Statistical methods for association analyses in the resequencing era

David Balding

Institute of Genetics
University College London

Journée thématique BILGWAS, Nantes, 28 Janvier 2010

GWAS: is the glass half full? or half empty?

Common Disease Common Variant hypothesis verified

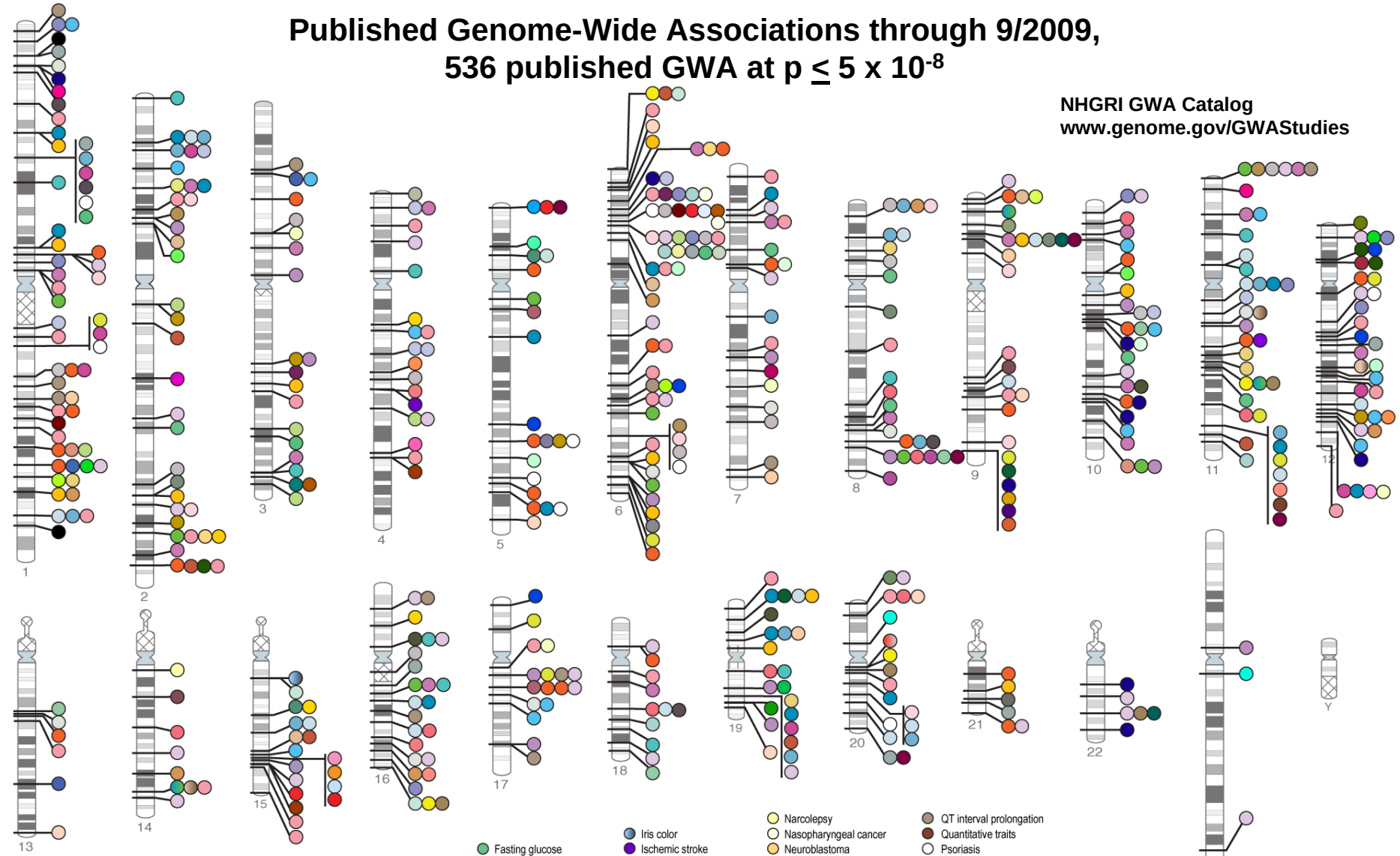
- many new common variants identified

But

- effect sizes small
- explain relatively little of genetic contribution to disease
- valid pointers to mechanisms but
  - weakly associated variants may play a minor role in complex mechanisms
  - perhaps many causal pathways
  - role of rarer variants probably easier to discern
- because of stringent significance thresholds, these form "tip of iceberg"
  - estimates of number of SNPs associated with some complex diseases are $\sim 10^4 - 10^5$.

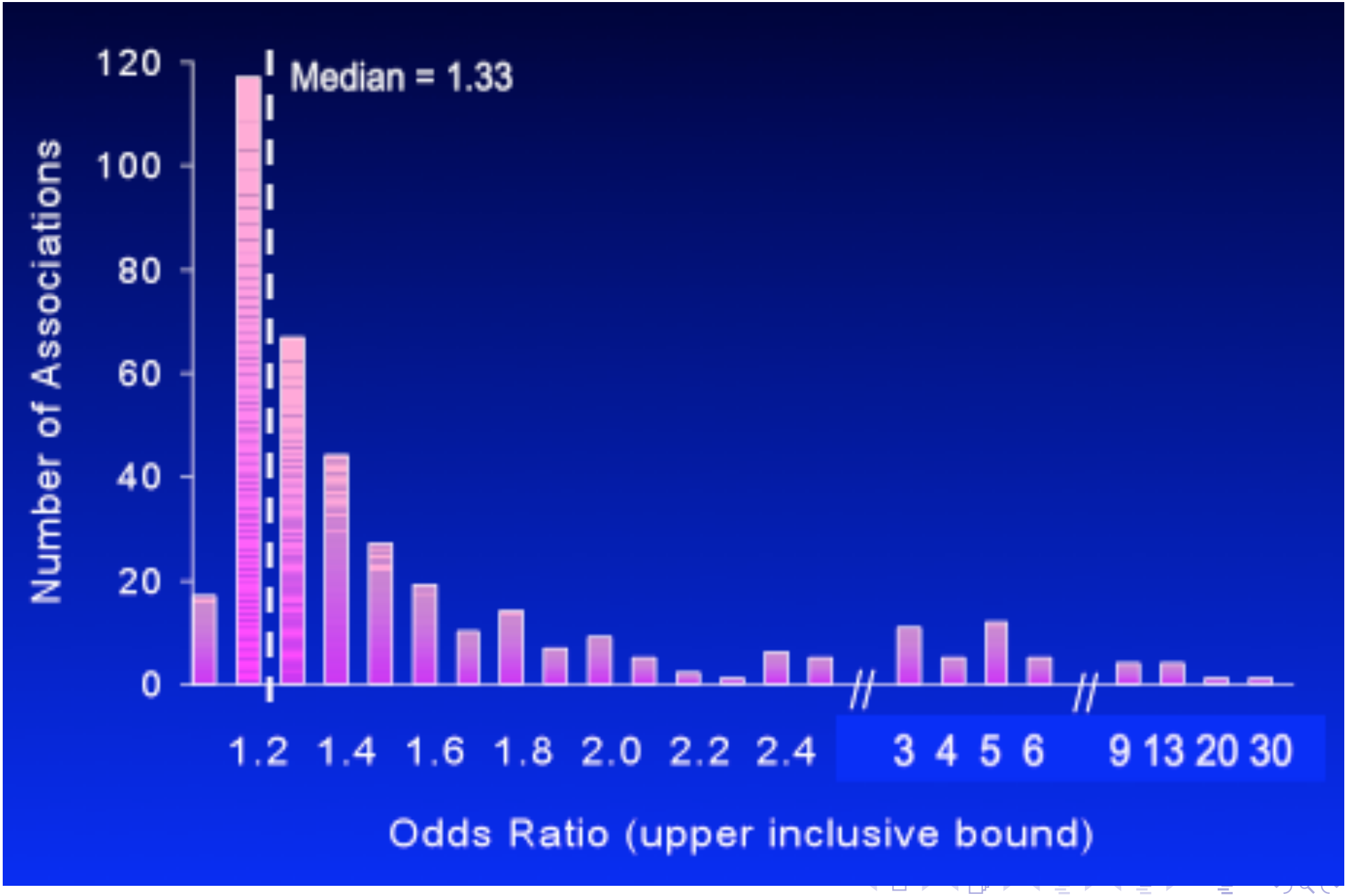
Published Genome-Wide Associations through 9/2009, 536 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies



- Acute lymphoblastic leukemia
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alzheimer disease
- Amyotrophic lateral sclerosis
- Asthma
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Bipolar disorder
- Bladder cancer
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Cardiac structure/function
- Carotenoid/tocopherol levels
- Celiac disease
- Chronic lymphocytic leukemia
- Cleft lip/palate
- Cognitive function
- Colorectal cancer
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Cutaneous nevi
- Dermatitis
- Drug-induced liver injury
- Eosinophil count
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- F cell distribution
- Fasting glucose
- Folate pathway vitamins
- Freckles and burning
- Gallstones
- Glioma
- Hair color
- Heart rate
- Height
- Hepatitis
- Hirschsprung's disease
- HDL cholesterol
- Idiopathic pulmonary fibrosis
- Inflammatory bowel disease
- Intracranial aneurysm
- Iris color
- Ischemic stroke
- Juvenile idiopathic arthritis
- Kidney stones
- LDL cholesterol
- Liver enzymes
- Lung cancer
- Malaria
- Male pattern baldness
- MCP-1
- Mean platelet volume
- Melanoma
- Menarche & menopause
- Multiple sclerosis
- Myeloproliferative neoplasms
- Narcolepsy
- Nasopharyngeal cancer
- Neuroblastoma
- Nicotine dependence
- Obesity
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pancreatic cancer
- Panic disorder
- Peripheral arterial disease
- Plasma LP (a) levels
- Primary biliary cirrhosis
- Prostate cancer
- Protein levels
- Pulmonary funct. COPD
- QT interval prolongation
- Quantitative traits
- Psoriasis
- Recombination rate
- Red vs.non-red hair
- Renal function
- Response to hepatitis C treat
- Restless legs syndrome
- Rheumatoid arthritis
- Schizophrenia
- Serum bilirubin
- Serum IgE levels
- Serum markers of iron status
- Serum metabolites
- Serum urate
- Skin pigmentation by reflectance spectroscopy
- Soluble E-selectin
- Soluble ICAM-1
- Speech perception
- Statin-induced myopathy
- Stroke
- Systemic lupus erythematosus
- Testicular germ cell tumor
- Thyroid cancer
- Total cholesterol
- Triglycerides
- Type 1 diabetes
- Type 2 diabetes
- Venous thromboembolism
- Vitamin B12 levels
- Warfarin dose
- Weight
- YKL-40 levels

Observed effect sizes: NHGRI Catalog of GWA Studies



Where is the missing genetic variation?

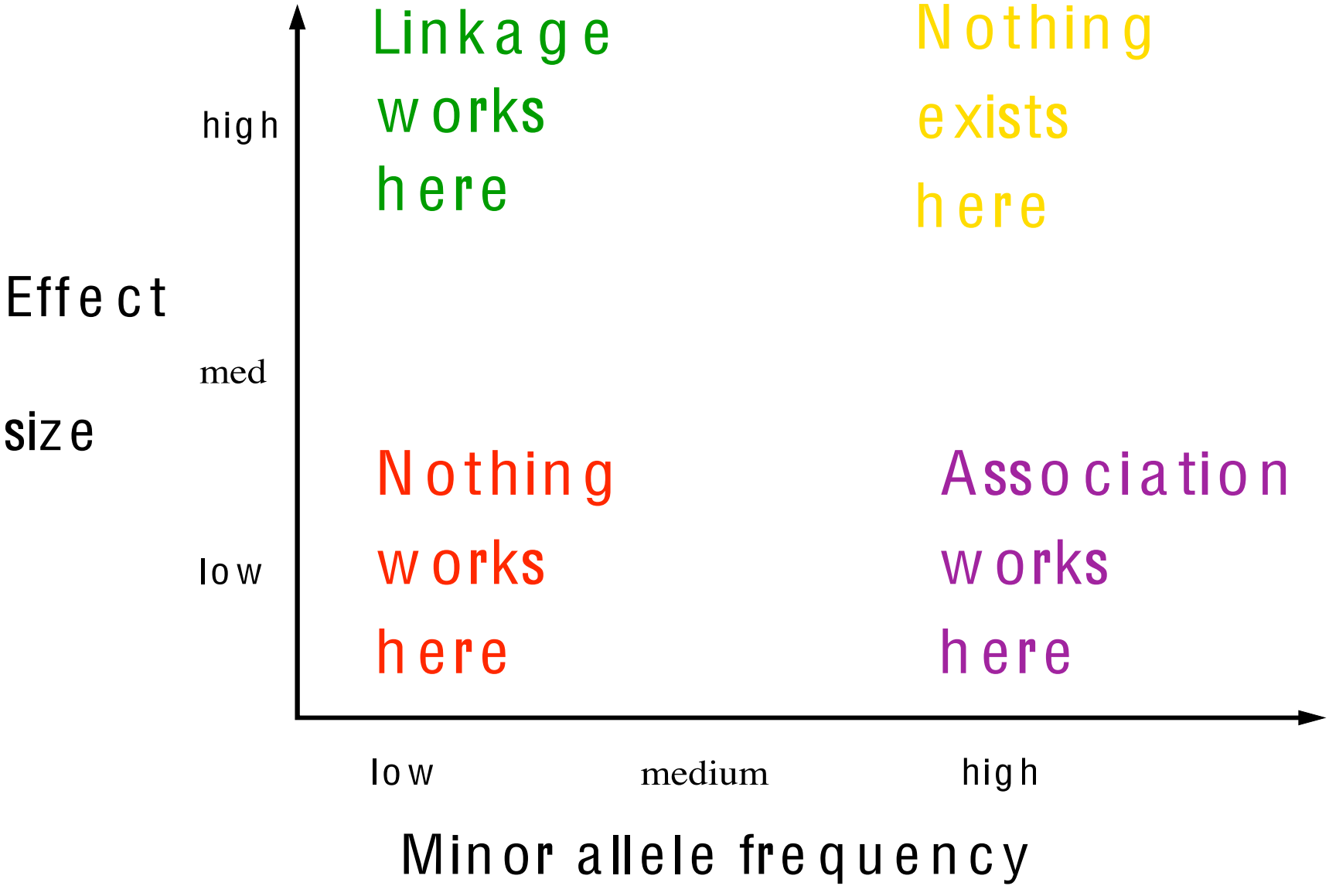
Weak, common variants form part of the story, but there could also be a big role for

- many rare variants of large effect sizes and/or intermediate-frequency alleles of intermediate effects
- copy-number variants
  - WTCCC2: common CNVs have little impact on disease
  - rare CNVs likely to be important
- epigenetic factors
- intermediate phenotypes (e.g. gene expression)

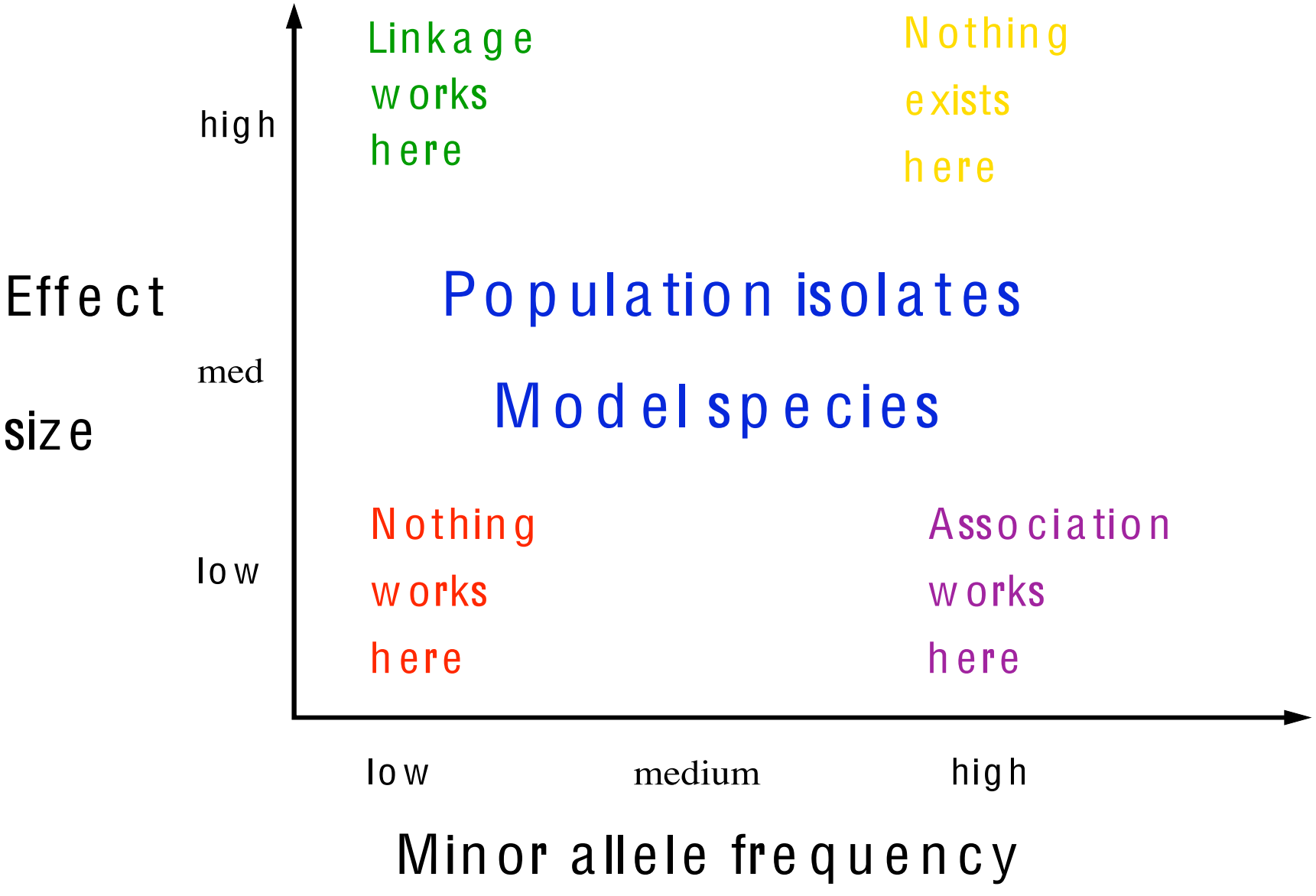
Some of the missing variation may come from ongoing GWAS - better designed and better analysed.

- multi-ethnic studies; admixed populations
- gene-by-environment interactions
  - prospective cohorts
- gene-by-gene interactions
  - pairwise or pathway-based analyses

Effectiveness of linkage and association

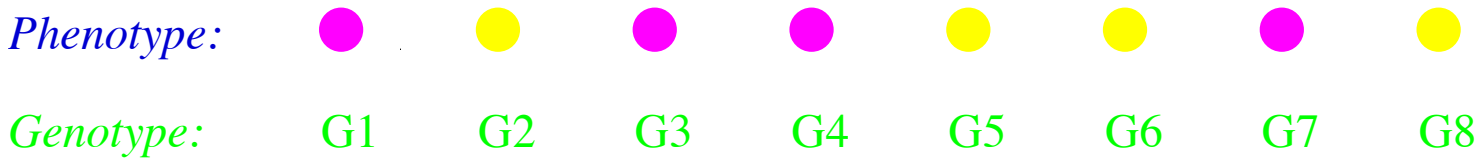


Intermediate variants?

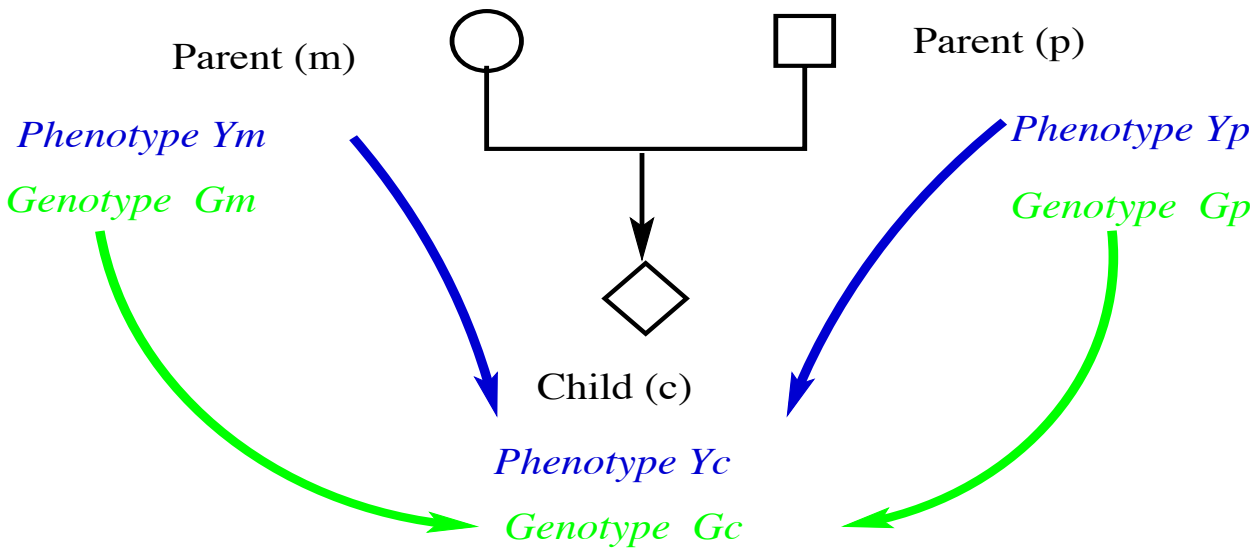


Genetic association studies

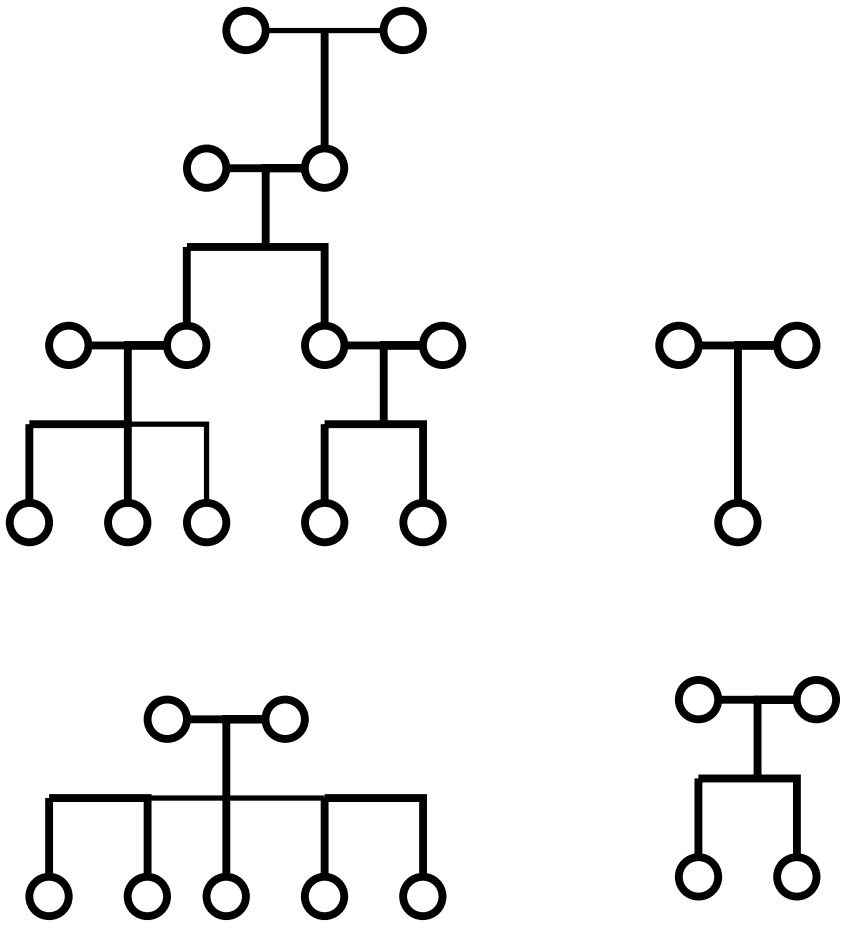
- Seek correlation between phenotype (e.g. disease state or drug response) and genotype, usually in "unrelated" individuals (the relationship is unknown and assumed to be distant).



- Contrast with linkage studies which look for correlation between phenotype and parent-child *transmissions* of alleles.

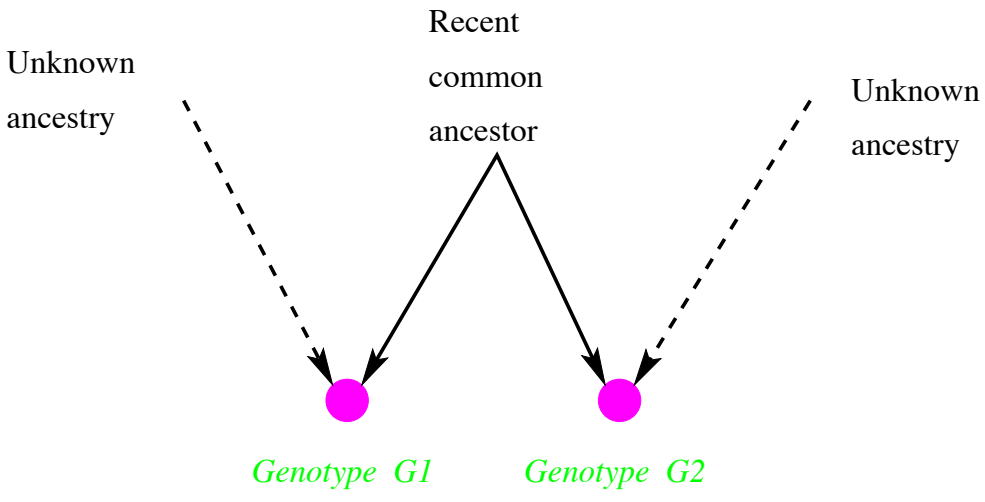


Rationale for a linkage study



- We can do linkage in all kinds of pedigrees

- it isn't necessary for transmissions over generations to be directly observed
- e.g. affected sib pairs (ASP):



Pros and cons of Linkage

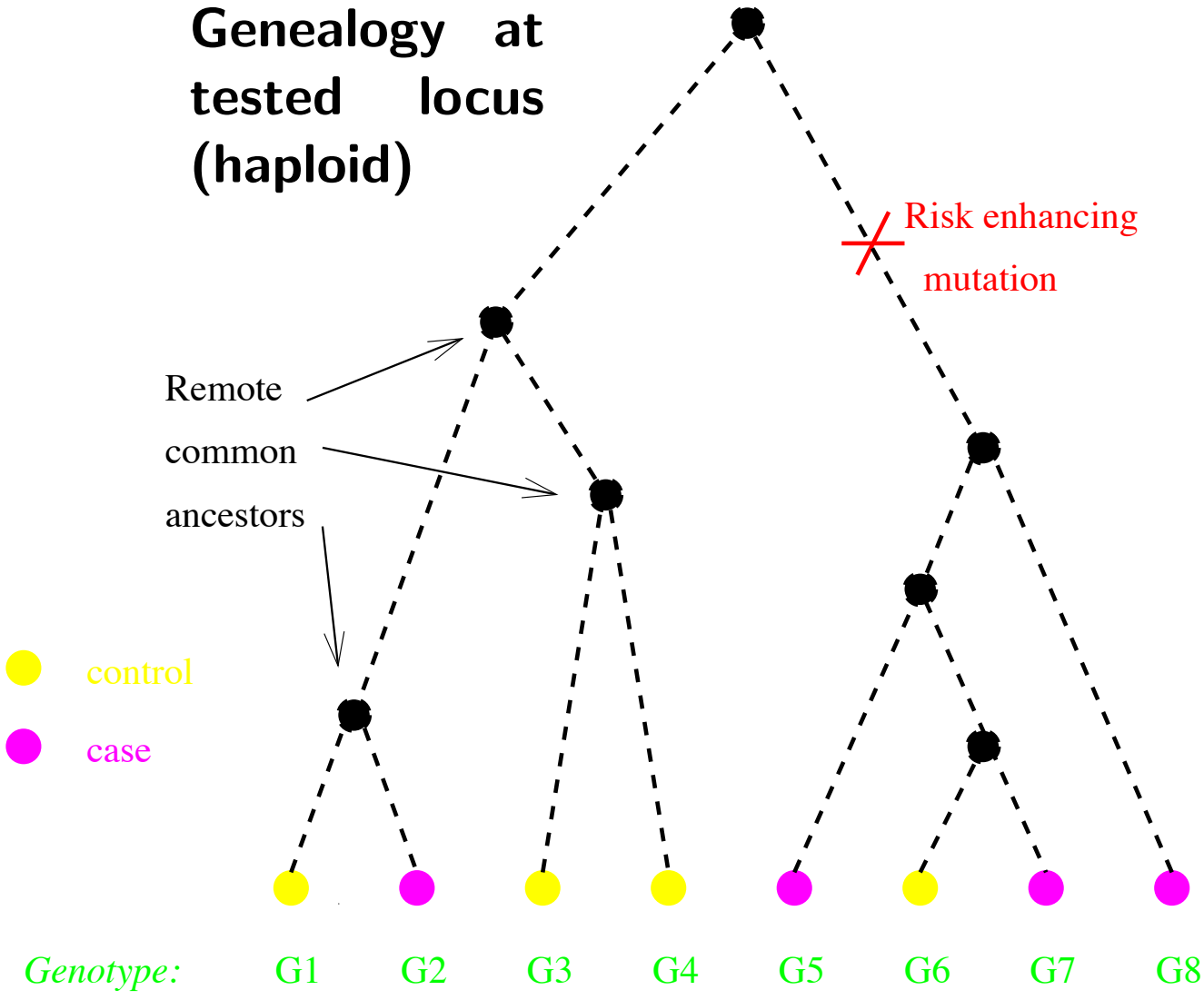
Pro:

- Conditioning on the known pedigree provides protection from confounding (we look for patterns of transmission that deviate from those expected given the pedigree).
- Need fewer markers.
- Larger linked region $\Rightarrow$ higher prior probability of linkage (less multiple testing problem) $\Rightarrow$ smaller sample sizes / greater power.
- Good for rare, high-penetrant causal alleles, since cases are then concentrated in families.

Con:

- Need to recruit study participants of known relatedness.
- Known pedigrees $\Rightarrow$ few meioses $\Rightarrow$ crude localisation $\Rightarrow$ greater fine-mapping problem.
- Ineffective for diseases of complex genetic etiology.

Rationale for an association study



Correlated transmissions of genotype and phenotype generate the associations that we seek in both linkage and association.

In effect we infer transmissions of a risk-enhancing allele from a remote unobserved ancestor in a case-rich group of study subjects with similar genotypes/haplotypes.

- similar to ASP study.

So what's the difference between linkage and association?

- linkage conditions on known pedigree when inferring IBD;
- in association, we infer shared ancestry at a locus ignoring the pedigree.

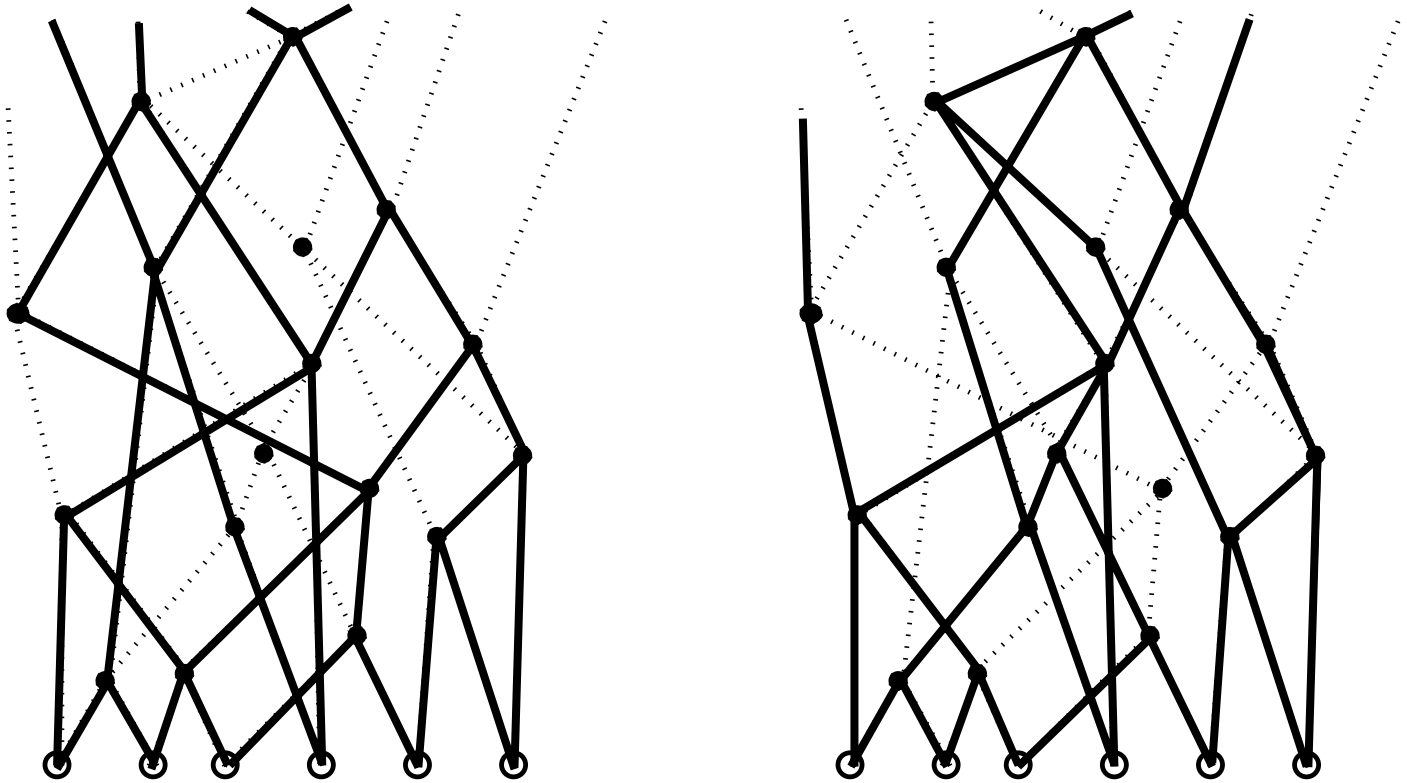
Inference without pedigree information brings **advantages**:

- can exploit remote shared ancestors $\Rightarrow$ many recombinations leading to fine-scale mapping;
- fewer constraints on ascertainment of study subjects
  - can enrich for rare phenotype, e.g. case-control design.

But also comes with **costs**

- allelic heterogeneity makes inference of ancestry difficult
  - best suited to common variants;
- lose possibility to track parent-of-origin and maternal effects;
- phenotype-related genotyping error can confound;
- unobserved pedigree may have structure that is associated with phenotype $\Rightarrow$ *confounding by population structure or cryptic relatedness*.

Unobserved pedigree is a confounder for association studies



solid: lineages of study subjects at two loci;

dashed: pedigree relationships not part of any lineage at *this* locus.

Two genealogies at distinct loci in the same individuals. They:

- are embedded in the same underlying pedigree;
- are independent at unlinked loci conditional on the pedigree;
- are not *unconditionally* independent.

How does the pedigree confound?

Polygenic inheritance:

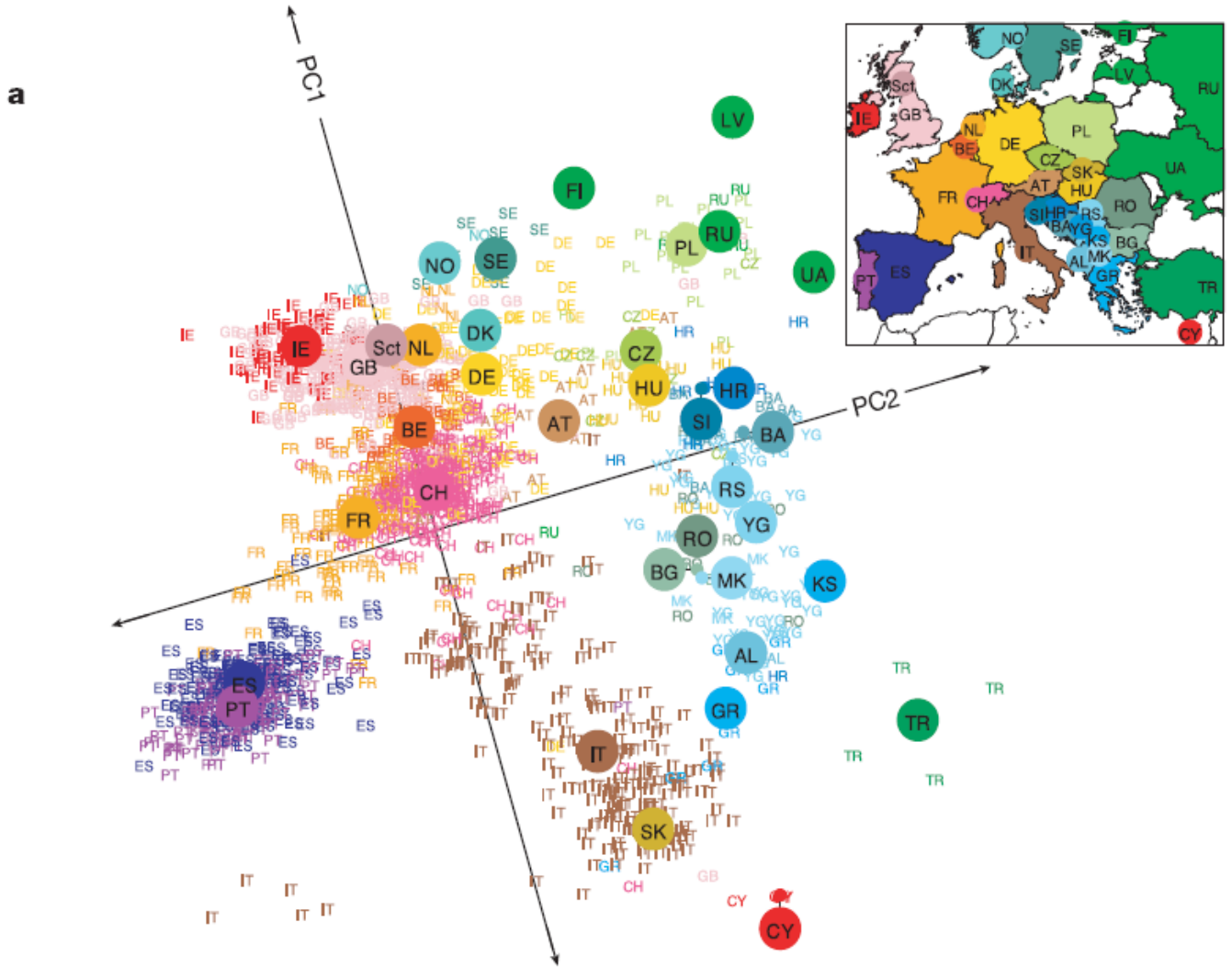
- Many loci distributed genome-wide each make a small contribution to phenotype;
- Genealogies at these loci are correlated because they are all constrained to follow the same pedigree;
- Pattern of association at a locus may be correlated with the pedigree and so signal may in part arise from polygenic inheritance.
  - e.g. if there is a North-South frequency gradient for an allele in Europe, it will be correlated with any N-S varying phenotype. But this also reflects the pedigree structure of Europeans $\Rightarrow$ confounding.

This is usually discussed as the problem of **population structure**

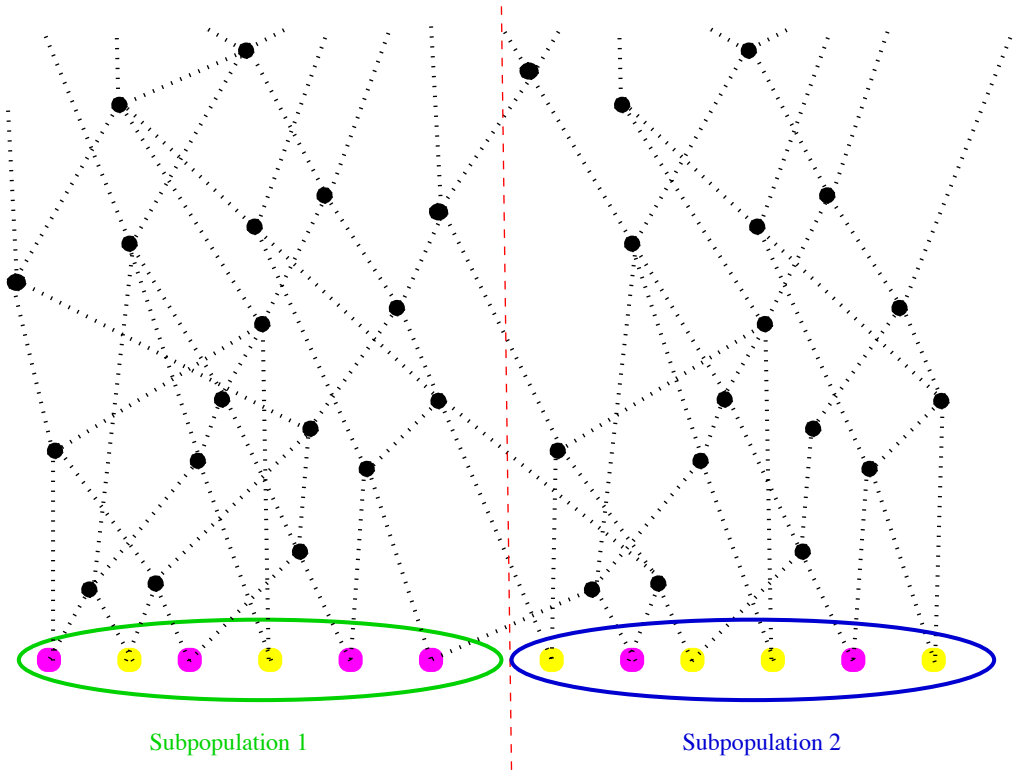
- but an "island" subpopulation model describes only one form of pedigree structure;
- another important special case is *cryptic relatedness*.

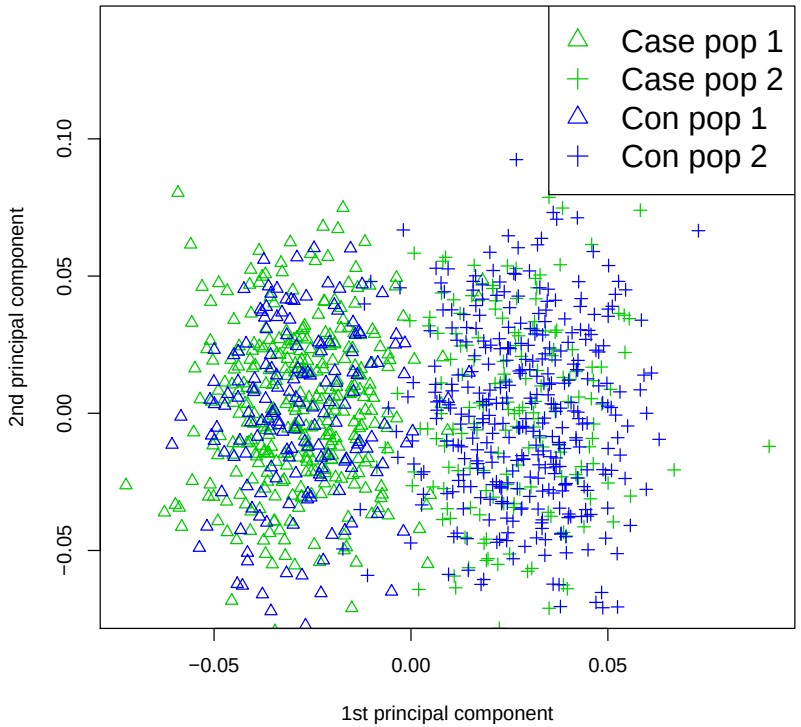
Principal Components Analysis of Europe

Genes Mirror Geography in Europe, Novembre *et al.* Nature 2008.



Principal components can measure major pedigree effects...





Major structure here is the two subpopulations: reflected in 1st PC.

.... and adjust for confounding

- Including leading PCs as regression predictors then removes from the test of SNP association any signal of phenotype that can be attributed to large-scale pedigree structure;
- generally an effective approach but
  - does not deal with finer-scale population structure or cryptic relatedness
  - can discard useful information

Solutions to pedigree confounding: 2 Linear mixed models

Tackle the pedigree confounding issue directly:

- the matrix of (estimated) pairwise kinship coefficients K gives a much richer description of the pedigree than leading PCs;
- many parameters to estimate, but OK for GWAS data.

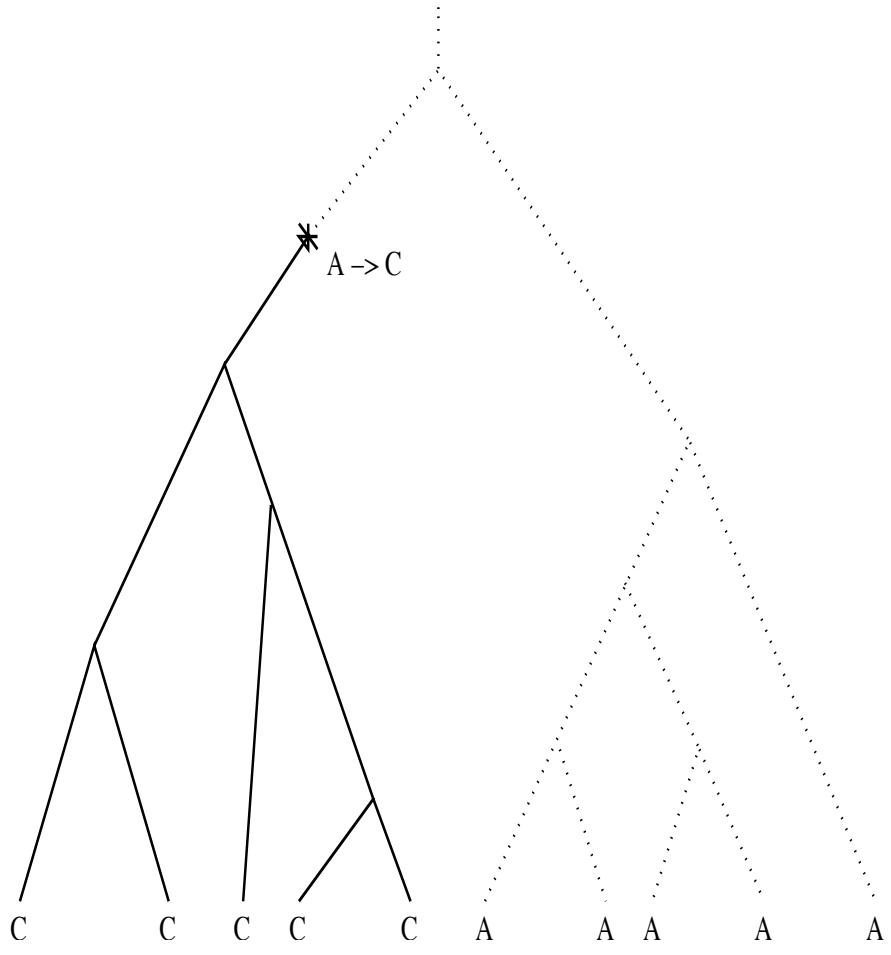
Include K in inference via a mixed model

$$y_i = \alpha + \beta x_i + u_i + e_i$$

- without u_i this is standard (prospective) regression model
  - β is the SNP effect parameter of interest;
- assume the random effects u_i have correlation $\propto K$ (the constant specifies the heritability).
  - **Intuition:** the correlation structure of y that can be attributed to additive polygenic effects is removed from inferences about β.

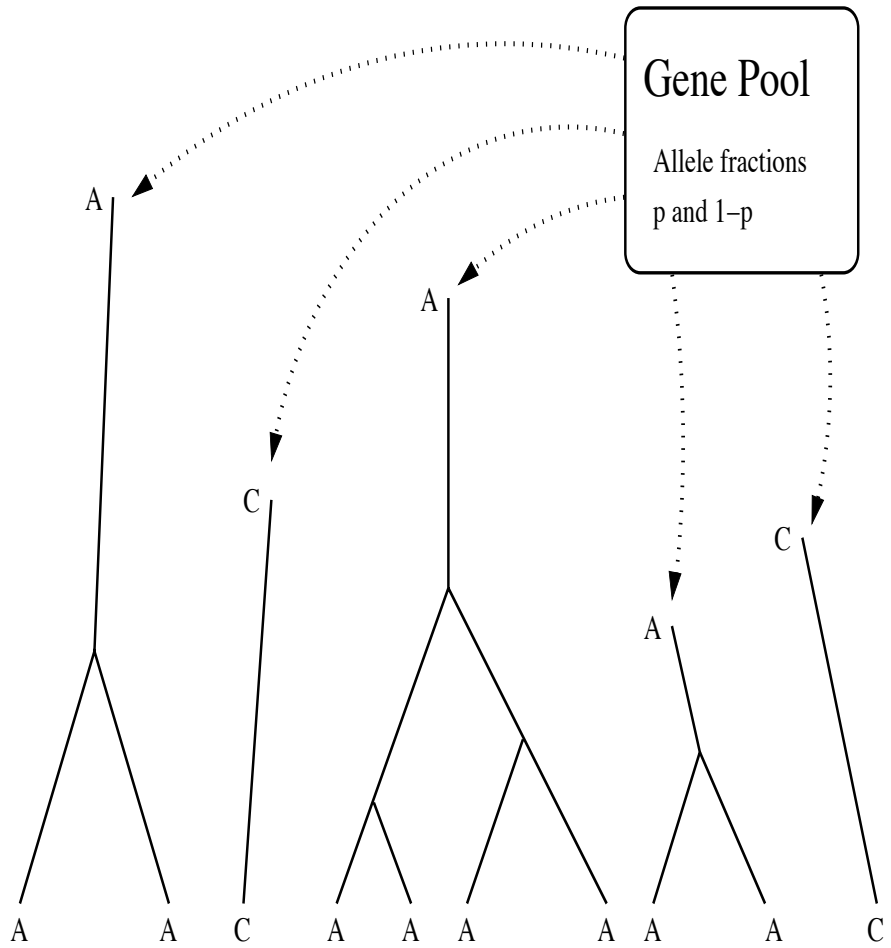
Estimating kinship from GWAS markers: 1. Total allele sharing (IBS = IBD).

- arises under assumption that IBS $\implies$ IBD; common assumption for SNPs, not always true;
- simple Method of Moment estimator;
- an allele counts the same whether it is rare or common;
- "Unrelated" individuals (all shared ancestors are remote) have kinship $\neq 0$



Estimating kinship from GWAS markers: 2. Kinship = allelic correlation.

- arises under population genetics models in which two alleles are *either* IBD *or* independent choices from a hypothetical gene pool with known allele fractions;
- simple Method of Moment estimator;
- interpretation as excess allele sharing: allele fraction p is important.
- estimation of p can lead to downward bias in kinships; negative values possible.



Allelic correlation estimator of K

- Allelic correlation estimator uses p and is more efficient than IBS estimator:
  - 40% lower s.d. in a small simulation study (Astle & B, *Statist Sci* 2010).
- Negative values are considered unpalatable by some but bias less important here than variance.
- Allelic correlation is natural to model phenotypic correlation.

We consider here only the 1-parameter kinships, defined in terms of IBD of alleles drawn at random, one from each individual.

Can also consider 2-parameter kinships that estimate probabilities of 1 and 2 IBD alleles — required to model dominant polygenic effects.

Principal components of $\hat{K}$

The allelic-correlation estimator $\hat{K}$ is the same matrix from which principal components are derived.

PC adjustment uses only the leading eigenvectors of K

- implies a prior that is diffuse for the first k eigenvector regression coefficients and concentrated at zero for remaining $n-k$;
  - gives more flexibility e.g. to adjust for SNP-specific selection effects correlated with a leading eigenvector
  - but truncation at k eigenvectors is unsatisfactory, can miss important structure;
- mixed model approach uses all the eigenvectors, with prior variance proportional to eigenvalues;
- better approaches possible: less shrinkage for leading e'vectors, truncation of trailing e'vectors to minimise noise.

See Astle & B (2010), also McVean paper in recent PLoS Genetics.

The current position

- The population is one big family but the pedigree is unknown.
- We can infer its principal features, and adjust association analyses for confounding, via kinship coefficients estimated from GWAS data using
  - principal components;
  - or linear mixed models
- studies of isolated populations are likely to be valuable to investigate variants that are globally rare
  - studies of model can play a similar role (e.g. dogs)
- little remaining rationale for recruiting nuclear families
  - except to look for parent-of-origin or maternal effects.

The near future

Even better analyses are now thinkable:

- The pedigree only gives expected genome-wide patterns of inheritance under Mendelian assumptions.
- Resequencing data provides the possibility to infer actual shared inheritance of genomic regions in two individuals separated by, say, up to 10 meioses.
  - $\sim 10^3$ maternal and $\sim 10^3$ paternal relatives within 10 meioses.
  - May share no genomic material with a 10-meiosis relative, but
  - if there is a shared haplotype, it is likely to be > 100Kb in length — detectable with confidence from resequencing data.

Some points to note:

- 10 meiosis is an arbitrary cut-off: in practice we look for sharing of extended haplotypes significantly beyond what is expected for "unrelateds" (random draw from an allele pool).
- We are all inbred: everyone is related to everyone else both paternally and maternally.

Population linkage analysis

- Given the genomic sequences of a 1% population sample (e.g. UK Biobank), expect ~ 10 maternal and ~ 10 paternal sequenced relatives of an individual. This permits
  - long-range phasing of much of the genomes (Kong 2008);
  - with some pedigree information, or mtDNA and Y data, can also distinguish maternal and paternal haplotypes (e.g. Iceland; Kong *et al.* 2010).
  - Detailed analysis of migration and demographic history.
  - **Population linkage analysis**.
- Purcell *et al.* (2007) propose a precursor method, using GWAS data and looking for excess shared ancestry (typically $>$ 1Mb).
- with resequencing data much finer analyses are possible:
  - can test linkage at a locus while controlling for (inferred) shared ancestry at all other loci genome-wide;
  - can accommodate allelic heterogeneity, as in linkage;
  - can even test for parent-of-origin effects in population data.

Multiple testing and genome-wide significance

- Even if a single test has a small probability to generate a false positive association result, the expected number of false positives increases linearly with the number of tests.
- When many SNPs are tested, we can get multiple false +ves.
- This is called the problem of **multiple testing**.

The usual solution to the problem is to employ the **Bonferroni correction**:

- Decide what expected number of false positives you are prepared to tolerate for the whole study *assuming no true association anywhere in the genome*;
- divide it by the number of tests to obtain α_{GW}.
- Example: we accept a 5% chance of a false positive under H_0 and 10^6 SNPs each undergo a single test, then $\alpha_{GW} = 0.05/10^6 = 5 \times 10^{-8}$.

False Discovery Rate

The Bonferroni correction remains popular, but it has drawbacks:

- the genome-wide H_0 is completely implausible;
- α_{GW} varies greatly with SNP density and distribution of MAF, LD (choice of population), sample size, and choice of test;
- correlation between the tests makes the resulting α_{GW} conservative;
- we should consider all SNPs, even if not genotyped;

An alternative is to control False Discovery Rate (FDR):

- FDR estimates the ratio: false +ves / all +ves;
  - doesn't assume H_0; based on an empirical estimate of the numbers of SNPs following H_0 and H_1, respectively;
  - an FDR of 5% is less stringent than $\alpha_{GW} = 5\%$;
- FDR has been successful for gene expression experiments but few, weak true +ves in GWAS, together with substantial LD, make FDR difficult to estimate.

Problems with p-values

- A small p-value is less convincing of a true association if the power of the test is low.
- The solution to this problem within the classical paradigm is to try to "ban" low-powered tests.
- But for association studies power is uneven across SNPs, as it can depend on
  - MAF
  - imputation quality (more later)
- Bayesian Posterior Probability of Association (PPA):
  - directly comparable across studies and across SNPs;
  - avoids multiple testing problem;
  - allows more quantitative and rational incorporation of background information, decision analysis, and meta-analysis;
  - Bayes vs frequentist can be viewed as "imperfect answer to the right question" vs "precise answer to the wrong question".
  - But there can be costs in terms of more sophisticated modelling and harder computations.

Computing the posterior probability of association (PPA)

Bayesian methods were not widely used for genetic association analyses until the WTCCC reported the Bayes Factor (BF):

$$\mathrm{BF} = \frac{P(\mathrm{data}|H_1)}{P(\mathrm{data}|H_0)}$$

under both strictly additive model and a general model that gives most weight to near-additive models. Then, to compute the PPA:

$$\mathrm{PPA} = \frac{\pi \mathrm{BF}}{1-\pi+\pi \mathrm{BF}} \qquad \text{where} \quad \pi = \frac{P(H_1)}{P(H_0)}.$$

π may vary across SNPs, depending on MAF, proximity to genes of interest, conservation across species,.... Typically $\pi \approx 10^{-4}$ (so *a priori* about 0.3 Mb of the genome has some true association).

The Bayesian solution to the problem of choosing the genetic model is to average the BF, weighted according to the plausibilities of different models.

Weighting additive and non-additive models in BF

Some results from WTCCC 07:

		p-value		BF	PPA	
Trait	SNP	Trend	General	($\log_{10}$)	$\pi = 10^{-4}$	$\pi = 10^{-5}$
BD	rs420259	2.2×10^{-4}	6.3×10^{-8}	4.1	0.56	0.11
CD	rs9858542	7.7×10^{-7}	3.6×10^{-8}	4.7	0.83	0.33
T2D	rs9939609	5.2×10^{-8}	1.9×10^{-7}	5.3	0.95	0.67
CD	rs17221417	9.4×10^{-12}	4.0×10^{-11}	8.9	0.99999	0.99987
T1D	rs17696736	2.2×10^{-15}	1.5×10^{-14}	12.5	1.00000	1.00000

Here, BF is computed as a 4:1 weighting of additive and general models (as defined by WTCCC).

- 1st row: $\log_{10}(\text{BF}) = 2.0$ (additive model); taking $\pi = 10^{-4}$, PPA $= 0.01$; likely to be ignored.
- Under general mode, $\log_{10}(\text{BF}) = 4.8$ and PPA $= 0.86$.
- But, general model often not tested – additive tests preferred.
- With 4:1 weighting, $\log_{10}(\text{BF}) = 4.1$, and PPA=0.56.
- Only 20% weight given to general model, but BF captures strong non-additive signal while still emphasising additivity.
- Don't calculate PPA for many models and pick the largest!

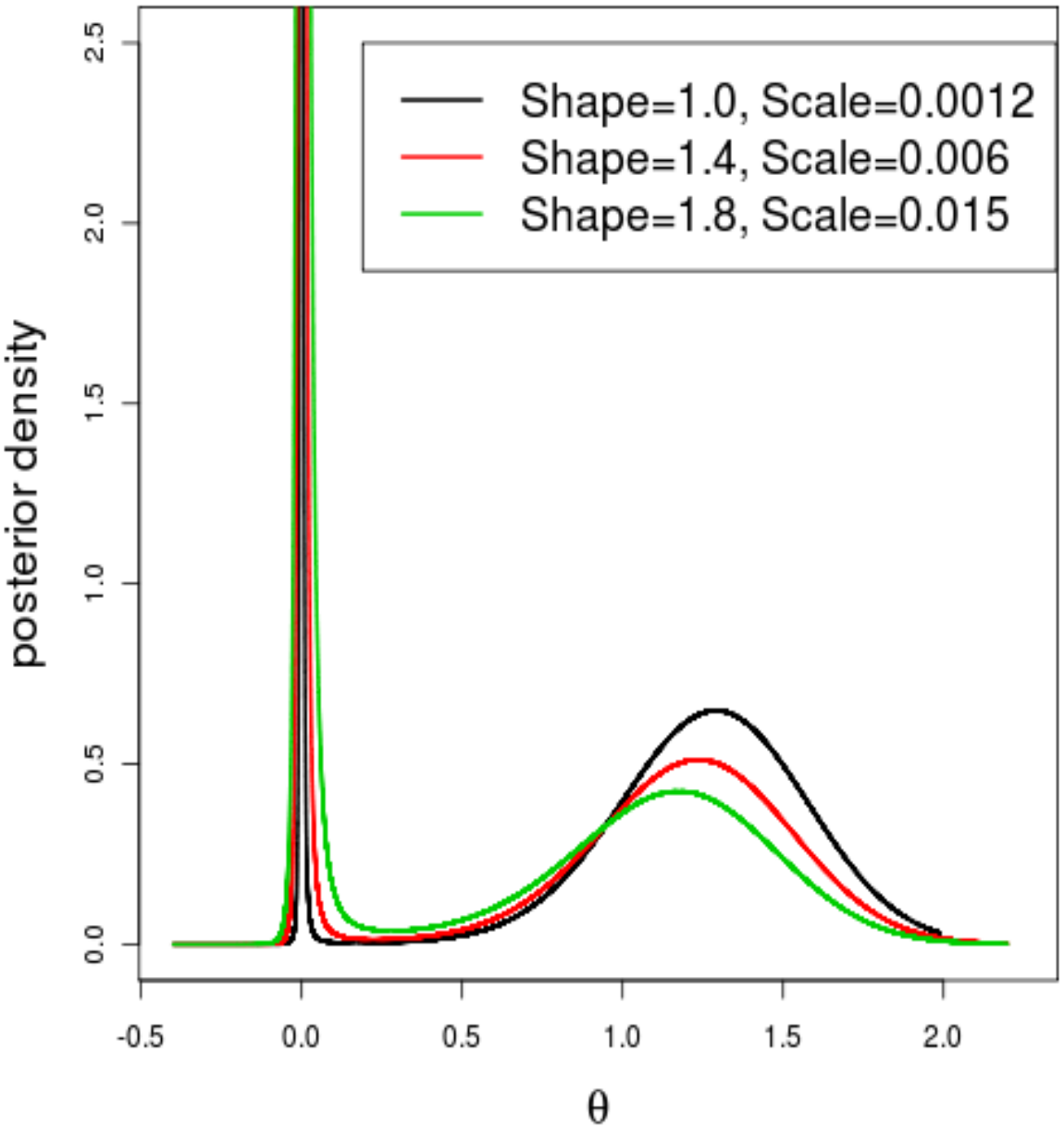
Effect sizes under an additive model

- Under the additive model, WTCCC assumed a $N(0, 0.2)$ prior on effect size (log odds).
- A drawback of this is rapid decay in the tails.

Example: effect of prior.

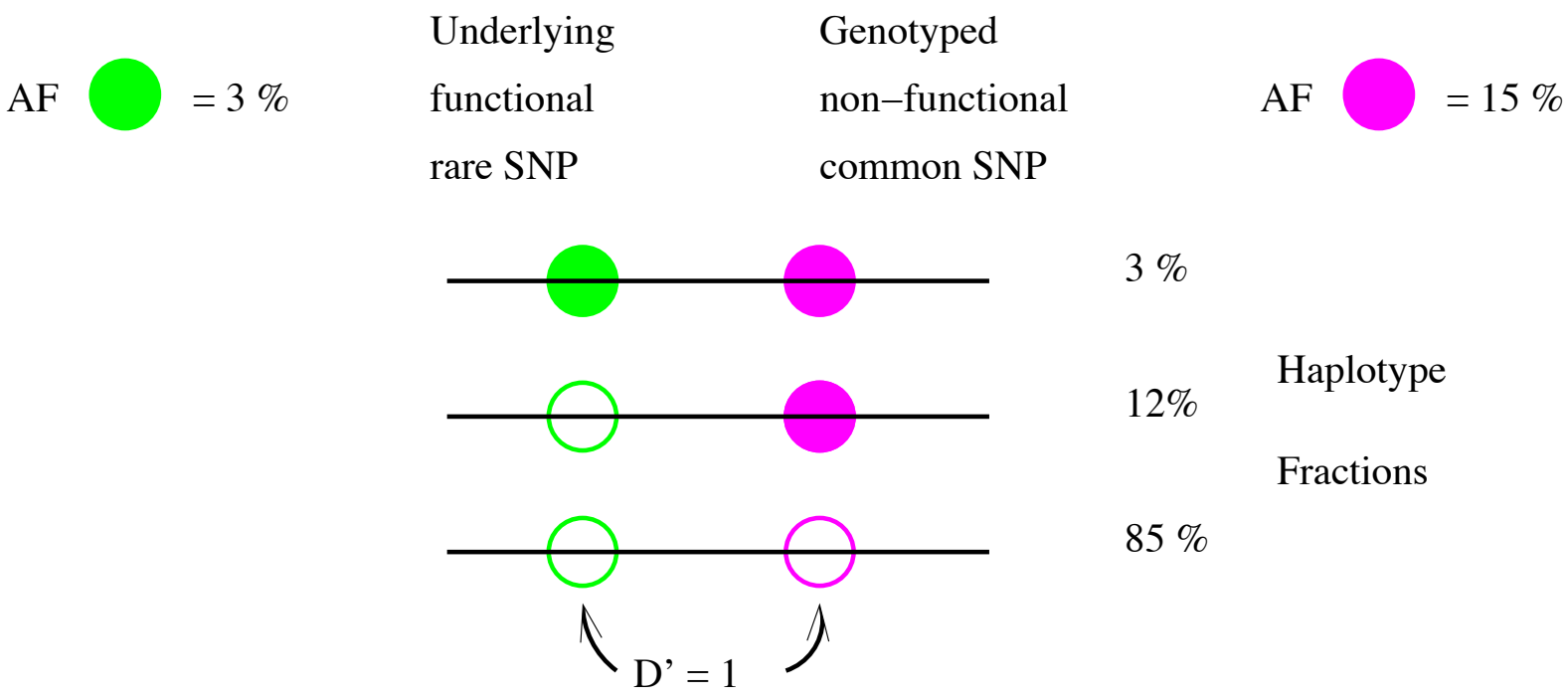
- the SEARCH collaborative group (08) reporting that variants in SLC01B1 are associated with statin-induced myopathy.
- most significant SNP is rs4363657, with $p = 4.1 \times 10^{-9}$.
- Using WTCCC prior, PPA ≈ 0.02
- Other Bayesian analyses with more plausible priors give (e.g. mixture of Gaussians, see also next slide) PPA ≈ 0.4.
- Big influence of prior, because data suggest very large effect size for a rare allele: WTCCC says this is *a priori* implausible.
- $p < 10^{-8}$ is conventionally regarded as highly significant, but Bayesian analysis says we should be far from convinced.

Do we need a null hypothesis?



- Division of SNPs into null or non-null is artificial;
- reality is a distribution of effect sizes that puts much weight near zero;
- can be modelled using Normal-Exponential-Gamma (NEG) prior, and posterior density obtained numerically;
- no BF in this approach, but posterior $P(|\theta| > 0.1) = 0.47$, 0.39 and 0.35.
- See Stephens & Balding 09.

Rare alleles: the dark matter of heritability?

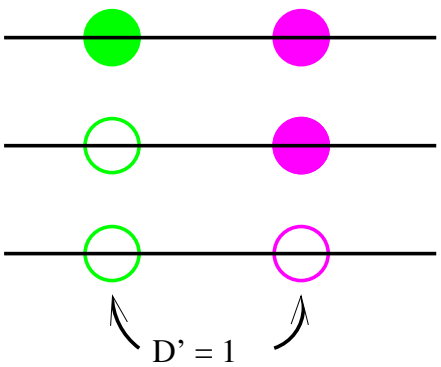


- Observed common genetic associations explain relatively little of the burden of disease even for highly-heritable disorders.
- Some may be due to a highly-penetrant, rare causal variant.
- Because rare haplotypes tend to be recent and therefore long, the rare causal may be very far from the observed association.

Rare alleles: the return of the family?

- Problem: rare alleles are hard to find.
- Currently much interest in exome resequencing to detect rare alleles of functional significance (e.g. Nickerson, Nature Genetics 2010).
- Widespread view that families will become important again because of concentration of rare alleles,
- e.g. use sibs with concordant phenotypic extremes.
- Better solution: isolated populations with high prevalence
  - role of pedigree (whether known or not) now even more important, but can be tackled e.g. using $\hat{K}$.

Rare causals change QT mean *and* variance



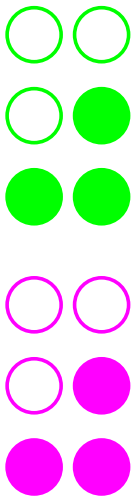
3 %

12%

85 %

Haplotype

Fractions



QT mean = 0, s.d. = 1

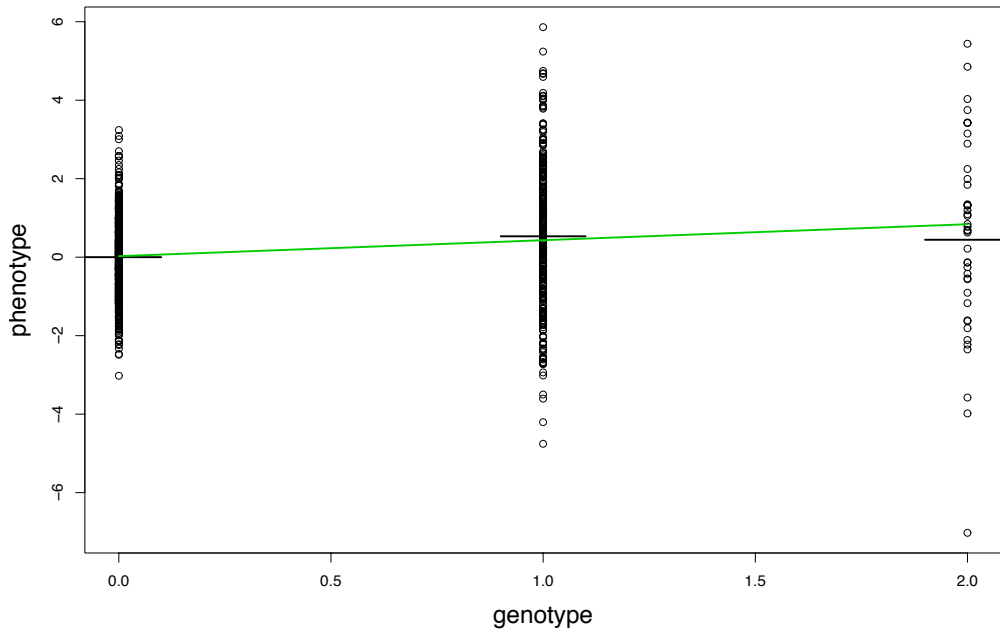
QT mean = 0.5, s.d. = 1

QT mean = 1, s.d. = 1

QT mean = 0, s.d. = 1

QT mean = 0.1, s.d. = 1.2

QT mean = 0.2, s.d. = 1.4



Need to extract signal from both mean and variance when testing for association at a QT.

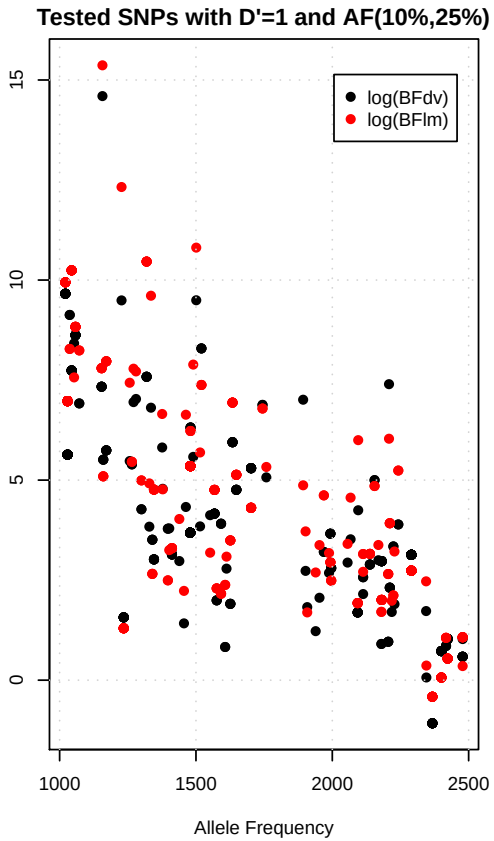
Bayes Factor for changes in QT mean and variance

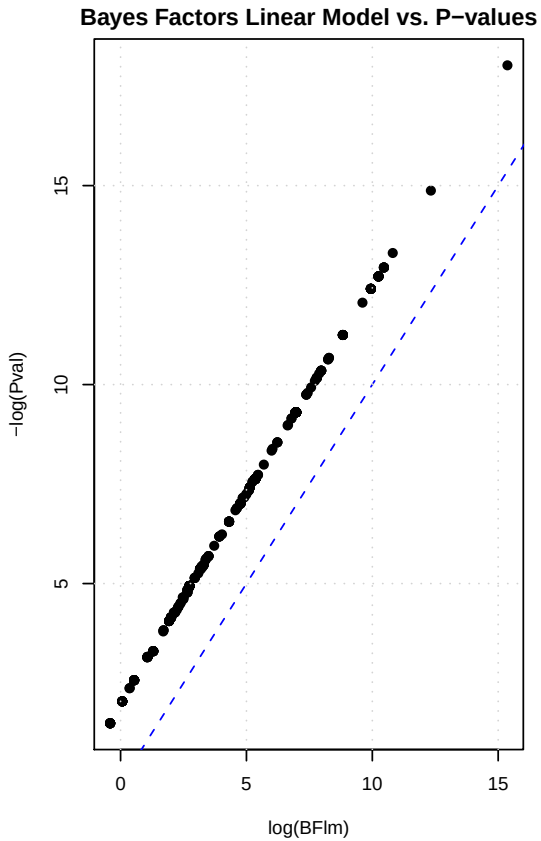
- Typical QT association analysis is to test for zero slope in linear regression on genotype (coded e.g. as 0, 1, 2). If phenotypes y_i are standardised, then standard Gaussian assumptions:
  - H_0: the y_i are i.i.d. $N(0, 1)$
  - H_1: the genotypes g_i are centred (mean $= 0$) and each y_i is $N(\beta g_i, 1/\gamma)$, for $\beta \sim N(0, \sigma^2)$ and $\gamma \sim \text{Gamma}(\alpha, \alpha)$, where σ^2 and α are known constants.
- An alternative approach is to assume under H_1 that for genotype j the y are i.i.d. $N(\mu_j, 1/\gamma_i)$ for $\mu_j \sim N(0, 1/\tau)$ and $\gamma_j \sim \text{Gamma}(\alpha, \alpha)$, where τ and α are known constants.

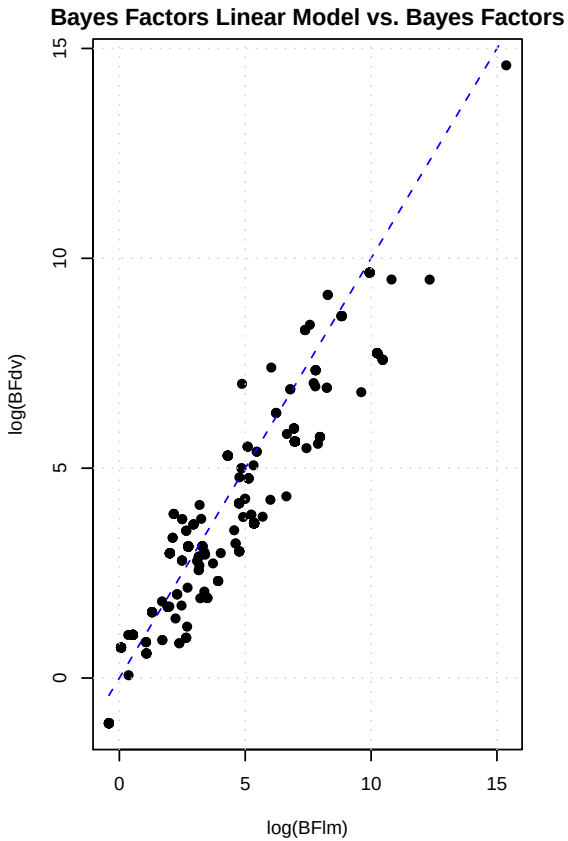
Each BF has a simple, exact formula. We tried out these two BFs in a small simulation study, in which the causal variants have fraction $\approx 4\%$ and the genotyped SNPs have $D' = 1$ with the causal variant and minor allele fraction between 10% and 25%.

Additive model at causal SNP

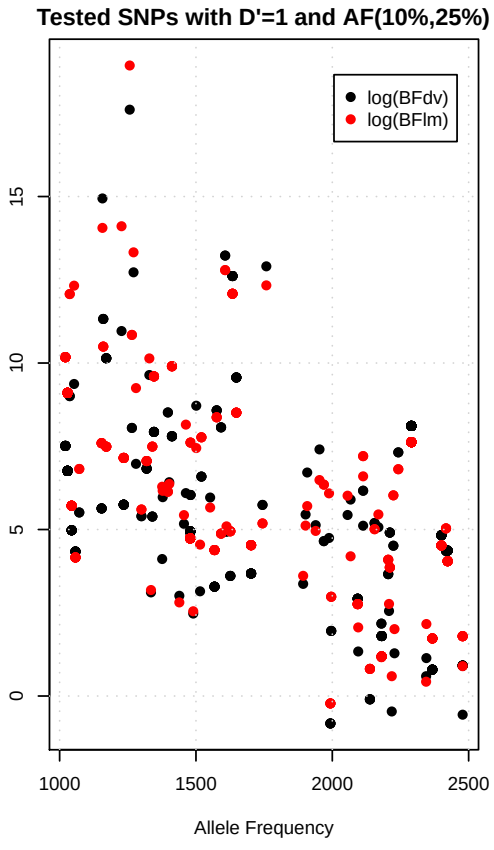
First we compare the BF that allows different variances (BFdv) with the linear model (BFlm) when the effect is additive at the causal variant $\Rightarrow$ at genotyped marker, the mean effect is linear in allele dose but with unequal variances:

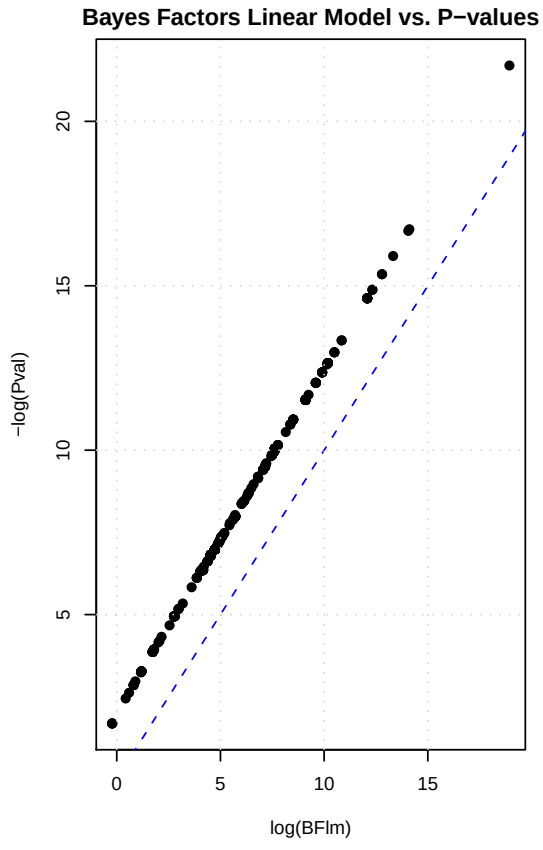


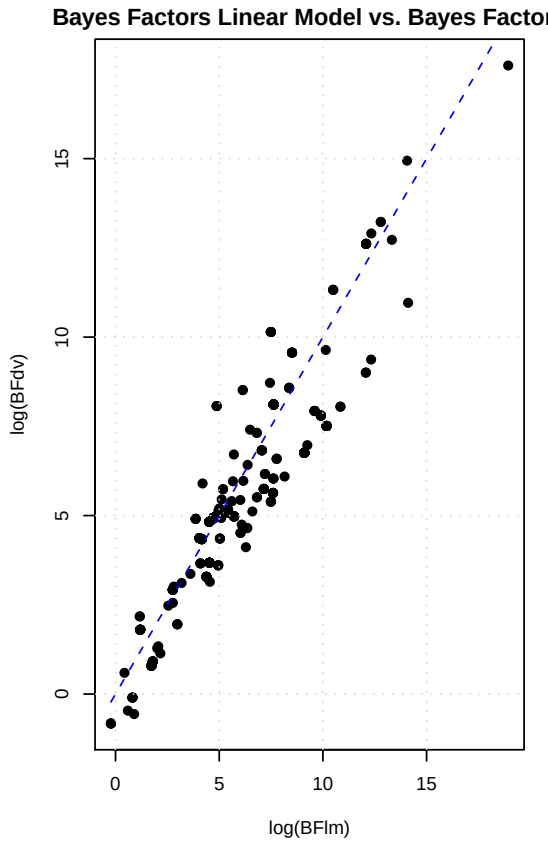




Near-recessive model at causal SNP



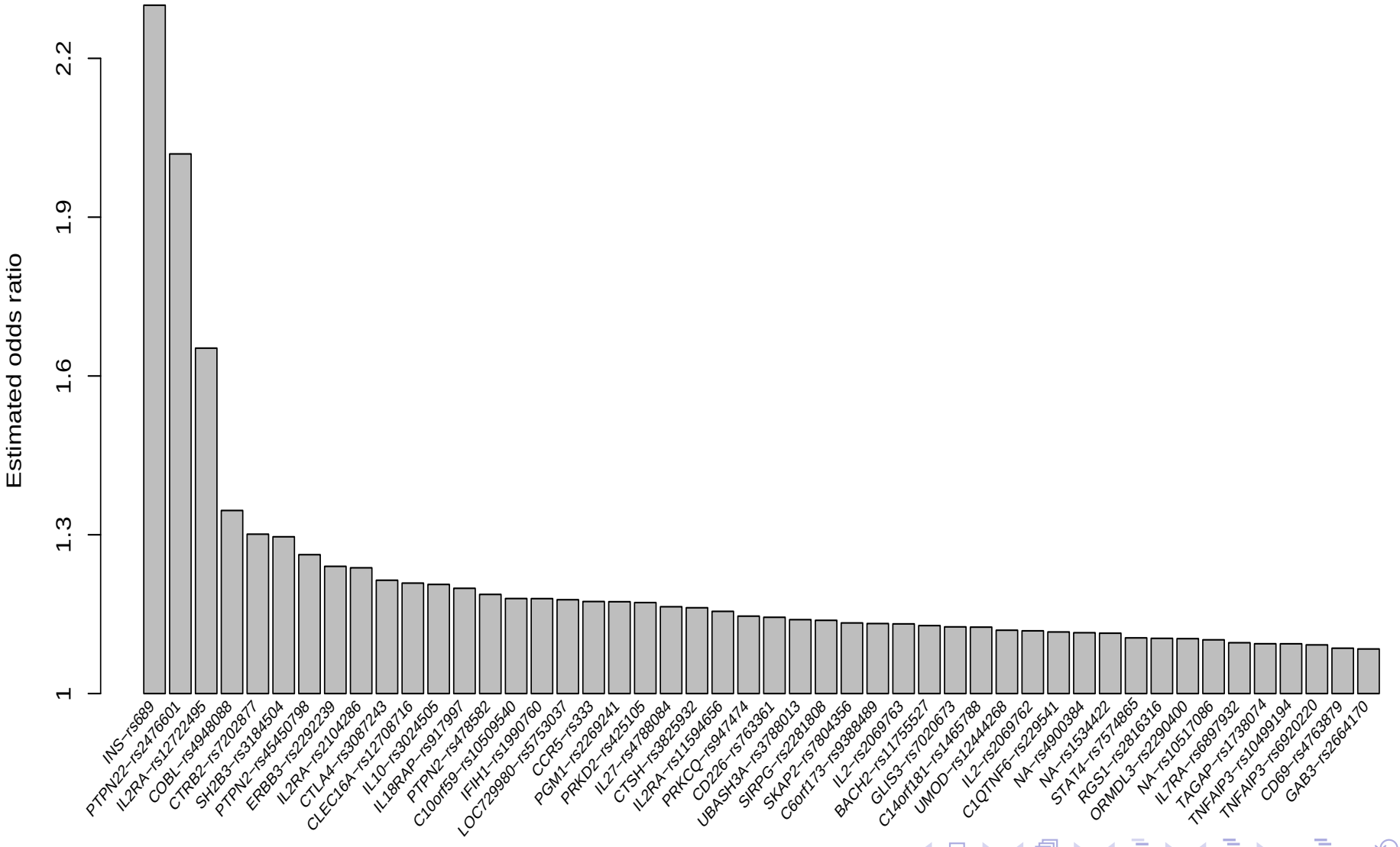




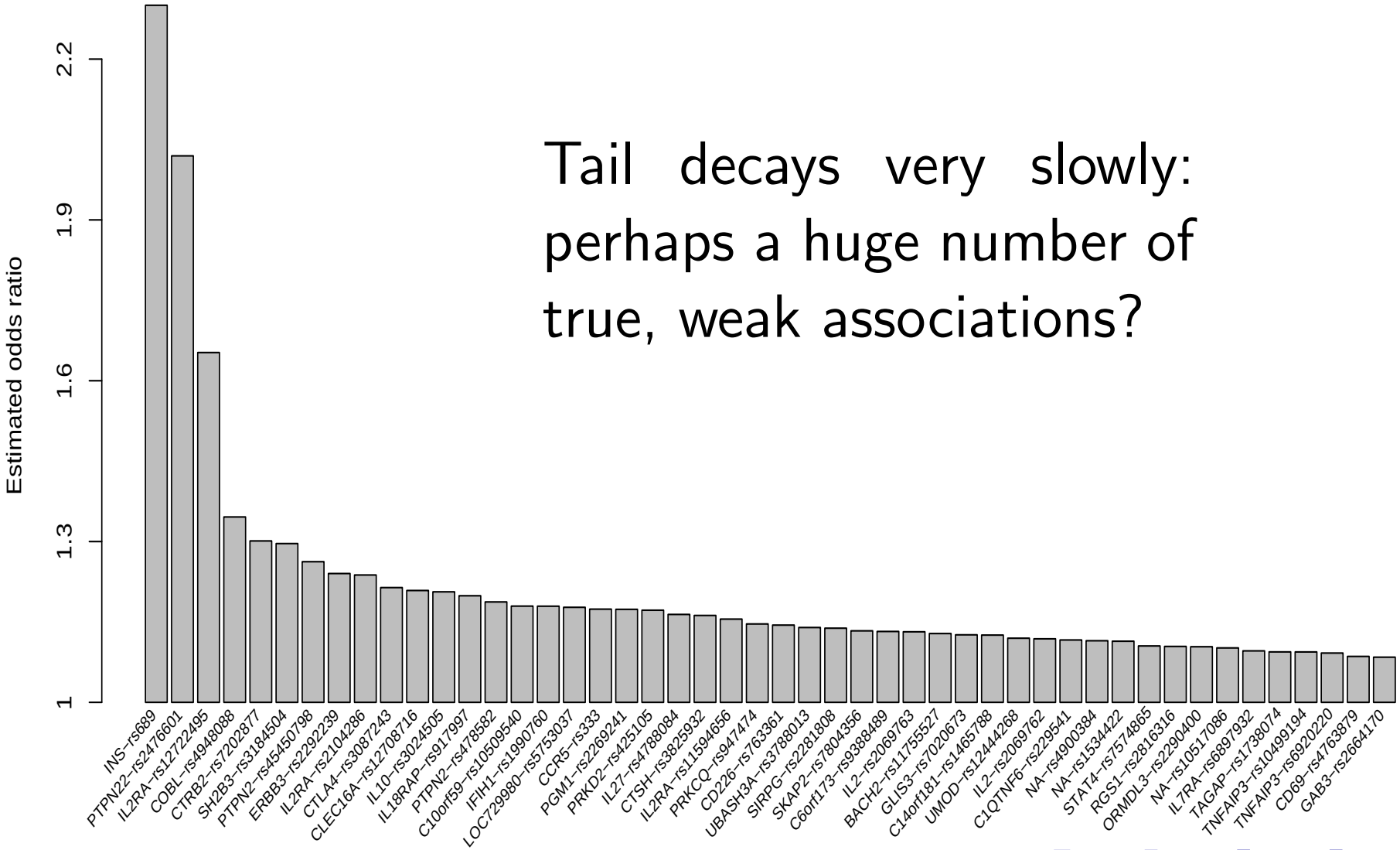
BFdv doesn't improve over BFlm; the sparse linear model (2 parameters) has an advantage over the full 6 parameter model even when it is strictly false.

However, BFdv does give a big advantage if the causal variant alters variance as well as the mean, and seems to find some interesting associations in real data.

T1D associations against effect size estimate (odds ratio, additive model)



T1D associations against effect size estimate (odds ratio, additive model)



Tail decays very slowly: perhaps a huge number of true, weak associations?

Prediction of Case-control status

Prediction of case-control status from confirmed variants is poor, but

- Prediction from many thousands of top SNPs from GWAS can achieve reasonably good prediction, even though many of the SNPs used in the prediction are false positives.
- Provides further evidence that many, common, low-effect-size SNPs may explain much of heritable variation.
  - Goldstein 2009: 93,000 SNPs required to explain 80% of the population variation in height
- Low *marginal* effects could reflect stronger underlying effects:
  - rarer untyped variants, including CNVs

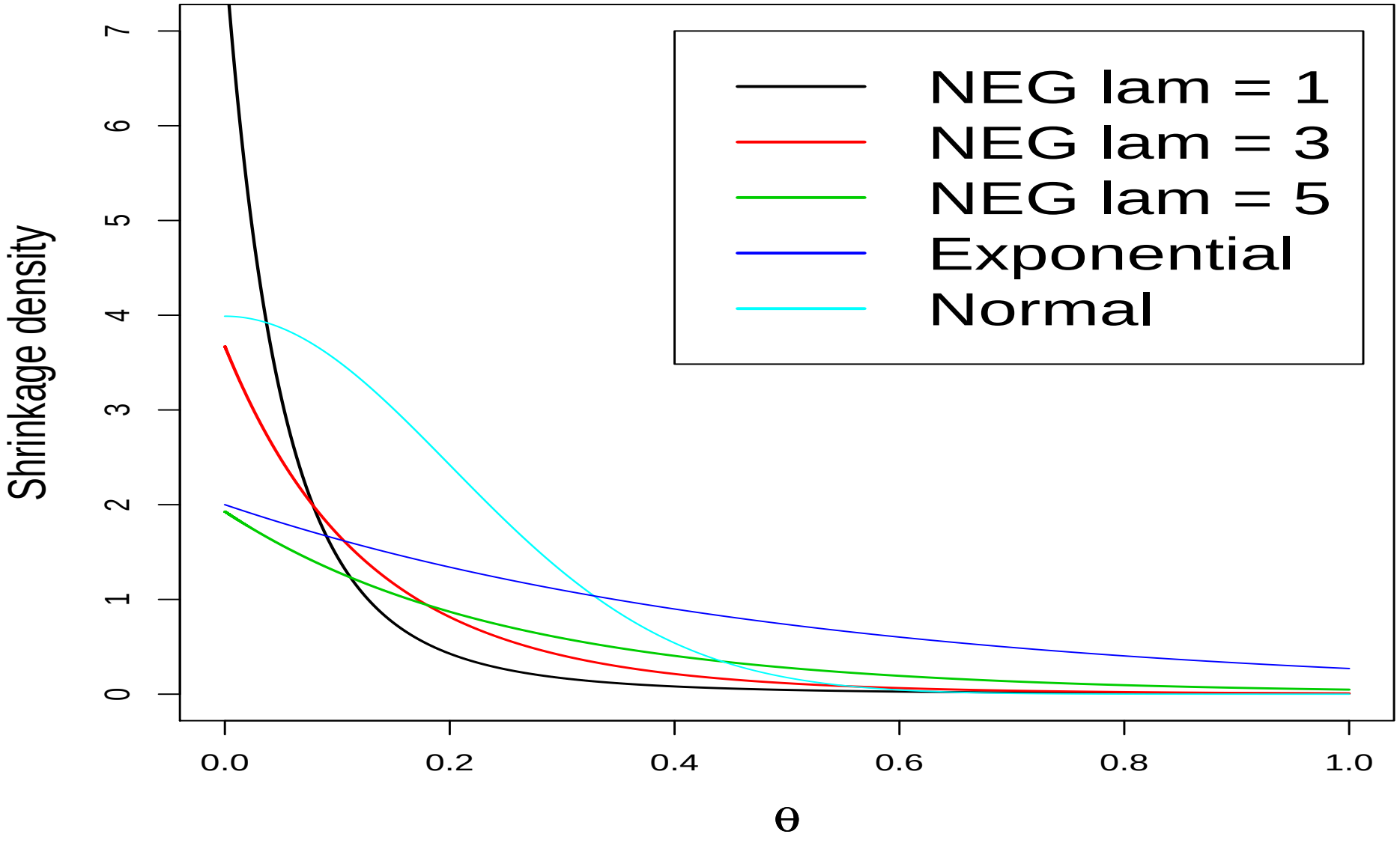
Shrinkage

- Standard of "significance" should be different for prediction than when identifying causal loci; your prior distribution of effect size may be the same but *utility* differs:
  - no-one wants to invest a lot of money in functional studies only to find that the locus isn't causal after all;
  - a few false +ves in a prediction model may do little harm, and the lower significance threshold can allow in true +ves that fail to reach genome-wide significance;
  - given some genotyping, cost of some extra SNPs is small.
- *shrinkage* (or *penalised*) regression is useful to minimise their false +ves and automatically select among SNPs in high LD.

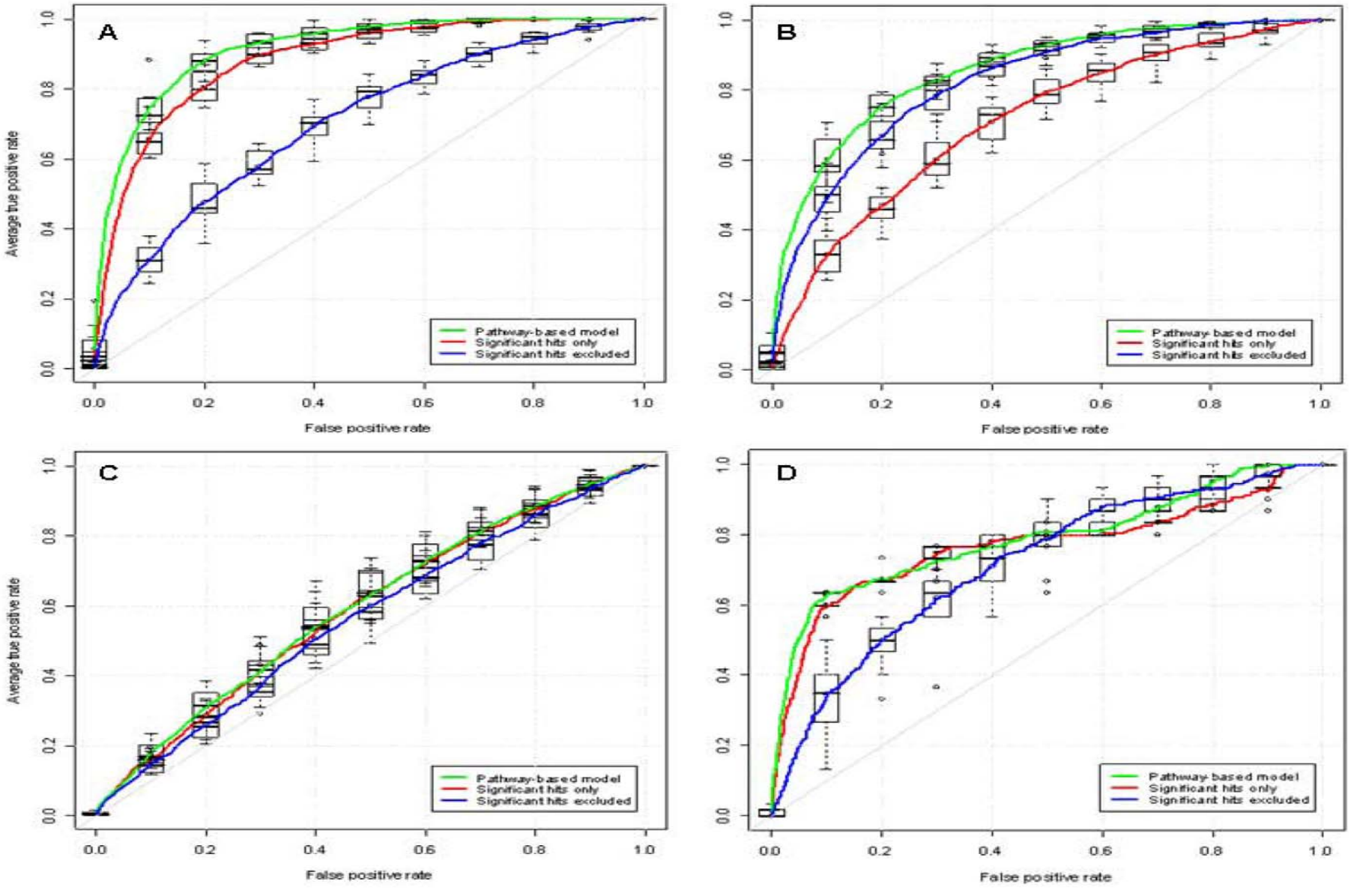
How to shrink?

- **Ridge regression**: Gaussian (normal) distribution;
- **LASSO**: Exponential distribution;
- **NEG** distribution (Hoggart *et al.* PLoS Genet 2008).

Shrinkage priors



Prediction of WTCCC autoimmune diseases



From Eleftherohorinou *et al.* PLoS1, 2009.

Prediction: some conclusions

- There is room for optimism about prediction, despite generally gloomy viewpoint of some authors (e.g. Clayton 09)
  - Best prospects are for common phenotypes.
  - Heterogeneous drug response can be reduced with even modest predictive accuracy.
- In addition to conventional risk factors, it may be effective to use tens or hundreds of markers, many unconfirmed as causal.
  - Some success from prediction based on genome-wide markers (Wray *et al.* 08, Purcell *et al.* 09)
- Further modelling developments feasible
  - some effects are non-additive; should allow for these in models;
  - further work to improve shrinkage priors;
  - plant and animal breeders are currently leading the way with sophisticated statistical modelling.

Acknowledgments

- Mixed models with kinship: Will Astle, Imperial College
  - **Review paper:** to appear *Statist. Sci.* Feb 2010 (available online).
- BFs for QTs with change in both variance and mean: Susana Perez Alvarez, Universitat Politècnica de Catalunya
- Vincent Plagnol and John Todd lab, Cambridge, for plot of T1D effect sizes.