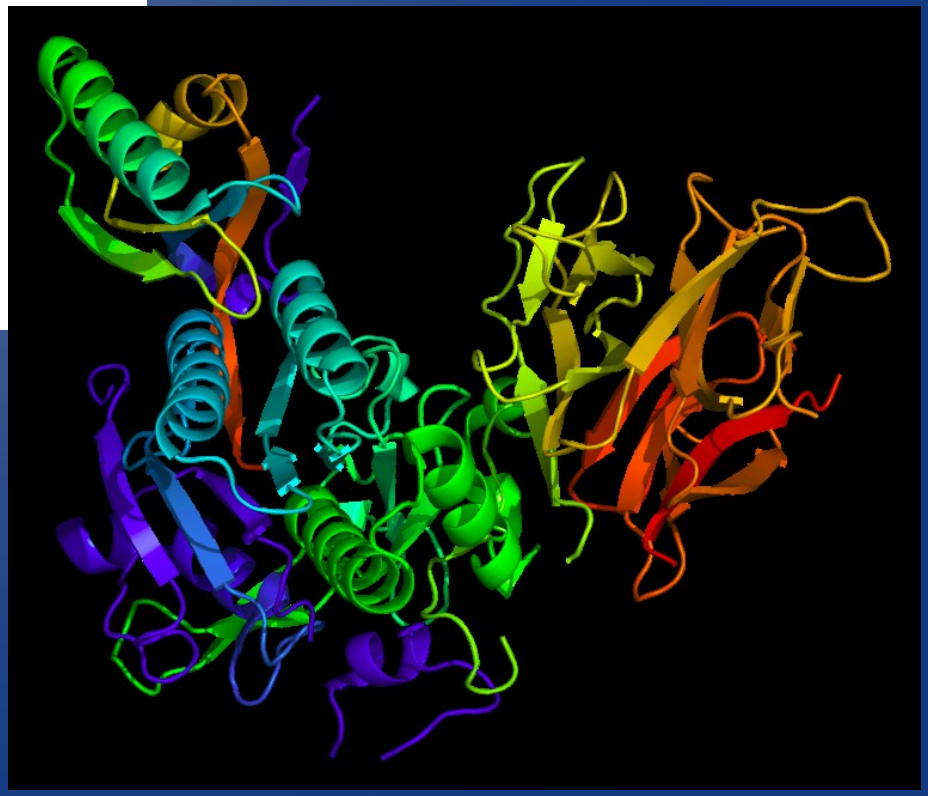
0: GG ~ 92.1%
1: GT ~ 7.7 %
2: TT ~ 0.2 %

SNP, alleles: G / T, minor allele frequency (MAF) = 4%

PCSK9 gene

- Location: Chromosome 1, 55.50 – 55.53 Mb

Codes protein



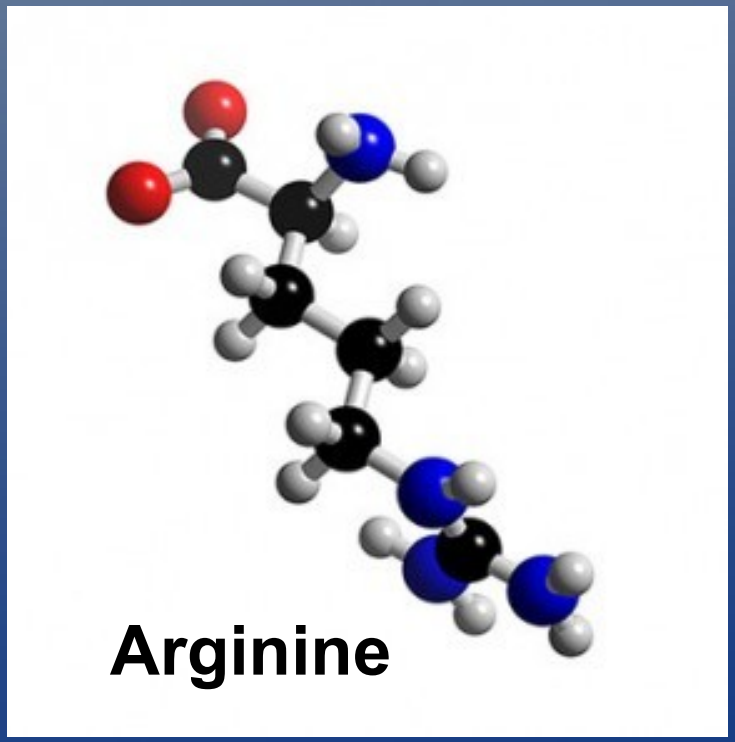
692 Amino acids

A SNP in PCSK9 gene

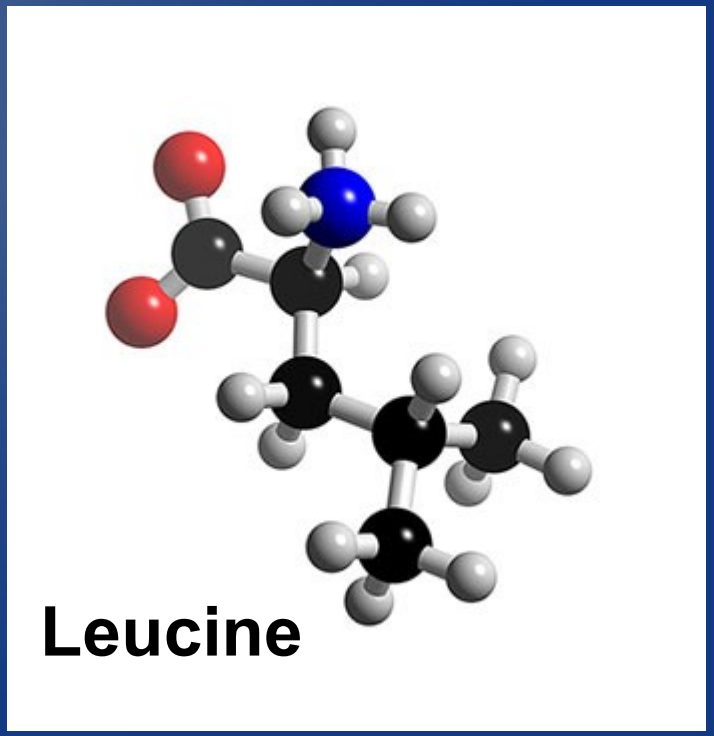
- Alleles: G / T , MAF=4% (Finland)
- Location: Chromosome 1, position 55,505,647

A SNP in PCSK9 gene

- Alleles: G / T , MAF=4% (Finland)
- Location: Chromosome 1, position 55,505,647
- Function: “missense”, “nonsynonymous”, changes 46th AA from Arginine to Leucine



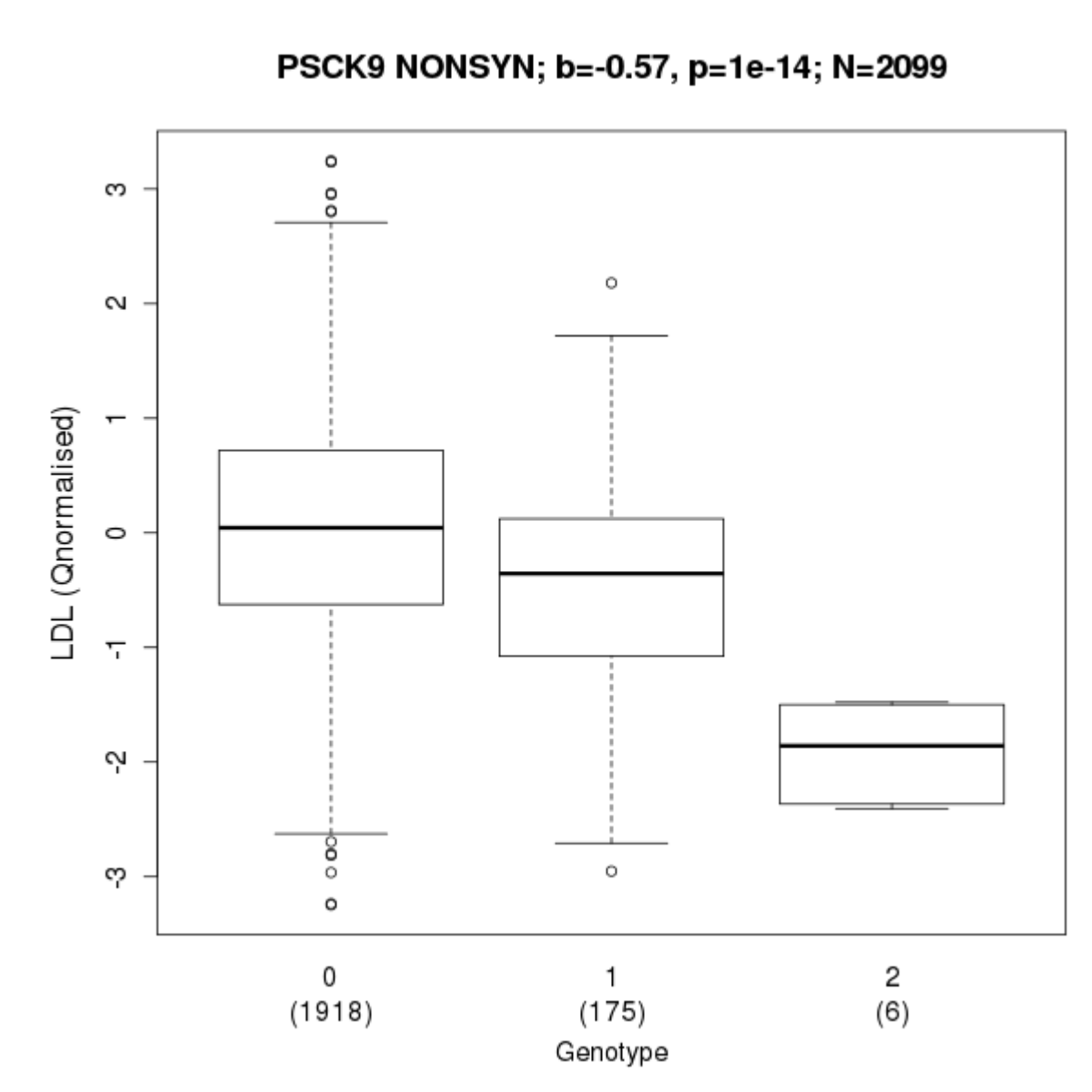
G T



A SNP in PCSK9 gene

- Alleles: G / T , MAF=4% (Finland)
- Location: Chromosome 1, position 55,505,647
- Function: “missense”, “nonsynonymous”, changes 46th AA from Arginine to Leucine
- Let's see if there is association with LDL cholesterol
  - LDL-C is a risk factor for heart disease

What is a “genetic association”?



GG **GT** **TT**

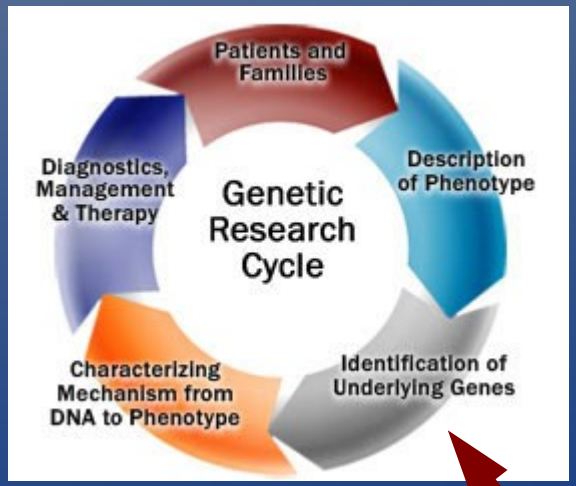
Finn-Metabo-Seq project: 2099 Finns exome seq'ed (Aug 2014)

Boxplot shows
(1) medians (thick lines),
(2) interquartile range (boxes)
(3) 1.5 x interquartile range (dotted segments)
(4) outliers (points)

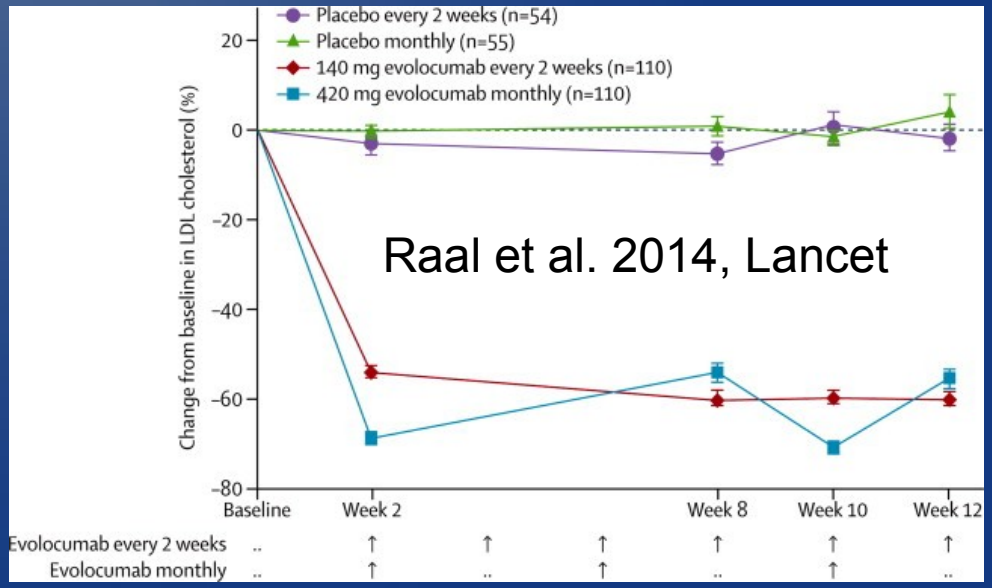
Carriers of allele T have lower LDL-C

Why are genetic associations important?

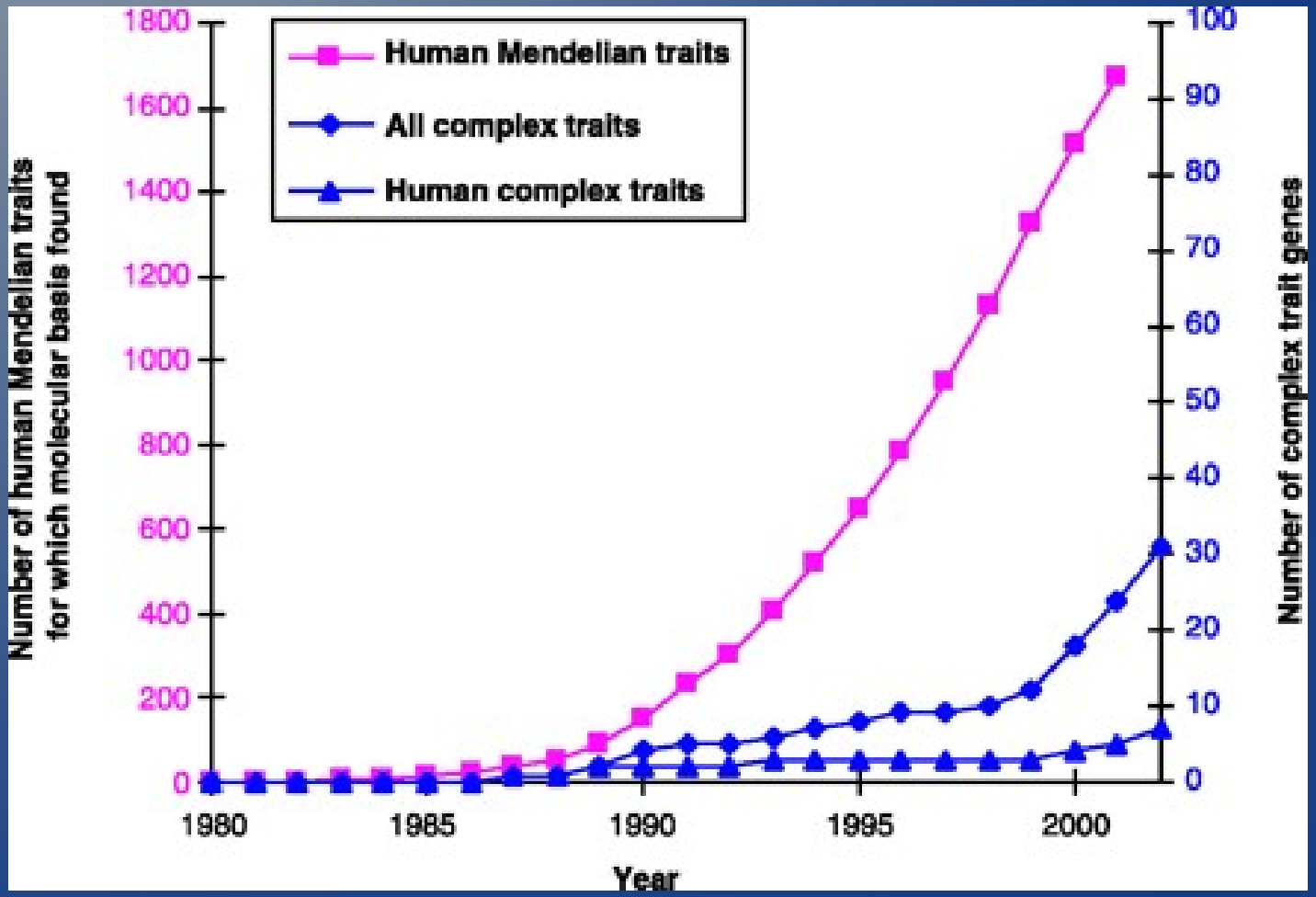
- Hints of biology behind the diseases and traits
  - Later: examples of Multiple sclerosis and schizophrenia
- Hints of targets for therapeutics
  - Inhibition of PCSK9 reduces LDL-C ?



www.chgr.org



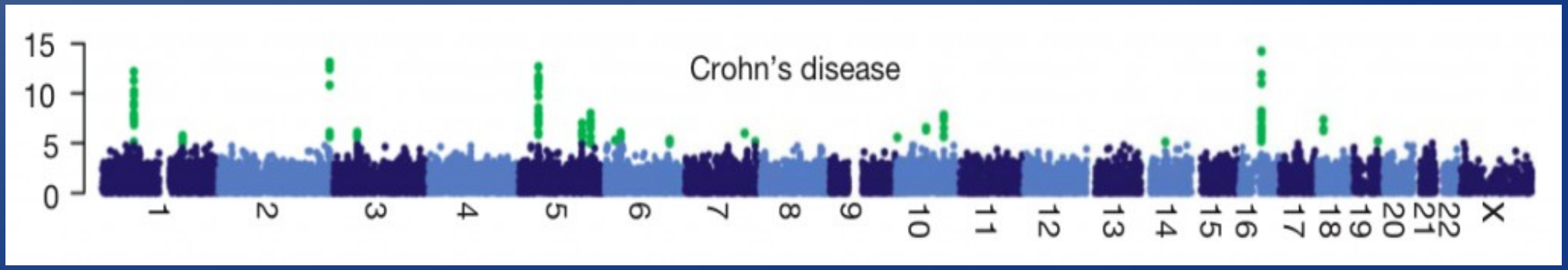
Large effect variants were found for many Mendelian traits but not for complex traits during 1985-2000



Glazier, Nadeau and Aitman, *Science* 2002

Genome-wide association studies (~2006 onwards)

- Idea: To look for associations in a detailed map of common variation across the genome
  - “Common disease – common variant”
- Became possible through
  - Technology (SNP arrays, later sequencing)
  - Collaboration (genetics + medicine + lab tech + bioinformatics + statistics)

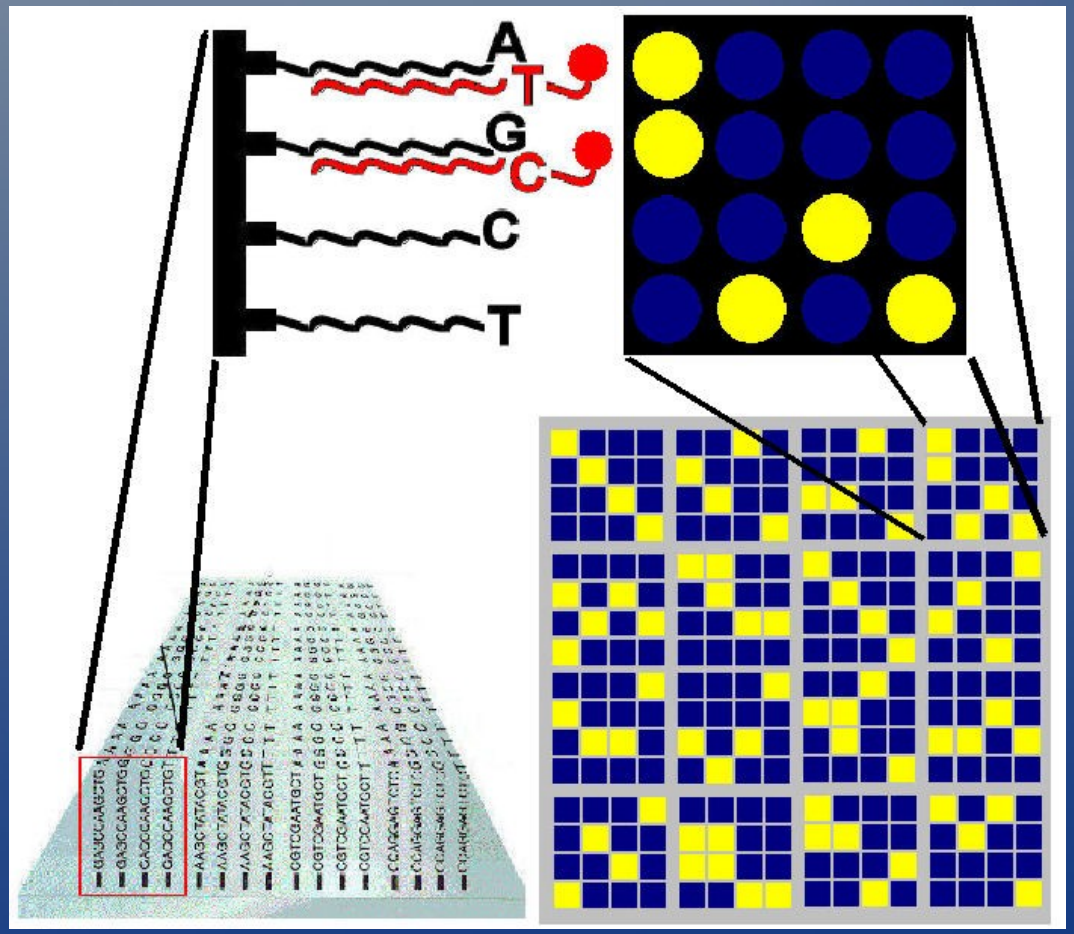


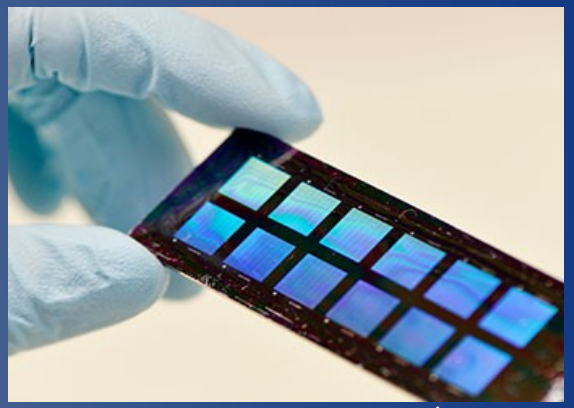
Part I: Common variants

- 1. Motivation
  - What is a “genetic association” ?
  - Why is this important ?
- 2. How to read the genome (for SNPs) ?

Human SNP arrays

- Contain probes for several million SNPs
- Price ~50-100 euros/sample

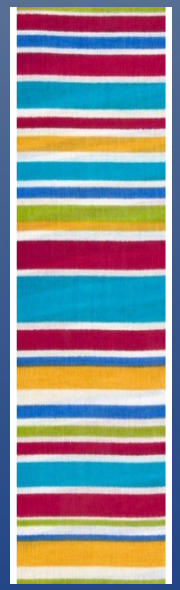




Steven M. Carr
www.mun.ca/biology/scarr/DNA_Chips.html

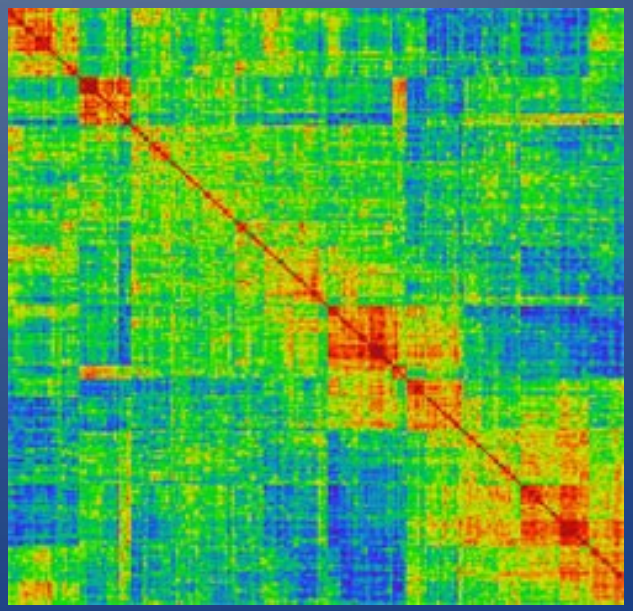
Does genetics affect trait? Is the trait *heritable*?

- With detailed genetic data, the question can be asked with “unrelated” individuals (program GCTA, Yang et al. 2010, Nat Gen)
  - Environment is not an (obvious) confounder



Y

~
?



G

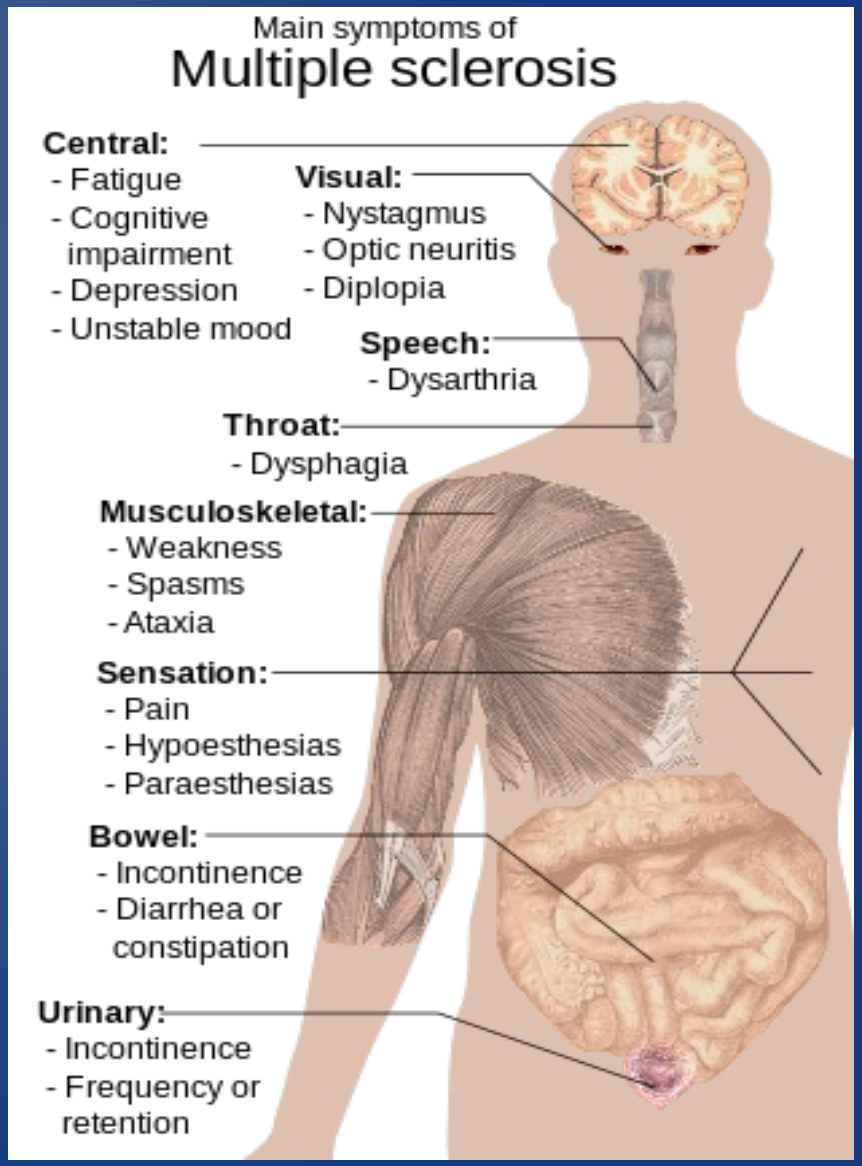
Example:
For height in 14,500 Finnish samples, genetic component of common SNPs explains 52% of variation.

Part I: Common variants

- 1. Motivation
  - What is a “genetic association” ?
  - Why is this important ?
- 2. How to read the genome (for SNPs) ?
- 3. An Example GWAS of Multiple Sclerosis

An example GWAS: Multiple Sclerosis by WTCCC2

- MS: nerve cells are damaged through inflammation reaction -> wide rang of severe symptoms
- A GWAS on 27,000 samples from 15 countries
  - 10,000 cases
  - 17,000 controls
- Nature, August 2011



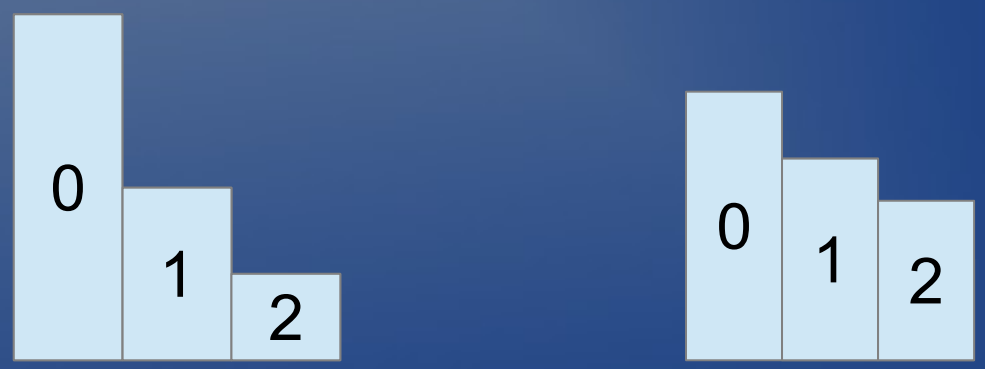
Case-control association study

Individuals

~10^4 Cases

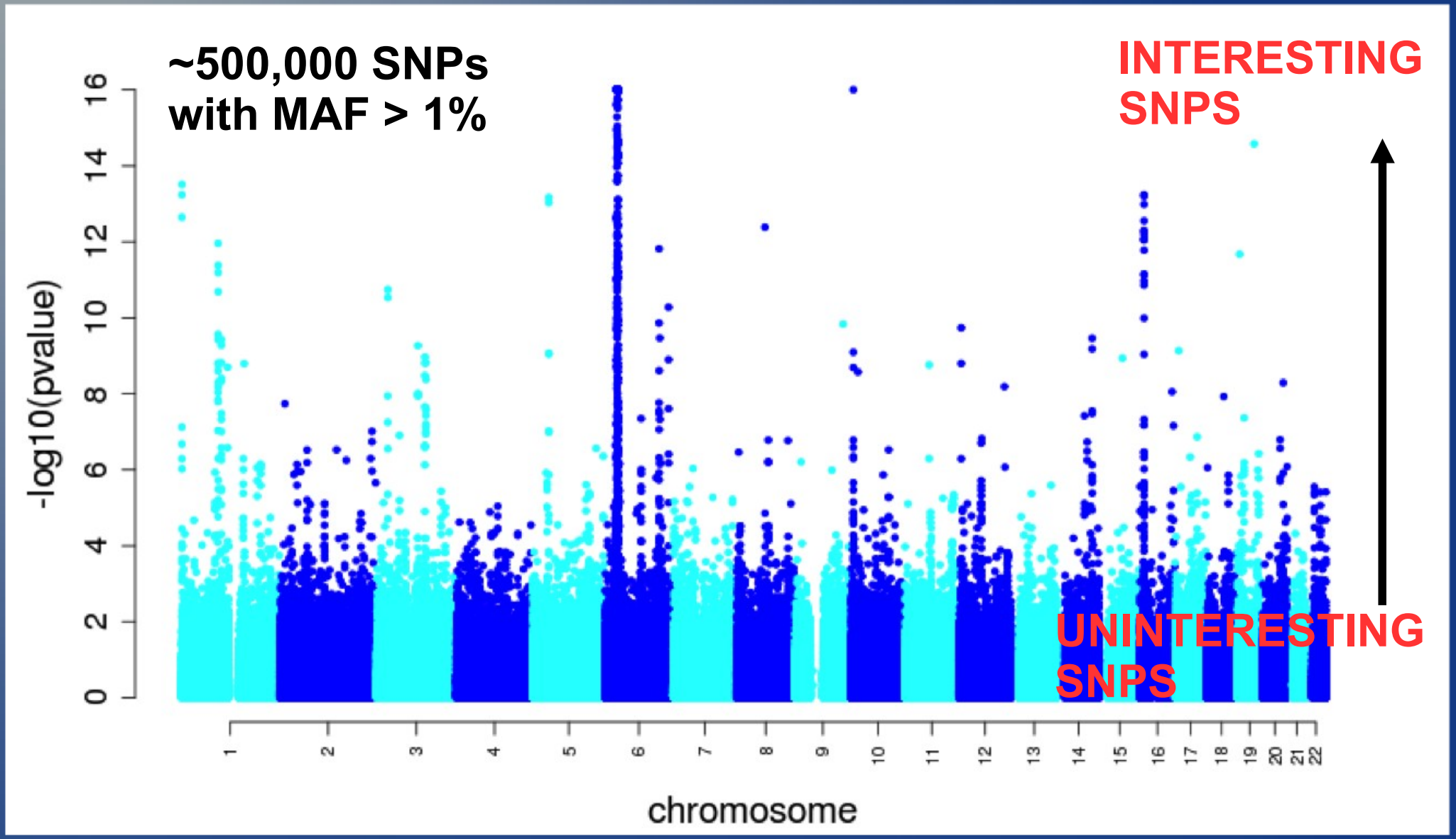
~10^4 Controls

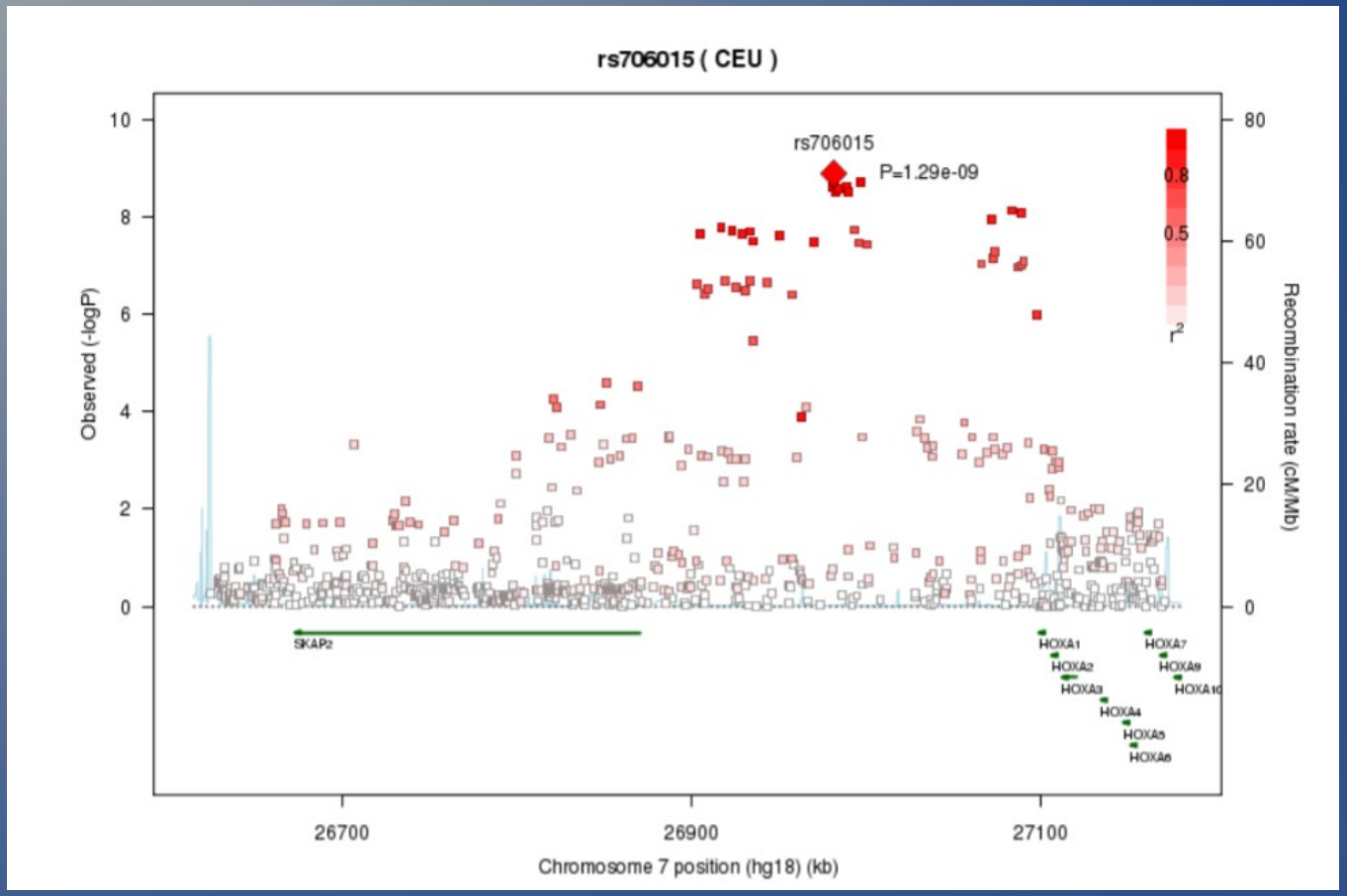
Genotypes ~10^6 Single nucleotide polymorphisms (SNPs)



Question Are the genotype distributions different between cases and controls?

A genome-wide scan



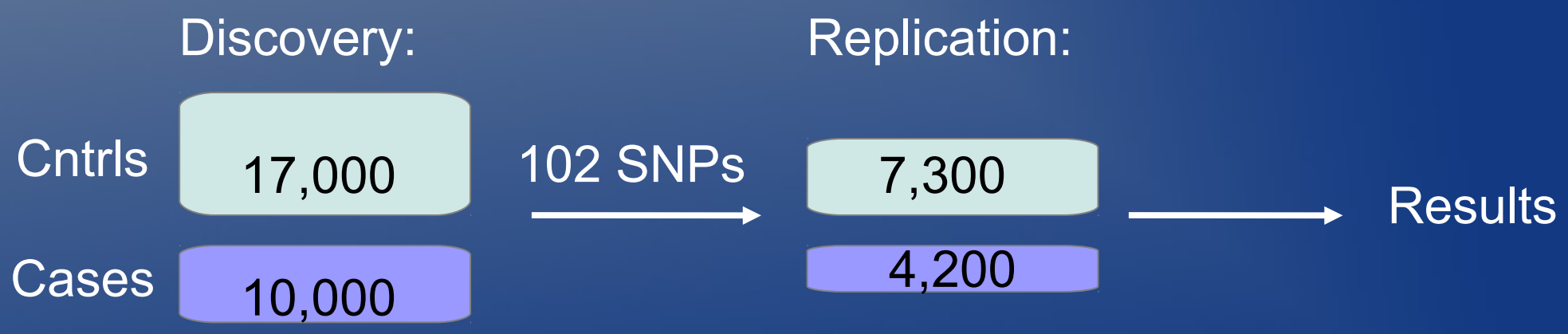


Each ”hit SNP” resides in a region with many correlated SNPs and possibly many (or none) genes

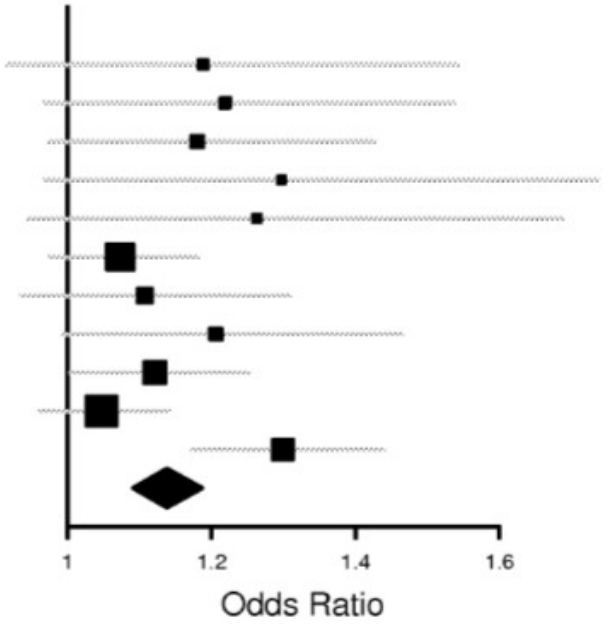
From Beecham 2013 Nat Genet

Replication

- To verify the most promising associations in an independent data set
- IN MS study, took 102 SNPs to replication



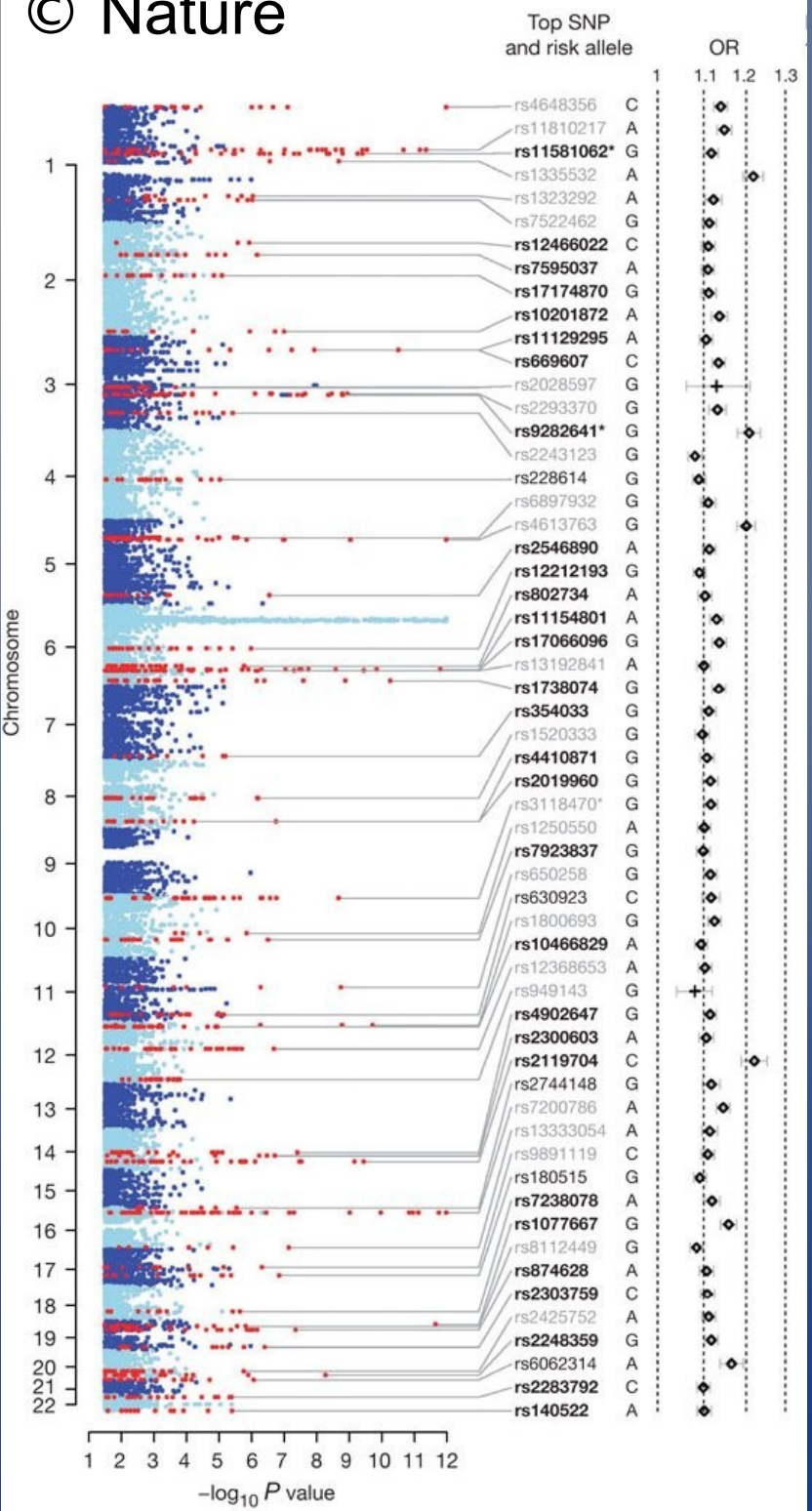
Stratum	OR	95% CI
AUSNZ	1.19	[0.92–1.54]
Belgium	1.22	[0.97–1.54]
Denmark	1.18	[0.98–1.43]
Finland	1.30	[0.97–1.74]
France	1.26	[0.94–1.69]
Germany	1.07	[0.98–1.18]
Italy	1.11	[0.94–1.31]
Norway	1.21	[0.99–1.46]
Sweden	1.12	[1.00–1.25]
UK	1.05	[0.96–1.14]
US	1.30	[1.17–1.44]
Summary	**1.14**	**[1.09–1.19]**



- Consistent effects across
  - populations
  - genotyping platforms

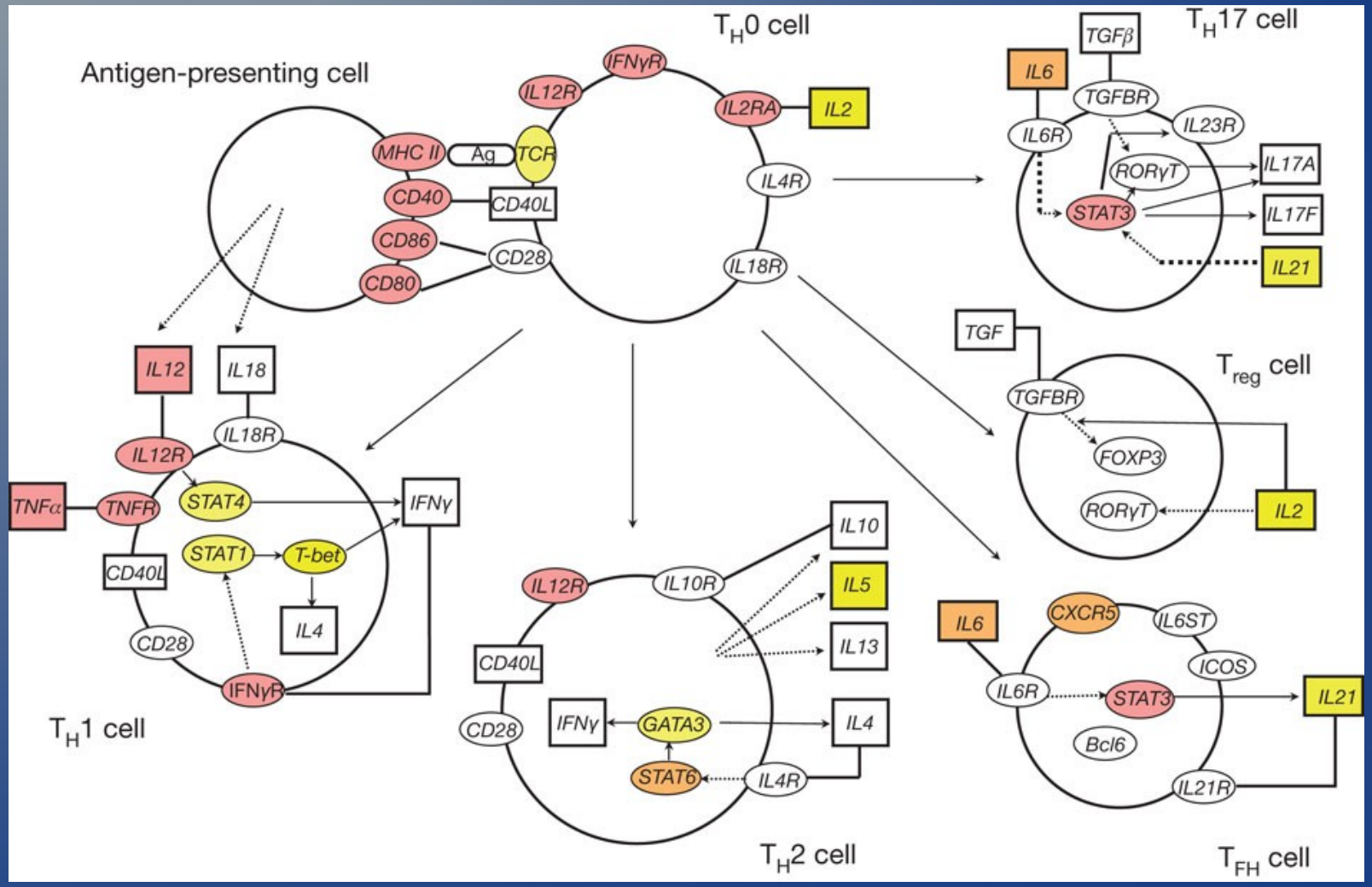
bring confidence

From Beecham 2013
Nat Genet



- 98 /102 SNPs have consistent effects in replication data
- Over 50 convincing associations with MS
- Immunological genes are over represented among the hits; in particular, T-helper cell differentiation pathway

T-helper cell pathway implicated by GWAS hits



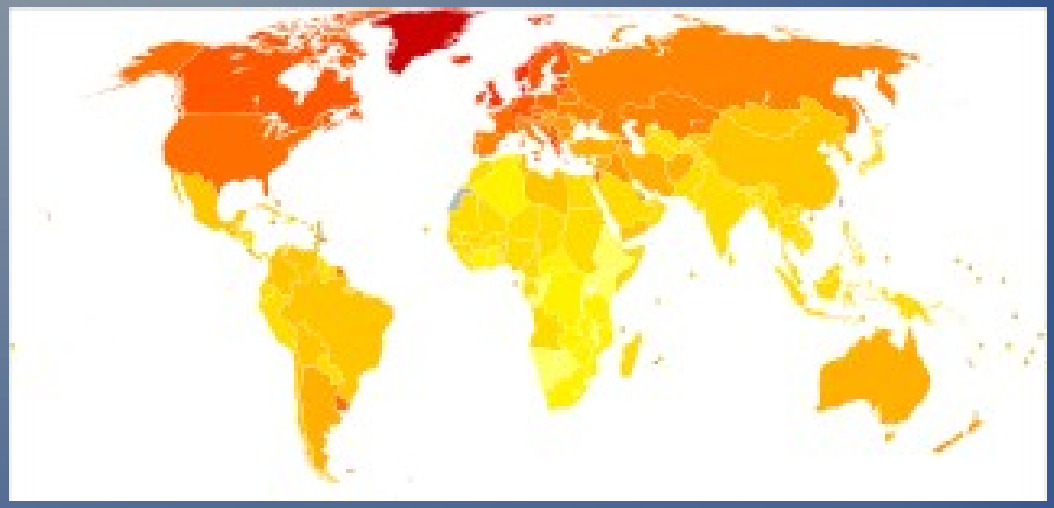
Red, strong signal ;

orange, medium signal

yellow, some signal

MS results

- Two implicated genes involved in vitamin D metabolism



Related to latitudinal variation in MS prevalence?

- Two other implicated genes are already targets of MS therapies
  - VCAM1 for 'natalizumab' and IL2RA for 'daclizumab'

Collaboration

- 245 authors with 140 affiliations around the world

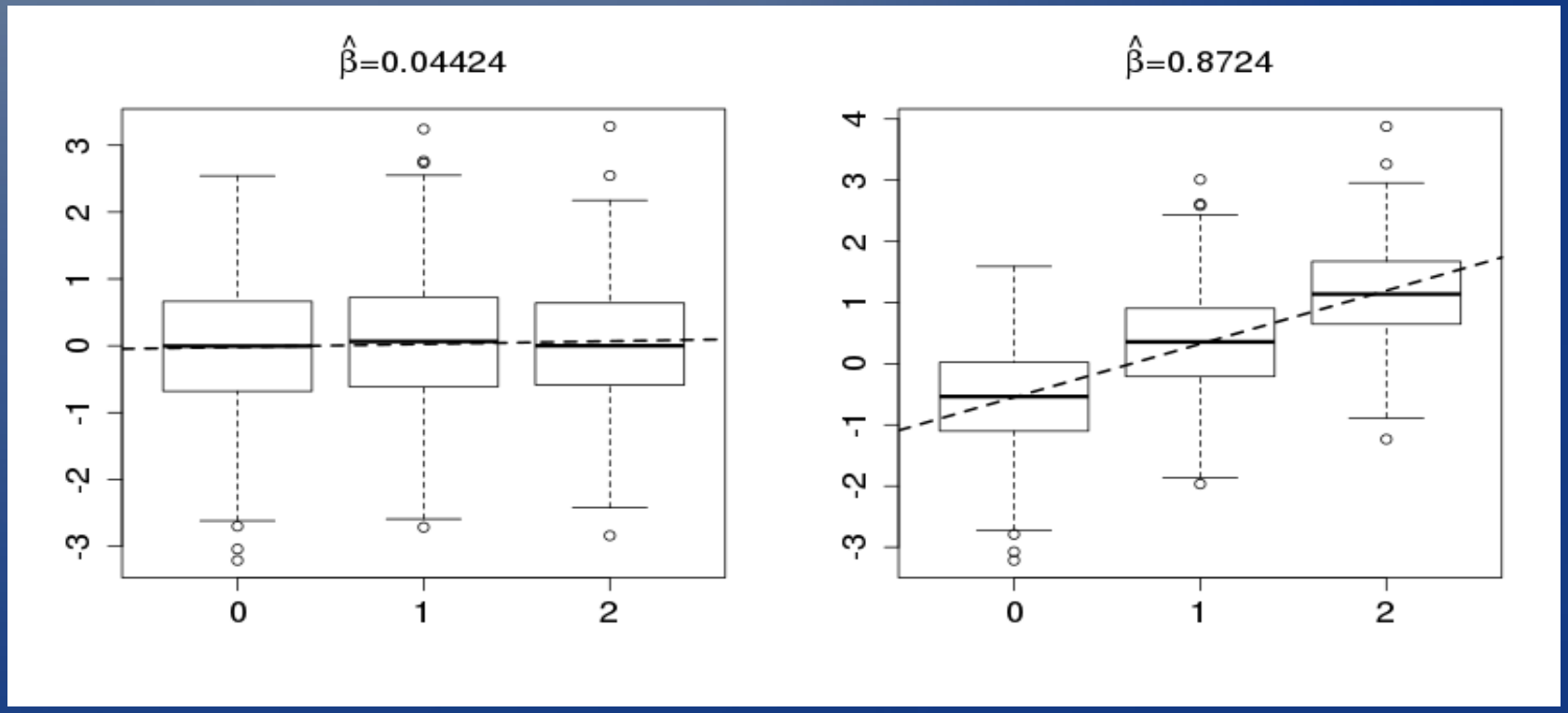
Stephen Sawcer[1*], Garrett Hellenthal[2*], Matti Pirinen[2*], Chris C. A. Spencer[2*], Nikolaos A. Patsopoulos[3,4,5], Loukas Moutsianas[6], Alexander Dilthey[6], Zhan Su[2], Colin Freeman[2], Sarah E. Hunt[7], Sarah Edkins[7], Emma Gray[7], David R. Booth[8], Simon C. Potter[7], An Goris[9], Gavin Band[2], Annette Bang Oturai[10], Amy Strange[2], Janna Saarela[11], Céline Bellenguez[2], Bertrand Fontaine[12], Matthew Gillman[7], Bernhard Hemmer[13], Rhian Gwilliam[7], Frauke Zipp[14,15], Alagurevathi Jayakumar[7], Roland Martin[16], Stephen Leslie[17], Stanley Hawkins[18], Eleni Giannoulatou[2], Sandra D'alfonso[19], Hannah Blackburn[7], Filippo Martinelli Boneschi[20], Jennifer Liddle[7], Hanne F. Harbo[21,22], Marc L. Perez[7], Anne Spurkland[23], Matthew J. Waller[7], Marcin P. Mycko[24], Michelle Ricketts[7], Manuel Comabella[25], Naomi Hammond[7], Ingrid Kockum[26], Owen T. McCann[7], Maria Ban[1], Pamela Whittaker[7], Anu Kemppinen[1], Paul Weston[7], Clive Hawkins[27], Sara Widaa[7], John Zajicek[28], Serge Dronov[7], Neil Robertson[29], Suzannah J. Bumpstead[7], Lisa F. Barcellos[30,31], Rathi Ravindrarajah[7], Roby Abraham[27], Lars Alfredsson[32], Kristin Ardlie[4], Cristin Aubin[4], Amie Baker[1], Katharine Baker[29], Sergio E. Baranzini[33], Laura Bergamaschi[19], Roberto Bergamaschi[34], Allan Bernstein[31], Achim Berthele[13], Mike Boggild[35], Jonathan P. Bradfield[36], David Brassat[37], Simon A. Broadley[38], Dorothea Buck[13], Helmut Butzkueven[39,40,41,42], Ruggero Capra[43], William M. Carroll[44], Paola Cavalla[45], Elisabeth G. Celius[21], Sabine Cepok[13], Rosetta Chiavacci[36], Françoise Clerget-Darpoux[46], Katleen Clysters[9], Giancarlo Comi[20], Mark Cossburn[29], Isabelle Cournu-Rebeix[12], Mathew B. Cox[47], Wendy Cozen[48], Bruce A. C. Cree[33], Anne H. Cross[49], Daniele Cusi[50], Mark J. Daly[4,51,52], Emma Davis[53], Paul I.W. de Bakker[3,4,54,55], Marc Debouverie[56], Marie Beatrice D'hooghe[57], Katherine Dixon[53], Rita Dobosi[9], Bénédicte Dubois[9], David Ellinghaus[58], Irina Elovaara[59,60], Federica Esposito[20], Claire Fontenille[12], Simon Foote[61], Andre Franke[58], Daniela Galimberti[62], Angelo Ghezzi[63], Joseph Glessner[36], Refujia Gomez[33], Olivier Gout[64], Colin Graham[65], Struan F. A. Grant[36,66,67], Franca Rosa Guerini[68], Hakon Hakonarson[36,66,67], Per Hall[69], Anders Hamsten[70], Hans-Peter Hartung[71], Rob N. Heard[8], Simon Heath[72], Jeremy Hobart[28], Muna Hoshi[13], Carmen Infante-Duarte[73], Gillian Ingram[29], Wendy Ingram[28], Talat Islam[48], Maja Jagodic[26], Michael Kabesch[74], Allan G. Kermode[44], Trevor J. Kilpatrick[39,40,75], Cecilia Kim[36], Norman Klopp[76], Keijo Koivisto[77], Malin Larsson[70], Mark Lathrop[72], Jeannette S. Lechner-Scott[47,78], Maurizio A. Leone[79], Virpi Leppä[11,80], Ulrika Liljedahl[81], Izaura Lima Bomfim[26], Robin R. Lincoln[33], Jenny Link[26], Jianjun Liu[82], Aslaug R. Lorentzen[22,83], Sara Lupoli[50,84], Fabio Macciardi[50,85], Thomas Mack[48], Mark Marriott[39,40], Vittorio Martinelli[20], Deborah Mason[86], Jacob L. McCauley[87], Frank Mentch[36], Inger-Lise Mero[21,83], Tania Mihalova[27], Xavier Montalban[25], John Mottershead[88,89], Kjell-Morten Myhr[90,91], Paola Naldi[79], William Ollier[53], Alison Page[92], Aarno Palotie[7,11,93,94], Jean Pelletier[95], Laura Piccio[49], Trevor Pickersgill[29], Fredrik Piehl[26], Susan Pobywajlo[5], Hong L. Quach[30], Patricia P. Ramsay[30], Mauri Reunanen[96], Richard Reynolds[97], John D. Rioux[98], Mariaemma Rodegher[20], Sabine Roesner[16], Justin P. Rubio[39], Ina-Maria Rückert[76], Marco Salvetti[99], Erika Salvi[50,100], Adam Santaniello[33], Catherine A. Schaefer[31], Stefan Schreiber[58,101], Christian Schulze[102], Rodney J. Scott[47], Finn Sellebjerg[10], Krzysztof W. Selmaj[24], David Sexton[103], Ling Shen[31], Brigid Simms-Acuna[31], Sheila Skidmore[1], Patrick M. A. Sleiman[36,66], Cathrine Smestad[21], Per Soelberg Sørensen[10], Helle Bach Søndergaard[10], Jim Stankovich[61], Richard C. Strange[27], Anna-Maija Sulonen[11,80], Emilie Sundqvist[26], Ann-Christine Syvänen[81], Francesca Taddeo[100], Bruce Taylor[61], Jenefer M. Blackwell[104,105], Pentti Tienari[106], Elvira Bramon[107], Ayman Tourbah[108], Matthew A. Brown[109], Ewa Tronczynska[24], Juan P. Casas[110], Niall Tubridy[40,111], Aiden Corvin[112], Jane Vickery[28], Janusz Jankowski[113], Pablo Villoslada[114], Hugh S. Markus[115], Kai Wang[36,66], Christopher G. Mathew[116], James Wason[117], Colin N. A. Palmer[118], H-Erich Wichmann[76,119,120], Robert Plomin[121], Ernest Willoughby[122], Anna Rautanen[2], Juliane Winkelmann[13,123,124], Michael Wittig[58,125], Richard C. Trembath[116], Jacqueline Yaouanq[126], Ananth C. Viswanathan[127], Haitao Zhang[36,66], Nicholas W. Wood[128], Rebecca Zuvich[103], Panos Deloukas[7], Cordelia Langford[7], Audrey Duncanson[129], Jorge R. Oksenberg[33], Margaret A. Pericak-Vance[87], Jonathan L. Haines[103], Tomas Olsson[26], Jan Hillert[26], Adrian J. Ivinson[51,130], Philip L. De Jager[4,5,51], Leena Peltonen†, Graeme J. Stewart[8], David A. Hafler[4,131], Stephen L. Hauser[33], Gil McVean[2], Peter Donnelly[2,6*] & Alastair Compston[1*]

Part I: Common variants

- 1. Motivation
  - What is a “genetic association” ?
  - Why is this important ?
- 2. How to read the genome (for SNPs) ?
- 3. An Example GWAS of Multiple Sclerosis
- 4. Interpreting a GWAS

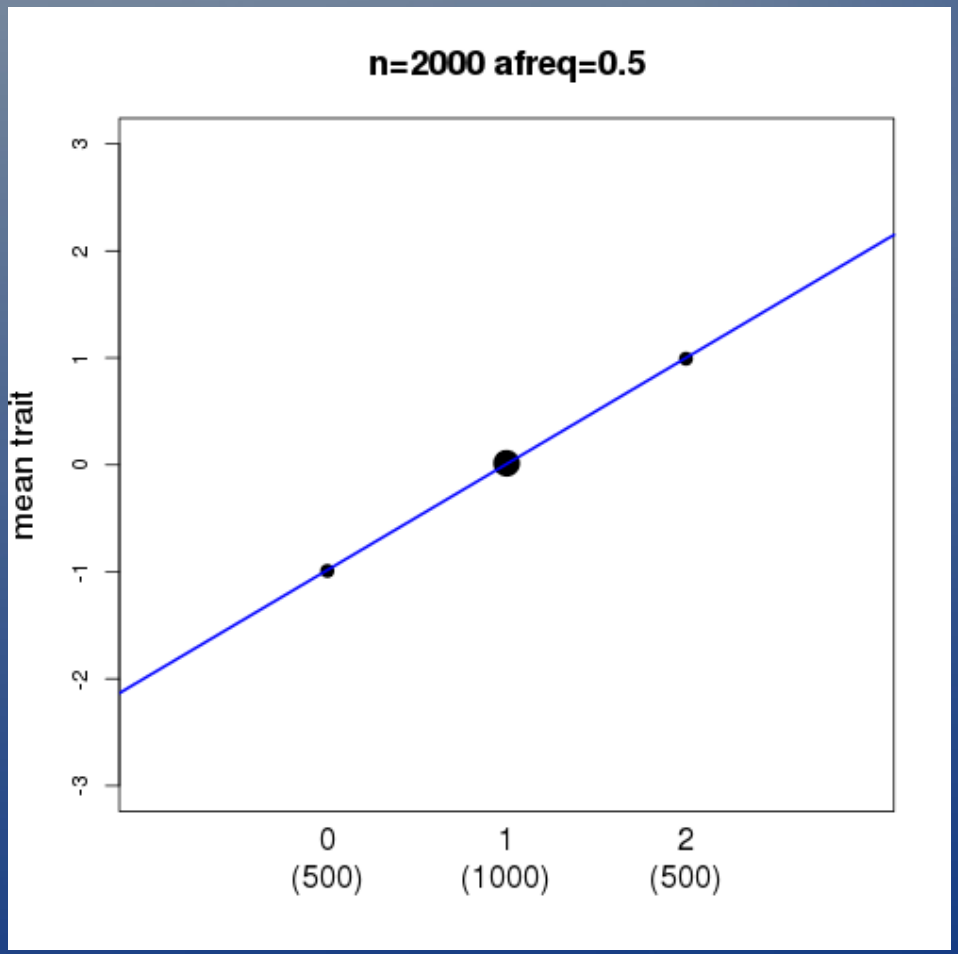
Linear model to measure association

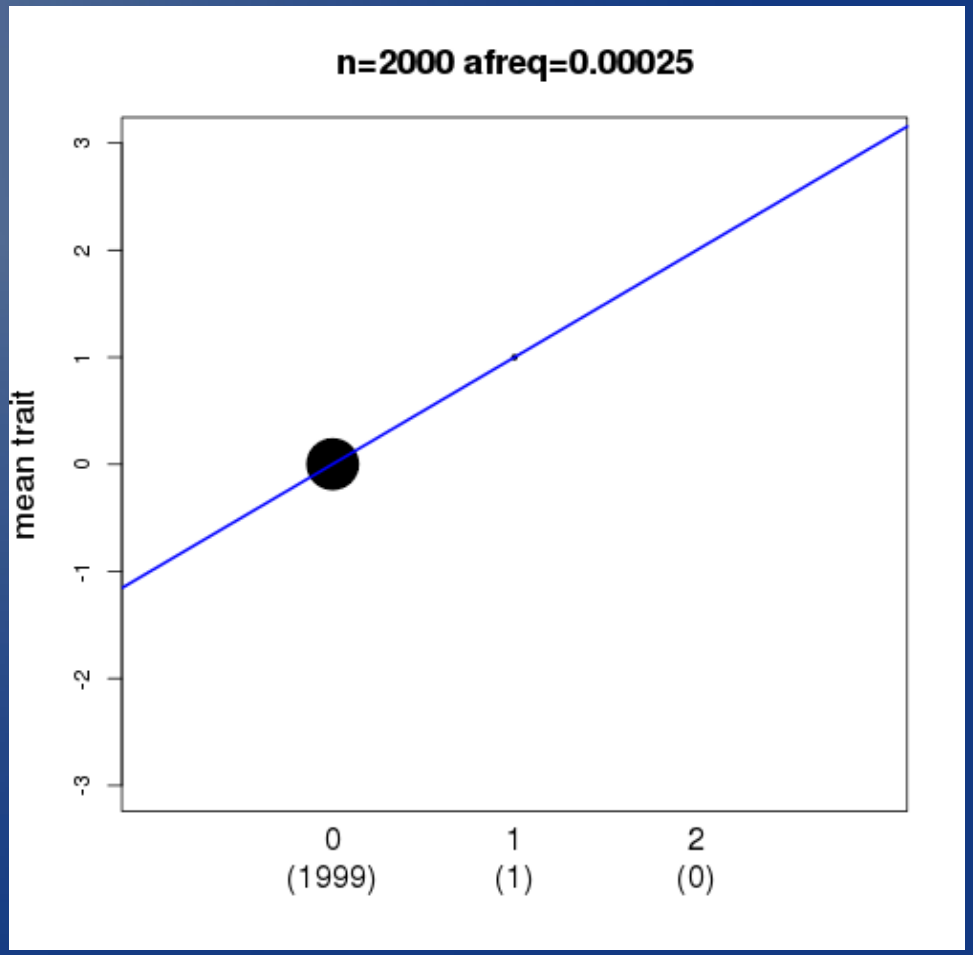
- Fit a line through 3 genotypes
  - Large slope = strong association (any prob's?)
  - Why is Manhattan plot not about |slope| but instead about “p-value”?



Why is slope not everything?

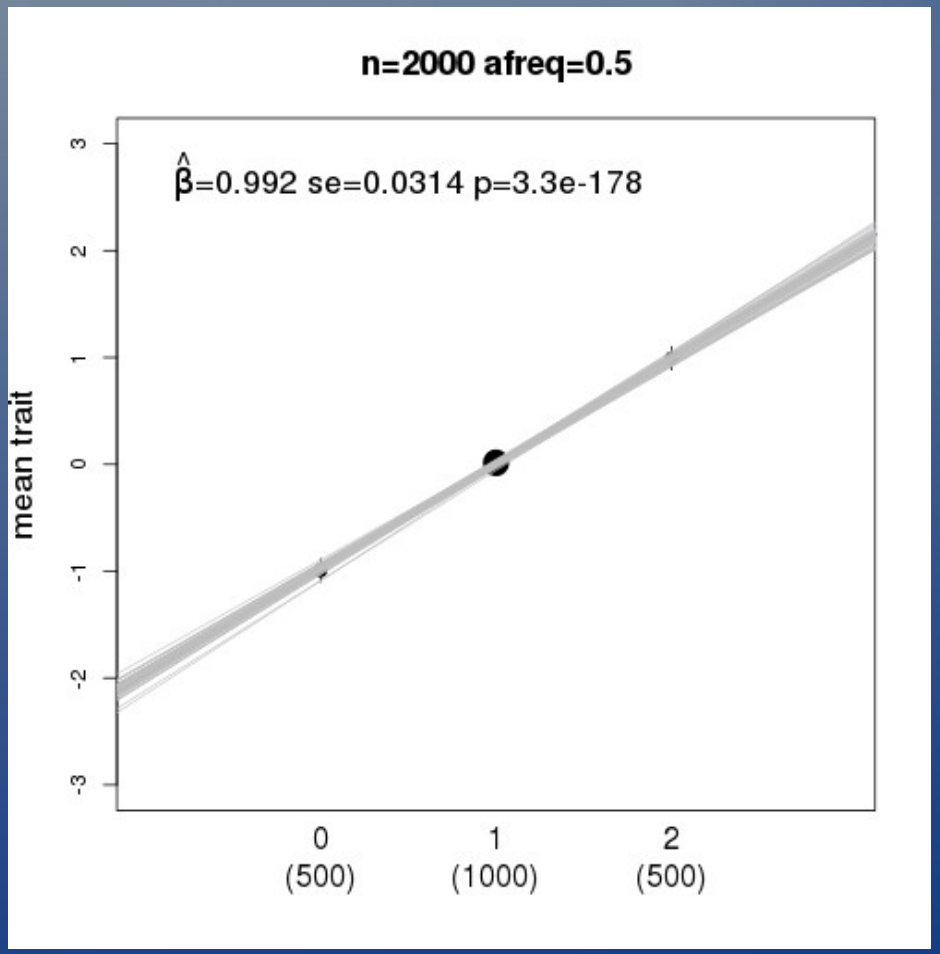
- Two SNPs with a slope of ~1.0 for n=2000

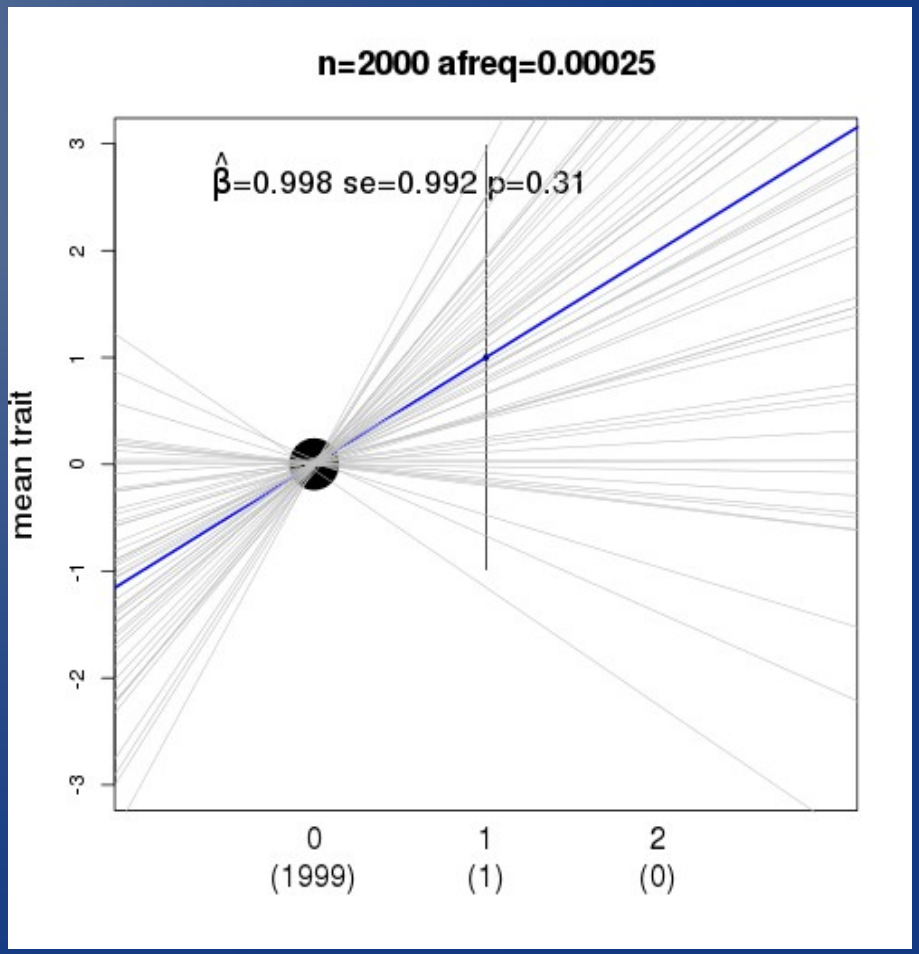




Why is slope not everything?

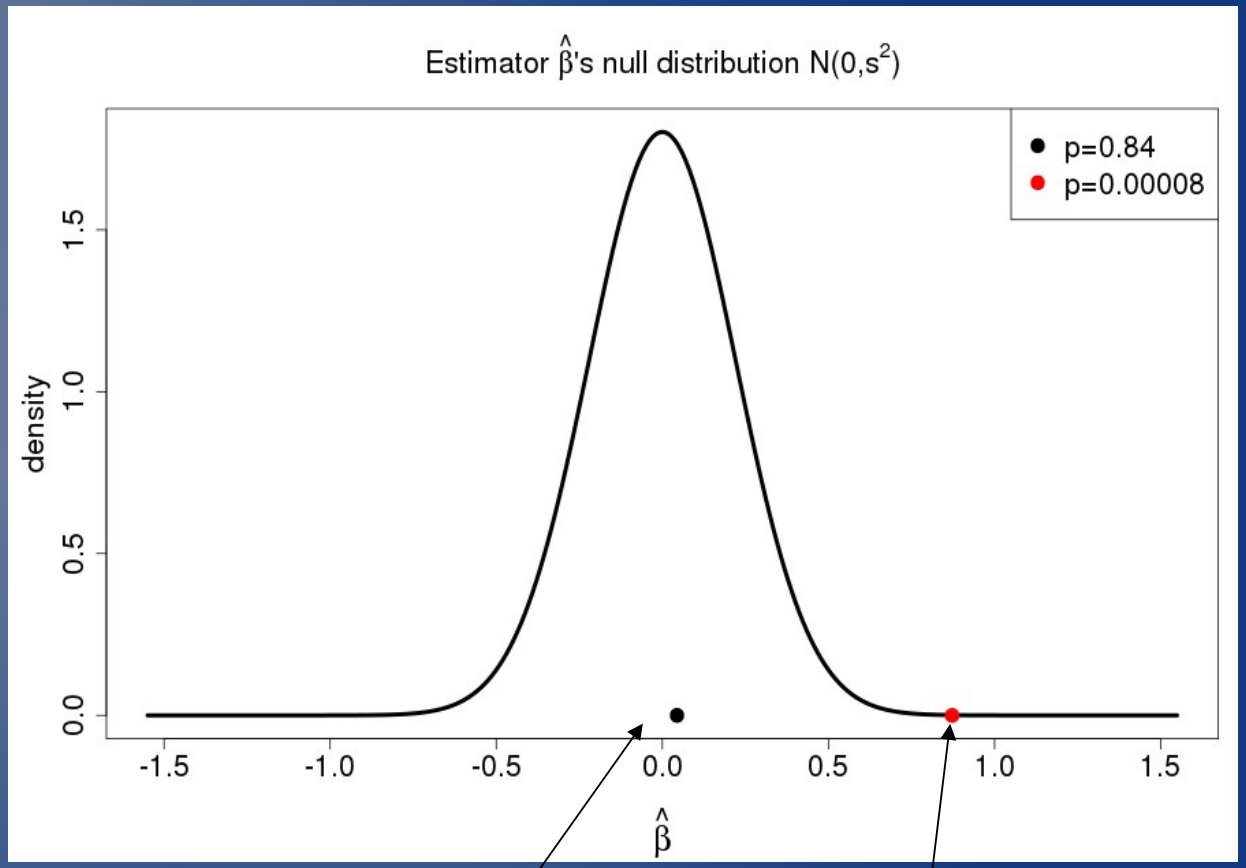
- Uncertainty about the slope
- Left 1.0 (0.97 ... 1.03); Right 1.0 (-1.0 ... 3.0)

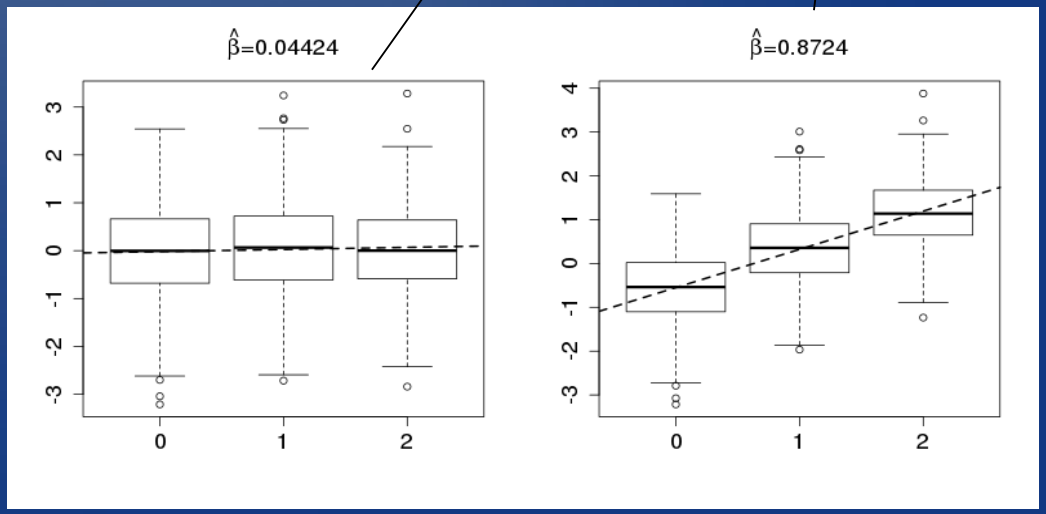




P-value

- Is the observed slope plausible if true slope=0?
- P-value: Probability that “by chance” we get as extreme value as observed
- P=0.84: No evidence for deviation from null
- P=8e-5: Unlikely under the null -> maybe not null





Crude inference procedure based on “statistical significance”

P-value cutoff

	NULL SNPs	TRUE EFFECTS
NON-SIGNIFICANT	MANY	?
SIGNIFICANT	(almost) none	?

- Label SNPs as “significant” if their p-value is small enough
- Use a stringent p-value threshold in order that there is (almost) no false positives
- Hope that there will be some true positives

Genome-wide significance threshold

- There are ~10^6 independent regions in the genome
  - Genome has block structure due to recombination process (linkage disequilibrium)
- If we use threshold $p = 0.05/10^6 = \underline{5 \times 10^{-8}}$ then, on average, 1 out of 20 GWAS reports a false positive association
- Very small p-value protects from false positives

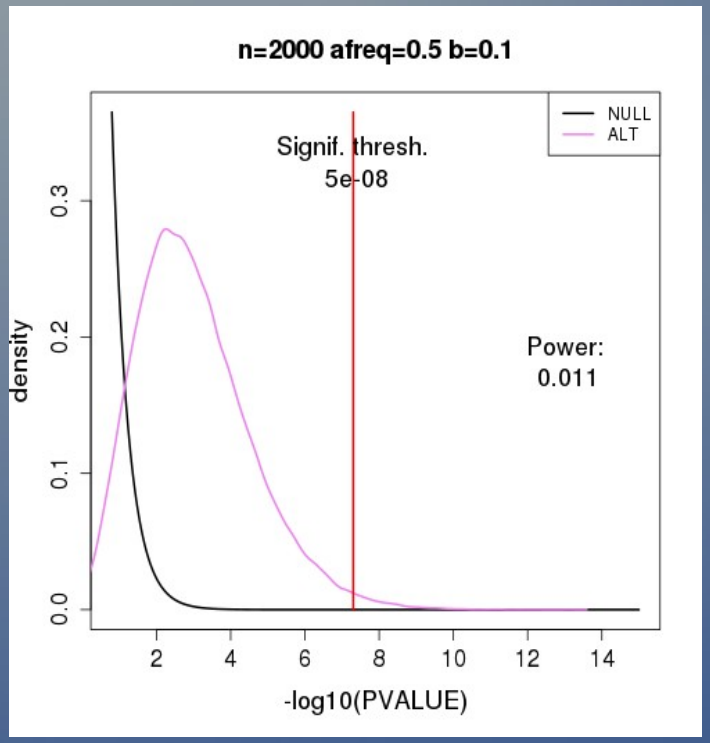
Genome-wide significance threshold

- What if I have data only on one SNP
  - Can I now use p-value threshold 0.05?
  - What if my SNP is known to truncate a protein? Is the same p-value threshold used as for a random SNP?

Genome-wide significance threshold

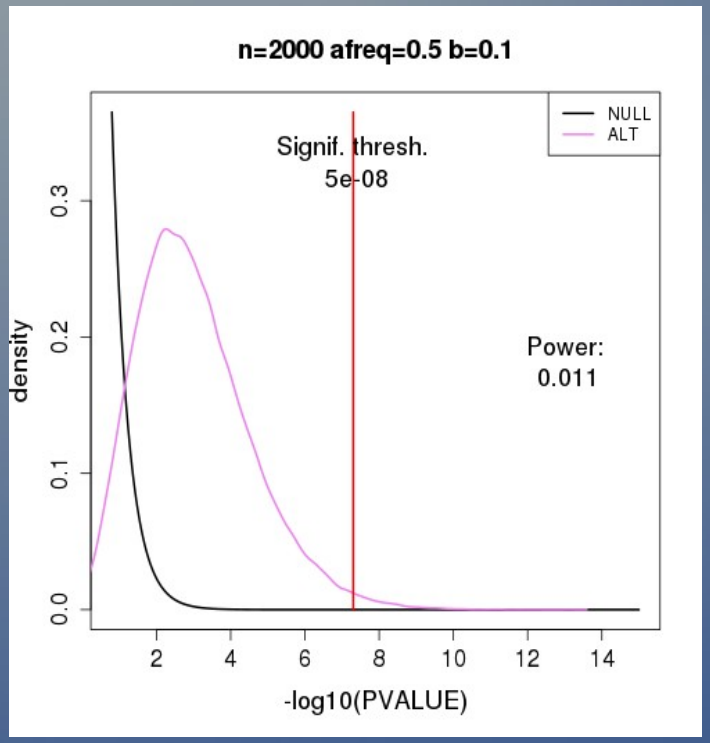
- What if I have data only on one SNP
  - Can I now use p-value threshold 0.05?
  - What if my SNP is known to truncate a protein? Is the same p-value threshold used as for a random SNP?
- Number of tests is NOT the general rule for defining consistent thresholds
  - We look at this more soon ... after statistical power

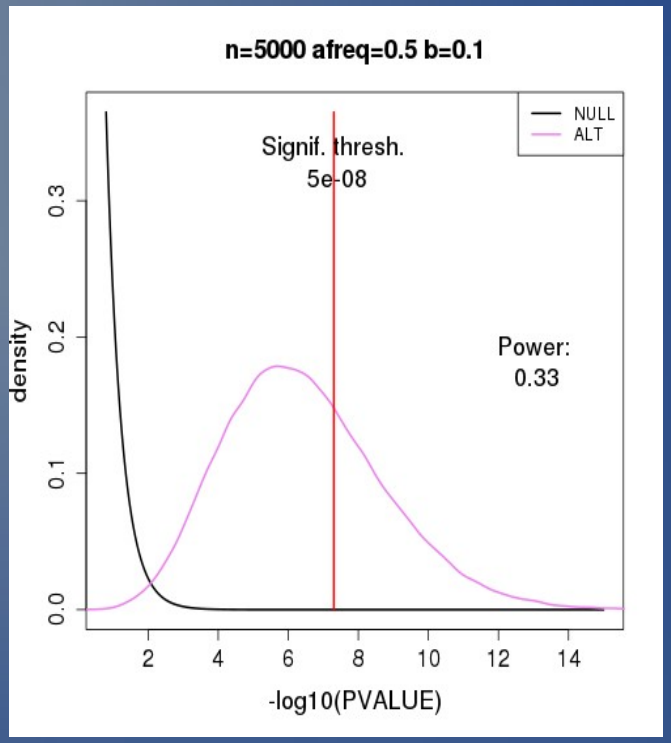
Power



- Power = Probability that a SNP will reach a given significance threshold
  - Depends on sample size, allele freq. and effect size

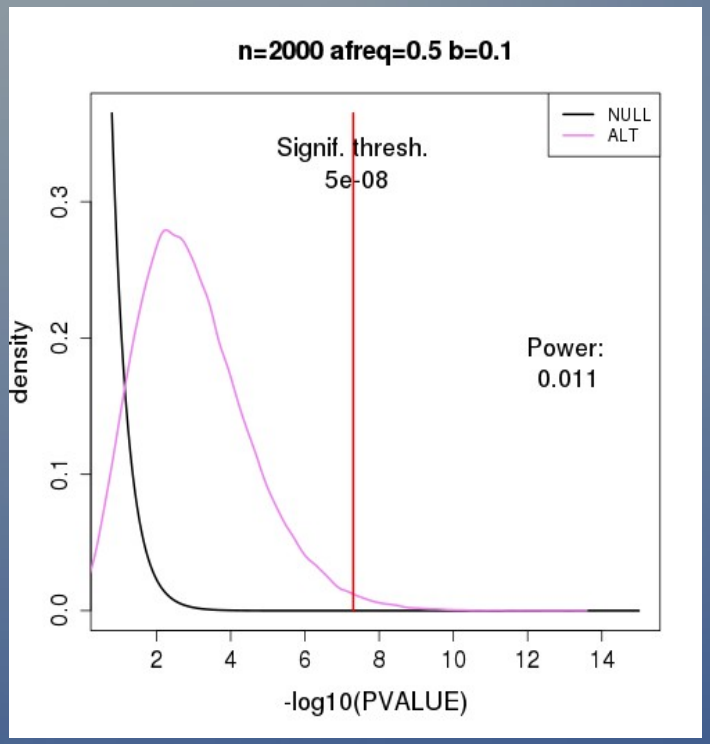
Power

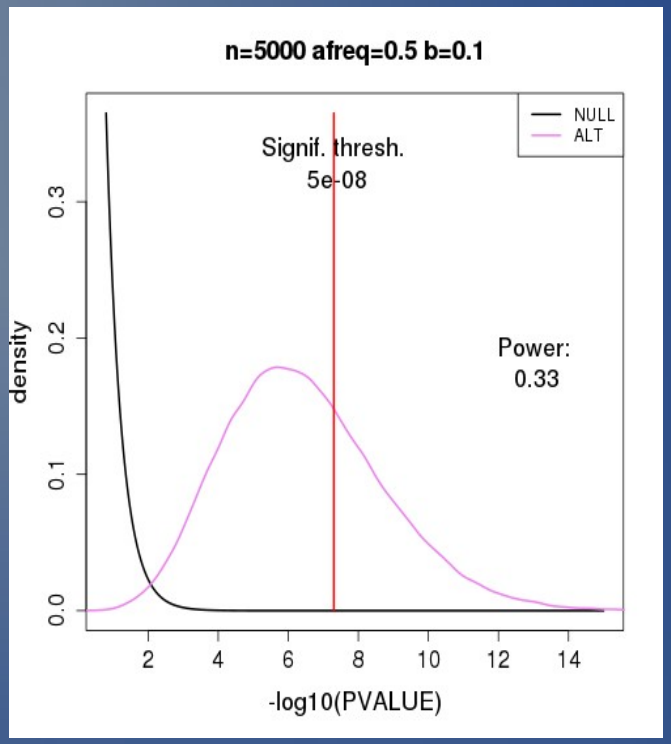


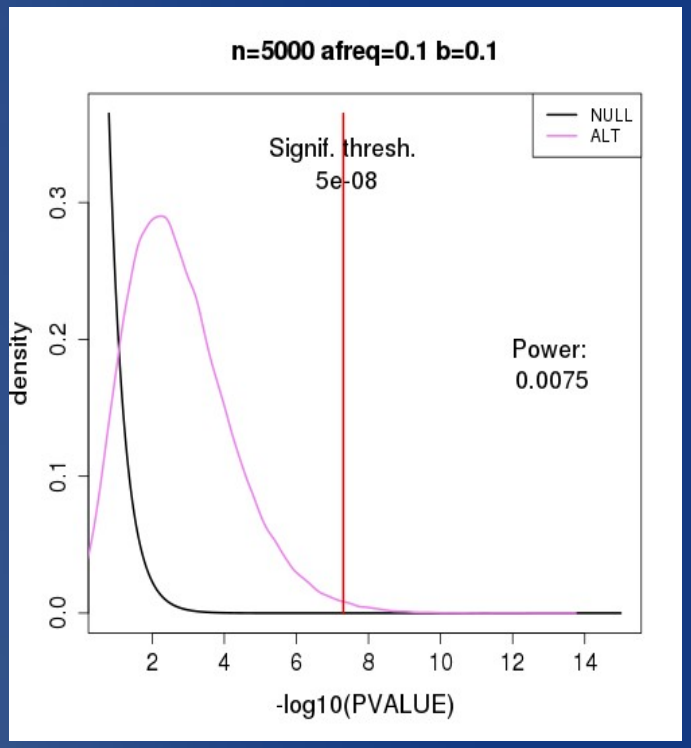


- Power = Probability that a SNP will reach a given significance threshold
  - Depends on sample size, allele freq. and effect size

Power

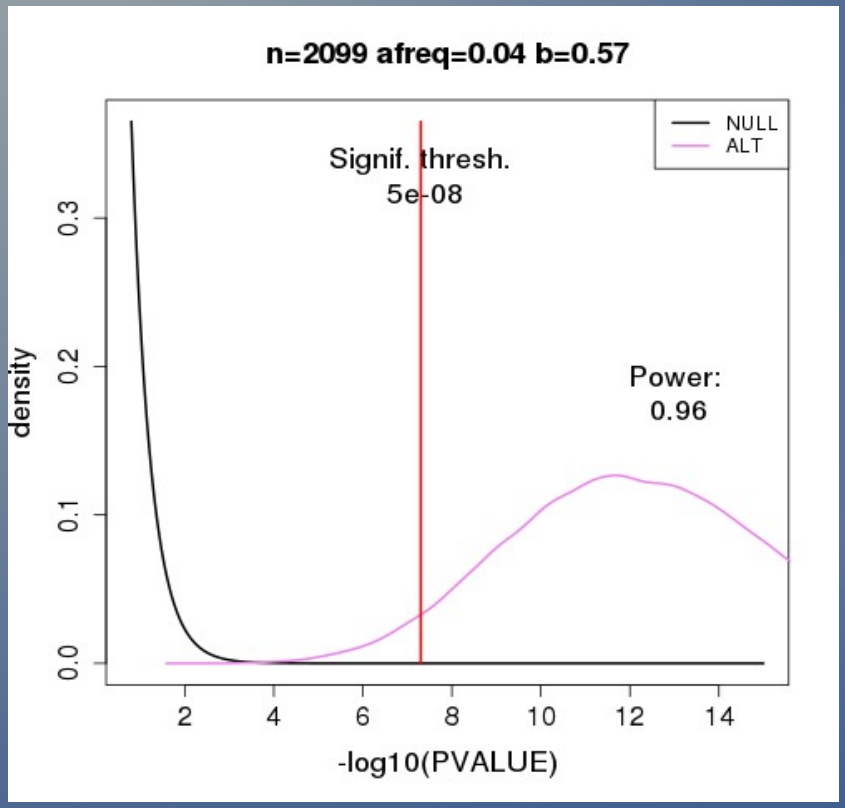






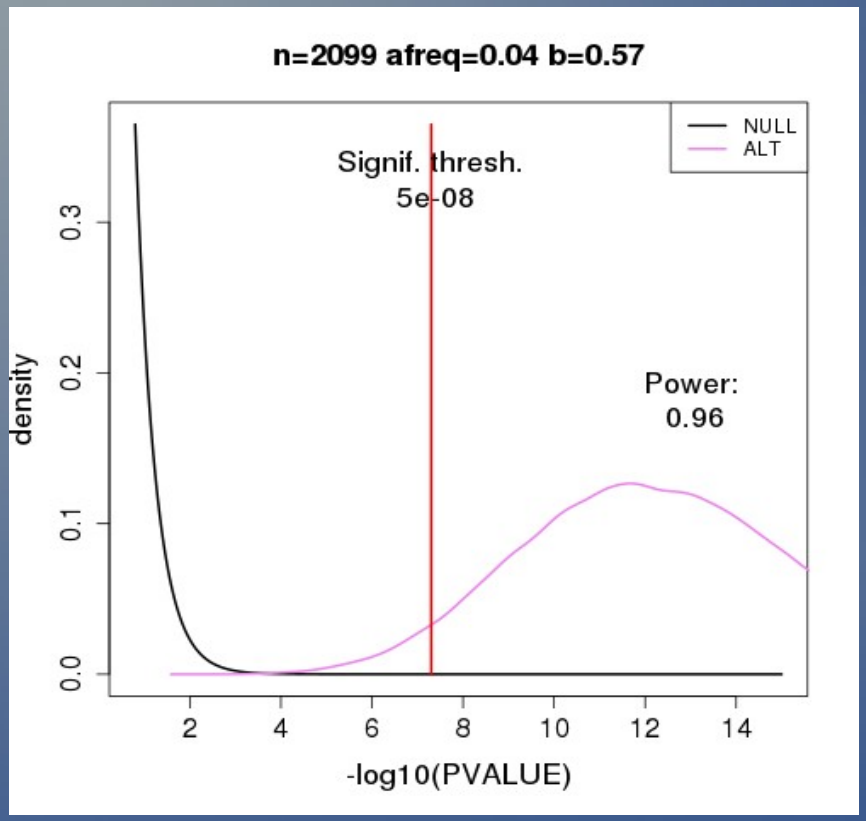
- Power = Probability that a SNP will reach a given significance threshold
  - Depends on sample size, allele freq. and effect size

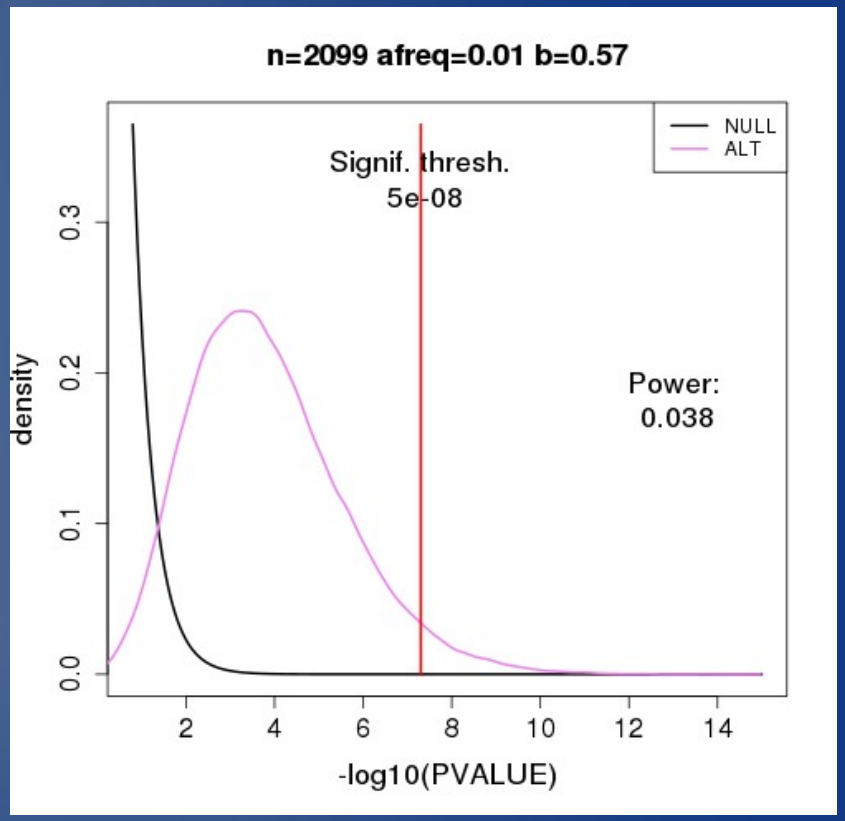
Power



- Our earlier PCSK9 variant was almost destined to be found from our Finnish data
- Power calculations needed in study design
  - Tell what kind of effects have we found / have not found

Power





- Our earlier PCSK9 variant was almost destined to be found from our Finnish data
- But would almost certainly not been found from European data where frequency is 0.01 (compared to 0.04 in FIN)

Ingredients of statistical power

- For quantitative traits power increases with

$$N\ f\ (1-f)\ b^2$$

  - N = sample size
  - f = minor allele frequency
  - b = effect size (“slope”) per one allele
- For case-control study power increases with

$$N\ t\ (1-t)\ f\ (1-f)\ b^2$$

  - t = proportion of cases in the study

Properties of power

- If, for a given SNP, pop1 has MAF=4% and pop2 has MAF=1%, how large a sample from pop2 is needed for same power as a sample of n=2,000 from pop1 ?

$$\mathbf{N\ f\ (1\text{-}f)\ b^2}$$

Properties of power

- If, for a given SNP, pop1 has MAF=4% and pop2 has MAF=1%, how large a sample from pop2 is needed for same power as a sample of n=2,000 from pop1 ?

$$\mathbf{N\ f\ (1\text{-}f)\ b^2}$$

N x 0.01 x (1-0.01) = 2000 x 0.04 x (1-0.04)
N=7758

Properties of power

- Which one of the following two options for collecting a case-control study should you choose?
  - 2,000 inds: 1,000 cases and 1,000 controls
  - 2,500 inds: 500 cases and 2,000 controls

$$\mathbf{N\ t\ (1\text{-}t)\ f\ (1\text{-}f)\ b^2}$$

Properties of power

- Which one of the following two options for collecting a case-control study should you choose?
  - 2,000 inds: 1,000 cases and 1,000 controls
  - 2,500 inds: 500 cases and 2,000 controls

$$N\ t\ (1-t)\ f\ (1-f)\ b^2$$

2000 x 1000/2000 x (1-1000/2000) = 500
2500 x 500/2500 x (1-500/2500) = 400

Option 1 has more power

From “significance” to probability of a true effect

P-value cutoff

	NULL SNPs	TRUE EFFECTS
NON-SIGNIF.	MANY	?
SIGNIF.	(almost) none	?

T = ”true effect”
N = ”null SNP”
S = ”significant p-value”

From “significance” to probability of a true effect

P-value cutoff

	NULL SNPs	TRUE EFFECTS
NON-SIGNIF.	MANY	?
SIGNIF.	(almost) none	?

T = ”true effect”
N = ”null SNP”
S = ”significant p-value”

$$\frac{\mathrm{P}(T|S)}{\mathrm{P}(N|S)} = \frac{\mathrm{P}(S|T)}{\mathrm{P}(S|N)} \times \frac{\mathrm{P}(T)}{\mathrm{P}(N)} = \frac{\text{power}}{\text{signif. cutoff}} \times \text{prior-odds of assoc.}$$

From “significance” to probability of a true effect

P-value cutoff

	NULL SNPs	TRUE EFFECTS
NON-SIGNIF.	MANY	?
SIGNIF.	(almost) none	?

T = ”true effect”
N = ”null SNP”
S = ”significant p-value”

$$\frac{P(T|S)}{P(N|S)} = \frac{P(S|T)}{P(S|N)} \times \frac{P(T)}{P(N)} = \frac{\text{power}}{\text{signif. cutoff}} \times \text{prior-odds of assoc.}$$

- Small p-value threshold is because of LOW PRIOR of association NOT because of the number of tests done
  - Often prior is not known and number of tests works OK in practice
- A “signif.” finding from a well-powered study is more likely to be true than that from a study with low power

(WTCCC. Nature 2007)

Part I: Common variants

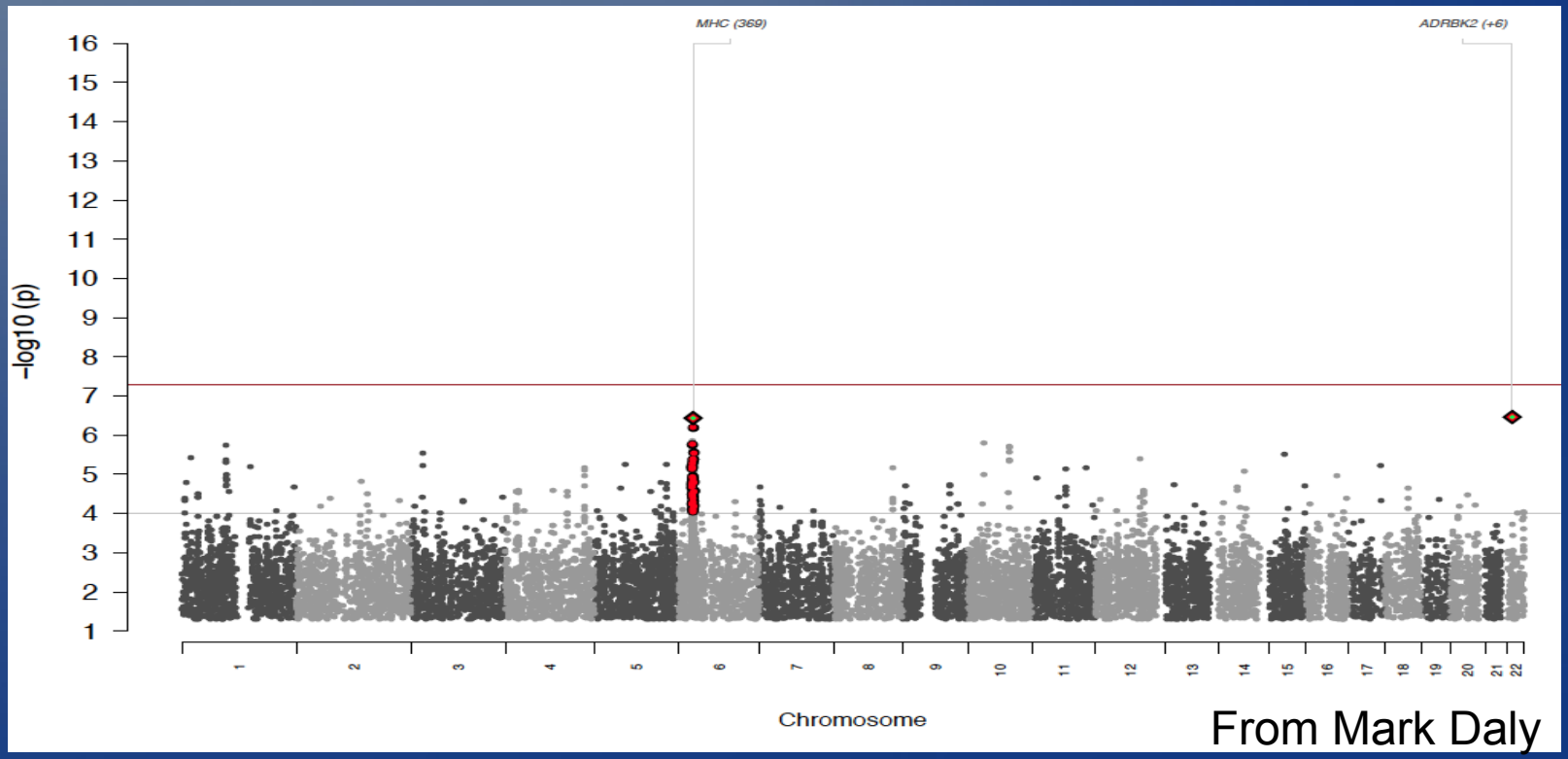
- 1. Motivation
  - What is a “genetic association” ?
  - Why is this important ?
- 2. How to read the genome (for SNPs) ?
- 3. An Example GWAS of Multiple Sclerosis
- 4. Interpreting a GWAS
- 5. Current state of GWAS

Schizophrenia (as an example of GWAS evolution)

- Mental disorder with abnormal social behavior and failure to recognize what is real
- Onset as young adults, prevalence 0.5%-1%
- High heritability, estimates up to 80%
- Linkage studies with families in 1980s-1990s were not successful
  - Unlikely to exist only a few “SZ-genes” which would account for most heritability

Int'l SZ Consortium, 2009, Nature

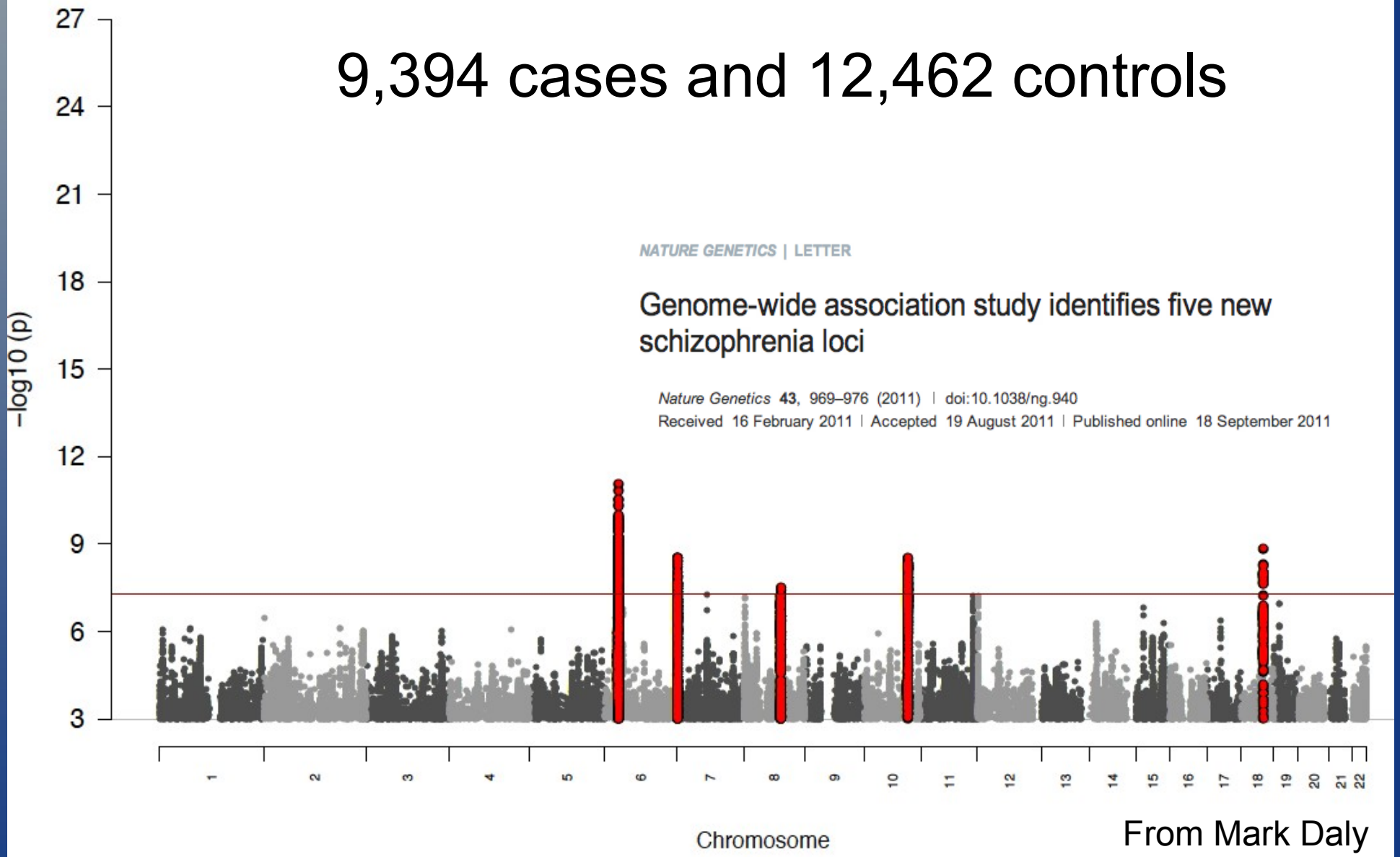
- 3,332 SZ cases and 3,587 controls at 1M SNPs
- Suggestive evidence for HLA-region on chr 6
- Evidence for highly polygenic basis for SZ
- But unable to identify “SZ genes”
  - GWAS is a failure ?



From Mark Daly

PGC 2011

9,394 cases and 12,462 controls



NATURE GENETICS | LETTER

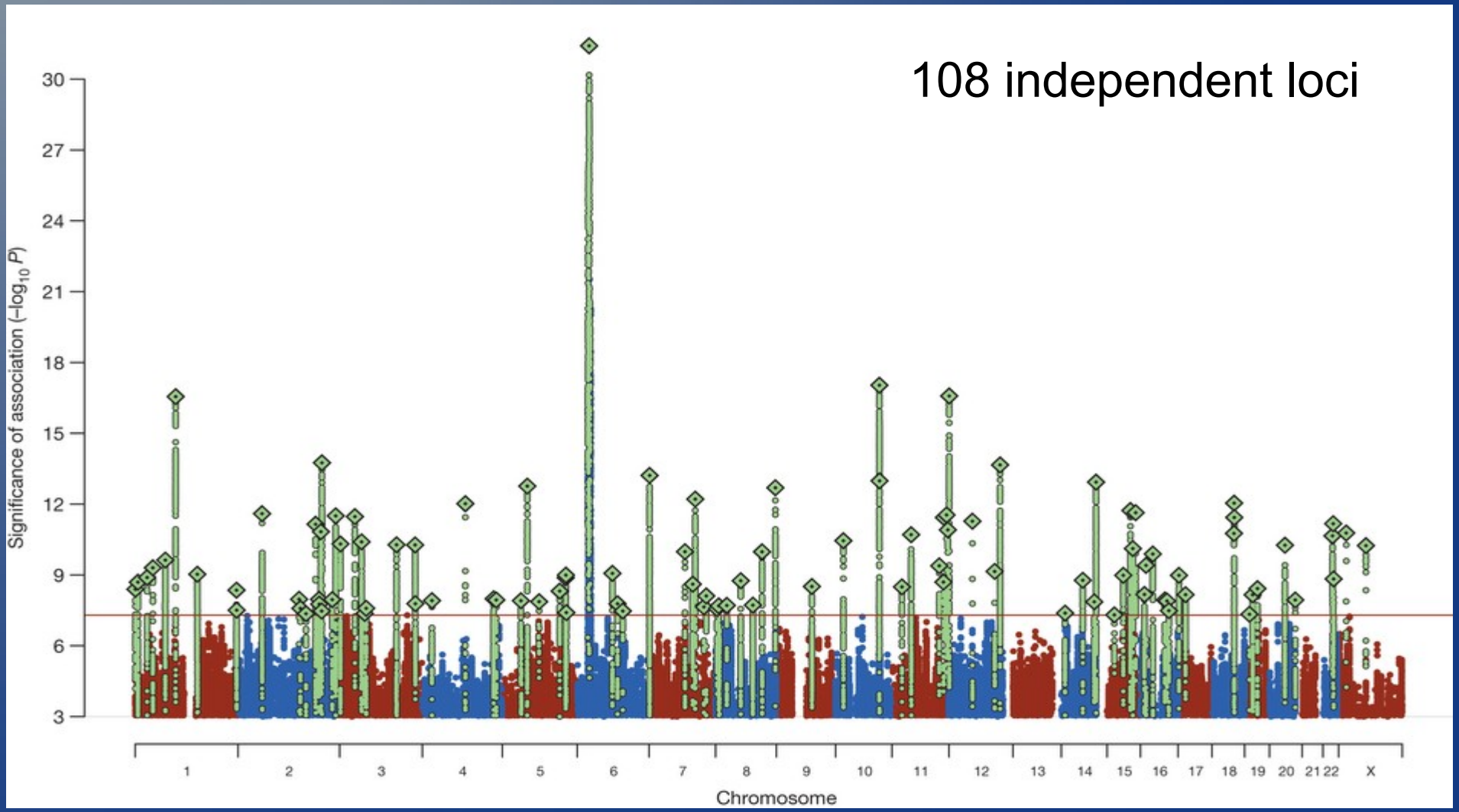
Genome-wide association study identifies five new schizophrenia loci

Nature Genetics **43**, 969–976 (2011) | doi:10.1038/ng.940
Received 16 February 2011 | Accepted 19 August 2011 | Published online 18 September 2011

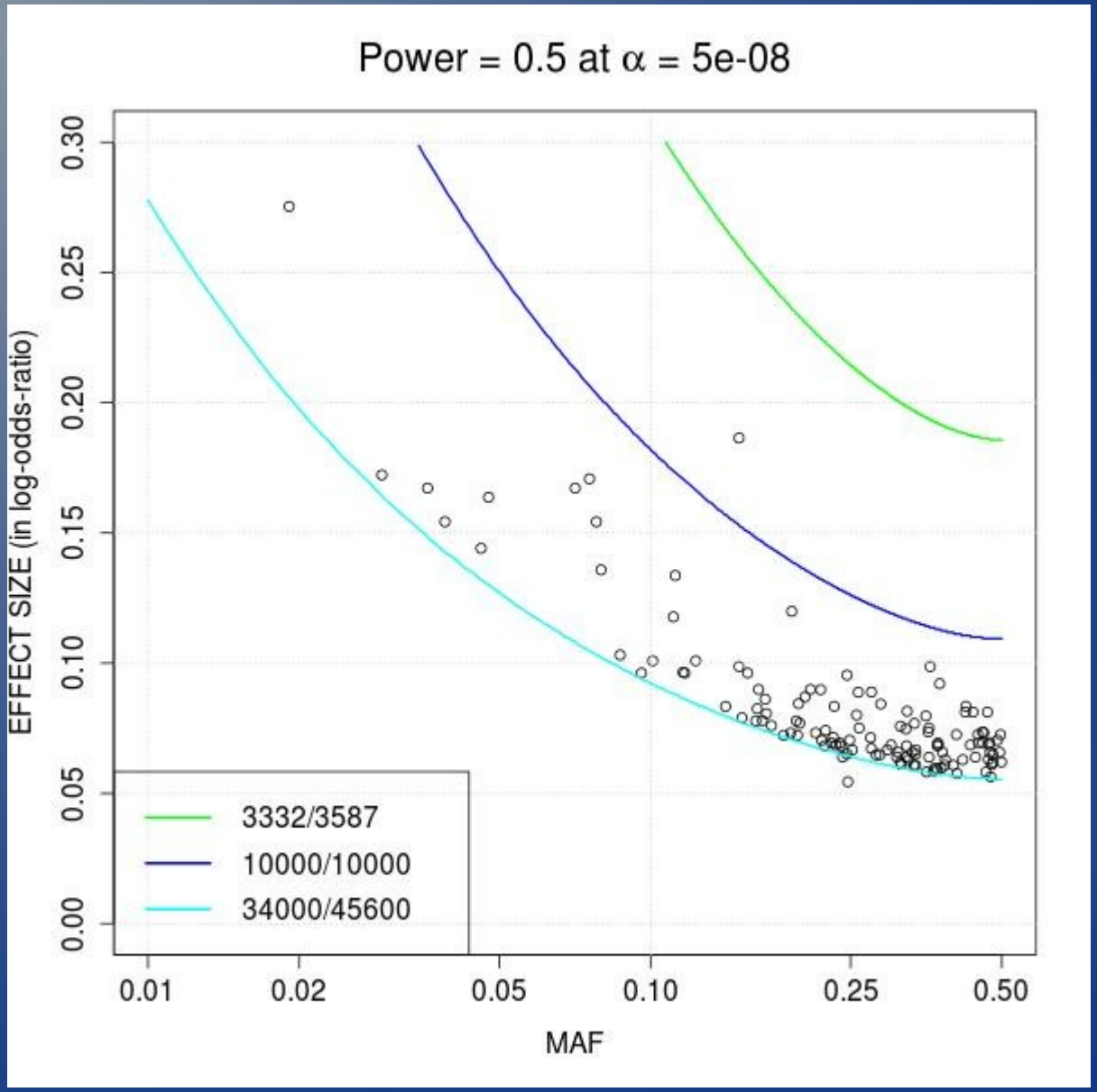
From Mark Daly

PGC 2014, Nature

- 34,000 SZ cases and 45,600 controls at 9.5M SNPs



Power for schizophrenia GWAS



Biological hints emerging...

- Finding many genes implies pathways and processes underlying schizophrenia
  - 4 voltage-gated calcium channels (CACNA1C, CACNA1D, CACNA1I, CACNB2)
  - Enricment of protein interacting with FMRP
  - Pathways highlighting postsynaptic density and dendritic spine heads
  - Enrichment of enhancer elements in specific brain regions (angular gynus, inferior temporal lobe) and immune system

From Mark Daly

A Pattern across common diseases and traits

	Adult height	Crohn's	Schizophrenia
Per N	10,000	1,000/1,000	3,000/3,000
1x	0	2	1
2x	2	4	2
3x	7	5	6
9x	68	51	62
18x	180	-	-

Regions with $p < 5e-08$

Schizophrenia is a heritable, medical disorder with a genetic architecture similar to non-brain diseases and traits

From Mark Daly

An era of Meta-analysis

- Meta-analysis is to combine results from several individual studies on the same question e.g. a particular disease
- Increases the sample size without re-analysing everything from the very beginning
- Large consortia have been formed to carry out meta-analyses for each disease

Published Genome-Wide Associations through 12/2013
at p≤5e-8 for 17 trait categories

- Digestive system disorder
- Cardiovascular disorder
- Metabolic disorder
- Immune system disorder
- Neurological disorder
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Chemical compound
- Biological process
- Cancer
- Other disease
- Other trait
- Trait mapping in progress

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

EMBL-EBI

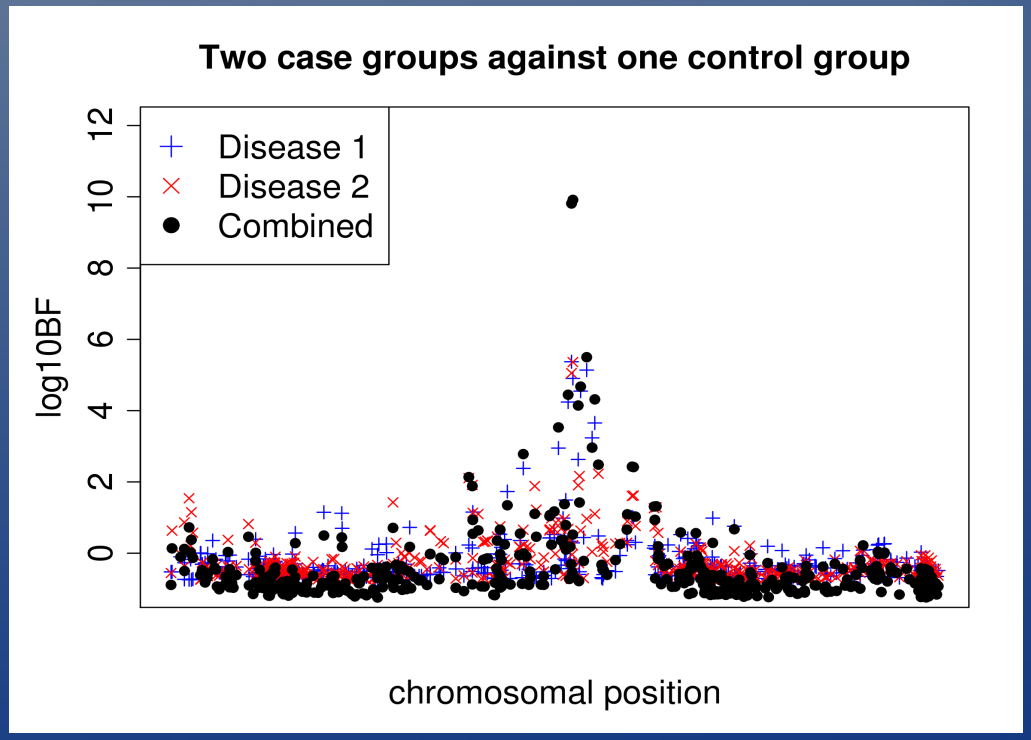
National Human Genome Research Institute

Picture emerging from GWAS

- A lot of common variants with small effects
  - Some are tagging rare variants

Picture emerging from GWAS

- A lot of common variants with small effects
  - How many tagging rare variants?
- Many shared effects across traits
  - Need joint analyses & phenotype refinement



Psoriasis and
Ankylosing spondylitis
around *IL23R*

Picture emerging from GWAS

- A lot of common variants with small effects
  - How many tagging rare variants?
- Many shared effects across traits
  - Need joint analyses & phenotype refinement
- Much to do on the biological side
  - Pathways
  - From association to function

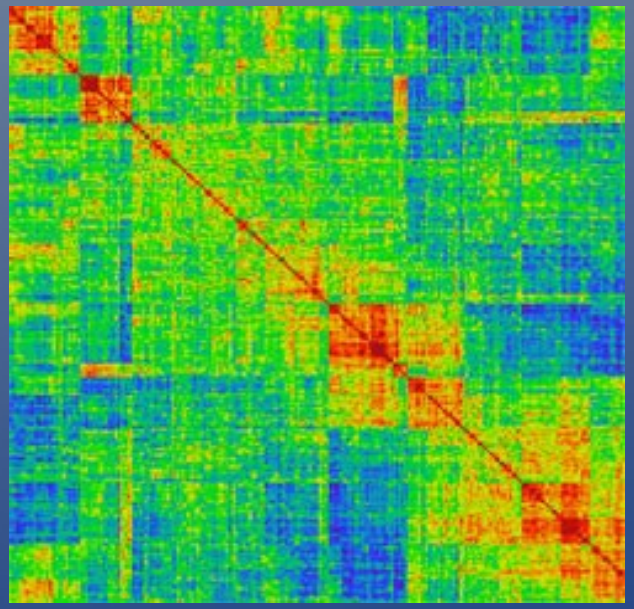
GWAS criticism

- Missing heritability
  - GWAS hits explain only a little of total genetic variance
- Missing mechanisms
- Small effect sizes
- Methodological flaws
  - Population structure
  - Separate handling of cases and controls

Weiss & Terwilliger
Nat Gen 2000

Criticism 1: Missing heritability

- GWAS SNPs typically explain only 10-20% of the estimated genetic variance (“heritability”)



For height in 14,500 Finnish samples, genetic component of common SNPs explains 52% of variation.

GWAS meta-analysis of 250,000 inds reports over 400 regions with top SNPs together explaining ~ 15% of the variance.

Where is the rest?

Criticism 1: Missing heritability

- GWAS SNPs typically explain only 10-20% of the estimated genetic variance (“heritability”)
  - We don't have power to pick out (myriad of?) still smaller effects (Yang et al. 2010)

Criticism 1: Missing heritability

- GWAS SNPs typically explain only 10-20% of the estimated genetic variance (“heritability”)
  - We don't have power to pick out (myriad of?) still smaller effects (Yang et al. 2010)
  - We haven't covered rare variants well (Gibson 2012)

Criticism 1: Missing heritability

- GWAS SNPs typically explain only 10-20% of the estimated genetic variance (“heritability”)
  - We don't have power to pick out (myriad of?) still smaller effects (Yang et al. 2010)
  - We haven't covered rare variants well (Gibson 2012)
  - Estimates of heritability may be biased (Zuk et al. 2012)

Criticism 2: Missing mechanisms

- GWAS hits pointers but we need time to sort out
- There are examples of mechanisms
  - "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus" Nature 2010

Recent genome-wide association studies (GWASs) have identified a locus on chromosome 1p13 strongly associated with both plasma low-density lipoprotein cholesterol (LDL-C) and myocardial infarction (MI) in humans. Here we show through a series of studies in human cohorts and human-derived hepatocytes that a common noncoding polymorphism at the 1p13 locus, rs12740374, creates a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site and alters the hepatic expression of the SORT1 gene. With small interfering RNA (siRNA) knockdown and viral overexpression in mouse liver, we demonstrate that Sort1 alters plasma LDL-C and very low-density lipoprotein (VLDL) particle levels by modulating hepatic VLDL secretion. Thus, we provide functional evidence for a novel regulatory pathway for lipoprotein metabolism and suggest that modulation of this pathway may alter risk for MI in humans. **We also demonstrate that common noncoding DNA variants identified by GWASs can directly contribute to clinical phenotypes.**

Criticism 3: Small effect sizes

- If an effect of an allele is 1 mm in height or relative risk of 1.1 for multiple sclerosis is that of any use?
- Can we make everything affect everything when sample size is large enough?

Small effect sizes

- Maybe individual common variants have small effects because nature does NOT tolerate large effects at those loci
- By therapies we can perturb those pathways with much larger effects than present in nature
- “The HMGCR locus has a common variant at 40% frequency that changes LDL by a modest 2.8 mg/dl and no known rare mutations of large effect, presumably because they would be lethal. Yet, the encoded protein is the target of statins, drugs taken by tens of millions of patients that can significantly reduce both LDL levels and myocardial infarction risk.” E Lander 2011

Criticism 4: Flawed design of GWAS

- Population structure
- Different genotyping errors and other properties in cases and controls if genotyped separately

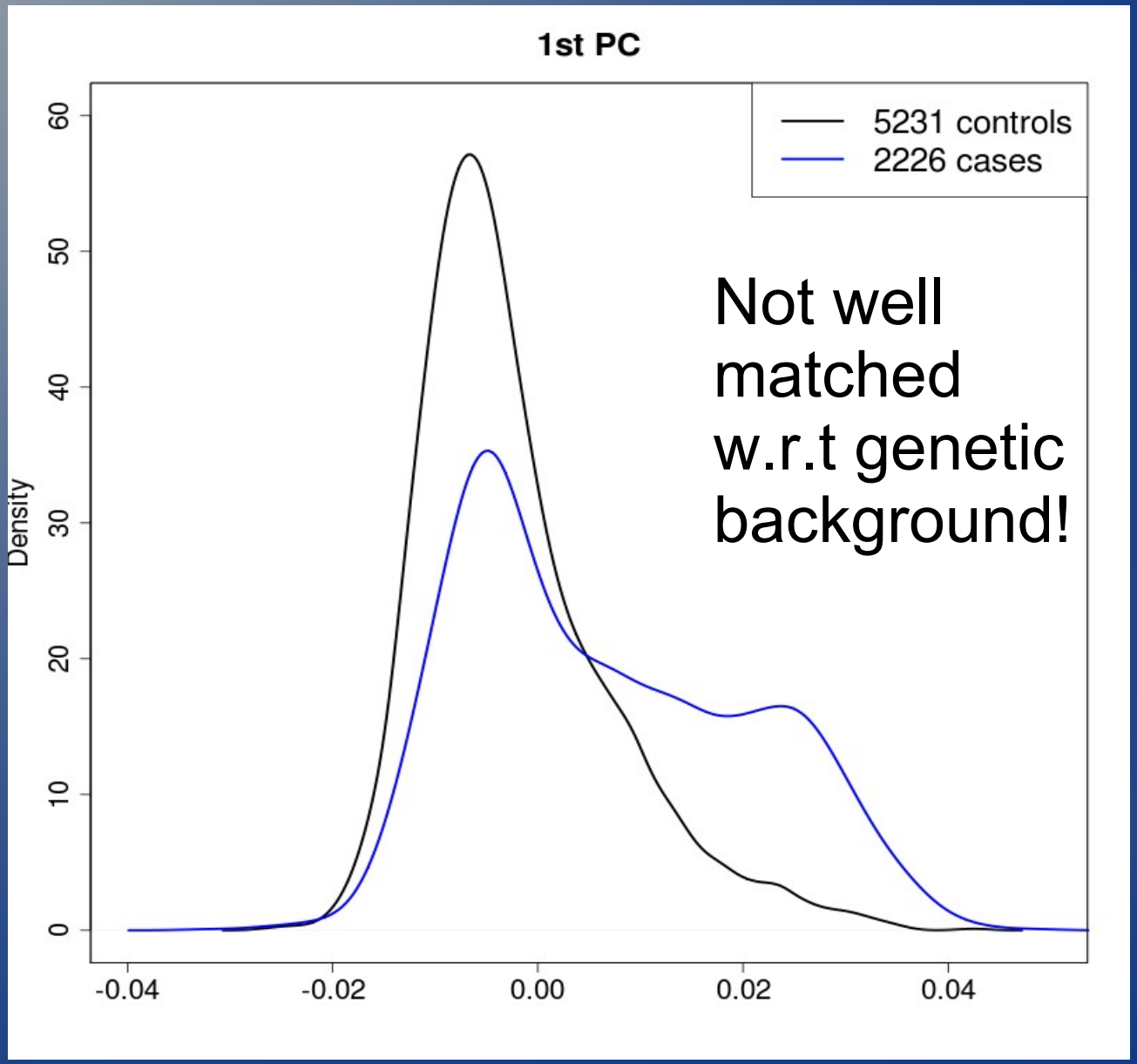
To tackle confounding in GWAS

- Good methods to warn about and correct for systematic biases: mixed models, covariates, careful quality control, qq-plots
- In large meta-analyses unlikely that biases of individual studies would multiply
- Replication in several cohorts provides convincing evidence
- Stringent statistical thresholds are applied (p-value 5e-8)

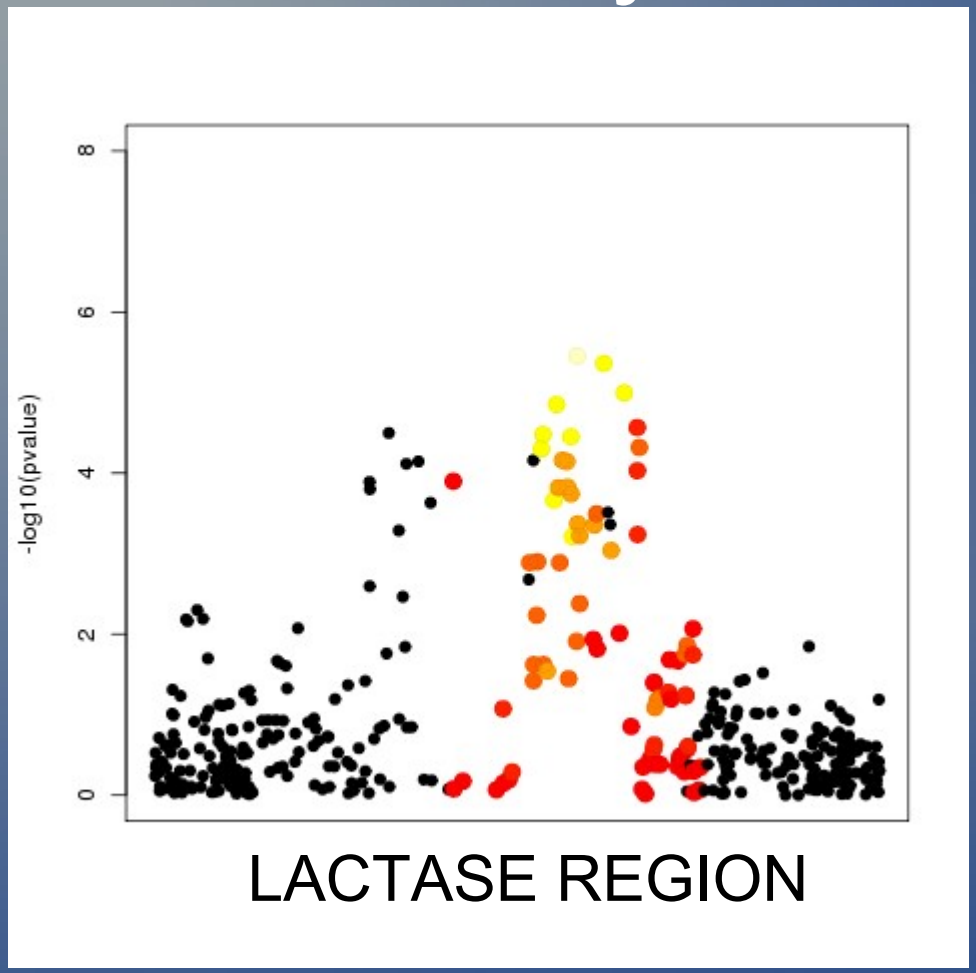
Population structure in case-control association studies

- Are there differences in genotype frequencies between cases and controls?
  - If yes, then locus is possibly interesting
  - But could also reflect ascertainment scheme if cases and controls are not well matched w.r.t. genetic background!
  - Example: Let's look at analysis where 5,000 UK controls are compared against 1,800 UK + 500 Irish Psoriasis cases

UK+Irish Psoriasis study



Basic analysis



First PC as covariate