# Emerging themes in genetics

Frank Dudbridge

London School of Hygiene and Tropical Medicine

# Missing heritability

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

## Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

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

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

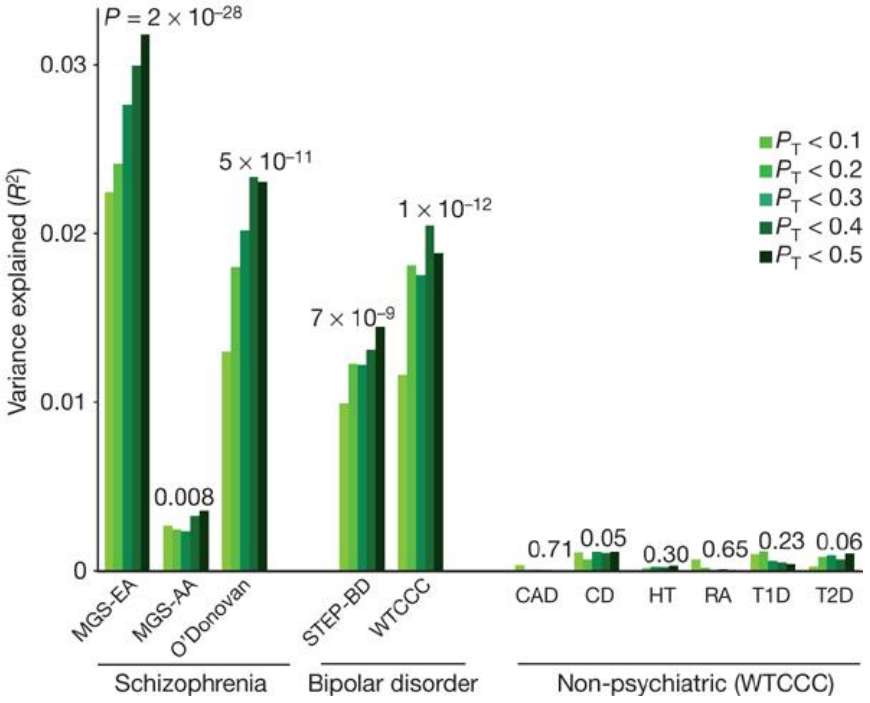
# What is missed by genomewide association scans

- SNPs with smaller effects
  - Strict p-values – larger samples/meta-analysis
- Rare variants
  - Included on newer chips – but low power
- Copy number variants
  - Mostly rare
- Parent of origin effects
- Gene-gene interaction
  - Gene-gene or SNP-SNP
- Heritability probably over-estimated originally

# Evidence for SNPs with smaller effects

Replication of the ISC-derived polygenic component in independent schizophrenia and bipolar disorder samples.

- Take all SNPs with $P<P_T$ in the ISC sample
- Calculate a score for each subject based on these SNPs
  - #risk alleles carried
- Associate this score with the phenotype




- Replicates "en-masse" SNPs in independent schizophrenia data
- Shows common genetic component between schiz and bipolar disorder

The International Schizophrenia Consortium *Nature* **000**, 1-5 (2009) doi:10.1038/nature08185

nature

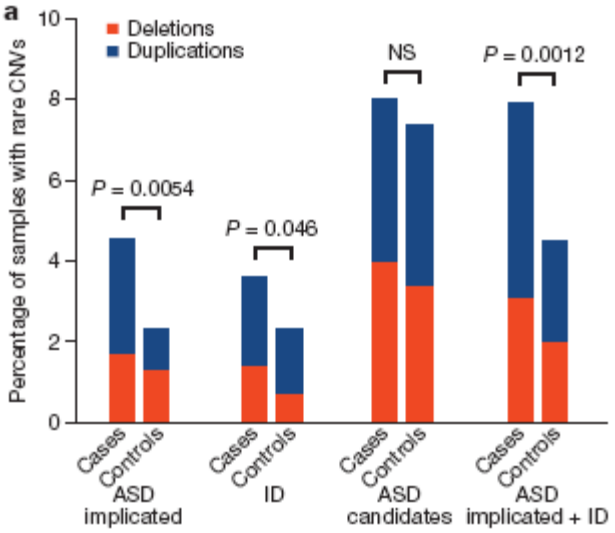
# Copy number variation

nature Vol 466|15 July 2010|doi:10.1038/nature09146

LETTERS

## Functional impact of global rare copy number variation in autism spectrum disorders

A list of authors and their affiliations appears at the end of the paper.




- Very few SNP associations in GWAS
- ASD cases carry more rare CNVs than do controls
- CNVs must cause ASD, but hard to identify individual variants

# Much heritability is tagged in GWAS

## Genome partitioning of genetic variation for complex traits using common SNPs

Jian Yang[1*], Teri A Manolio[2], Louis R Pasquale[3], Eric Boerwinkle[4], Neil Caporaso[5], Julie M Cunningham[6], Mariza de Andrade[7], Bjarke Feenstra[8], Eleanor Feingold[9], M Geoffrey Hayes[10], William G Hill[11], Maria Teresa Landi[12], Alvaro Alonso[13], Guillaume Lettre[14], Peng Lin[15], Hua Ling[16], William Lowe[17], Rasika A Mathias[18], Mads Melbye[8], Elizabeth Pugh[16], Marilyn C Cornelis[19], Bruce S Weir[20], Michael E Goddard[21,22] & Peter M Visscher[1]

NATURE GENETICS VOLUME 43 | NUMBER 6 | JUNE 2011 519

The SNPs are there, but are not statistically significant

**Table 1 Estimates of the variance explained by all autosomal SNPs for height, BMI, vWF and QTi**

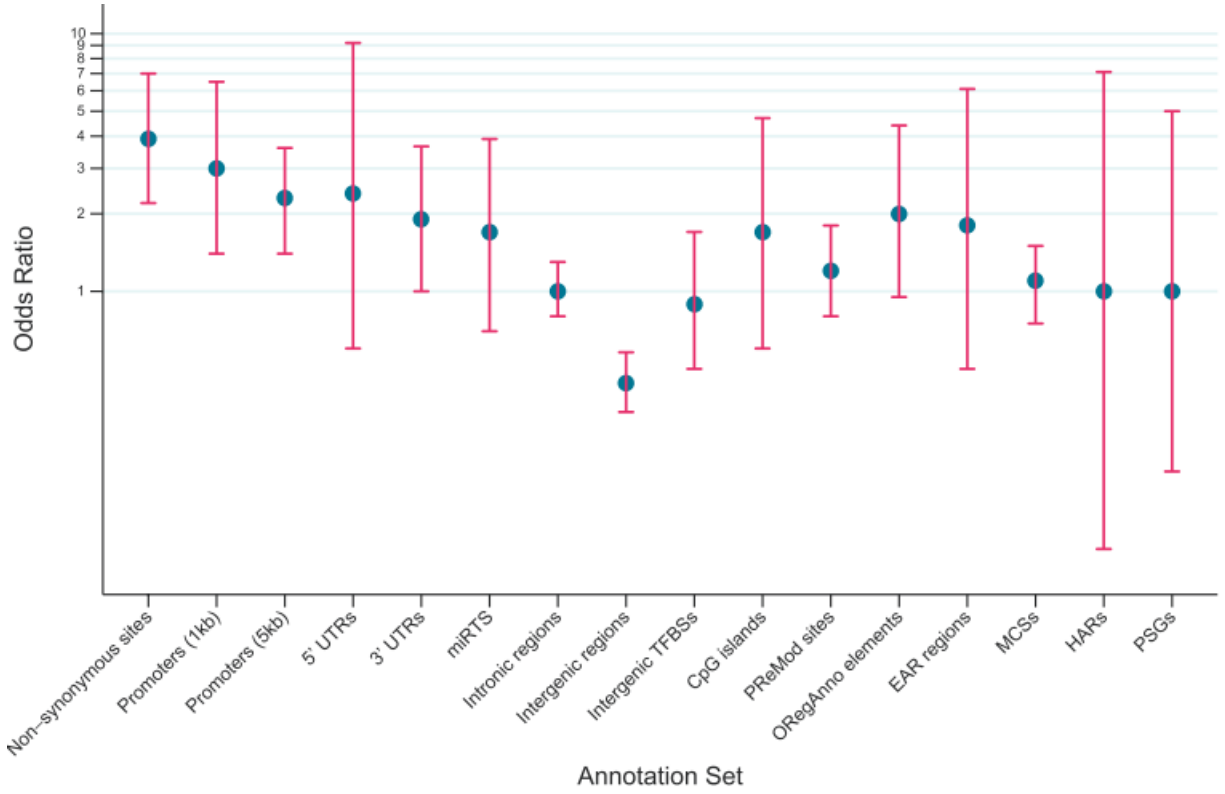
| | | No PC[a] | | 10 PCs[b] | | | |
|---|---|---|---|---|---|---|---|
| Trait | *n* | $h^2_G$ (s.e.)[c] | *P* | $h^2_G$ (s.e.) | *P* | Heritability[d] | GWAS[e] |
| Height | 11,576 | 0.448 (0.029) | $4.5 \times 10^{-69}$ | 0.419 (0.030) | $7.9 \times 10^{-48}$ | 80–90%[32] | ~10%[23] |
| BMI | 11,558 | 0.165 (0.029) | $3.0 \times 10^{-10}$ | 0.159 (0.029) | $5.3 \times 10^{-9}$ | 42–80%[25,26] | ~1.5%[14] |
| vWF | 6,641 | 0.252 (0.051) | $1.6 \times 10^{-7}$ | 0.254 (0.051) | $2.0 \times 10^{-7}$ | 66–75%[33,34] | ~13%[15] |
| QTi | 6,567 | 0.209 (0.050) | $3.1 \times 10^{-6}$ | 0.168 (0.052) | $5.0 \times 10^{-4}$ | 37–60%[35,36] | ~7%[16] |

# GWAS hits are enriched in functional regions

## Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

Lucia A. Hindorff[a,1], Praveen Sethupathy[b,1], Heather A. Junkins[a], Erin M. Ramos[a], Jayashri P. Mehta[c], Francis S. Collins[b,2], and Teri A. Manolio[a,2]

9362–9367 | PNAS | June 9, 2009 | vol. 106 | no. 23



# Pathway analysis

- Aims to find association of a group of functionally related genes
  - Individual SNPs may not be significant
  - Can enhance the results of GWAS by finding associated pathways
- Various databases of biological pathways
  - eg GO, KEGG, BioCarta, Panther, Reactome
- Test whether more SNPs in the pathway are significant at some threshold (eg $P<0.01$) than SNPs outside the pathway
- Complex methods allow for network structure

# Example – Alzheimer's Disease

## Genetic Evidence Implicates the Immune System and Cholesterol Metabolism in the Aetiology of Alzheimer's Disease

Lesley Jones[19], Peter A. Holmans[19], Marian L. Hamshere[1], Denise Harold[1], Valentina Moskvina[1], Dobril Ivanov[1], Andrew Pocklington[1], Richard Abraham[1], Paul Hollingworth[1], Rebecca Sims[1], Amy Gerrish[1], Jaspreet Singh Pahwa[1], Nicola Jones[1], Alexandra Stretton[1], Angharad R. Morgan[1], Simon Lovestone[2], John Powell[3], Petroula Proitsi[3], Michelle K. Lupton[3], Carol Brayne[4], David C. Rubinsztein[5], Michael Gill[6], Brian Lawlor[6], Aoibhinn Lynch[6], Kevin Morgan[7], Kristelle S. Brown[7], Peter A Passmore[8], David Craig[8], Bernadette McGuinness[8], Stephen Todd[8], Clive Holmes[9], David Mann[10], A. David Smith[11], Seth Love[12], Patrick G. Kehoe[12], Simon Mead[13], Nick Fox[14], Martin Rossor[14], John Collinge[13], Wolfgang Maier[15], Frank Jessen[15], Britta Schürmann[15], Hendrik van den Bussche[16], Isabella Heuser[16], Oliver Peters[16], Johannes Kornhuber[17], Jens Wiltfang[18], Martin Dichgans[19,20], Lutz Frölich[21], Harald Hampel[22,23], Michael Hüll[24], Dan Rujescu[23], Alison M Goate[25], John S. K. Kauwe[26], Carlos Cruchaga[25], Petra Nowotny[25], John C. Morris[25], Kevin Mayo[25], Gill Livingston[27], Nicholas J. Bass[27], Hugh Gurling[27], Andrew McQuillin[27], Rhian Gwilliam[28], Panos Deloukas[28], Ammar Al-Chalabi[29], Christopher E. Shaw[29], Andrew B. Singleton[30], Rita Guerreiro[30], Thomas W. Mühleisen[31,32], Markus M. Nöthen[31,32], Susanne Moebus[33], Karl-Heinz Jöckel[33], Norman Klopp[34], H.-Erich Wichmann[34–36], Eckhard Rüther[37], Minerva M. Carrasquillo[38], V. Shane Pankratz[39], Steven G. Younkin[38], John Hardy[40], Michael C. O'Donovan[1], Michael J. Owen[1]*, Julie Williams[1]*

# More enriched pathways than expected

| SNP list criterion | #genes | enrichment p<0.05 | | enrichment p<0.01 | | enrichment p<0.001 | |
|---|---|---|---|---|---|---|---|
| | | **#cat** | **p** | **#cat** | **p** | **#cat** | **p** |
| **p<1e-4** | **72** | **115** | **0.009** | **50** | **0.006** | **16** | **0.008** |
| **p<1e-3*** | **589** | **254** | **0.005** | **127** | **<0.001** | **57** | **<0.001** |
| **p<0.005** | **2212** | **291** | **0.006** | **76** | **0.006** | **18** | **<0.001** |
| **p<0.01** | **3703** | **282** | **0.023** | **64** | **0.031** | **8** | **0.110** |
| **p<0.05** | **10709** | **228** | **0.078** | **44** | **0.096** | **4** | **0.295** |

# Top GO categories

| GO process | category total | Genes in GWAS | expected | p-value | expected hits/study | Process |
|---|---|---|---|---|---|---|
| GO:0008203 | 83 | 11 | 2.54 | 0.00E+00 | 0.06 | cholesterol metabolic process |
| GO:0016125 | 92 | 12 | 2.85 | 0.00E+00 | 0.06 | sterol metabolic process |
| GO:0032488 | 4 | 3 | 0.25 | 2.00E-05 | 0.10 | Cdc42 protein signal transduction |
| GO:0006958 | 29 | 6 | 0.72 | 2.00E-05 | 0.10 | complement activation, classical pat |
| GO:0002455 | 29 | 6 | 0.72 | 2.00E-05 | 0.10 | humoral immune response mediate |
| GO:0008202 | 175 | 16 | 5.65 | 4.00E-05 | 0.13 | steroid metabolic process |
| GO:0033700 | 8 | 4 | 0.45 | 6.00E-05 | 0.16 | phospholipid efflux |
| GO:0002253 | 60 | 8 | 1.43 | 6.00E-05 | 0.16 | activation of immune response |
| GO:0045087 | 118 | 11 | 2.74 | 1.00E-04 | 0.25 | innate immune response |
| GO:0006956 | 37 | 6 | 0.81 | 1.60E-04 | 0.40 | complement activation |
| GO:0002541 | 39 | 6 | 0.85 | 1.80E-04 | 0.44 | activation of plasma proteins during |
| GO:0050746 | 6 | 3 | 0.18 | 2.20E-04 | 0.52 | regulation of lipoprotein metabolic p |
| GO:0030001 | 423 | 36 | 20 | 3.00E-04 | 0.71 | metal ion transport |
| GO:0006812 | 499 | 38 | 21.69 | 3.60E-04 | 0.84 | cation transport |
| GO:0015672 | 298 | 25 | 12.59 | 5.40E-04 | 1.30 | monovalent inorganic cation transpo |
| GO:0048583 | 237 | 16 | 6.45 | 5.80E-04 | 1.37 | regulation of response to stimulus |
| GO:0034447 | 3 | 2 | 0.06 | 6.40E-04 | 1.50 | very-low-density lipoprotein particle |
| GO:0022411 | 50 | 6 | 1.24 | 6.80E-04 | 1.59 | cellular component disassembly |
| GO:0016064 | 47 | 6 | 1.12 | 6.80E-04 | 1.59 | immunoglobulin mediated immune r |
| GO:0019724 | 49 | 6 | 1.17 | 8.80E-04 | 2.05 | B cell mediated immunity |
| GO:0002526 | 73 | 7 | 1.59 | 9.20E-04 | 2.14 | acute inflammatory response |

# Risk prediction

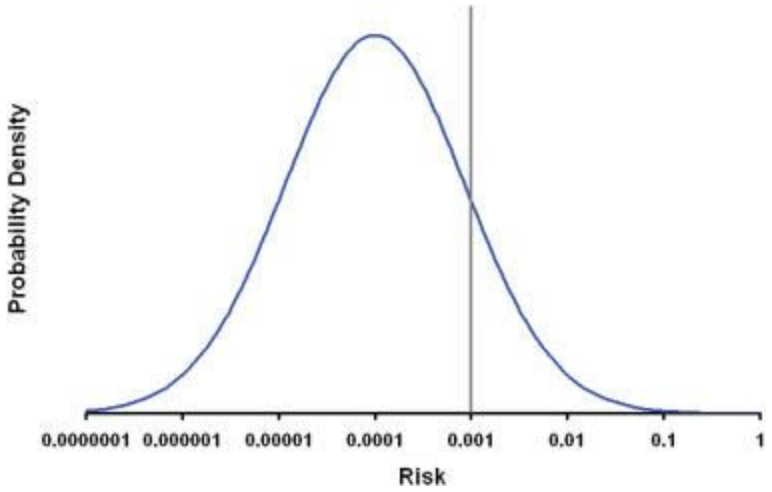
- A great hope of genetics is that we can accurately predict an individual's risk of disease
  - Early intervention
  - Enrolment in screening programmes, eg mammography
  - Selection of individuals for research trials
- In Mendelian monogenic disease, such prediction is already in clinical use
  - Disease mutations have very high penetrance
  - Genetic counselling, IVF screening
- Not so clear-cut in complex disease: genes have small effects

# Issues in genetic risk prediction

- Sensitivity
  - How many cases of disease are correctly predicted by genetics?
- Specificity
  - How many non-cases of disease are falsely predicted as cases?
- Area under ROC curve (AUC)
  - Summarises sensitivity and specificity over the range of risk scores
  - 0.99 for population screening, 0.75 for screening “at-risk” subjects
  - 0.5 is no better than random prediction
- Positive predictive value
  - If disease is predicted, what’s the chance of actually developing it?
  - Particularly challenging in rare diseases
- Does genetics improve on established risk prediction models?

# Distribution of risk

- Assuming that risk is due to many genes of small effect, the log-risk is normally distributed in the population
- The variance of the risk depends on the heritability

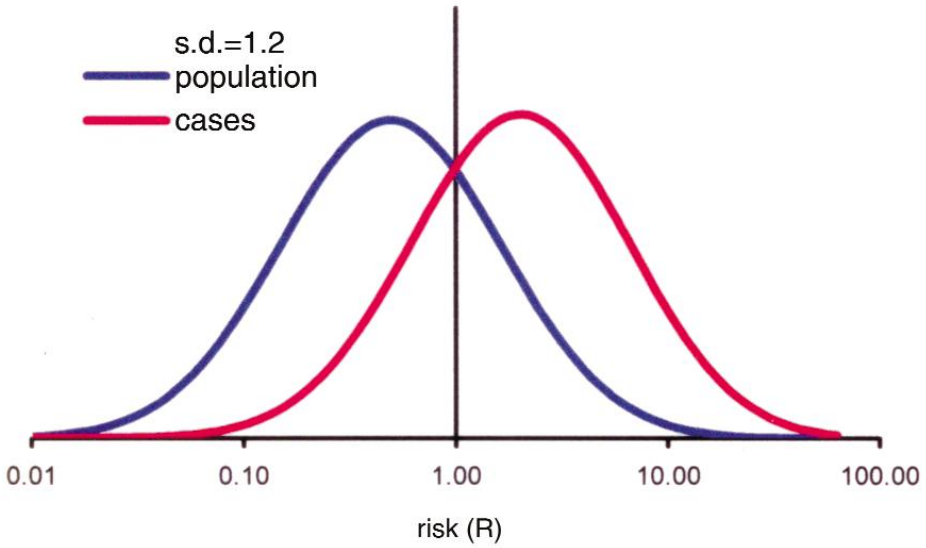



Distribution of risk for multiple sclerosis, $\lambda_s$=10

Sawcer et al, Ann Neurol 2010

# Risk in the cases

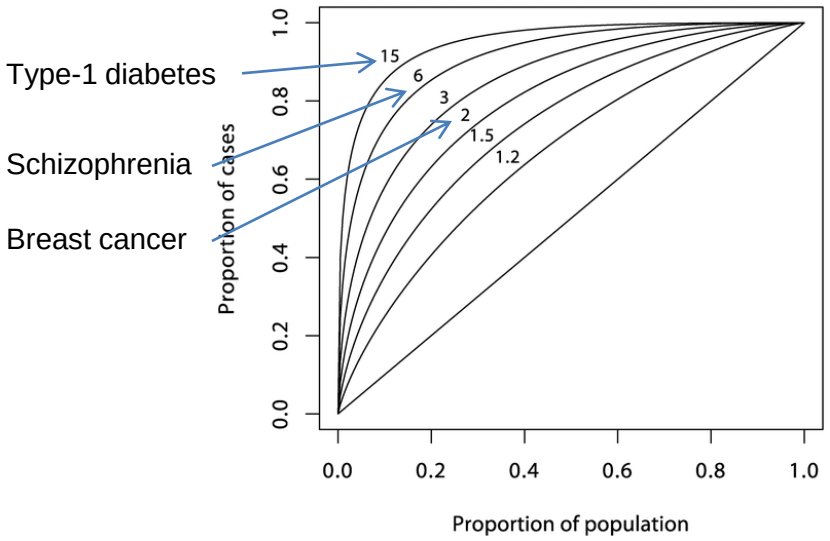
- Looking at those who did develop disease, we can look at what risk they had carried
- This is also normally distributed, but with a higher mean




Pharoah et al, Nat Genet 2002

# Limits on predictive power of genetics

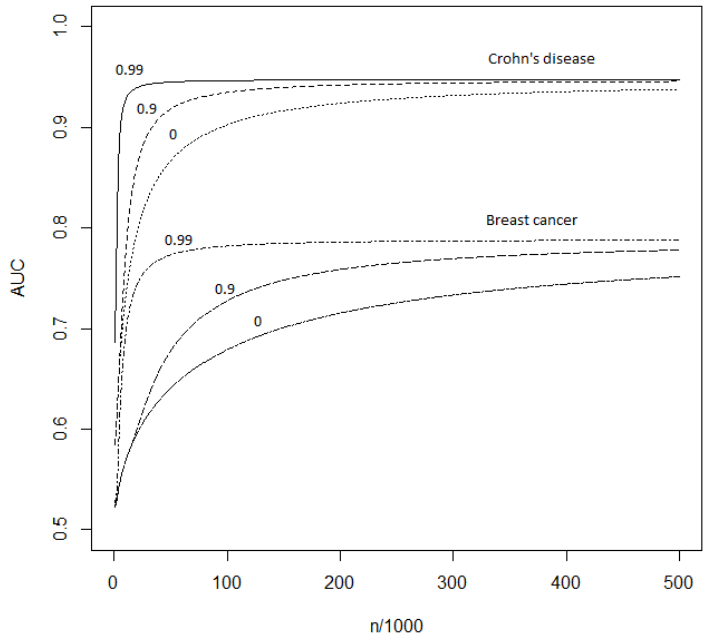
- The distributions of risk in the population and in the cases imply a limit on the ROC curve
- The heritability places a limit on how good prediction can ever be, *even if we knew all the genes that cause disease*
  - For many common diseases, genetics can never give a great AUC




Clayton, PLoS Genetics 2009

# Large samples needed to derive accurate predictors

- If disease is explained by hundreds or thousands of SNPs, very large samples are needed to minimise the sampling error in the risk score
- Possibly 100,000's of cases and controls




100k SNPs explaining half the heritability
Varying proportions of null SNPs

OPEN ACCESS Freely available online

PLOS | GENETICS

## Power and Predictive Accuracy of Polygenic Risk Scores

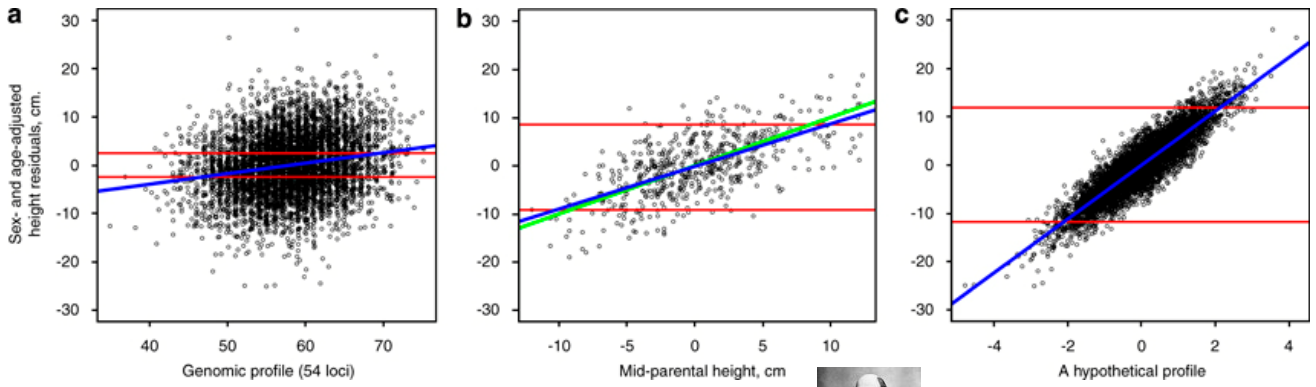
**Frank Dudbridge***

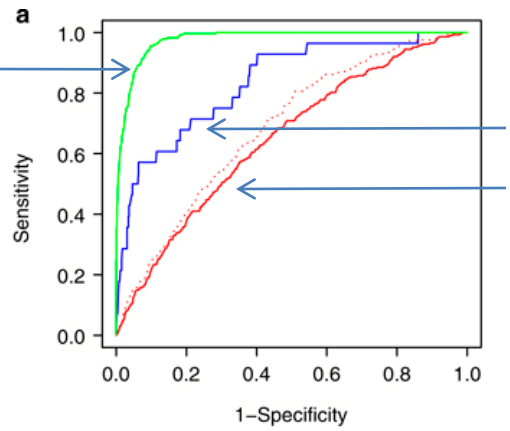
Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9(3): e1003348.

# Family history

- Family history of a trait can predict almost all well as the individual genetics




Genes explaining 100% heritability

Mid-parental height

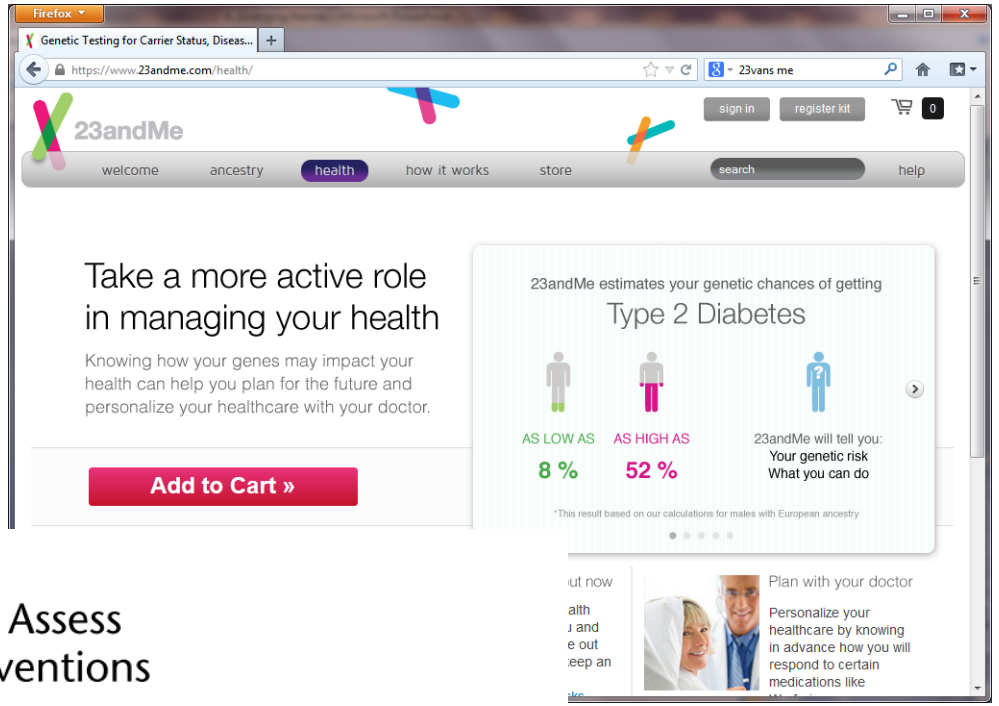
54 known genes

Aulchenko et al, Predicting human height by Victorian and genomic methods, Eur J Hum Genet 2009

# Utility of genetic risk prediction

- Genetics offers real benefits over family history and lifestyle risk prediction at the *population level* only when:
  - Heritability is fairly high
  - Genes explain a high proportion of the heritability
- Genetic prediction more useful in specific contexts
  - Identifying individuals for more intensive, non-invasive screening
  - Predicting molecular biomarkers rather than disease endpoints
  - Predicting time-to-disease, allowing early intervention
  - Predicting adverse drug reactions
  - Selecting at-risk individuals for trials

# Commercial prediction kits – caveat emptor!




## A Critical Appraisal of the Scientific Basis of Commercial Genomic Profiles Used to Assess Health Risks and Personalize Health Interventions

A. Cecile J.W. Janssens,[1,*] Marta Gwinn,[2] Linda A. Bradley,[2] Ben A. Oostra,[3] Cornelia M. van Duijn,[4] and Muin J. Khoury[2]

control group. The seven companies tested at least 69 different polymorphisms in 56 genes. Of the 56 genes tested, 24 (43%) were not reviewed in meta-analyses. For the remaining 32 genes, we found 260 meta-analyses that examined 160 unique polymorphism-disease associations, of which only 60 (38%) were found to be statistically significant. Even the 60 significant associations, which involved 29 different polymorphisms and 28 different diseases, were generally modest, with synthetic odds ratios ranging from 0.54 to 0.88 for protective variants and from 1.04 to 3.2 for risk variants. Furthermore, genes in cardiogenomic profiles were more frequently associated with

Firefox

Genes, patents, and big business: at 23a...

www.theverge.com/2012/12/12/3759198/23andme-genetics-testing-50-million-data-mining

verge 23andme

LOG IN or JOIN

Big Story: Major security hole found in Apple password system

THE VERGE FEATURES REVIEWS PRODUCTS SECTIONS SHOW PODCAST ABOUT TIP US FORUMS Search articles & products

PREVIOUS STORY
Game or be gamed: Douglas Rushkoff on prototyping democracy through play

NEXT STORY
Good deal: 'Duke Nukem 3D' free for Mac and PC on GOG.com

CULTURE POLICY & LAW WEB & SOCIAL REPORT

37 COMMENT

# Genes, patents, and big business: at 23andMe, are you the customer or the product?

*Ethical questions swirl as the personal genetics company starts scaling up*

By Adrianne Jeffries on December 12, 2012 02:56 pm Email @adrjeffries

www.theverge.com/2012/12/12/3759464/good-deal-gog-duke-nukem-3d-free

Firefox

Genetic testing service deCODEme shutt...

www.geneticliteracyproject.org/2012/12/13/genetic-testing-service-decodeme-shutting-down-in-wake-of-amg

decodeme

Enter Your Email Address

GENETIC LITERACY PROJECT

WHERE SCIENCE TRUMPS IDEOLOGY

Search

About Human Agriculture Policy Gene-ius Resources Browse

## Genetic testing service deCODEme shutting down in wake of Amgen/deCODE deal

Dan Vorhaus | December 13, 2012 | Genomics Law Report

Print

The big biotech news of the day is the $415 million sale of deCODE Genetics to Amgen. Coverage of the deal is everywhere, including a typically excellent overview from Matthew Herper of *Forbes*.

### Browse By

Authors
Tags
Sources

or try our Advanced Search

### More from this Author

Dan Vorhaus

- Amgen buys deCODE: What are the implications for genetic testing customers?

### More from this Source

Genomics Law Report

- Amgen buys deCODE: What are the implications for genetic testing customers?

### Related Articles

Genetic testing growing sharply, but will it help cut medical costs?

Breast cancer genetic testing gets covered by health care reform

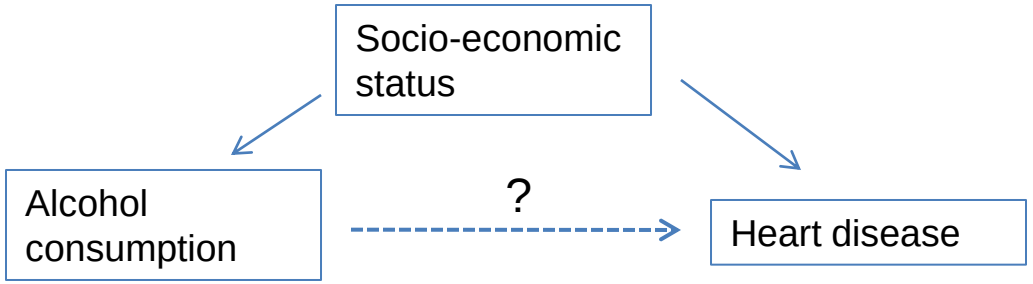
New prenatal testing can detect more genetic problems

Genetic testing of Newtown shooter will provide few answers

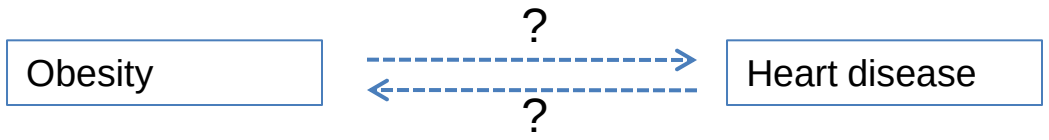
# Mendelian Randomisation

- An application of genetics to reduce problems in traditional observational epidemiology
- "Association does not equal causation"

1. Confounding




2. Reverse causality

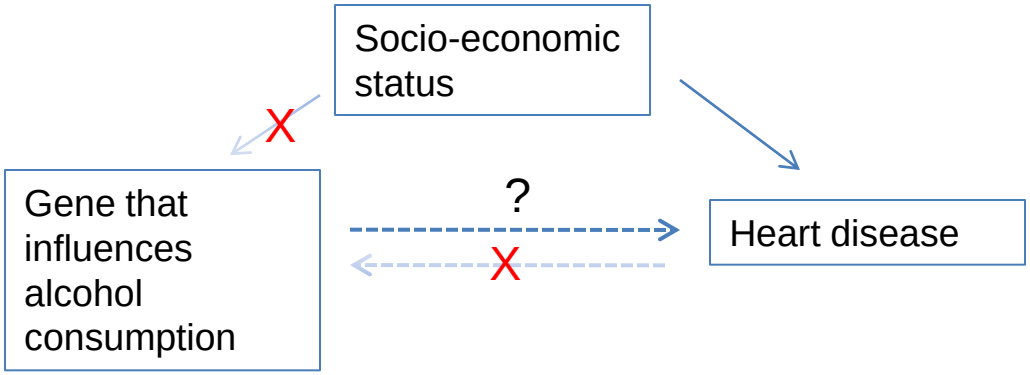


# Observation vs randomisation

- In randomised clinical trials, confounding and reverse causation are reduced or even eliminated
  - "Treatment" and "Placebo" groups differ only in the treatment received, and by no other characteristics
  - Events that follow treatment (in time) are more likely to be caused by the treatment
- Some experiments cannot be performed with randomisation
  - Effects of smoking or alcohol
  - Socio-economic effects
- Observational studies are the only ethical option

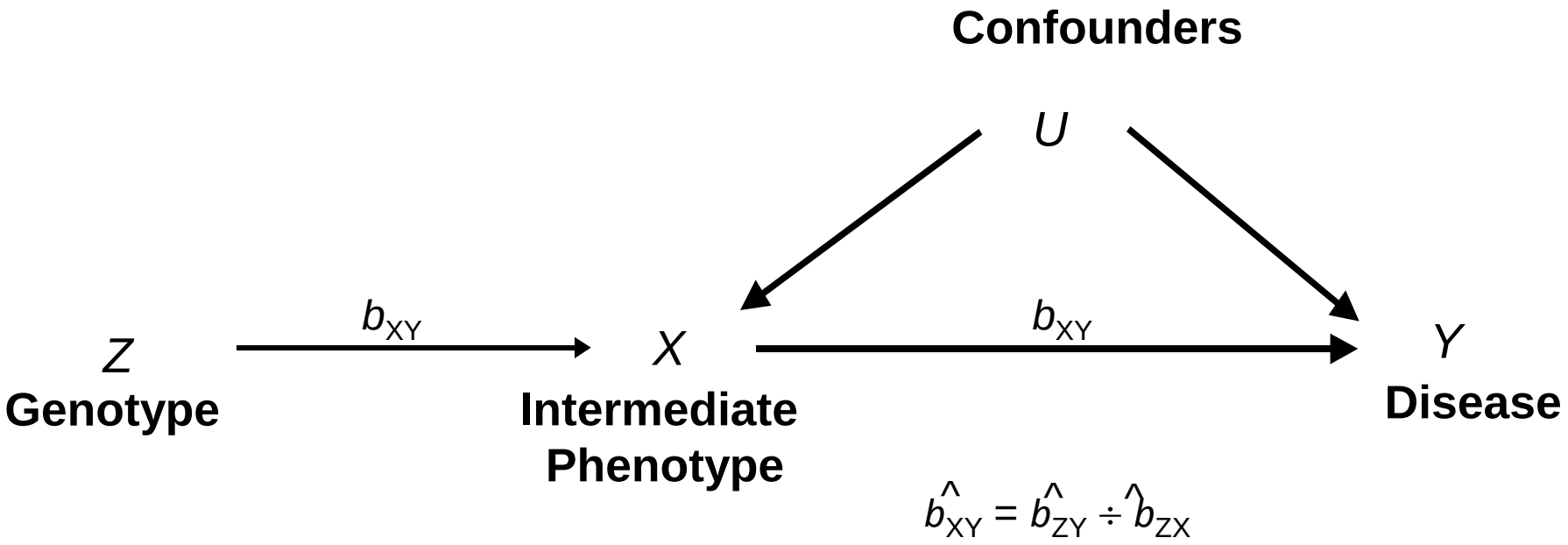
# Mendelian Randomisation principle

- Genes are randomly allocated, independent of confounders
- Genes cause phenotypes, but phenotypes do not cause genes
- Therefore, a gene that causes the exposure of interest can be a proxy for that exposure, without confounding or reverse causation



# Mendelian randomisation analysis

## "Instrumental variable" technique




**Assumptions of IV technique for Mendelian randomisation:**

1. The Instrumental variable (IV) *Z* is associated with exposure of interest *X*
2. *Z* is independent of the confounding factors *U* (that confound X-*Y* association)
3. Genotype is related to the outcome *only* via its association with the modifiable exposure

- For more on Mendelian Randomisation, and other aspects of causal inference:

Causal Inference in Epidemiology: recent methodological developments
November 2013
Duration: One week

http://www.lshtm.ac.uk/study/cpd/causal_inference.html

# Epigenetics

- Heritable information that is not encoded in the DNA sequence
- NB “heritable” often refers to regeneration of cells *within one organism*
- For genetic epidemiology, *transgenerational epigenetics* may be more relevant

# Epigenetic modifications

- Most common: DNA methylation
  - Chemical alteration to DNA molecule, usually a CpG dinucleotide
  - Has a number of effects on gene expression and regulation
  - Can be induced by environment and inherited transgenerationally
  - Allele-specific methylation: SNP associated with methylation status
- Also: histone modification
  - Changes the 3D “wrapping” of DNA inside the nucleus

- These mechanisms provide a way to explain the molecular effects of environmental exposures, and to map the path from genotype to phenotype
  - Rakyan et al, Nat Rev Genet 2011; Daxinger & Whitelaw, Nat Rev Genet 2012

# Prospects

- Most current activity is still in finding genes that cause disease
  - So far, few traits have more than 25% heritability explained
  - Account for the “missing heritability”
- Even where associations have been validated, the causal variants have not been identified
  - Only have SNPs in linkage disequilibrium with causal variant
  - Needs genotyping in multiple populations, sequencing, and functional biology
- However, as we have seen, the identification of disease genes is leading to increased interest in applications to public health