Heritability-based models for prediction of complex traits

David Balding
(with much help from Doug Speed, funding: UK MRC)

Schools of BioSciences and of Maths & Stats
University of Melbourne
and UCL Genetics Institute London

Summer School, European Network on Machine Learning for Personalised Medicine, Manchester, 21 September 2015

SNP-based heritability analysis: born 2010

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

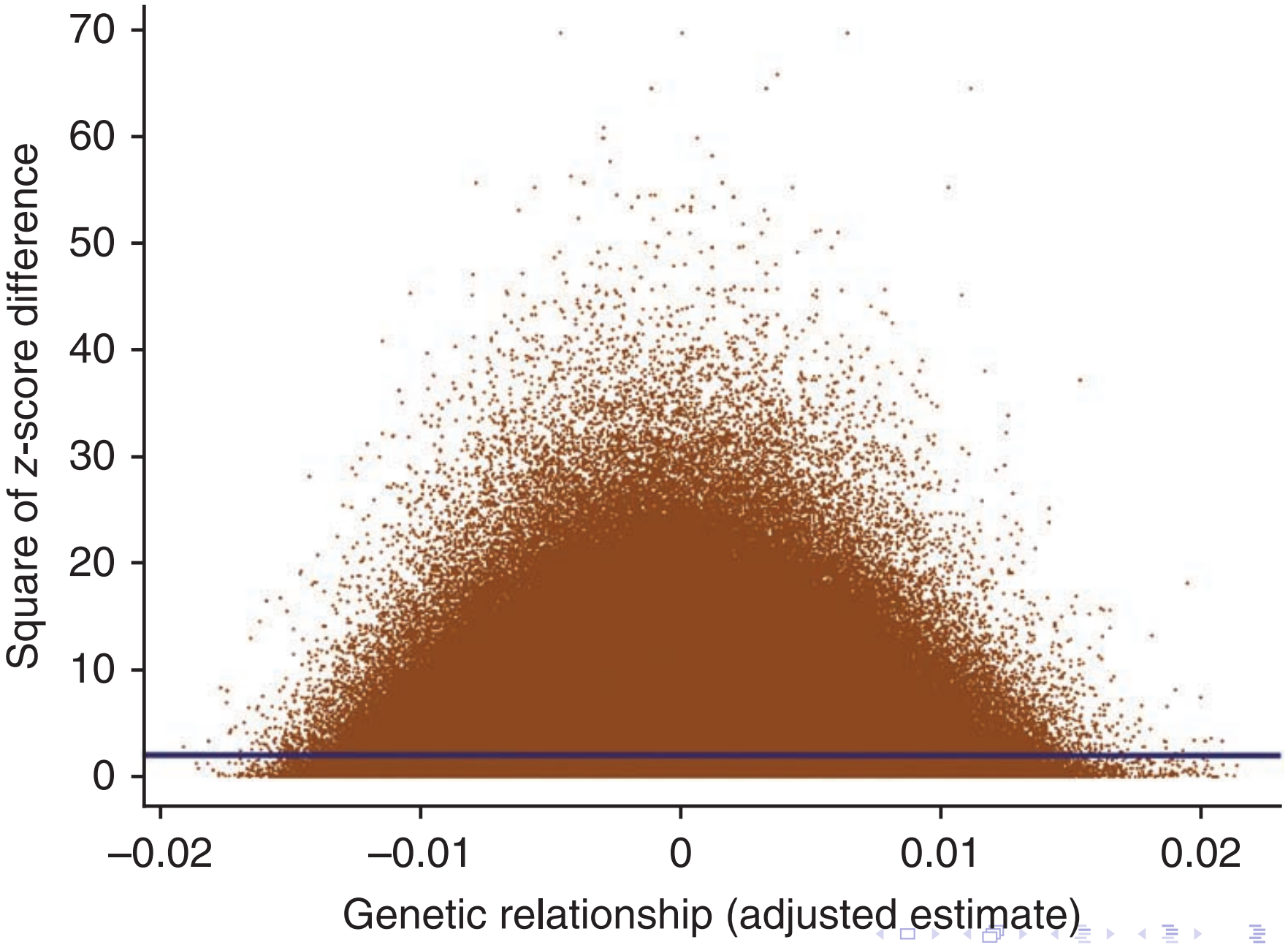
SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[9,10]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

NATURE GENETICS VOLUME 42 | NUMBER 7 | JULY 2010

SNP-based heritability analysis: slope of regression line



Heritability: key ideas

Heritability is the fraction of phenotypic variance that can be explained by genetics. Related individuals have correlated genotypes: heritability measures the extent to which this implies correlated phenotypes.

- It measure how "genetic" a trait is, relative to "environmental" causes, so it is environment-specific.
- We are mostly concerned with "narrow sense" heritability or h^2 and so only additive genetics. h^2 is the variance explained by a linear regression

$$\mathrm{E}[Y] = \beta_0 + \sum_j \beta_j X_j = \mathbf{X}\beta$$

  where Y is phenotype, X_j is genotype (additive coding; standardised) at jth locus, and the sum is over **causal** loci.
- **Problem:** we don't know the causal variants or effect sizes.

Clever idea: mixed model approach

Assuming a Gaussian model, the linear regression can be formulated as a mixed regression model:

$$Y = \gamma + \epsilon$$

where $\text{Var}[\epsilon] = \sigma_e^2 \mathbf{I}$ and γ is a latent genetic "random effect" with $\text{Var}[\gamma] = \sigma_g^2 \mathbf{K}$. Then $h^2 = \sigma_g^2/(\sigma_g^2+\sigma_e^2)$.

- Estimation of σ_g^2 and σ_e^2 usually done via REML.
- Ideally we want $\mathbf{K} = (\mathbf{X}\beta)(\mathbf{X}\beta)^T$ but we don't know β or $\mathbf{X}$.
- Traditional approach has been to approximate $\mathbf{K}$ by kinship coefficients computed from pedigrees.

What's wrong with pedigree-kinship?

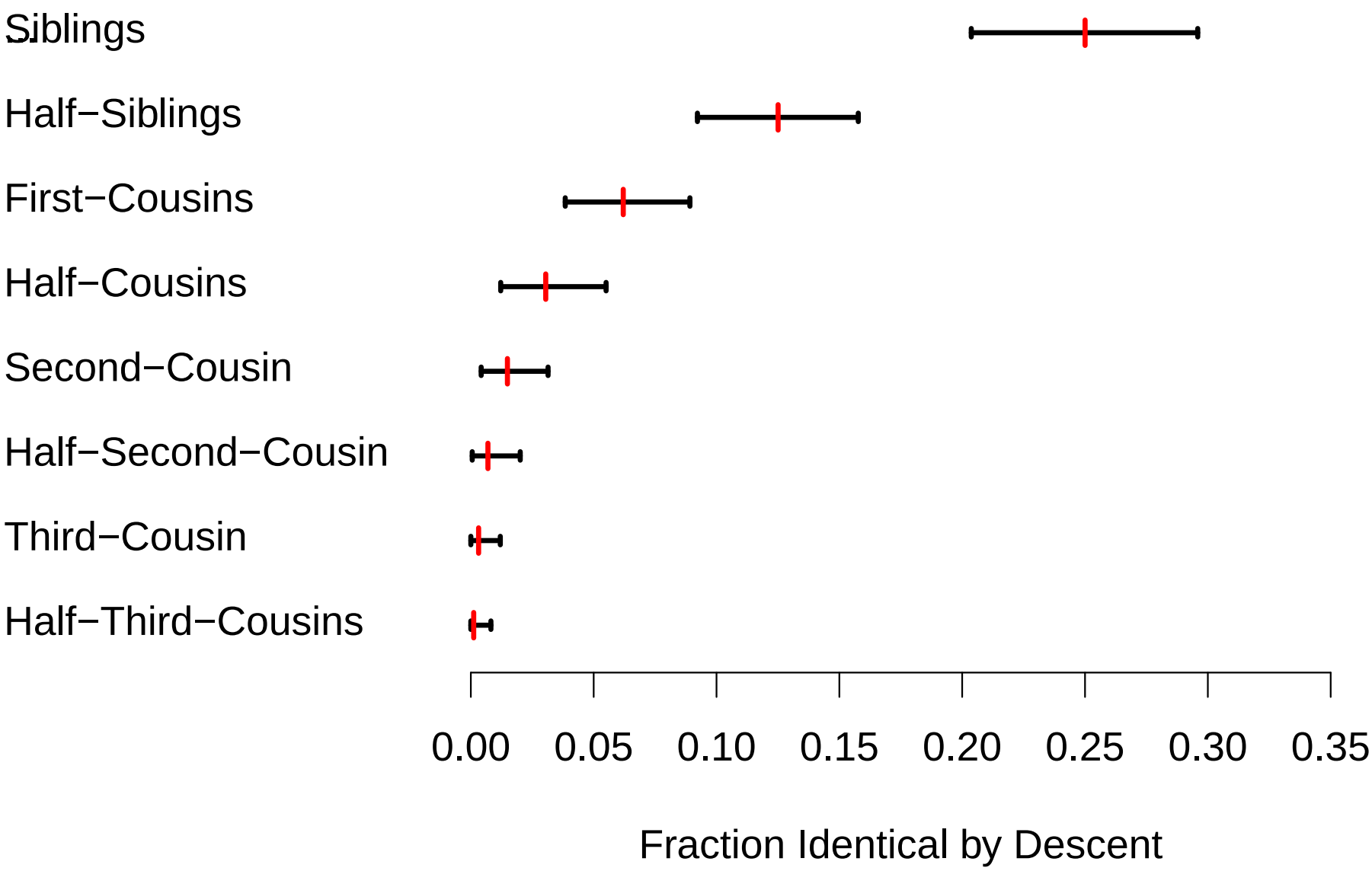
Through familiarity, pedigree-based kinships came to be seen as the canonical measures of relatedness, but they aren't very good.

- They depend on the pedigree that happens to be available: there is no such thing as a complete or ideal pedigree.
- What matters is allele sharing at causal loci, but pedigrees only specify expected, genome-wide allele sharing;
  - The fraction of genome shared by sibs from their parents can be < 0.4 or > 0.6.

In fact, there is no definitive way to measure the kinship of two individuals, and it is better to speak of genomic similarity, which can be measured from genome-wide SNPs or sequences.

See: Speed & Balding "Relatedness in the post-genomic era: is it still useful?" *Nat Rev Genet* Jan 2015

Genome sharing between pairs of "regular" relatives



Statistics of IBD sharing (update of Donnelly 1983)

Relationship	# G	# A	$\theta(A,B)$ $\mathbb{E}[\text{IBD}]/4$	95% CI	$\mathbb{P}[\text{IBD}>0]$	$\mathbb{E}[\#\text{sr}]$	$\mathbb{E}[\text{rl}]$ (Mb)
Sibling	1	2	0.250	(0.204,0.296)	1.000	85.3	31.3
1/2-sib	1	1	0.125	(0.092,0.158)	1.000	42.6	"
Cousin	2	2	0.063	(0.039,0.089)	1.000	37.1	18.0
1/2-cuz	2	1	0.031	(0.012,0.055)	1.000	18.5	"
2nd-cuz	3	2	0.016	(0.004,0.031)	1.000	13.2	12.6
1/2-2nd-cuz	3	1	0.008	(0.001,0.020)	0.995	6.6	"
3rd-cuz	4	2	0.004	(0.000,0.012)	0.970	4.3	9.7
	5	2	0.001	(0.000,0.005)	0.675	0.7	7.9
	7	2	$(1/2)^{14}$	(0.000,0.001)	0.098	0.1	5.5
	9	2	$(1/2)^{18}$		0.009	0.0	4.4

G: # generations: we consider a single lineage path of 2G steps;

A: ancestors; sr = shared regions; rl = region length

SNP-based measures of genomic similarity

There are many ways to measure genetic similarity of two individuals from genome-wide genetic markers (SNPs),

- which one is the best?

One difficulty in humans is that we are all closely related:

- Any two haploid human genomes share over 99.9% sequence identity due to shared ancestry.
- This isn't evident for SNPs because they are highly polymorphic, but
  - measures of similarity can depend sensitively on the Minor Allele Fraction (MAF) spectrum.
  - more low-MAF sites $\Rightarrow$ more similarity.
  - MAF spectrum depends on SNP chip **and** QC.

SNP-based kinships

Two approaches:

- **Average haplotype sharing.** Useful in some settings, but small (e.g. $<$ 1Mb) shared fragments are informative yet hard to exploit.
- **Genome-wide average of a single-SNP measure.**

Single-SNP approach 1: Average allele-sharing

- Given two individuals, code the SNP genotypes of each as 0,1 and 2, where 1 = heterozygote. Average the following scores:

$$\begin{array}{rcl} (0,0) \text{ or } (2,2) & \rightarrow & 1 \\ (0,1), (1,1) \text{ and } (1,2) & \rightarrow & 1/2 \\ (0,2) & \rightarrow & 0 \end{array}$$

- Disagreement about how to code heterozygotes: PLINK codes (1,1) as 1, rather than 0.5.

single-SNP approach 2: Average allelic correlation

Write G_{ij} for genotype of i at the jth SNP (allele count), then for i and i' use genome-wide average of single-SNP sample-size-1 correlation estimates:

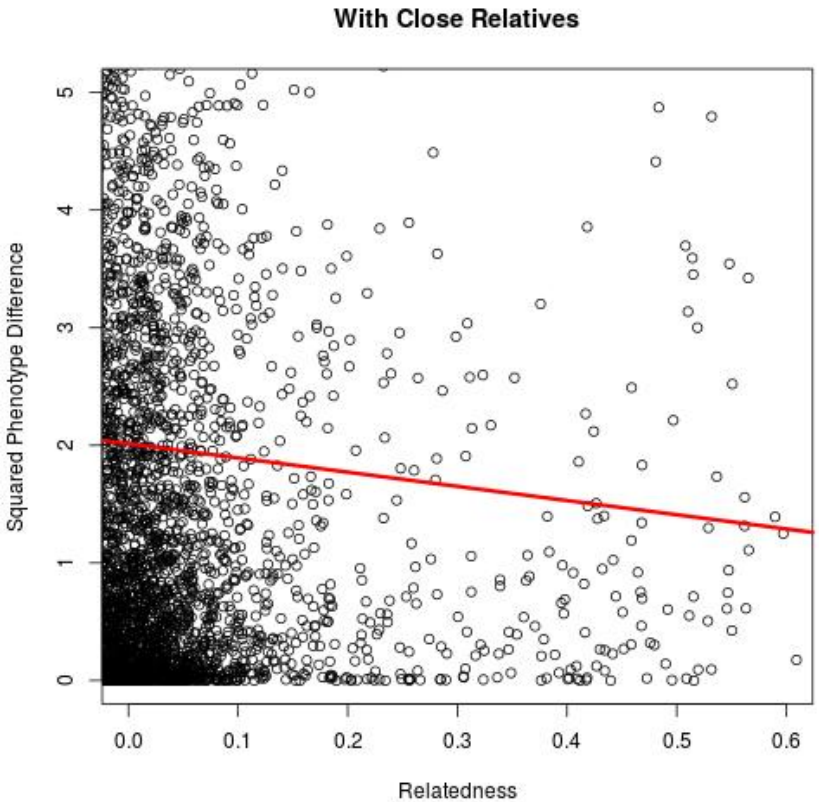
$$\frac{1}{m}\sum_{j=1}^{m}\frac{(G_{ij}-2p_j)(G_{i'j}-2p_j)}{2p_j(1-p_j)} \qquad \text{so} \quad \mathbf{K}=\frac{1}{m}\mathbf{X}\mathbf{X}^T$$

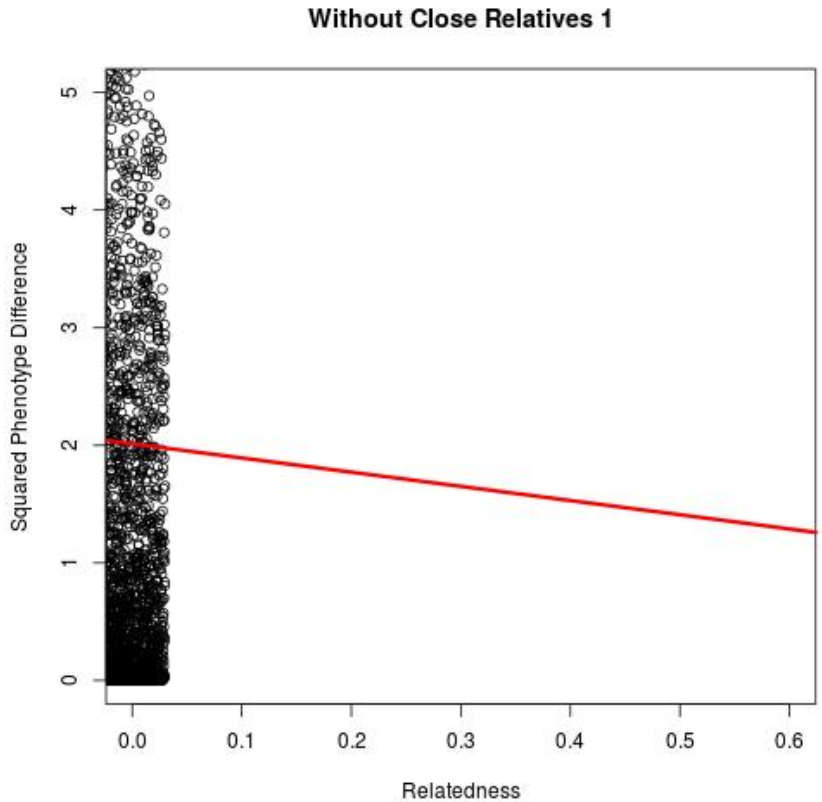
Now $\mathbf{K}$ has contributions from genome-wide SNPs:

- better than pedigrees: actual allele sharing
- worse: causal variants contribute only if tagged by SNPs (biased towards common variants)

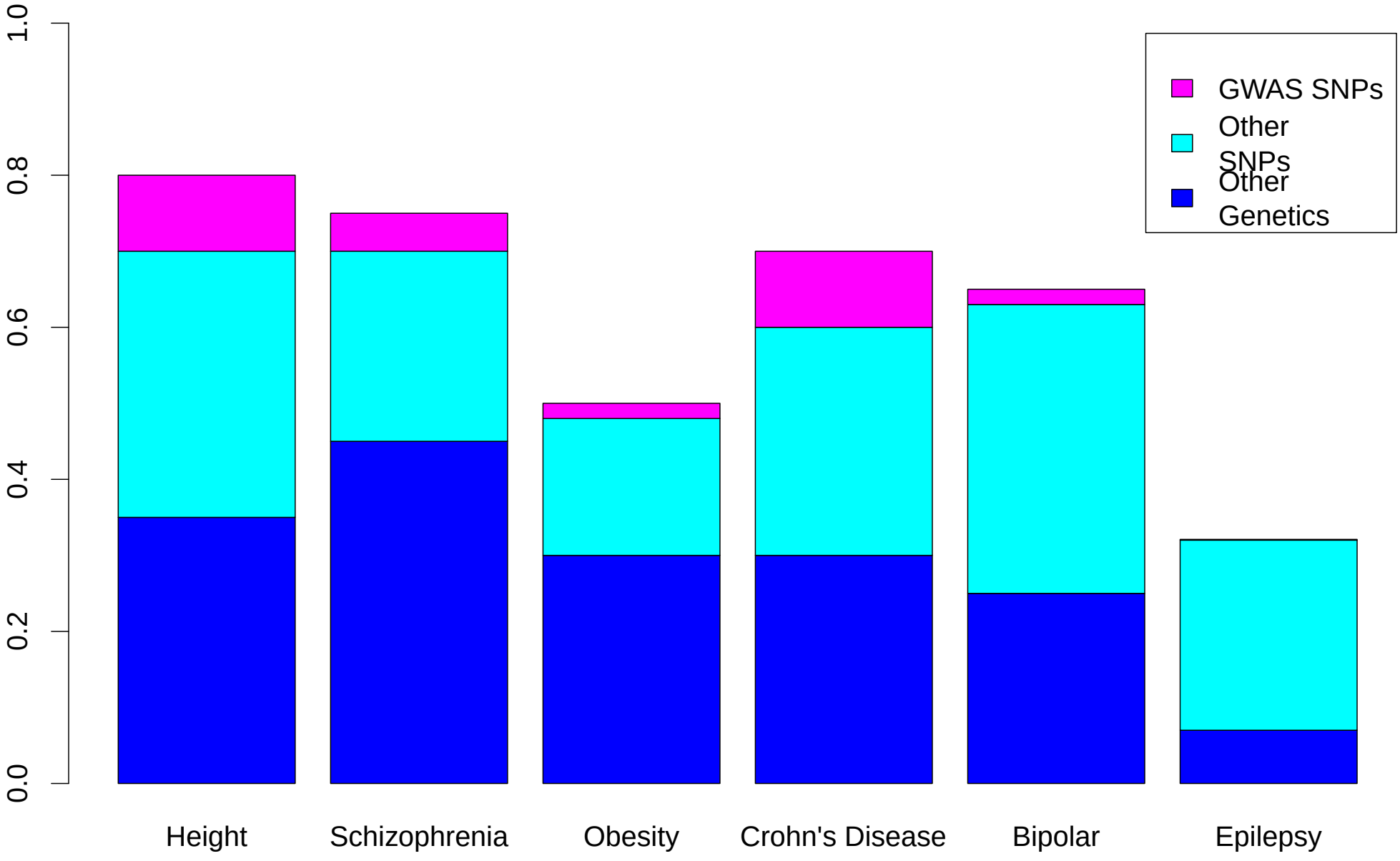
Yang *et al.* (2010) estimated h^2 using mixed-model with average-allelic-correlation **K** from genome-wide SNPs.

- A key feature is the use of unrelated pairs of individuals.





Heritabilities of some human traits



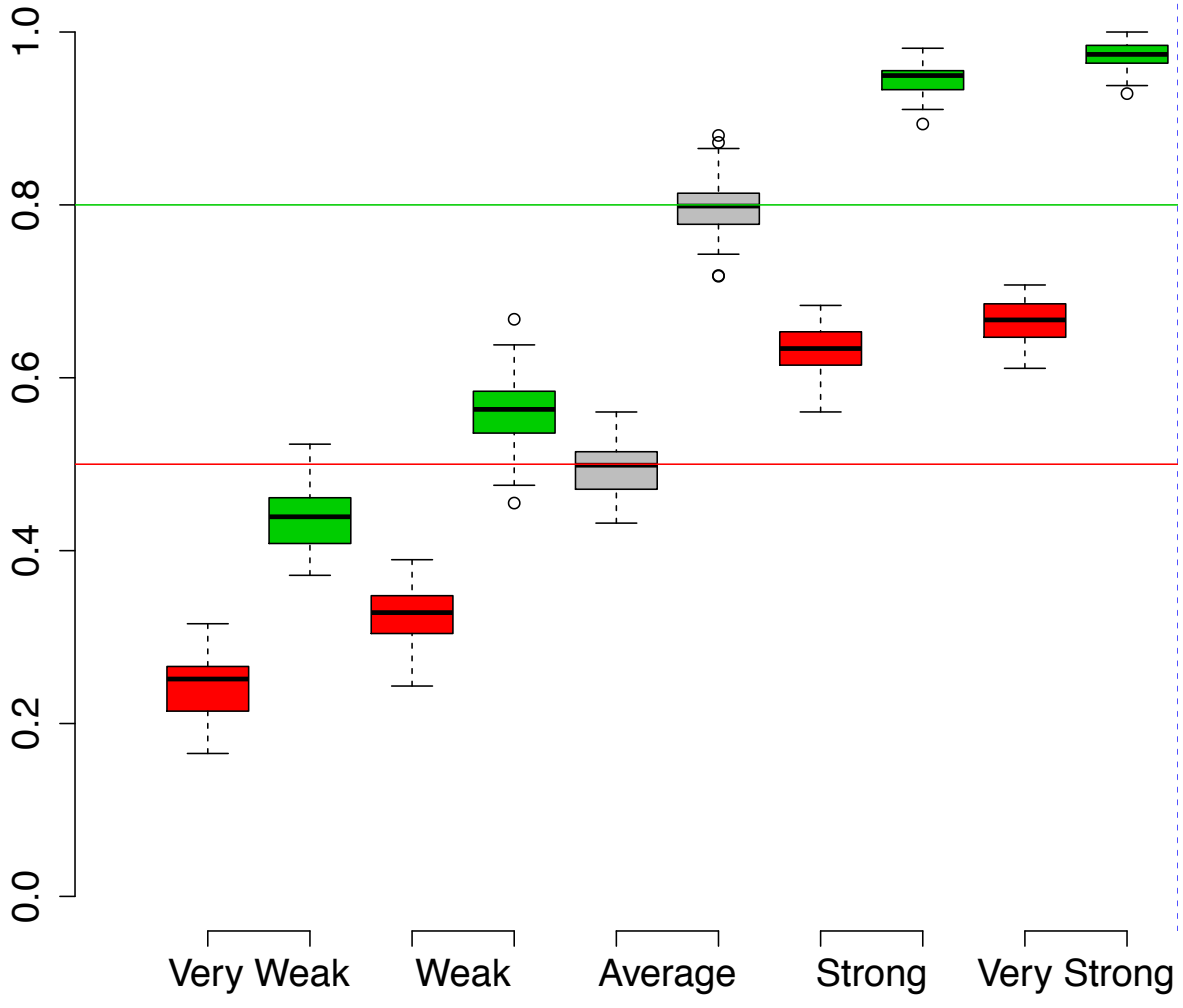
Mixed-model estimation of SNP-h^2 works well

We conducted a simulation study to investigate the robustness of $\hat{h^2}$ based on this method. See Speed *et al.* Am J Hum Genet (2012) for details. We found the method to be remarkably robust to

- number of causals,
- causal MAF spectrum,
- effect size distribution

But ...

$\hat{h}^2$ estimates not robust to LD

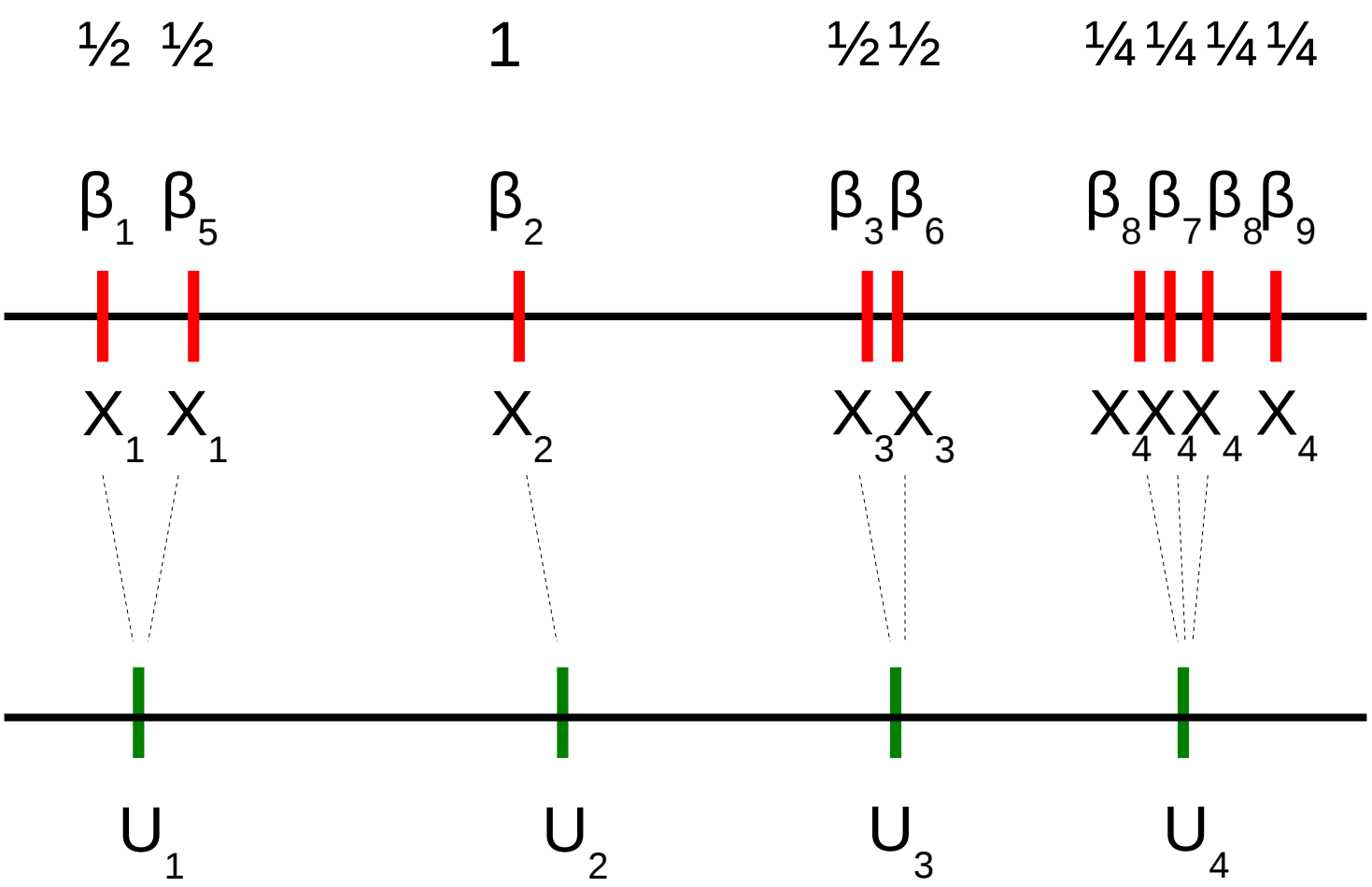


Local level of LD

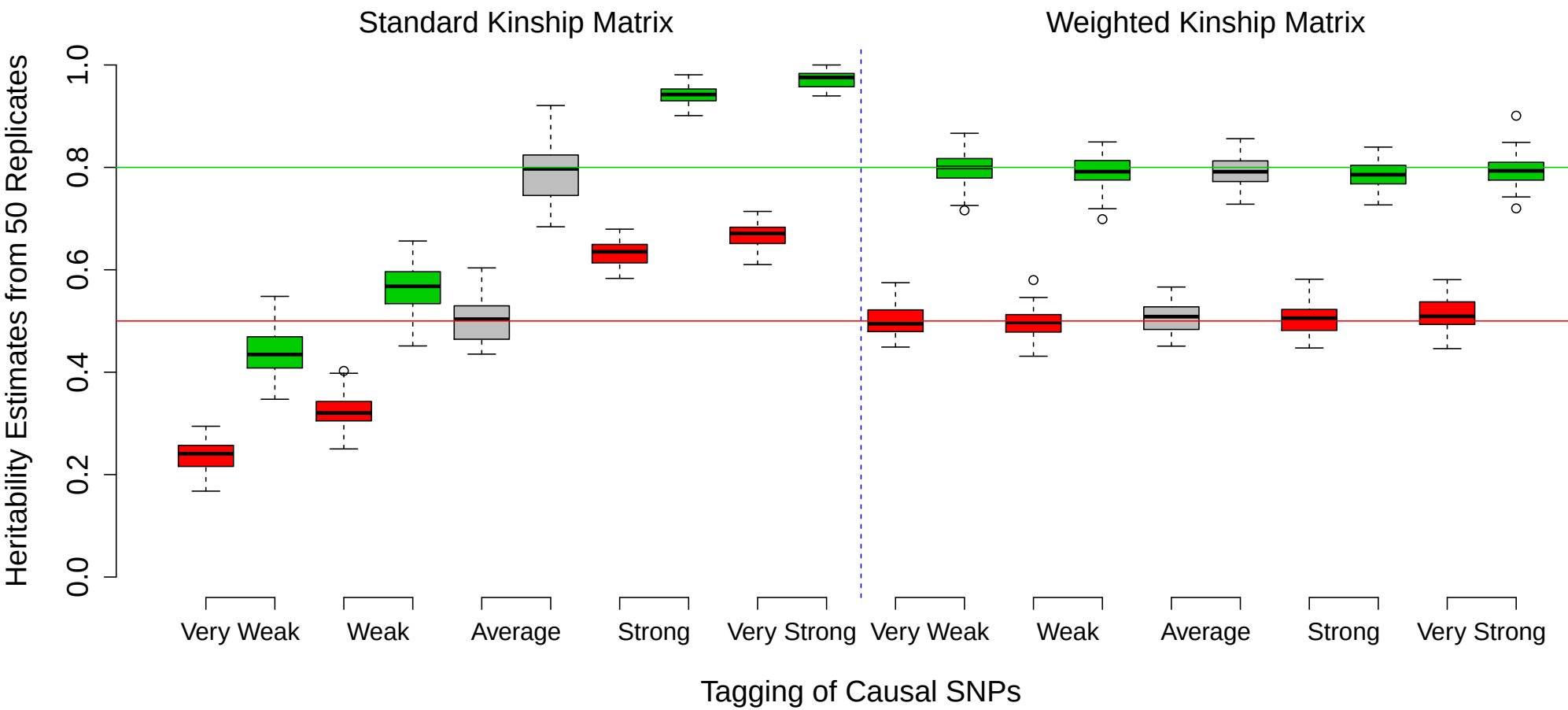
If an extra SNP is genotyped in high LD with an existing SNP tagging a causal, then part of the contribution to $\hat{h}^2$ from that causal is double-counted: "over-tagging".

This can occur whether or not the causal is itself genotyped.

Reweighting to reduce the problem of uneven tagging

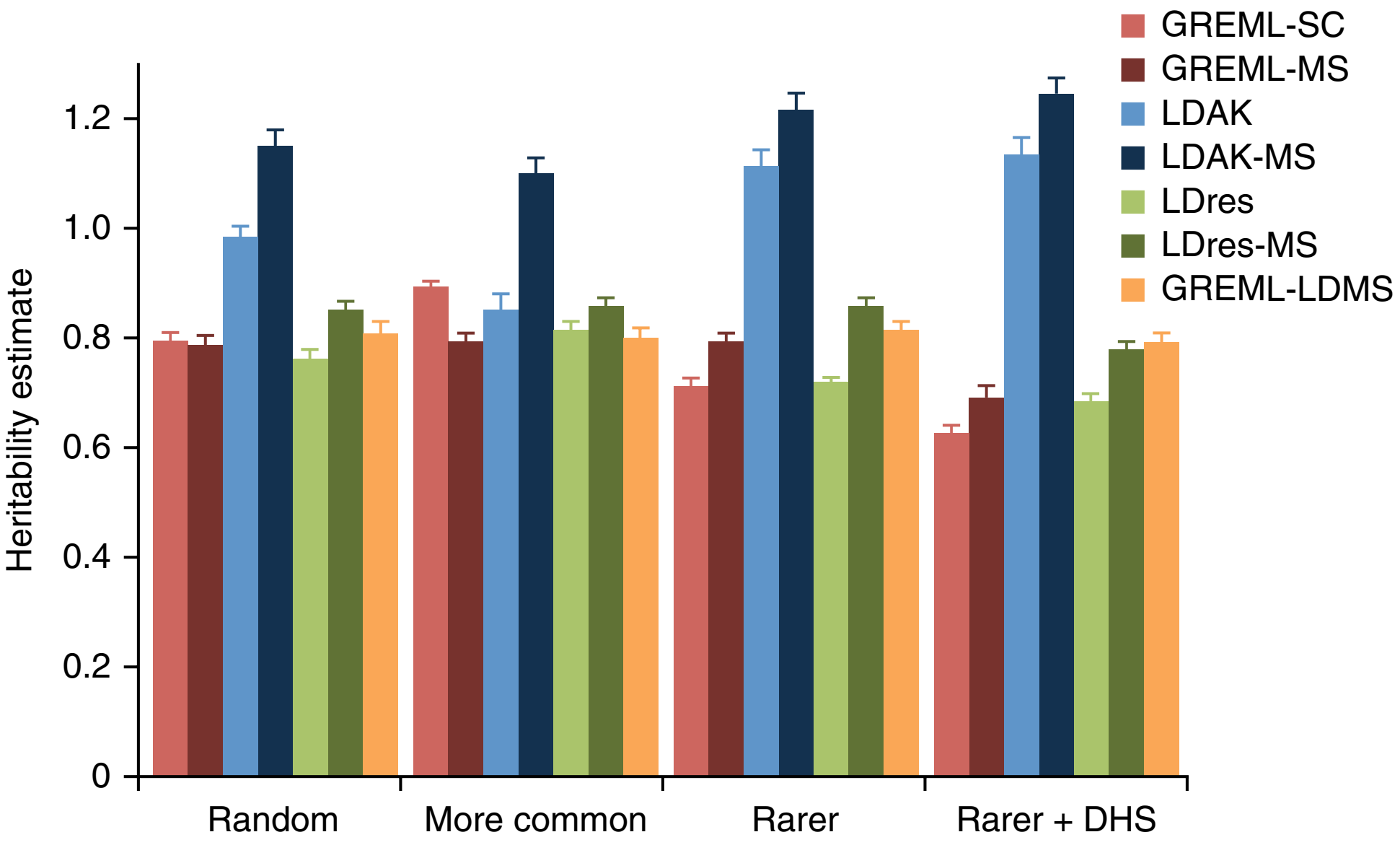


Reweighting improves estimation of h^2



This reweighting is implemented in Doug Speed's software for h^2 estimation and prediction, LDAK (LD-Adjusted Kinships) http://dougspeed.com/ldak/

How are causal variants distributed with respect to LD?



How are causal variants distributed with respect to LD?

- Yang *et al. Nat Genet* 2015 claimed that LDAK over-estimates h^2 for sequence data.
- But their simulations distributed causal variants across SNPs ignoring LD
  - So their simulations assume that the problem LDAK is designed to solve doesn't exist!
- LDAK is based on the idea of downweighting apparent contributions to h^2 when LD is high:
  - For SNP data, this makes sense, as SNPs in high LD are likely to be tagging the same causal variant (if any).
  - For sequence data, it also seems likely that two SNPs in high LD tag less causal variation in total than two SNPs in low LD.
  - This is an empirical question that can be checked (not easy).

New flexibility in heritability analysis

The mixed-model h^2 analysis brings with it useful computational tools, but is now unnecessary and with SNP data better to go back to the defining linear regression model $\mathrm{E}[Y] = \mathbf{X}\beta$ except now

- use genotyped SNPs as proxies for causal variants;
- apply a Gaussian "shrinkage" distribution on the β (ridge regression)

By restricting to SNPs in particular genomic regions, we can now investigate the distribution of h^2 across the genome.

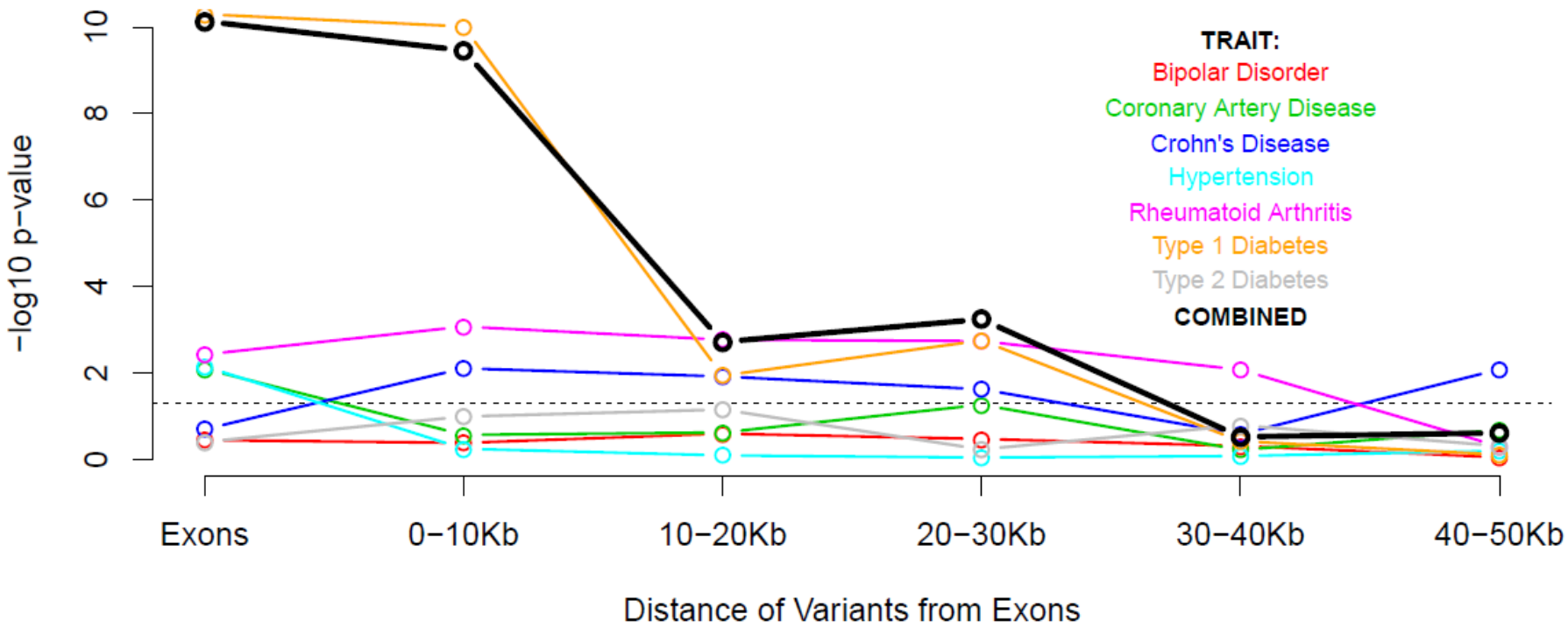
- Pioneered by Yang *et al.* (2011) but we've made several improvements.

h^2 intensity over genomic regions

For larger genomic regions we need to compare the heritability with that expected given the region size.

- "Intensity of heritability" is the heritability per unit genetic variance of the region.

Apply first to genes and their flanking regions:



h^2 intensity of exons and non-genic regions

		Intensity of heritability (h^2/1000 "SNPs")		
Trait	**Total h^2**	**Exons**	**Intergenic**	**P**
Bipolar Disorder	68%	1.7	1.3	0.37
Coronary Artery Disease	44%	3.1	0.6	**0.008**
Crohn's Disease	62%	1.6	0.7	0.21
Hypertension	54%	3.6	1.1	**0.007**
Rheumatoid Arthritis	52%	3.1	0.3	**0.004**
Type 1 Diabetes	76%	7.5	0.3	**5e-11**
Type 2 Diabetes	47%	0.9	0.6	0.40

Intensity of heritability for breast cancer eQTLs

$\sim$3K SNPs associated with expression of any gene in tumour tissue, corrected for somatic effects (from Curtis *et al.*, 2012).

		h^2 intensity		
Trait	h^2	eQTLs	other SNPs	p
Control-Control	21	0	0.3	0.55
CD	60	5	1	0.055
BD	64	5	1	0.074
CAD	37	1	0.6	0.47
T2D	46	3	0.8	0.16
Hypertension	48	2	0.8	0.30
Schizophrenia	62	0	1	0.76
RA	45	40	0.5	5e-32
T1D	63	70	0.6	2e-88

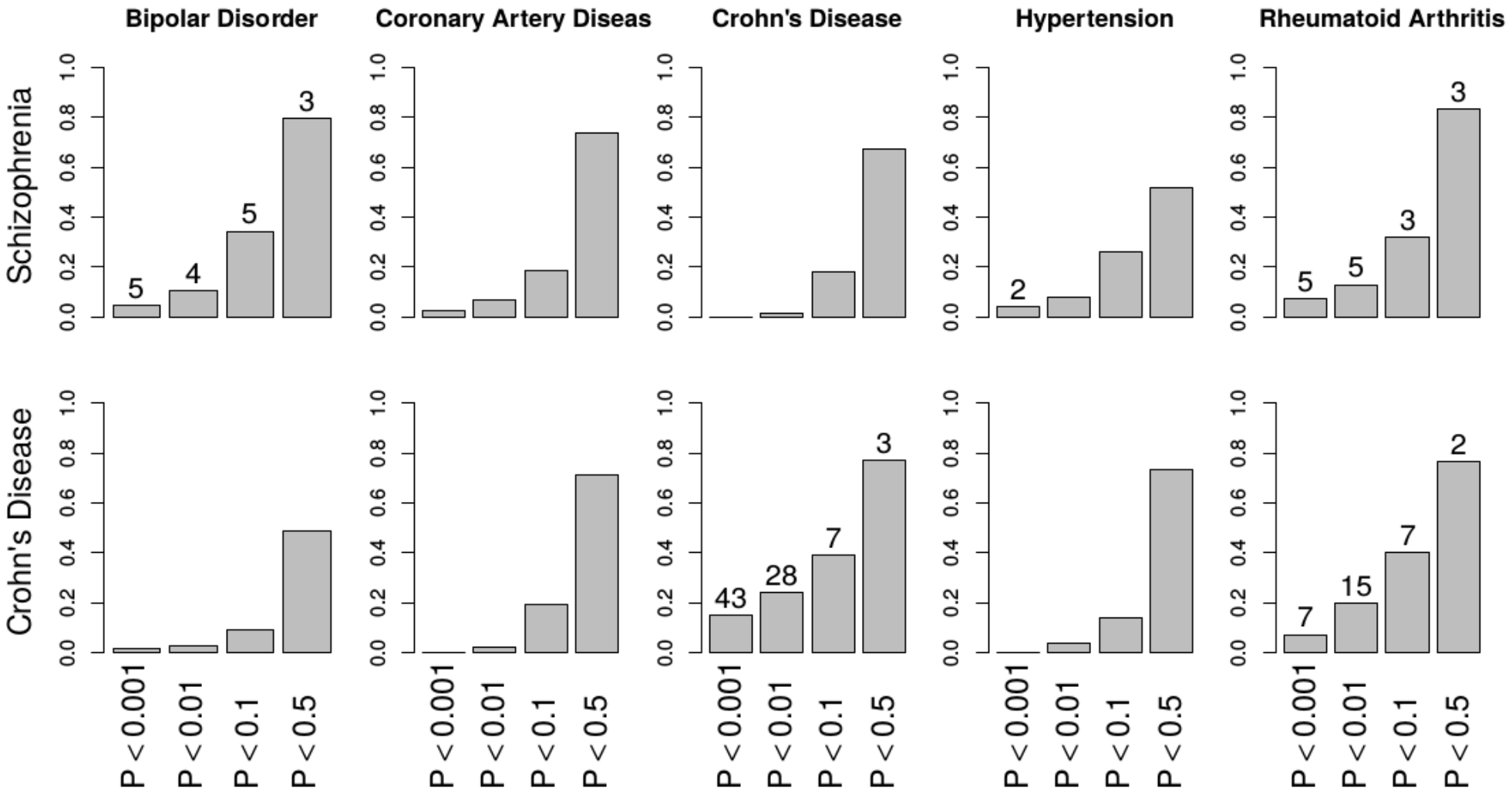
Many of the tumours have significant lymphocytic infiltration which could explain the large effect on the auto-immune diseases RA and T1D.

p-values for h^2 intensity of eQTLs in different tissue types

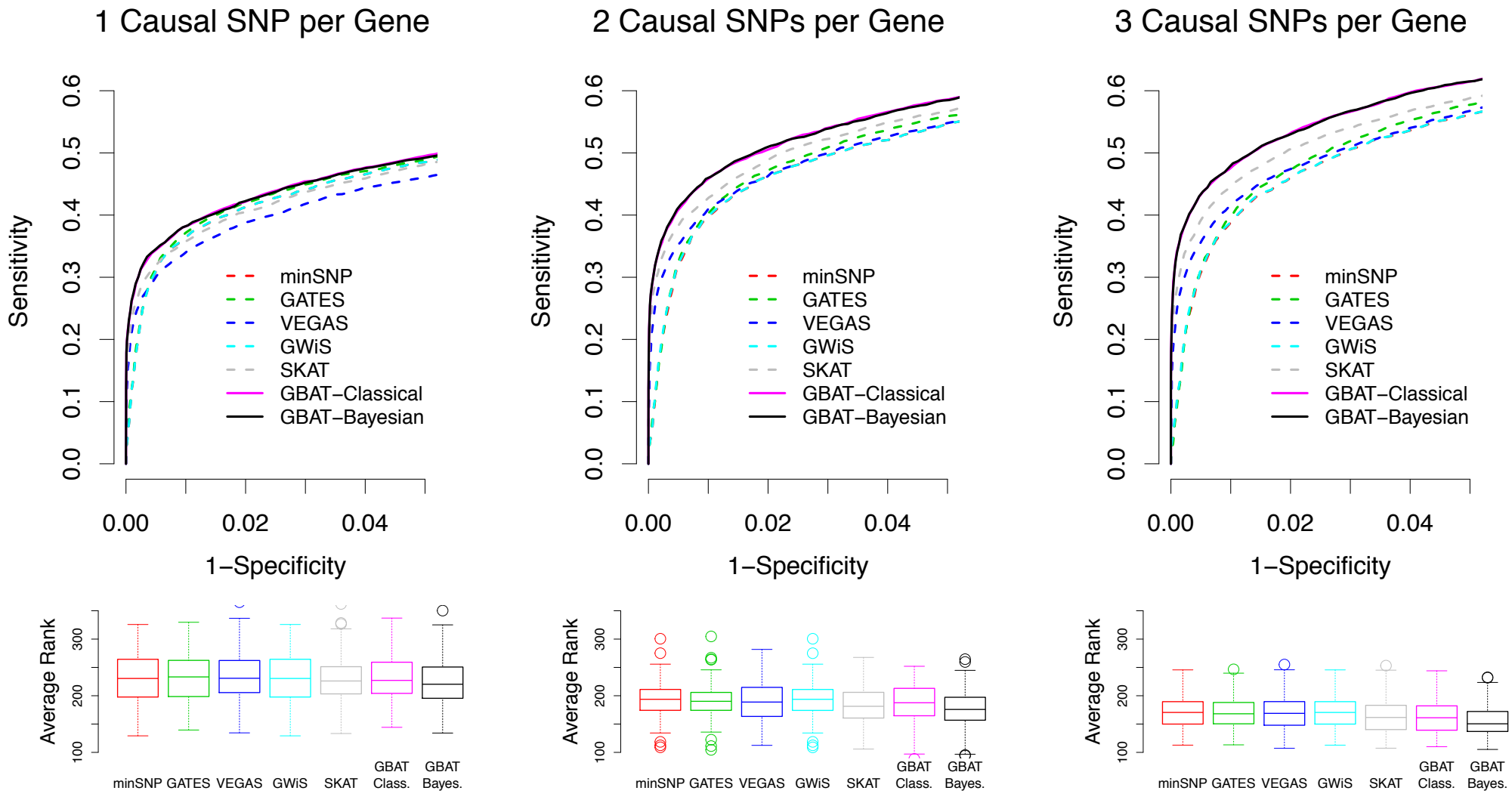
Trait	BC	Monocytes	EB-Lympho	Hap Map	Brain
C-C	0.55	0.60	0.59	0.38	0.72
CD	0.055	0.014	0.082	0.24	0.86
BD	0.074	0.078	0.79	0.50	0.92
CAD	0.47	0.27	0.44	0.30	0.71
T2D	0.16	0.20	0.70	0.46	0.56
Hyp	0.30	0.00027	0.39	0.74	0.93
Schiz	0.76	0.54	0.25	0.84	0.66
RA	5e-32	0.081	0.0070	0.044	0.257
T1D	2e-88	0.00021	7e-16	3e-5	0.75

- Human monocytes, Zeller 2010, PLoS1, n=1500, cis or trans.
- Epstein-Barr-transformed lymphoblastoid cell lines, Dixon 2007, Nat Genetics. n=400, cis or trans eQTLs.
- HapMap lymphoblastoid cell lines: Choy, 2008, Dimas 2009, Mongomery 2010, Pickrell 2010, Price 2008, Spielman 2008, Stranger 2007. n=1400, cis ONLY
- Human cortical gene expression, neuropathologically normal human brain samples (Myers, 2007). n=200, cis ONLY

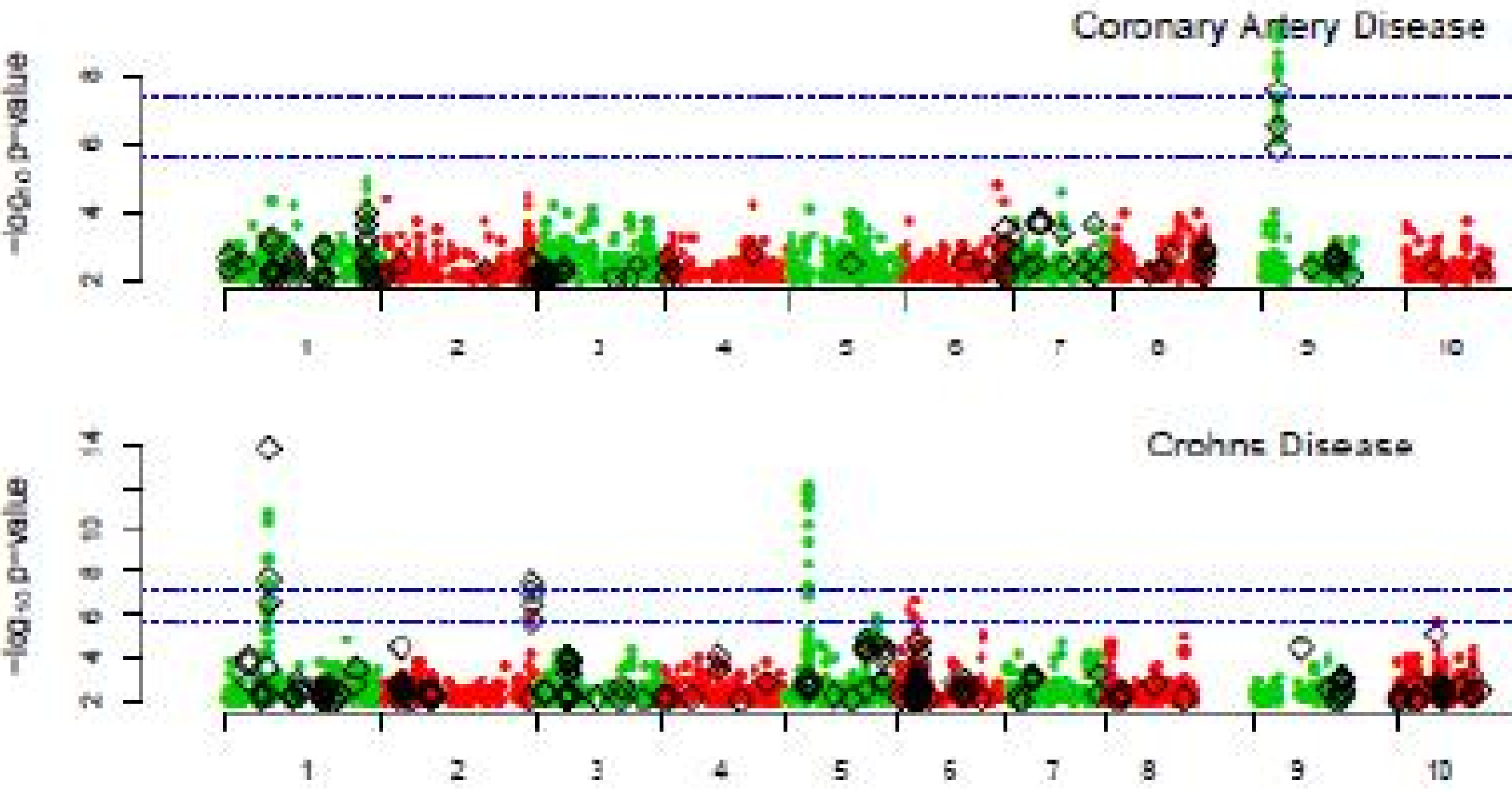
h^2 of SNPs associated with another trait



Gene-based tests of association using local h^2

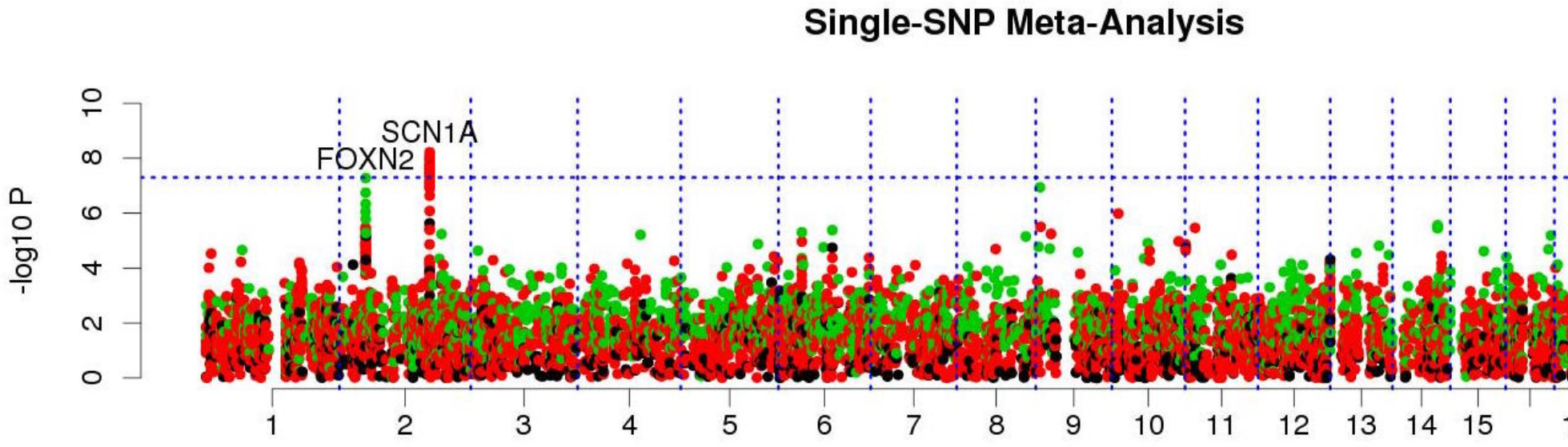


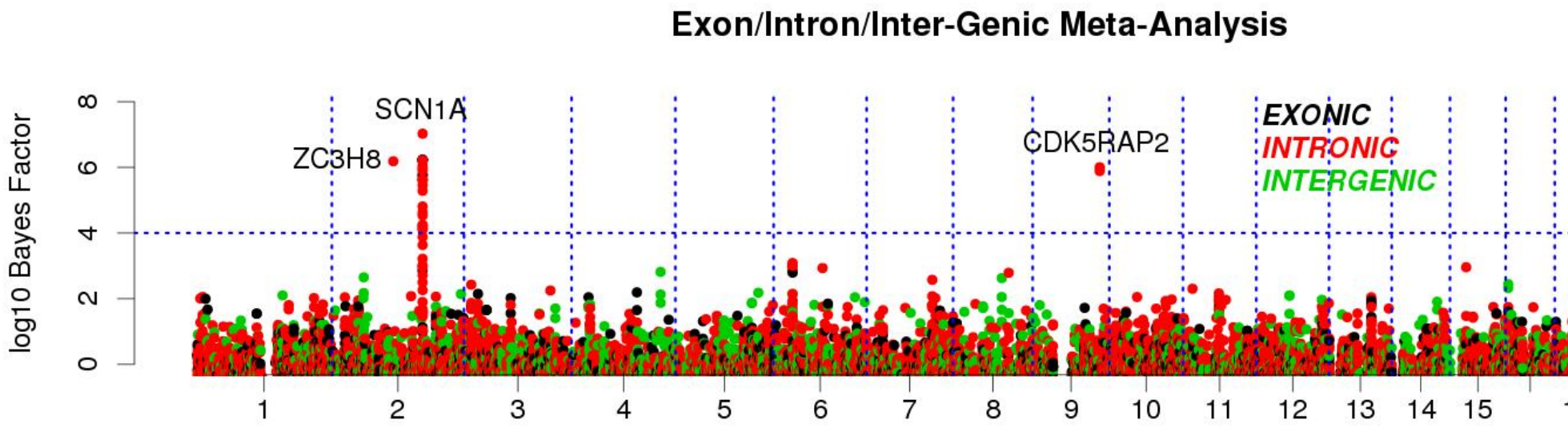
Gene-based association applied to three CCC traits



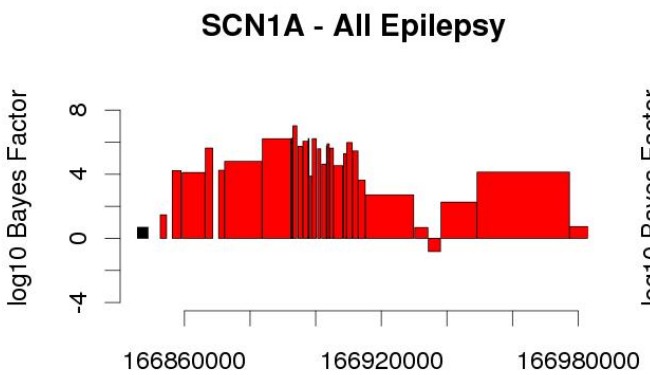
Gene-based tests (diamonds) complement single-SNP tests (red/green)

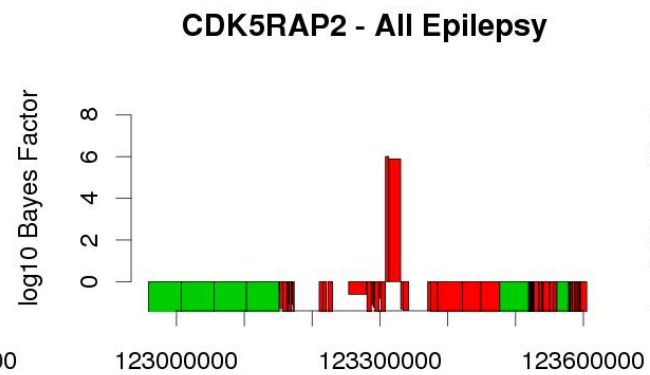
Extend gene-based tests to meta-analysis and subdivide genes into exons: Epilepsy consortium 12 cohorts

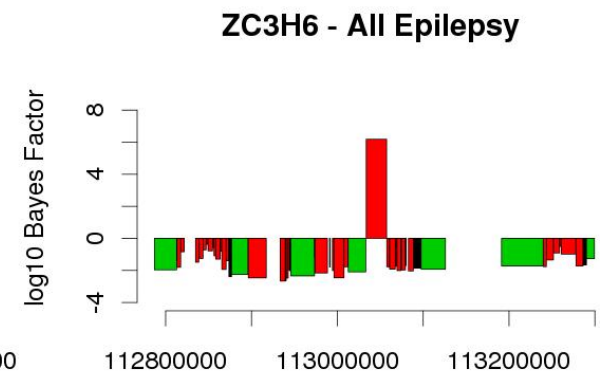


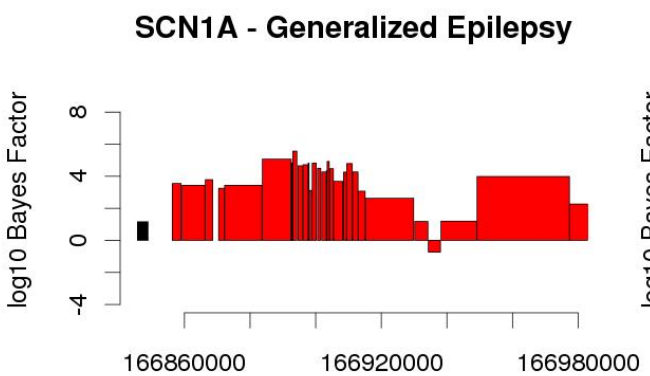


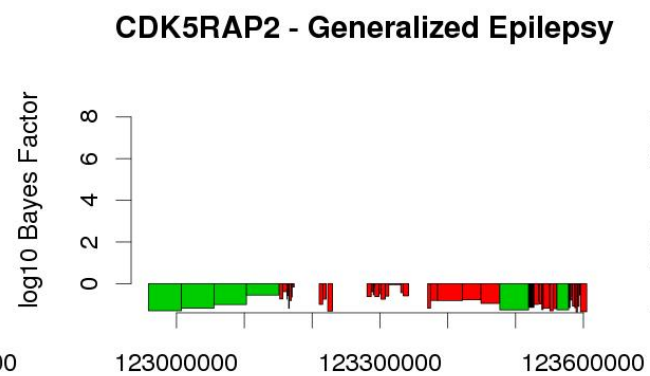
Closer look at top 3 hits: by genomic region

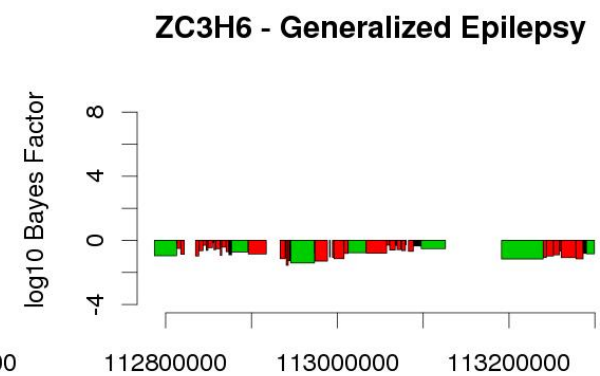


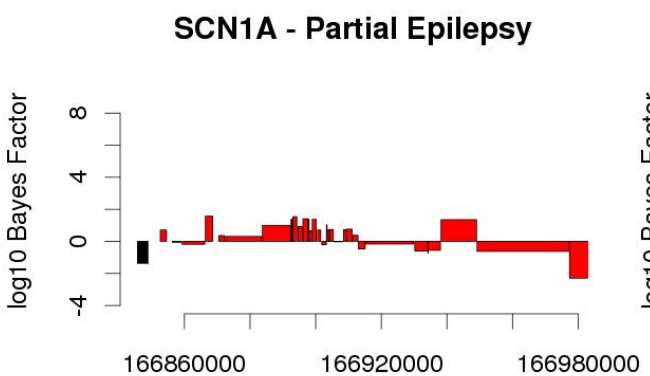


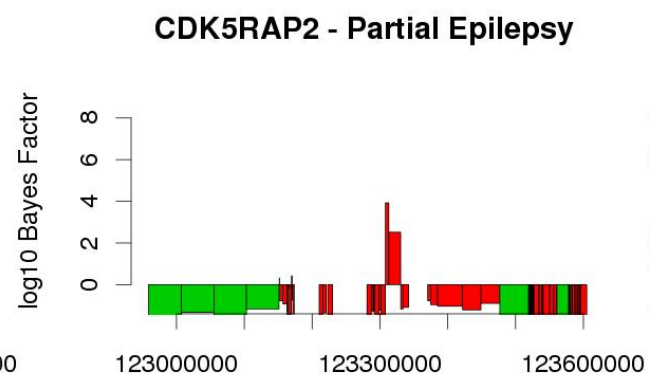


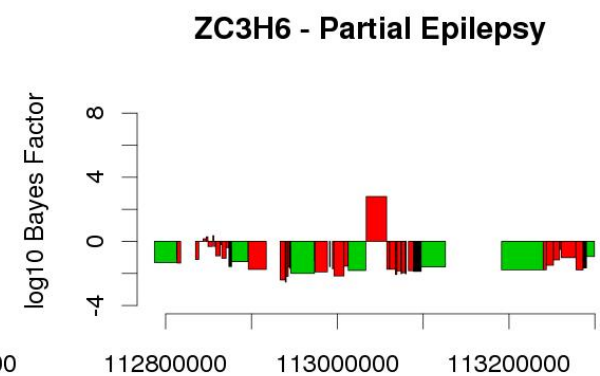




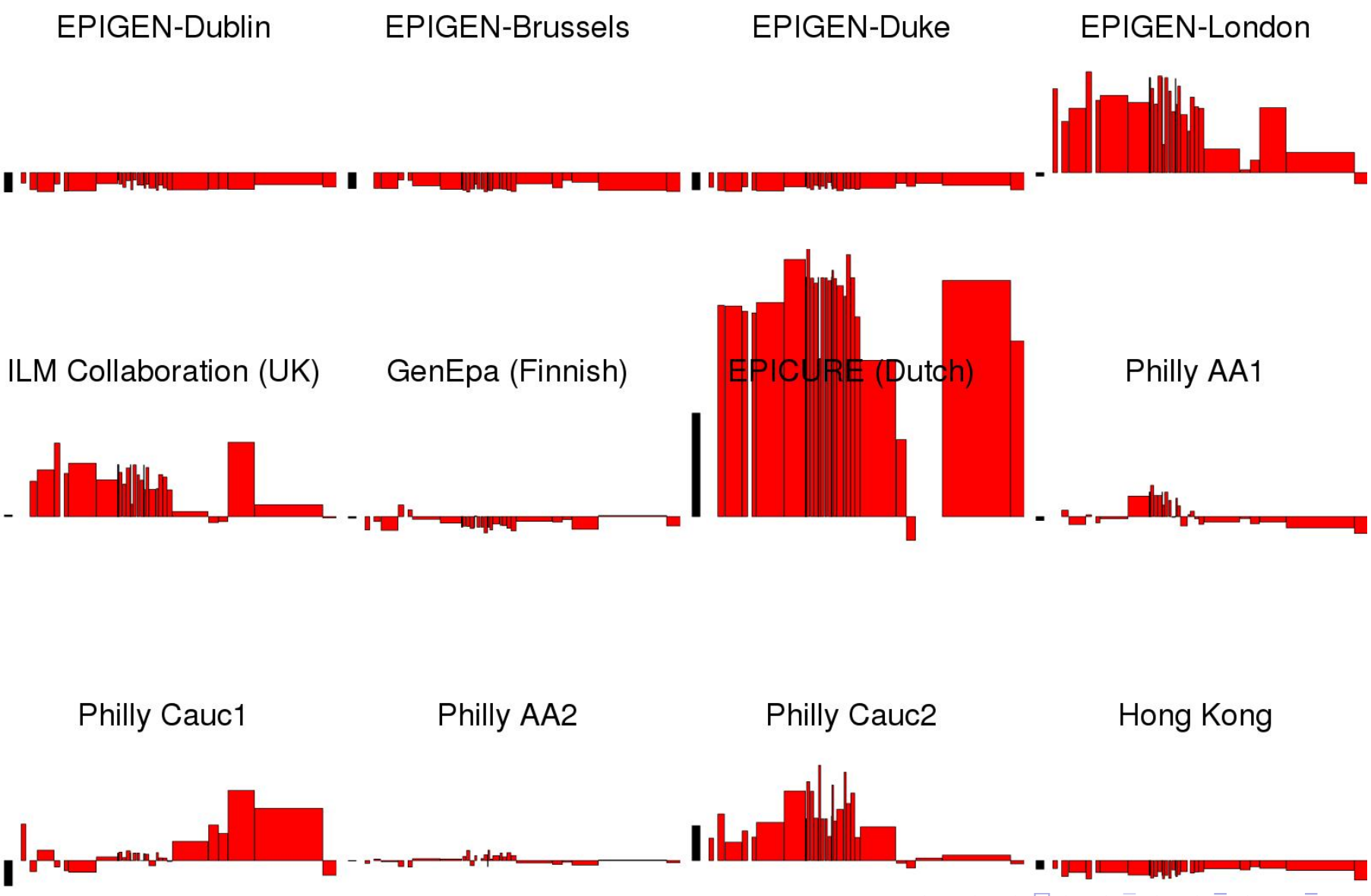








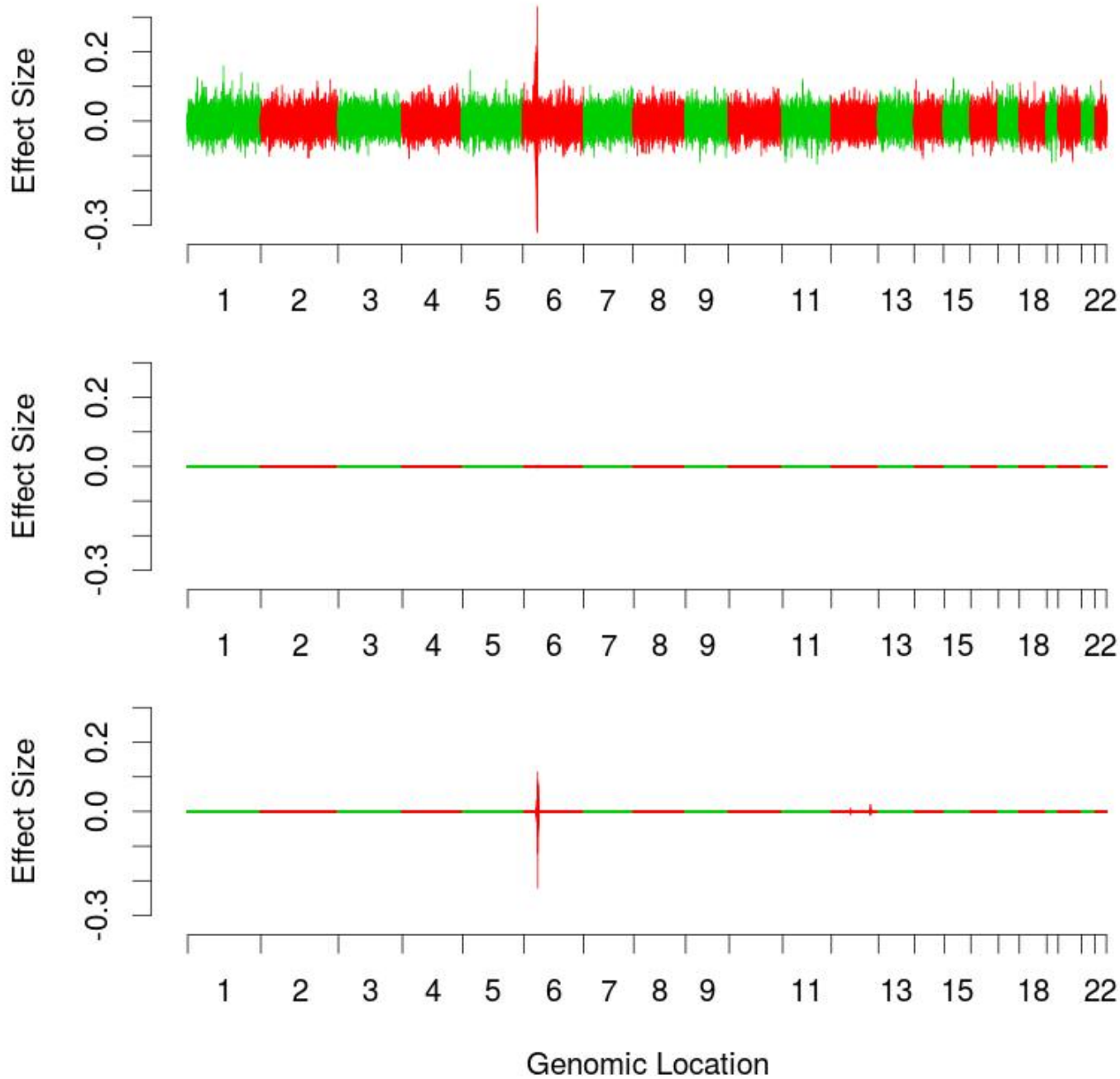
Closer look at SCN1A: by cohort



Prediction of phenotype from genome-wide SNPs

- The new ideas about heritability are having an impact on prediction of traits.
- BLUP is a long-established "shrinkage regression" technique for phenotype prediction, much used in animal/plant breeding.
- It uses a matrix of kinship coefficients to describe phenotype correlations due to (polygenic) inheritance.
- In the past, **pedigree** kinships, now SNP **allelic correlations**.
- **MultiBLUP** (Speed & Balding, *Genome Res*, Dec 2014) extends BLUP by allowing reduced shrinkage in promising genomic regions.
- Model $Y = \sum_{m=1}^{M} \gamma_m + \epsilon$ where $\text{Var}[\gamma_m] = \sigma_m^2 \mathbf{K}_m$ with K_m computed from SNPs in mth region and $\text{Var}[\epsilon] = \sigma_e^2 \mathbf{I}$.
- The M regions can be pre-specified or chosen by MultiBLUP.

MultiBLUP is incorporated in the LDAK software.



Genetic Profile Risk Scores

(no shrinkage)

BLUP

(uniform shrinkage)

Adaptive MultiBLUP

(flexible shrinkage, here $M = 3$)

Prediction for Crohn's Disease with 5 *a priori* regions: 3 pathways + 2 genes

Random Effect	Region h^2	Region r^2
IL-9 Signalling	0.006	0.003
IL-2 Receptor Beta Chain	0.003	0.001
IL12 Pathway	0.019	0.016
Gene NOD2	0.012	0.012
Gene IL23R	0.008	0.007
Background Region	0.96	0.09

Correlation of predicted and true values in cross-validation improves from 0.10 (BLUP) to 0.12 (MultiBLUP with 5 regions).

Adaptive MultiBLUP for WTCCC 1 disease traits

	Current methods				
Trait	BLUP	Risk Score ($-\log_{10}(P)$)	Stepwise Regression	BSLMM	Adaptive MultiBLUP
BD	**0.27**	0.25 (1)	0.02	0.27	0.27
CAD	0.13	0.12 (1)	0.08	0.15	**0.16**
CD	0.32	0.28 (1)	0.18	0.34	**0.36**
Ht	0.15	0.14 (1)	0.00	0.14	**0.17**
RA	0.21	0.28 (3)	0.32	0.33	**0.37**
T1D	0.25	0.34 (5)	0.54	0.57	**0.59**
T2D	0.16	0.14 (1)	0.10	0.17	**0.18**
Av.	0.21	0.22	0.18	0.28	**0.30**

Entries are correlations, **bold** indicates highest predictive accuracy.

Compute times: Risk score / BLUP: $<$ 1 hr, Stepwise Regression: 2 hrs to 5 days, MultiBLUP: 2-3 hrs, BSLMM: 8-30 hrs.

Some larger datasets

Stepwise Regression and BSLMM not feasible.
Performance, measured as correlation (AUC):

Irritable Bowel Disease (12,678 individuals, 1.5M SNPs):

- BLUP: 0.15 (0.58)
- Risk Score: 0.21 (0.63)
- MultiBLUP: 0.34 (0.68)

Celiac Disease (15,283 individuals, 200k SNPs):

- BLUP: 0.40 (0.76)
- Risk Score: 0.44 (0.78)
- MultiBLUP: 0.54 (0.84)

Speed et al. *Brain* 2014: "heritability" analysis of epilepsy

- Estimated 26% of variance of the liability to "all epilepsy" is attributable to 4 million genotyped and imputed SNPs (after correction for population structure effects and genotyping errors).
  - SNPs near previously-reported epilepsy loci explain only about 4% of variance.
  - Can similarly attribute heritability to various functional classifications (up to a margin of error).
  - Contribution from different large-scale genomic regions approximately uniform.
- From lack of genome-wide significant SNPs, inferred 100s and probably 1,000s of causal variants.
- Common genetic basis of focal and non-focal epilepsy estimated around 50% of total
  - imprecise estimate, but significantly different from both 0 and 100%.
- Showed potential for useful prediction of disease progression in single-seizure cases.