

The design and specification of biobanks

Paul Burton

Professor of Genetic Epidemiology
University of Leicester



P³G Consortium

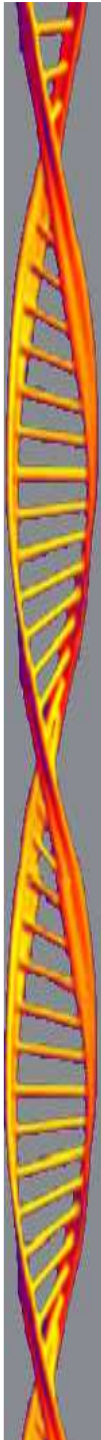


PHOEBE

PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

UK Biobank

biobank^{uk}
Improving the health of future generations





Thanks!

- Anna Hansell and Paul Elliott
 - Imperial College
- Isabel Fortier
 - P³G



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations



Structure of talk

- What determines the size and shape of a genetic epidemiology study?
- The statistical power of case-control studies
- Expected event rates in large cohort studies
- International biobank harmonization



What determines “size” and “shape”?

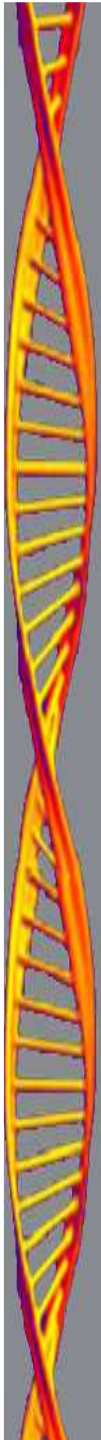


PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

What determines “shape”?

- The scientific question
 - Unrelated individuals v families
 - Association v linkage
 - CDCV v rare alleles with large effects
 - Case-control v cohort designs
 - Classical public health research
 - Special populations
- Pragmatic opportunities and challenges
 - Record linkage
 - Special approaches to recruitment
- Ethico-legal considerations



What determines “size”?

- The scientific question
 - Statistical power
 - Type of end-point
 - Main effects v interactions
 - Time required to generate enough cases
- Cost and resources
- Pragmatic restrictions



How large is “LARGE”?



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations



Two classes of cohort biobanks

- Very large
 - Primary focus on binary end-points
 - Nested case-control studies
 - Hundreds of thousands of recruits
 - e.g. UK Biobank, LifeGene, Kadoorie Study
- Large
 - Primary focus on quantitative end-points
 - Tens of thousands of recruits
 - e.g. ALSPAC, Generation Scotland, CARTaGENE



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE



The £59,000,000 question!!!

- Just how big does a cohort-based biobank have to be?
 - Interest in binary disease related events and (some) binary exposures
 - Middle aged recruits (40-69 years)
 - Population-based recruitment



The statistical power of case-control studies

- Contemporary pre-eminence of genetic association studies rather than genetic linkage studies
- Covers **both** stand-alone case-control studies, **and** nested case-control studies in large cohorts. Main issue is the number of cases.
- Sample size determining in **both** settings



Simulation-based power calculations

- Work with the least powerful (**common**) setting
 - Disease outcome and exposures all binary
- Logistic regression; interactions = departure from a multiplicative model
- Four controls per case
- Complexity (arbitrary but realistic)

Formal power calculations

- Realistic bio-analytic complexity
 - Logistic regression
 - Assessment errors, frailty, $p < 10^{-4}, 10^{-7}, 10^{-10}$
 - ≈ 4 controls per case

ESPRESSO: (**E**stimating **S**ample-size and **P**ower in **R** by **E**xploring **S**imulated **S**tudy **O**utcomes).

<http://www.p3gobservatory.org/powercalculator.htm>

See also: Paul R Burton; Anna L Hansell; Isabel Fortier;
Teri A Manolio; Muin J Khoury; Julian Little; Paul Elliott.
International Journal of Epidemiology 2008;
doi: 10.1093/ije/dyn147



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

How small is “small”?

Most in range:
1.1 – 1.5

Disease	Gene	Polymorphism	Approximate frequency of the disease associated allele	Approximate odds ratio for disease associated allele	Ref
Thrombophilia	<i>F5</i>	Leiden Arg506Gln	0.03	4	12
Crohn's disease	<i>CARD15</i>	3 SNPs	0.06(composite)	4.6	67
Alzheimer's disease	<i>APOE</i>	$\epsilon 2/3/4$	0.15	3.3	13,68
Osteoporotic fractures	<i>COL1A1</i>	Sp1 restriction site	0.19	1.3	69,70
Type 2 diabetes	<i>KCNJ11</i>	Glu23Lys	0.36	1.23	71
Type 1 diabetes	<i>CTLA4</i>	Thr17Ala	0.36	1.27	72,73
Graves' Disease	<i>CTLA4</i>	Thr17Ala	0.36	1.6	74
Type 1 diabetes	<i>INS</i>	5' VNTR	0.67	1.2	75
Bladder Cancer	<i>GSTM1</i>	Null (gene deletion)	0.70	1.28	76
Type 2 diabetes	<i>PPARG</i>	Pro12Ala	0.85	1.23	11

Hattersley AT, McCarthy MI. Lancet 2005;366:1315-1323
Examples of some polymorphisms or haplotypes that have shown consistent association with complex disease



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

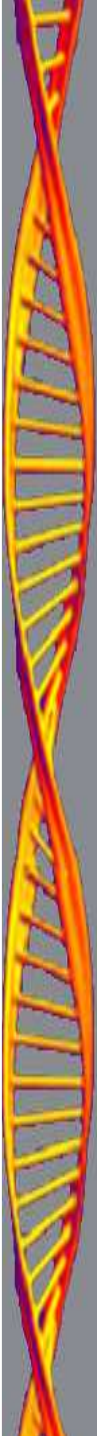
biobank^{uk}
Improving the health of future generations



Recent findings***

Type 1 diabetes^{1,2}
Type 2 diabetes^{2,6}
Coronary heart disease^{2,7-9}
Breast cancer^{10,11}
Colorectal cancer¹²⁻¹⁴
Prostate cancer^{15,16}
Age-related macular degeneration¹⁷⁻¹⁹
Crohns disease^{2,20}

***See full reference list in reserve slides



An example:
Diabetes mellitus defined by
Hba1C > 97.5 percentile



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

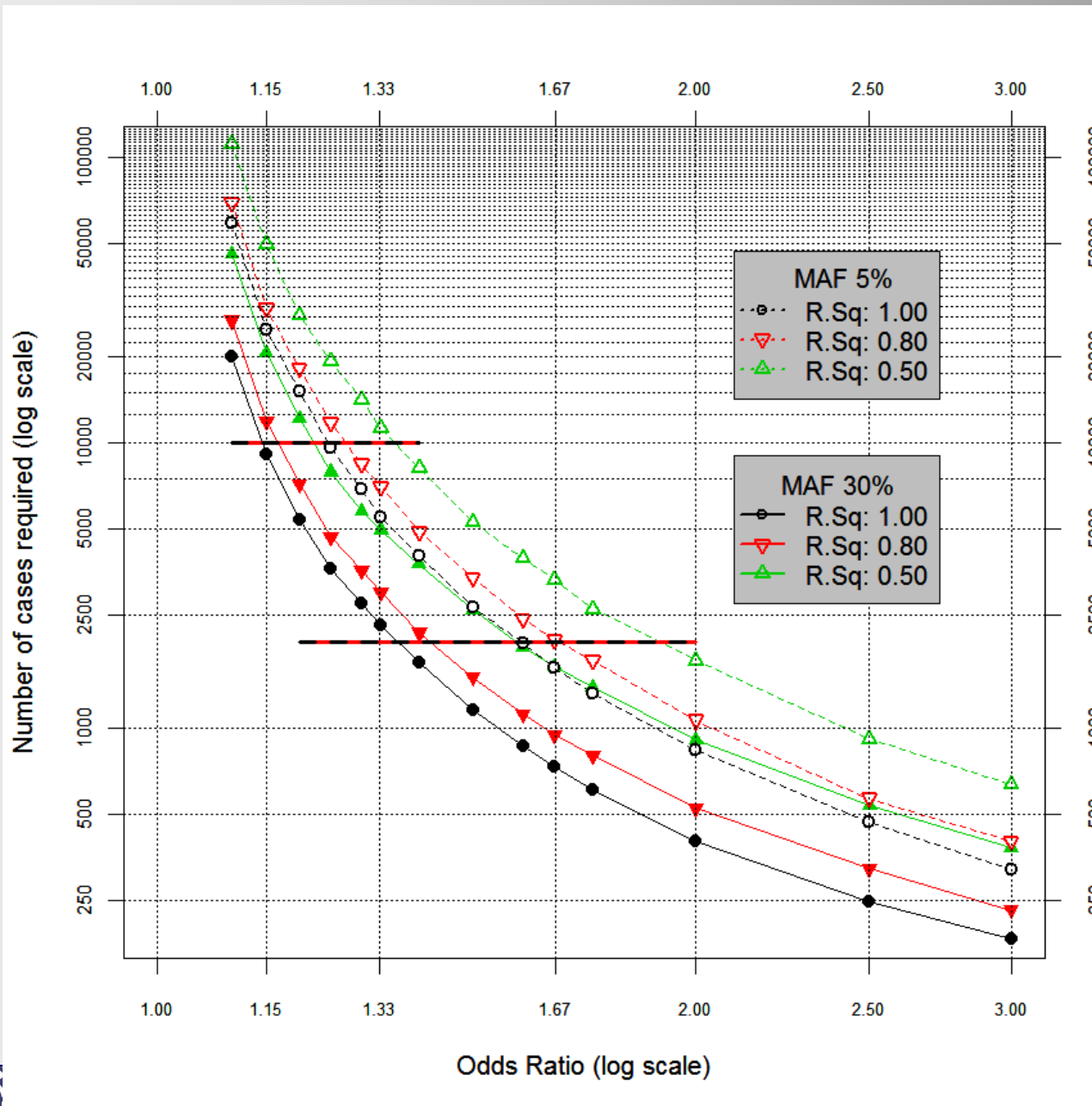
biobank^{uk}
Improving the health of future generations



Simulation-based power calculations

- Complexity (arbitrary but realistic).

Frailty variance	Genotyping error	Environmental error	Sensitivity disease phenotype	Specificity disease phenotype	Critical P-value	Power
10 fold	$R^2 = 0.5, 0.8$	Reliability = 0.3-1.0	89%	97.4%	10^{-4} 10^{-7}	80%



Genetic main effects

Vague candidate:
 $p < 10^{-4}$

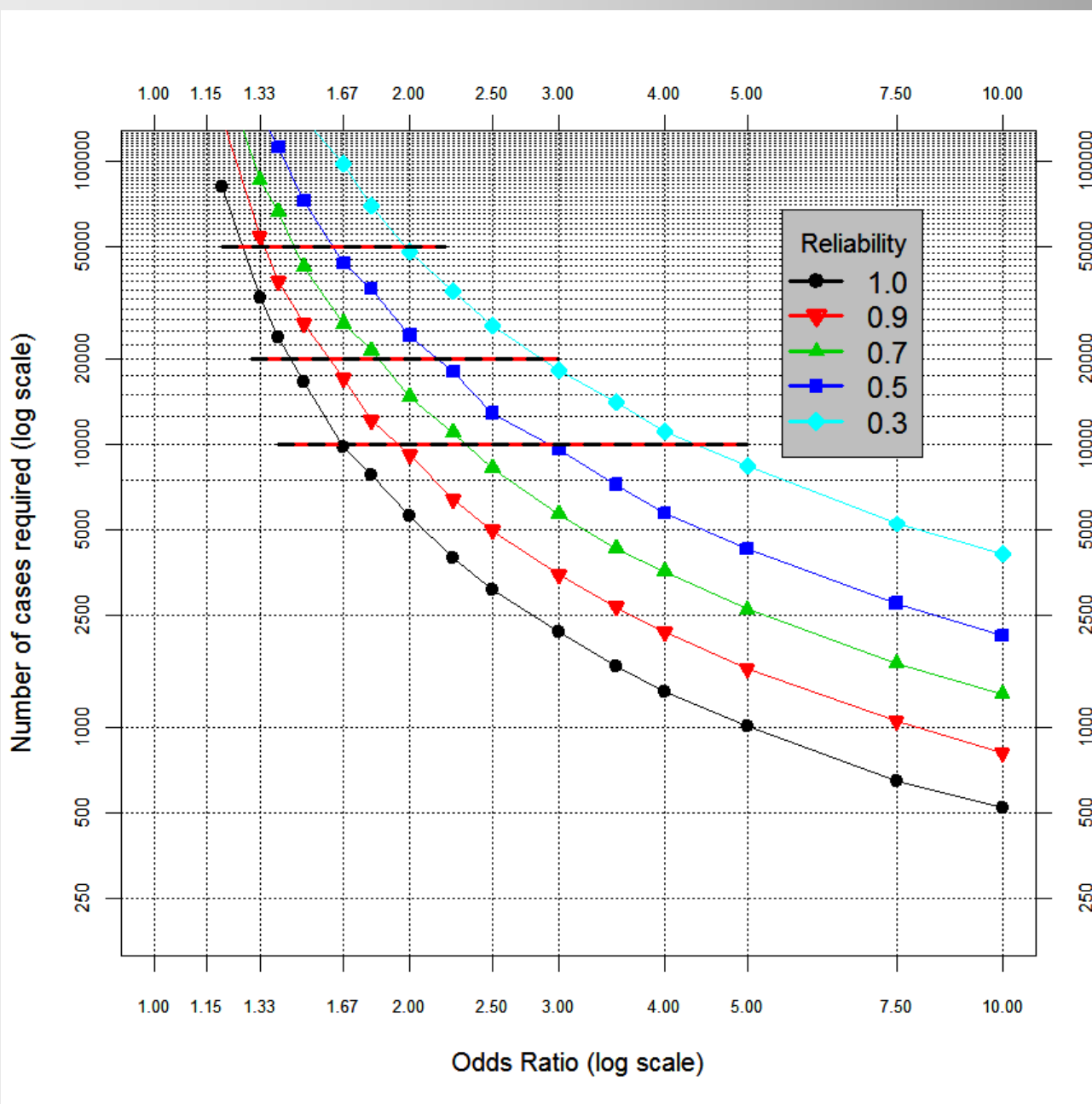
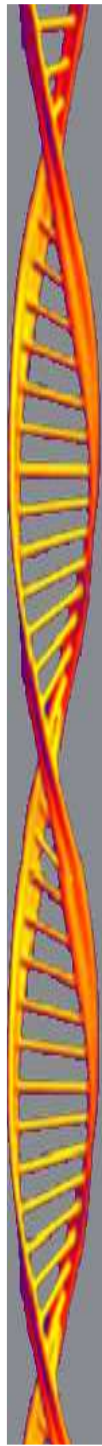
GWA: $p < 10^{-7}$
 $\rightarrow \approx 1.7 \times N$

Valid additive genetic model
 $\rightarrow \approx 0.8-0.9 \times N$

Gene-lifestyle interactions

Table 2 Formal estimates of test-retest reliability for a number of exemplar lifestyle/environmental determinants that are widely studied

Reliability of measurement	Lifestyle/environmental factor
≥ 0.95	Body mass index (BMI) calculated from measured height and weight in various studies ⁷⁶
~ 0.9	Measured hip or waist circumference ^{76,77}
~ 0.7	Blood pressure measurement in the Intersalt Study ⁷⁸
~ 0.5	Many nutritional components in a dietary recall study, mean of four 24 h assessments ⁷⁹
~ 0.3	Many nutritional components in a dietary recall study, a single 24 h assessment ⁷⁹



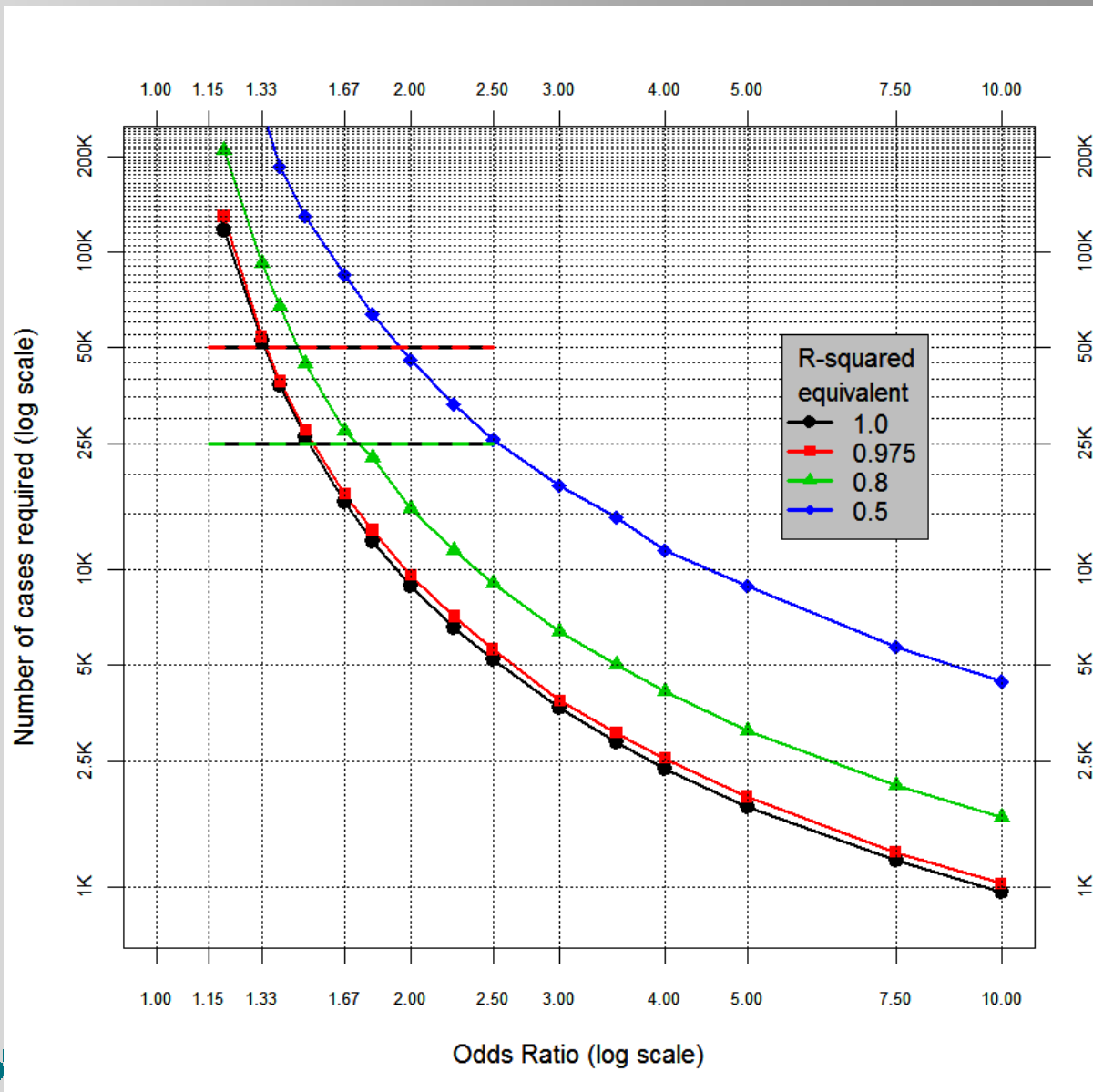
Gene-lifestyle interactions

MAF for 'at-risk' genotype = 5%
 $R^2=0.8$

Prevalence of 'at-risk' life-style factor = 20%

Gene:gene interactions

MAF 5% : 25%
4 controls/case
GWA: $p < 10^{-10}$



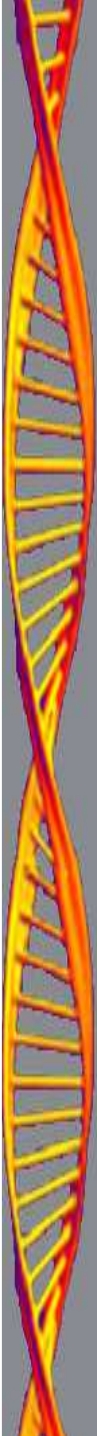
PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank
Improving the health of future generations



Additive genetic model

- Binary or additive genetic model?
- If truly additive, additive model could add *substantial* power
- If truly binary, binary model is *slightly* more powerful
- But, the gain in power is greater when MAF is high
 - When MAF is low, very few subjects are homozygote for MA, and so the locus is almost binary, and the fall in required sample size is small
 - But when MAF is high, power is not such an issue



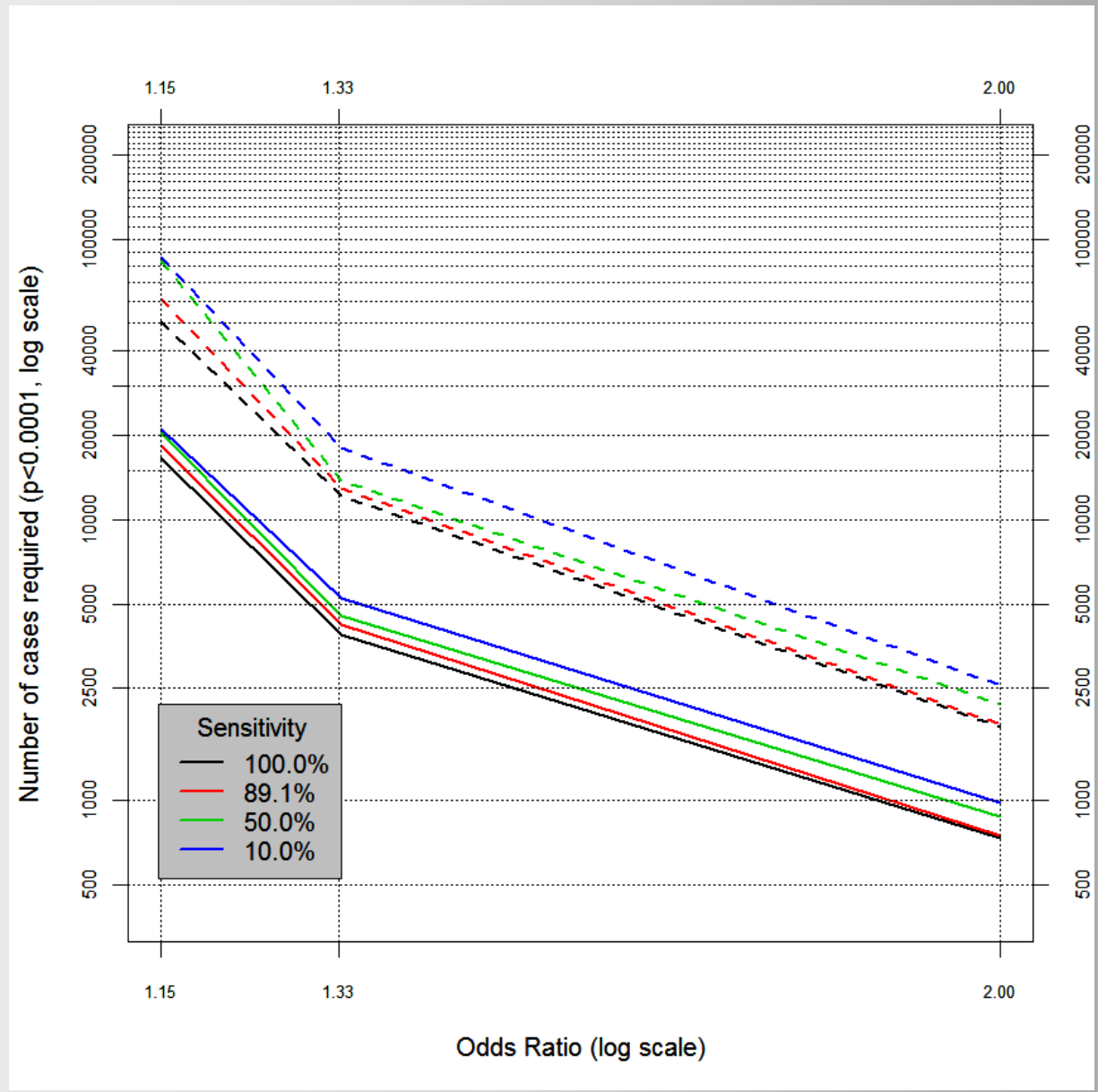
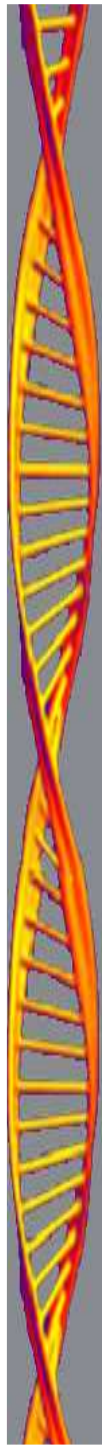
Be specific, not sensitive

but only if you have time!!



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

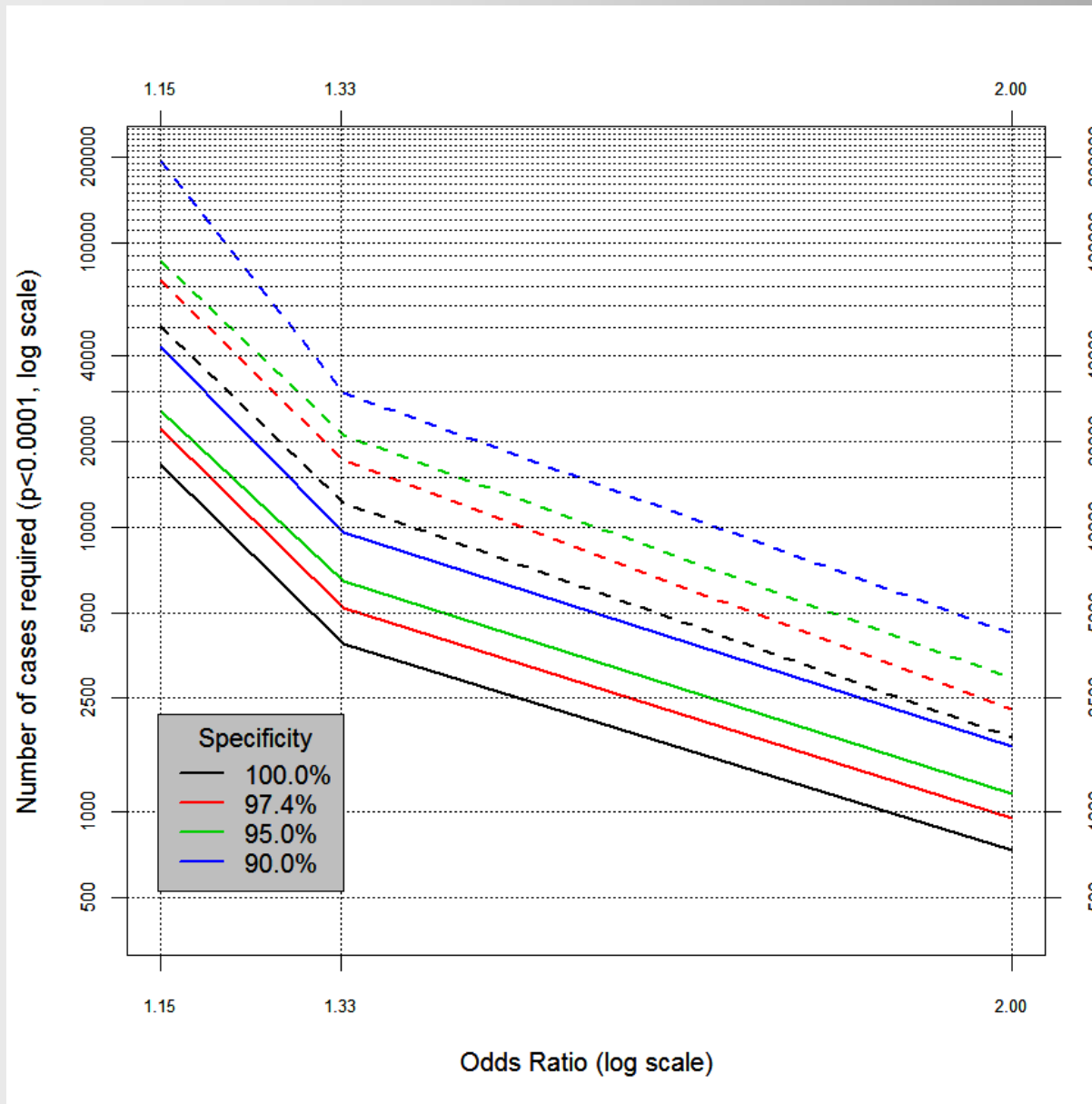
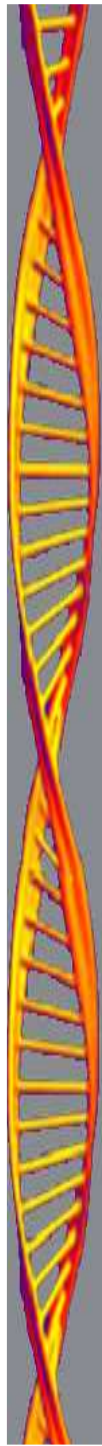
biobank^{uk}
Improving the health of future generations



Specificity
= 100%
low sensitivity

All cases are true cases

Even if sensitivity = 10%,
most controls are from
the large pool of non-
diseased subjects



Sensitivity= 100%
low specificity

All controls are true controls

Even if specificity is as high as 90%, *many* “cases” are from the large pool of truly non-diseased subjects that have been misclassified

How many cases?

- Genetic main effects
 - 2,000 minimum, 5,000 better
- Lifestyle main effects
 - 2,000-20,000
- Gene-lifestyle “interactions”
 - Absolute minimum 10,000, often need at least 25,000, a comprehensive platform needs at least 50,000
- Pooling and replication!!

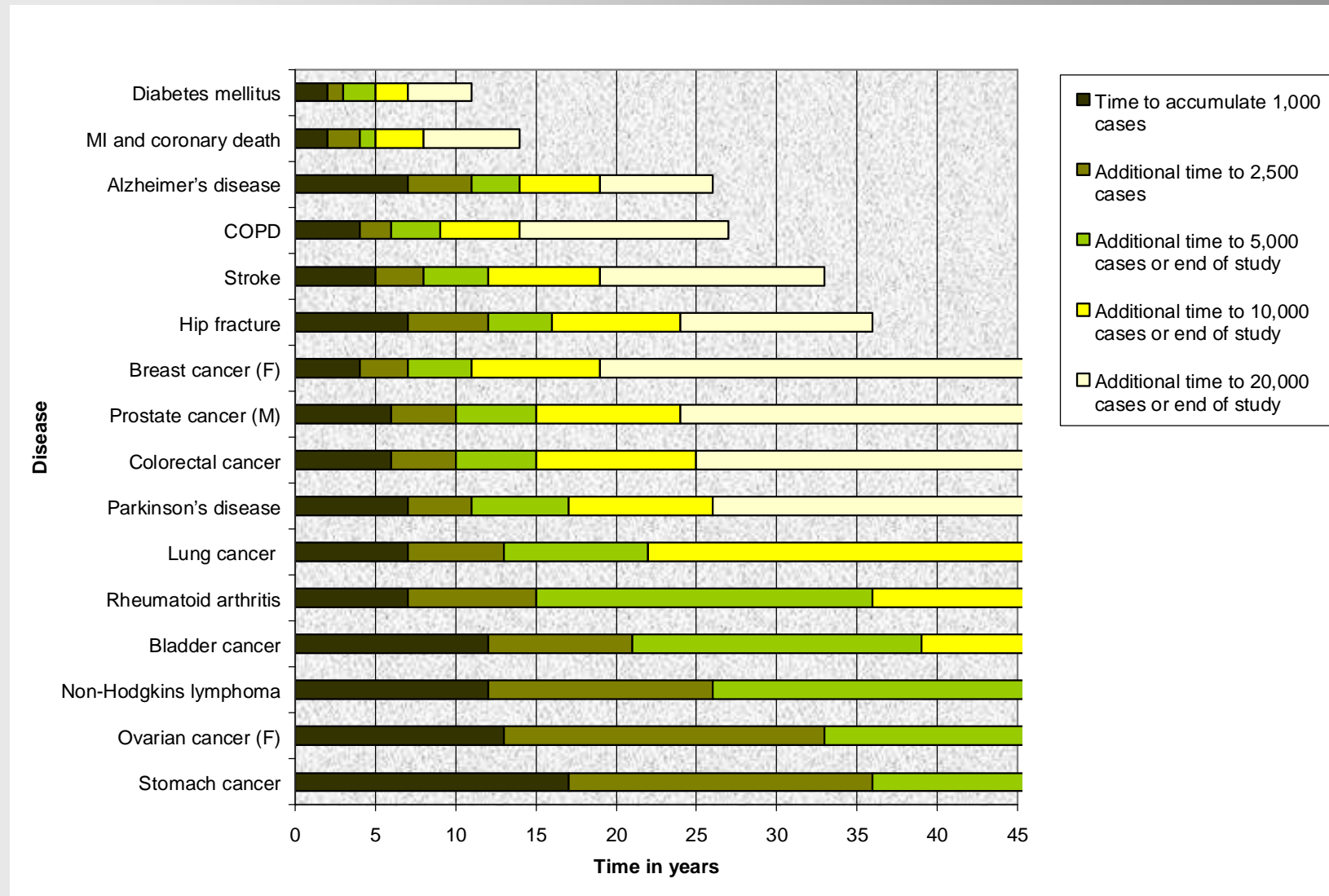




How long is “LONG”?

- Age range at recruitment 40-69 years
- Recruitment over 5 years
- All cause mortality
- Disease incidence (“healthy cohort effect”)
- Migration overseas
- Withdrawal from the study

How long is "LONG"?





So how can we get enough power?

- Can it be achieved at all?
 - Recent successes
- Is genetic epidemiology beyond its limits?
 - Taubes, *Science*, 1995
 - Protection afforded by 'Mendelian Randomization'
- Large disease-based biobanks
- Very large cohort-based biobanks
- CDCV v rare alleles with large effects

So how can we get enough power?

- Conduct studies with optimal designs
 - Enhance the quality of individual studies.
 - When relevant, use continuous disease-related traits and health determinants.
- Increase the size of individual studies
- Promote the conduct of meta-analysis
 - **Sharing results** (traditional meta-analysis): Ideally based on published and unpublished results.
 - **Sharing raw data and samples**: Need to promote harmonization between biobanks to enable pooling of raw information.



International biobank harmonization programs

- P³G
 - Public Population Program in Genomics (P³G)
- PHOEBE
 - Promoting Harmonization Of Epidemiological Biobanks in Europe
- BBMRI
 - Biobanking and BioMolecular Resources Research Infrastructure
- ISBER
 - International Society for Biological and Environmental Repositories
- HuGENet
 - Human Genome Epidemiology Network



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

Biobanks associated with P³G



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

Number of participants targeted (recruited or to be recruited) (N=87)

Number of participants	Number of studies	Number of participants TARGETED
Less than 49 000	47	900,000
50 000 to 99 000	14	1,000,000
100 000 to 499 000	18	2,800,000
500 000 and more	8	4,900,000
		Total: 9,600,000

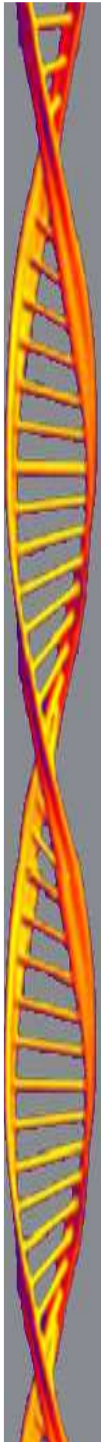


Thanks!!



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations



PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

Reserve slides



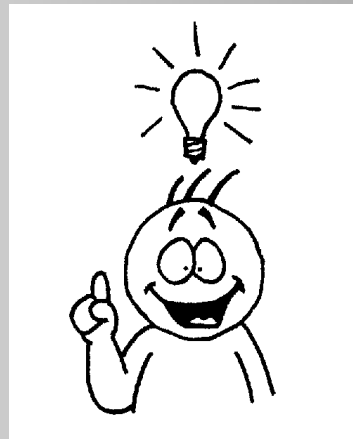
PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

Biobank harmonization

- “A set of procedures that promote, both now and in the future, the effective interchange of valid information and samples between a number of studies or biobanks, accepting that there may be important differences between those studies”

“I understand”



“I UNDERSTAND”



“I comprehend”



“vx jhmkaiwb”

“Je comprends”

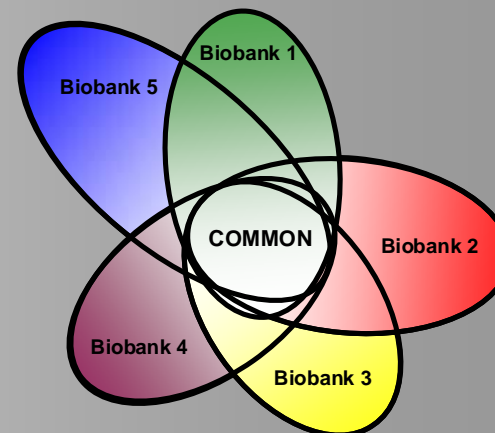


PROMOTING
HARMONISATION OF
PHOEBE
EPIDEMIOLOGICAL
BIOBANKS IN EUROPE

biobank^{uk}
Improving the health of future generations

Information common to all the cohorts and information specific to some of them

- Questionnaire
- Physical and cognitive measures
- Environmental measures
- Biochemical measures
- Governmental databases
- ***DATASHaPER***
 - Template to facilitate harmonization between pre-existing biobanks and support the design of emerging ones.



**At recruitment
and during the follow-up**

References for slide 16

- 7. Wellcome_Trust_Case_Control_Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661-678.
- 32. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007;39(7):857-864.
- 33. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of Genome-Wide Association Signals in U.K. Samples Reveals Risk Loci for Type 2 Diabetes. *Scienceexpress* 2007;10.1126:1-4.
- 34. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316(5829):1331-6.
- 35. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316(5829):1341-5.
- 36. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Analysis of two genome-wide association studies identifies and validates novel gene loci for myocardial infarction. *New England Journal of Medicine* 2007:in press.
- 37. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007;316(5830):1491-3.
- 38. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007;316(5830):1488-91.

- 
- 39. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;advanced online publication.
 - 40. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007;39(7):865-9.
 - 41. Haiman CA, Le Marchand L, Yamamoto J, Stram DO, Sheng X, Kolonel LN, et al. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* 2007.
 - 42. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007.
 - 43. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007.
 - 44. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 2007.
 - 45. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 2007;39(5):631-7.
 - 46. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;308(5720):385-9.
 - 47. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 2005;308(5720):419-21.
 - 48. Edwards AO, Ritter R, 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. *Science* 2005;308(5720):421-4.
 - 49. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007;39(5):596-604.