Fun with Mixed Models

Overview

The Linear Mixed Model

Suppose our GWAS data comprise

Phenotype Y (vector of length n)
SNP calls S (matrix of size $n \times N$)
Plus any covariates Z.

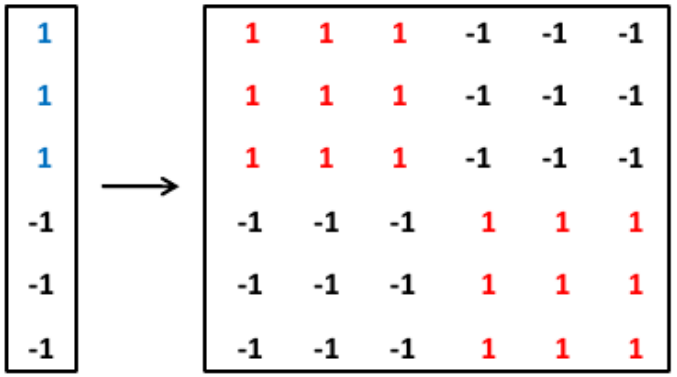
$$Y = Z\alpha + g + e \quad \text{with} \quad g \sim \mathbb{N}(0, K\sigma_g^2) \quad \text{and} \quad e \sim \mathbb{N}(0, I\sigma_e^2)$$

α is a vector of fixed effects, g and e are the genetic and environmental random effects (with corresponding components of variance σ_g^2 and σ_e^2).

Typically, use kinship matrix $K = \frac{XX^T}{N}$, where $X_{ij} = \frac{S_{ij} - \bar{S}_j}{SD(S_j)}$.

Traditionally Used for Heritability Estimation

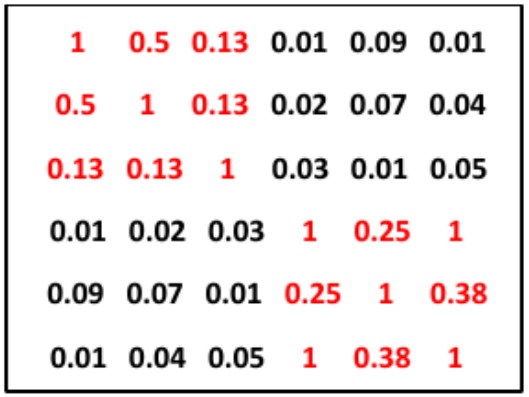
We are deciding how to divide the phenotypic correlation
into a genetic component K σ_g^2
and a noise component I σ_e^2



Phenotype Y

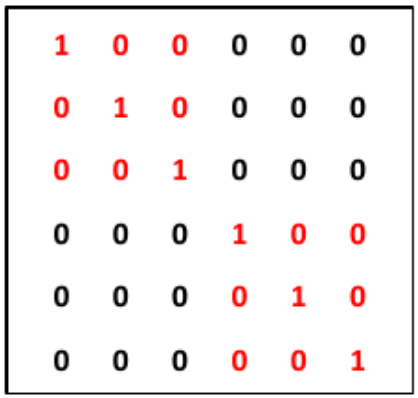
Phenotypic covariance Y Y^T

σ_g^2



Genetic Similarities K

σ_e^2



Identity (Noise) Matrix

Divide the phenotypic variance:
σ_g^2 to the genetics
σ_e^2 to the noise term

Then the heritability is $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$

Solved Via REML

$Y = Z\alpha + g + e \quad$ with $\quad g \sim \mathbb{N}(0, K\sigma_g^2) \quad$ and $\quad e \sim \mathbb{N}(0, I\sigma_e^2)$

The raw model likelihood follows from assuming

$Y \sim \mathbb{N}(Z\alpha, V) \quad$ where $\quad V = K\sigma_g^2 + I\sigma_e^2$:

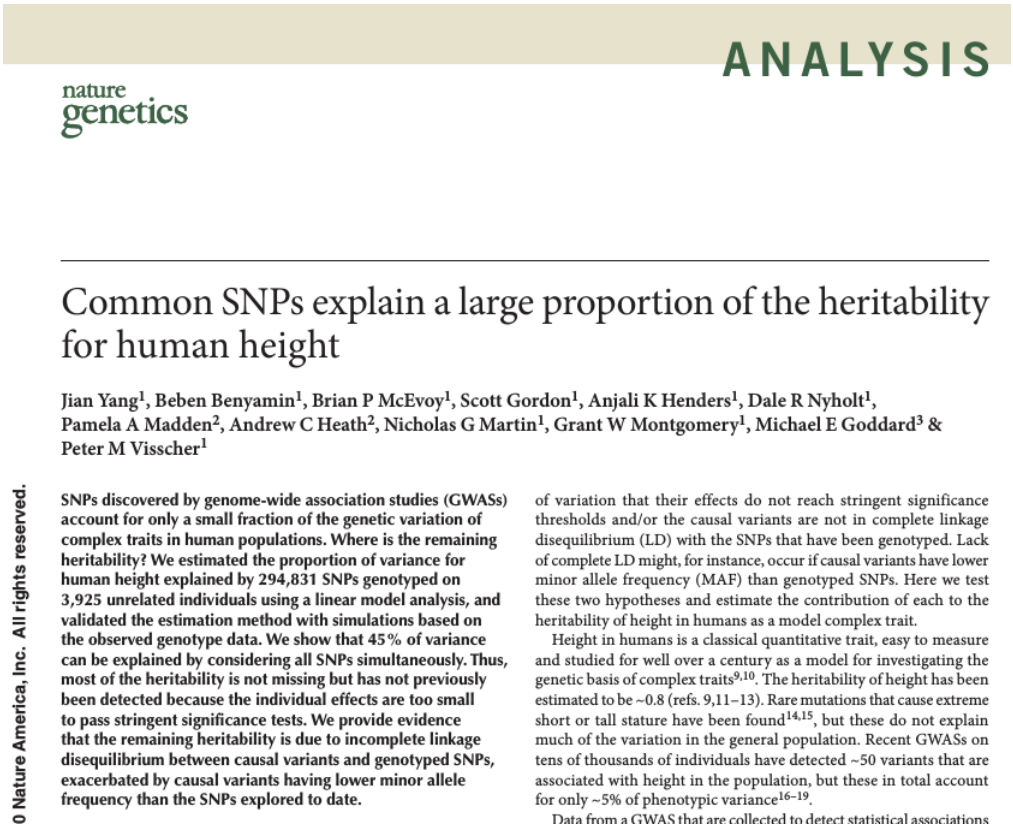
$$l(Y|\alpha, K, \sigma_g^2, \sigma_e^2) = -\frac{n}{2}\log(2\pi\sigma_e^2) - \frac{1}{2}(Y - Z\alpha)^T V^{-1}(Y - Z\alpha) - \frac{1}{2}\log|V|.$$

The restricted likelihood is obtained by "integrating across" α.

$Y \sim \mathbb{N}(0, P) \quad$ where $\quad P = V^{-1} - V^{-1}Z(Z^T V^{-1} Z)^{-1} Z^T V^{-1}$:

$$l(Y|K, \sigma_g^2, \sigma_e^2) = -\frac{n\text{-}p}{2}\log(2\pi\sigma_e^2) - \frac{1}{2}Y^T PY - \frac{1}{2}\log|V| - \frac{1}{2}\log|Z^T V^{-1} Z|.$$

Estimating Total SNP Heritability



ANALYSIS

nature genetics

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]

0 Nature America, Inc. All rights reserved.

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].

Data from a GWAS that are collected to detect statistical associations

Jian Yang *et al.* realised by applying to "unrelated individuals", could estimate total proportion of phenotypic variance explained by all SNPs.

Linear Random Effects Regression Model

Suppose we assume the following relationship:

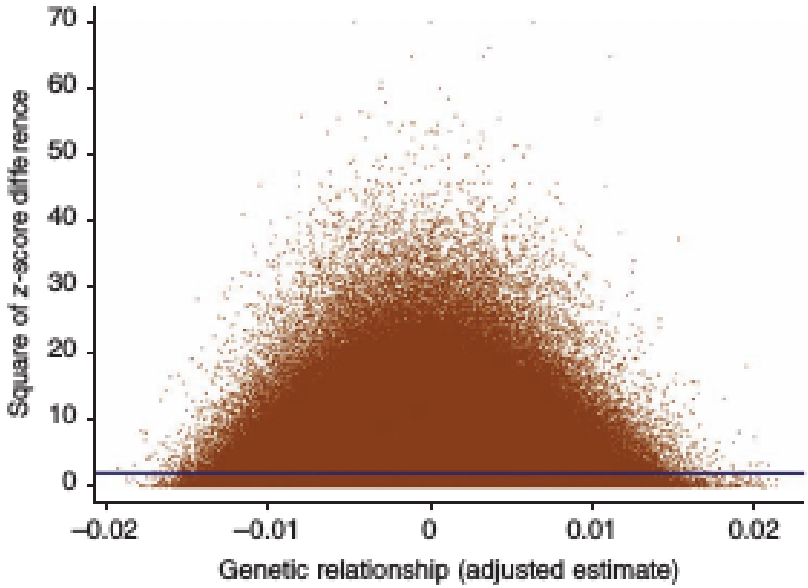
$$
\begin{aligned}
Y = \alpha & \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
& + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
& + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
& + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
& + \ldots + \beta_{500\,000} X_{500\,000} \\
& + e,
\end{aligned}
$$

where $\beta_j \sim \mathbb{N}(0, \sigma_g^2/N)$ and $e \sim \mathbb{N}(0, \sigma_e^2)$.

Then $g = \sum_{j=1}^{N} \beta_j X_j = X\boldsymbol{\beta} \sim \mathbb{N}(0, \frac{XX^T}{N}\sigma_g^2)$ and $Y \sim \mathbb{N}(\alpha, K\sigma_g^2 + I\sigma_e^2)$

where $K = \frac{XX^T}{N}$

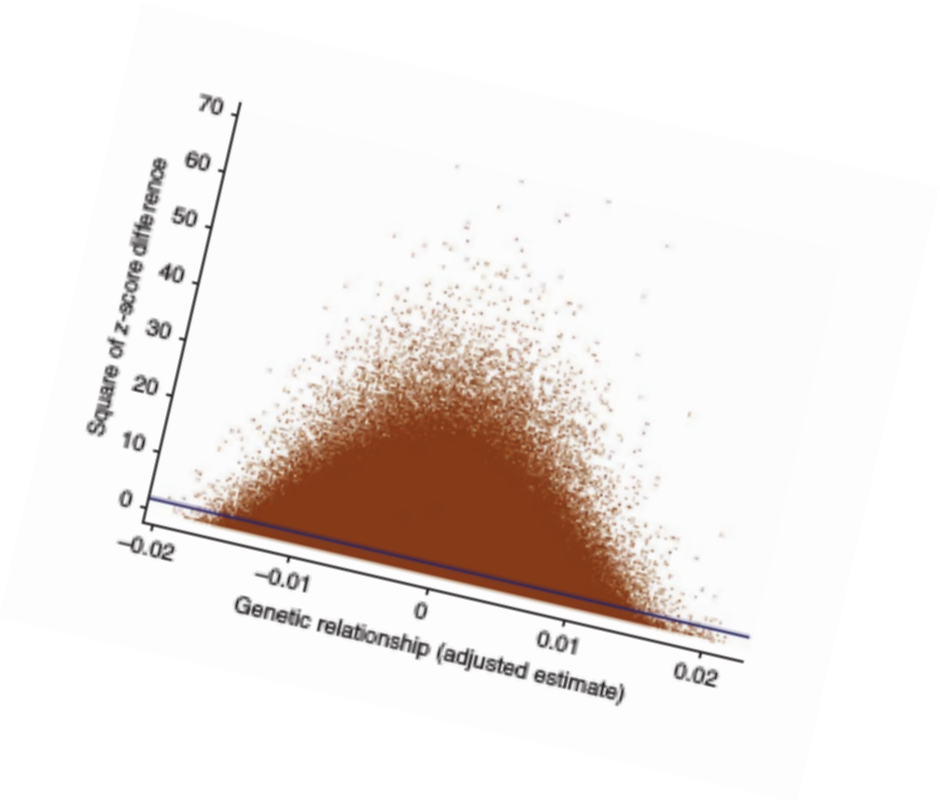
Estimating Total SNP Heritability



The heritability of human height is 80%.

Jian Yang *et al.* calculated that 45% of phenotypic variance could be explained by common SNPs.

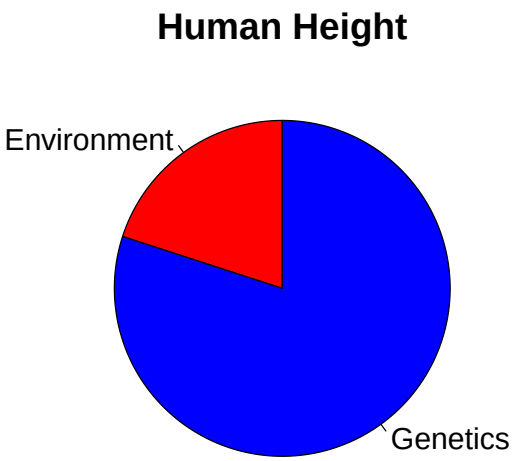
Estimating Total SNP Heritability



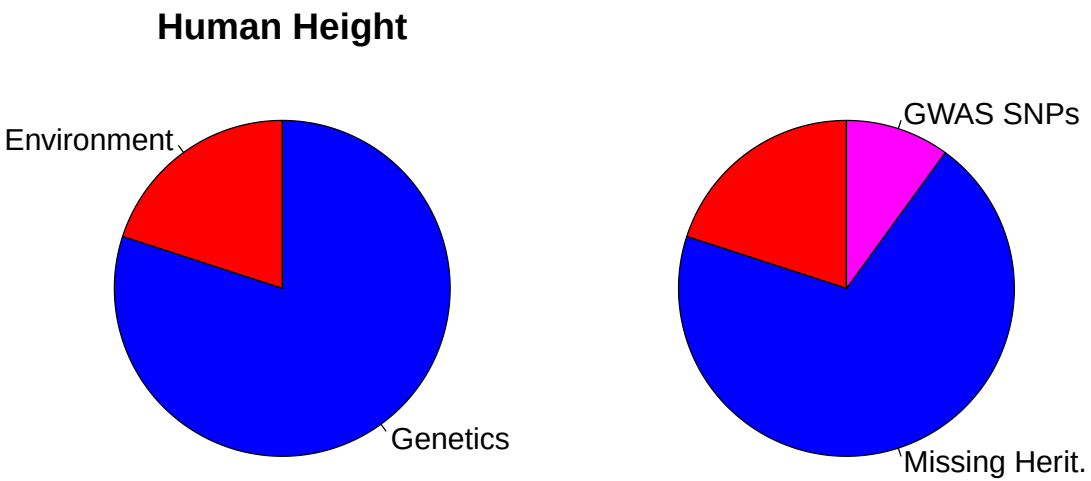
The heritability of human height is 80%.

Jian Yang *et al.* calculated that 45% of phenotypic variance could be explained by common SNPs.

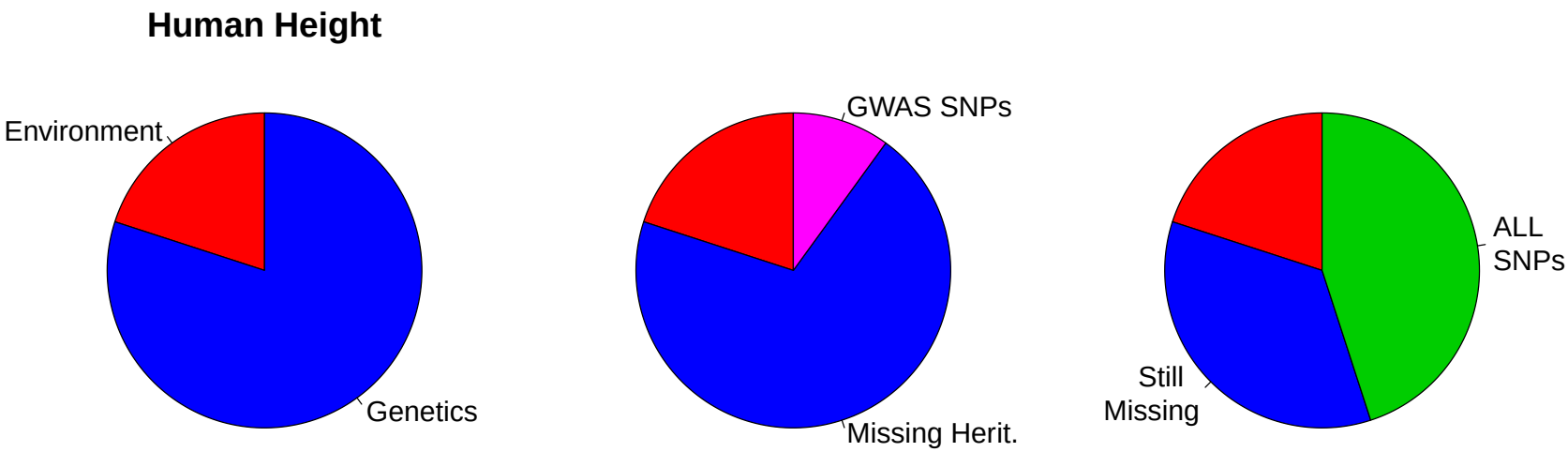
Solves the "Missing Heritability" Problem



Solves the "Missing Heritability" Problem

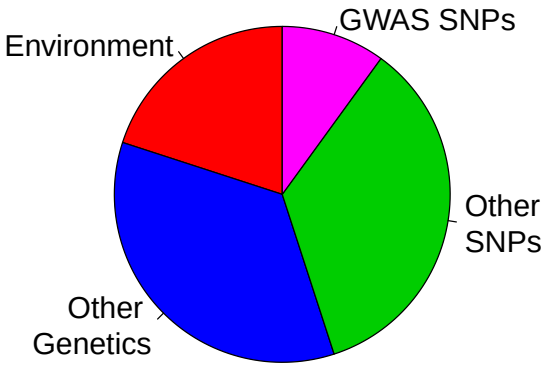


Solves the "Missing Heritability" Problem

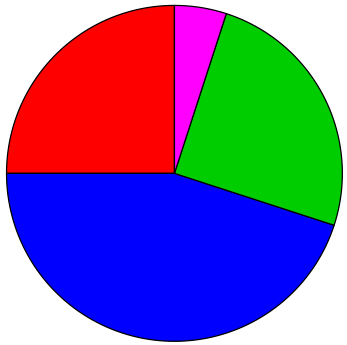


Solves the "Missing Heritability" Problem

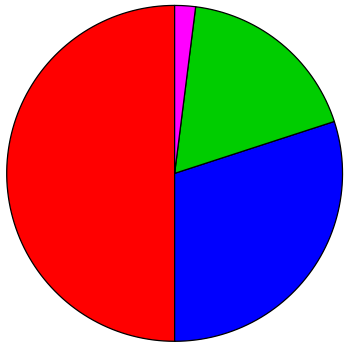
Human Height



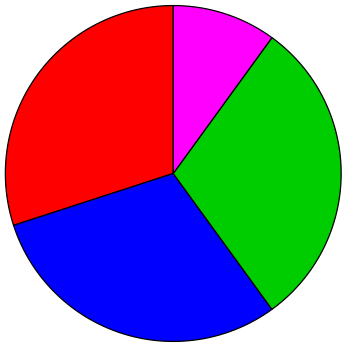
Schizophrenia



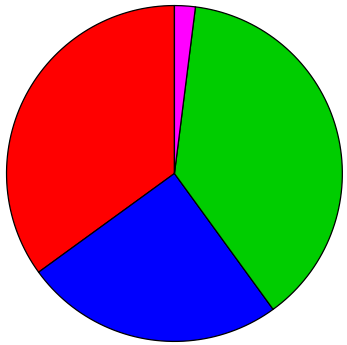
Obesity



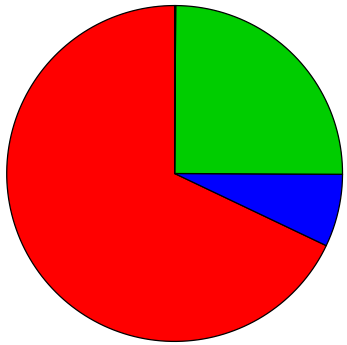
Crohn's Disease



Bipolar Disorder



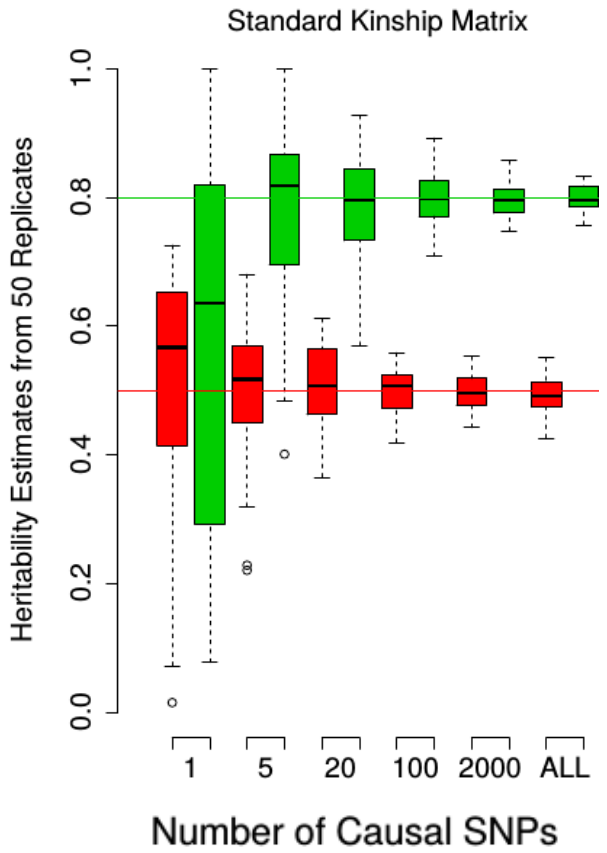
Epilepsy



The Method Generally Works

REML Assumptions:

- All SNPs are Causal
- Gaussian Effect Sizes
- Gaussian Noise Terms
- Inverse Relationship between MAF and Effect Size



Overall, we found the approach amazingly robust to misspecification

Computing $K = \frac{XX^T}{N}$

K =

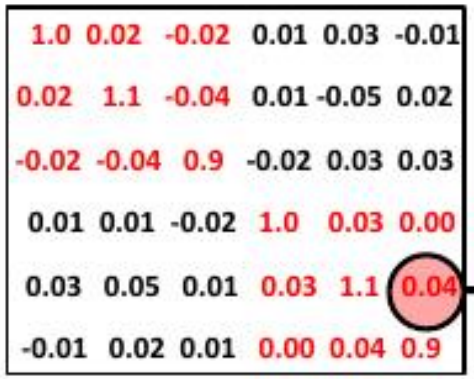
1.0	0.02	-0.02	0.01	0.03	-0.01
0.02	1.1	-0.04	0.01	-0.05	0.02
-0.02	-0.04	0.9	-0.02	0.03	0.03
0.01	0.01	-0.02	1.0	0.03	0.00
0.03	0.05	0.01	0.03	1.1	0.04
-0.01	0.02	0.01	0.00	0.04	0.9

Each entry of K represents genetic similarities between a pair of individuals.

e.g., this (single) value, summarizes all the similarities and differences between the genomes of Individuals 5 and 6.

Individual 5 SNPs	0	1	2	1	2
Individual 6 SNPs	2	1	2	1	0
Contribution to K_{56}	−	+	+	+	−

Impact of Uneven Tagging



Each entry of K represents genetic similarities between a pair of individuals.

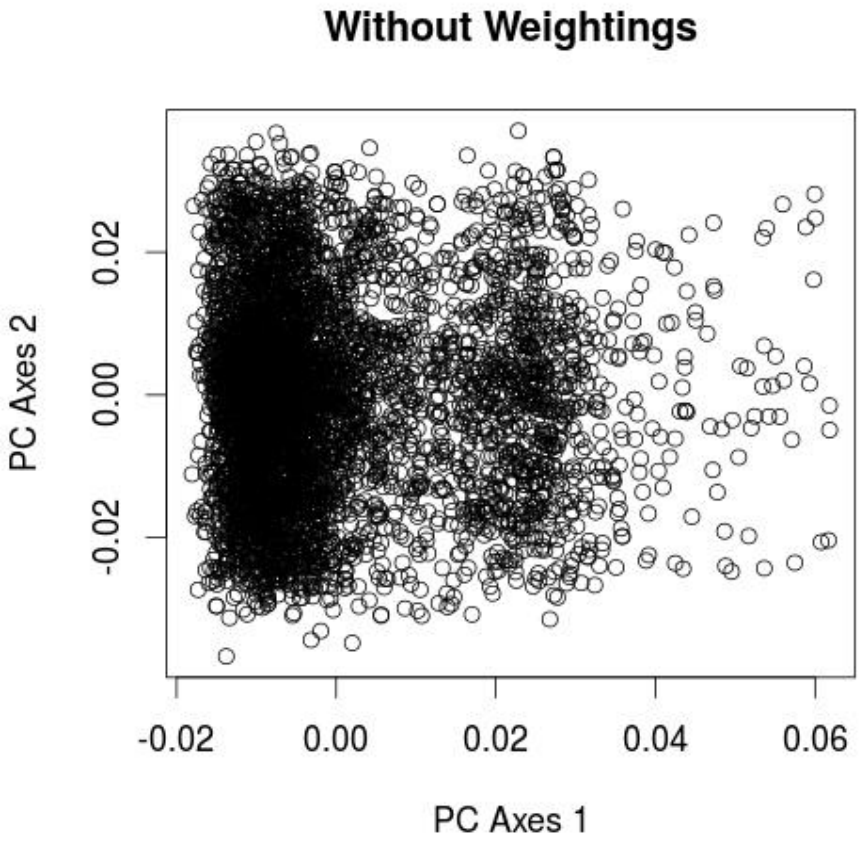
e.g., this (single) value, summarizes all the similarities and differences between the genomes of Individuals 5 and 6.

Individual 5 SNPs	0	1	0 2 2 2 0	1	2
Individual 6 SNPs	2	1	0 2 2 2 0	1	0
Contribution to K_{56}	−	+	+ + + + +	+	−

HIGH LD REGION

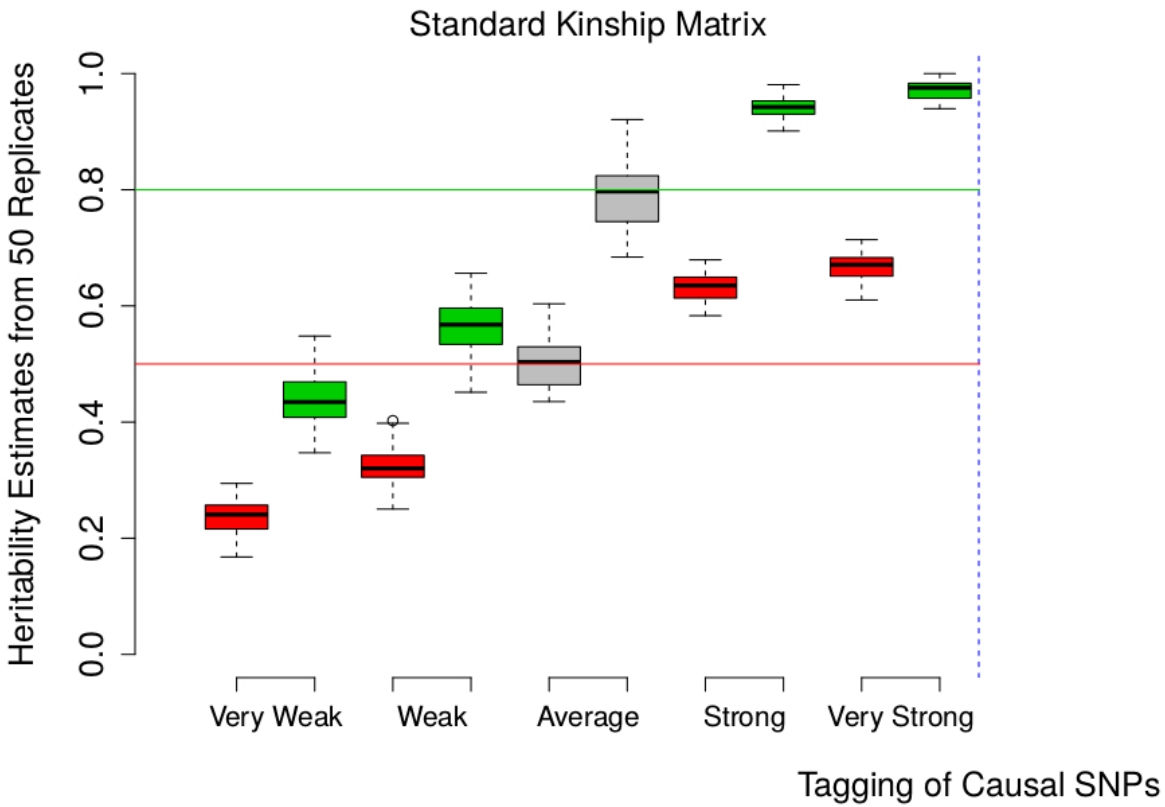
Regions of high LD have disproportionately large contribution

Impact of Uneven Tagging



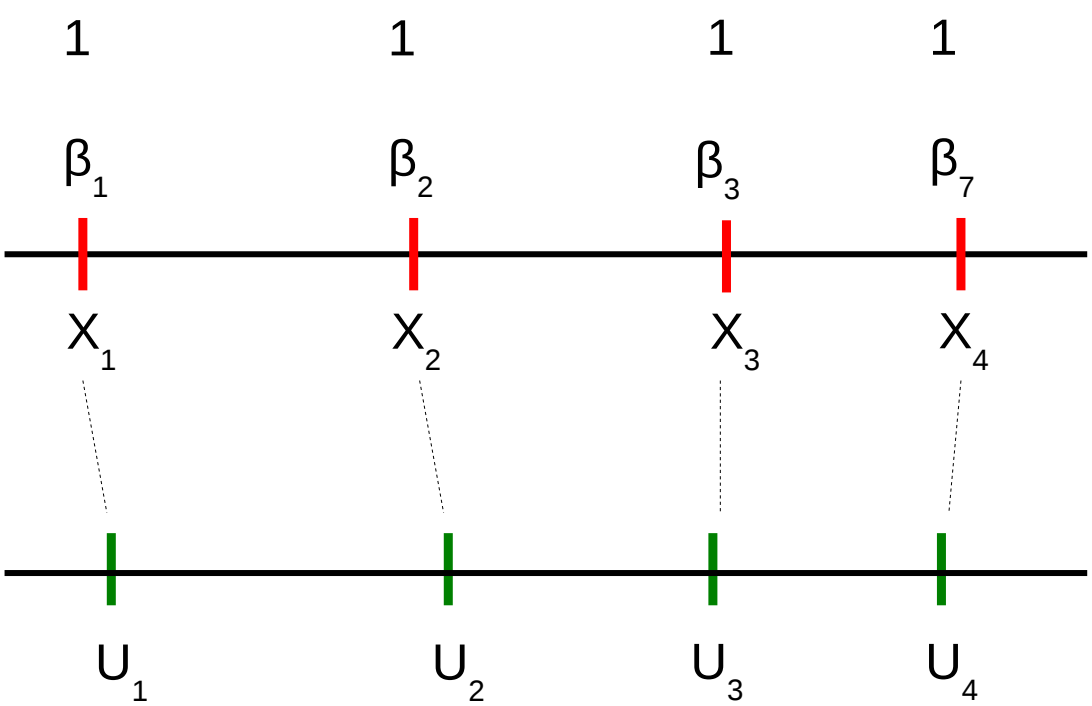
A common problem when performing principal component analysis.

Estimates Can be Sensitive to LD of Causal Variants

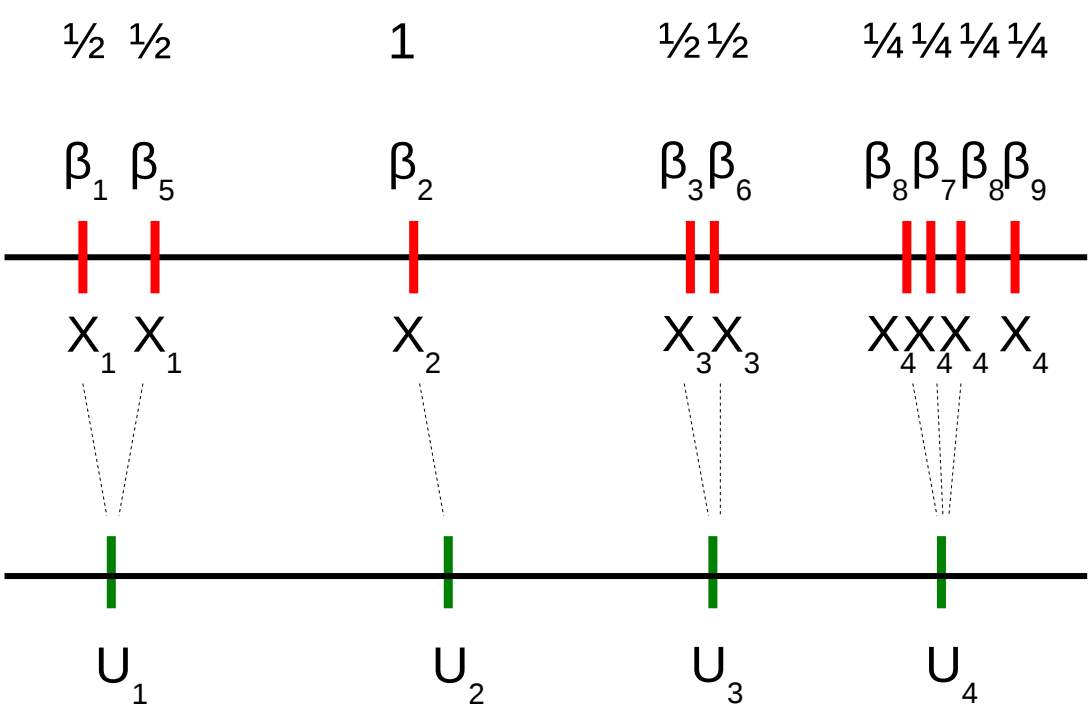


Causal variants in high LD areas $\Rightarrow$ over-estimation of h^2_{SNP}
Causal variants in low LD areas $\Rightarrow$ under-estimation of h^2_{SNP}

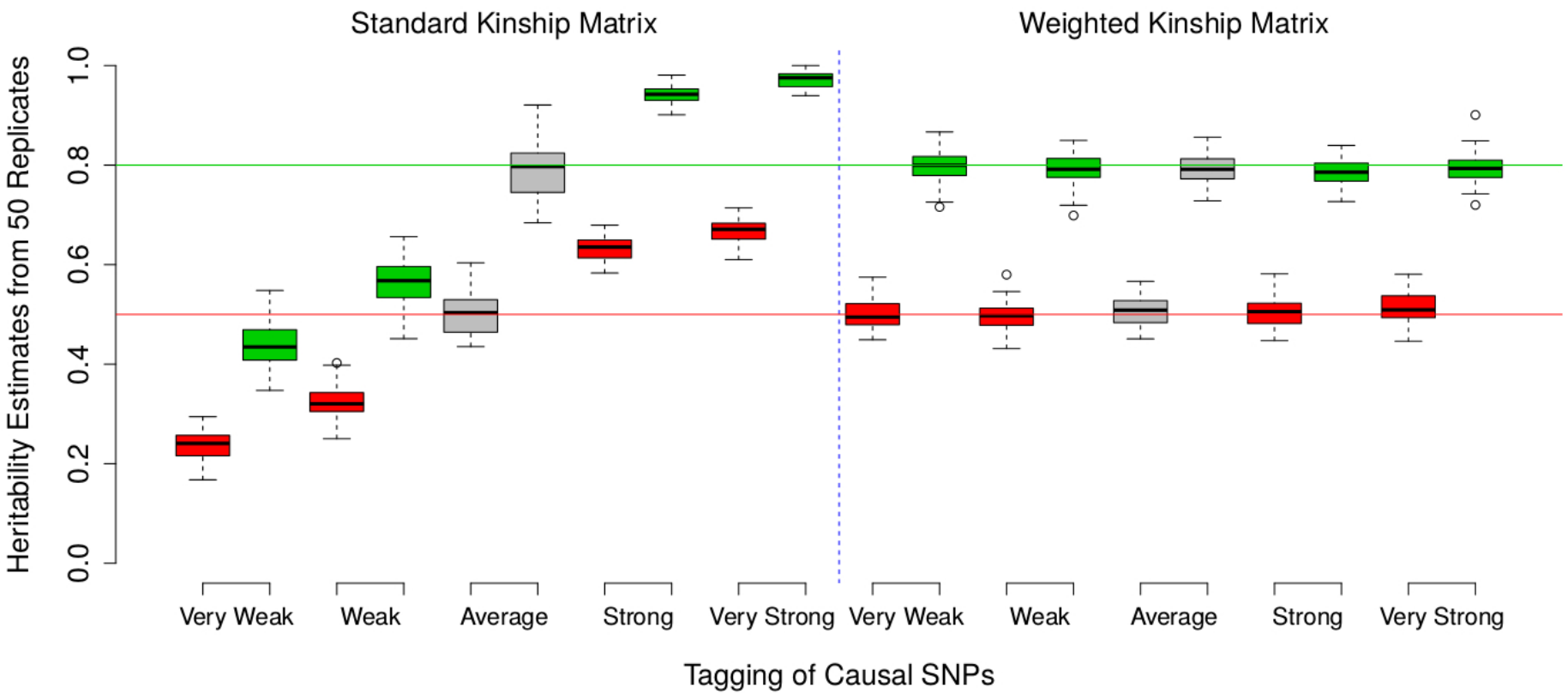
Adjusting for Uneven Tagging



Adjusting for Uneven Tagging

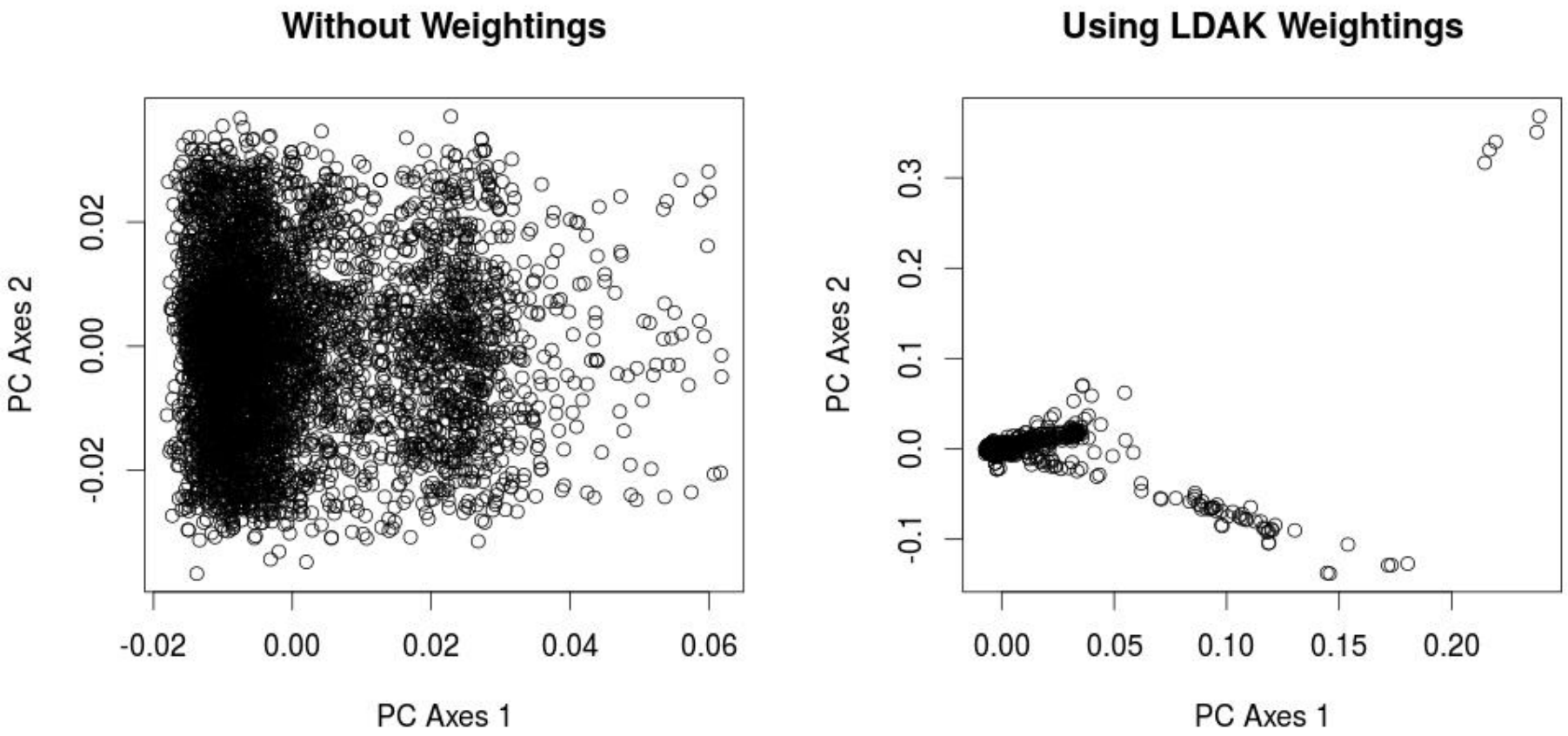


Weightings Reduce the Biases



LDAK weightings down-weight SNPs well-tagged by neighbours and up-weight SNPs poorly-tagged by neighbours

LDAK: Linkage Disequilibrium Adjusted Kinships



LDAK weights offer an alternative to pruning.
e.g., when computing genetic profile risk scores.

GCTA vs LDAK

In the end, whether SNPs explain 50% or 60% of heritability not a big deal.

Basic Model

$$
\begin{aligned}
Y = \alpha & \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
& + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
& + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
& + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
& + \ldots + \beta_{500\,000} X_{500\,000} \\
& + e.
\end{aligned}
$$

Assume $\beta_j \sim \mathbb{N}(0, \sigma_g^2/N)$ and $e \sim \mathbb{N}(0, \sigma_e^2)$.

Then $Y \sim \mathbb{N}(\alpha, K\sigma_g^2 + I\sigma_e^2)$, where $K = \frac{XX^T}{N}$

Bivariate Analysis

Trait 1: $$\begin{aligned} Y_1 &= Z\alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{500\,000} X_{500\,000} + e_1 \\ &= Z\alpha_1 + g_1 + e_1 \end{aligned}$$

Trait 2: $$\begin{aligned} Y_2 &= Z\alpha_2 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_{500\,000} X_{500\,000} + e_2 \\ &= Z\alpha_2 + g_2 + e_2 \end{aligned}$$

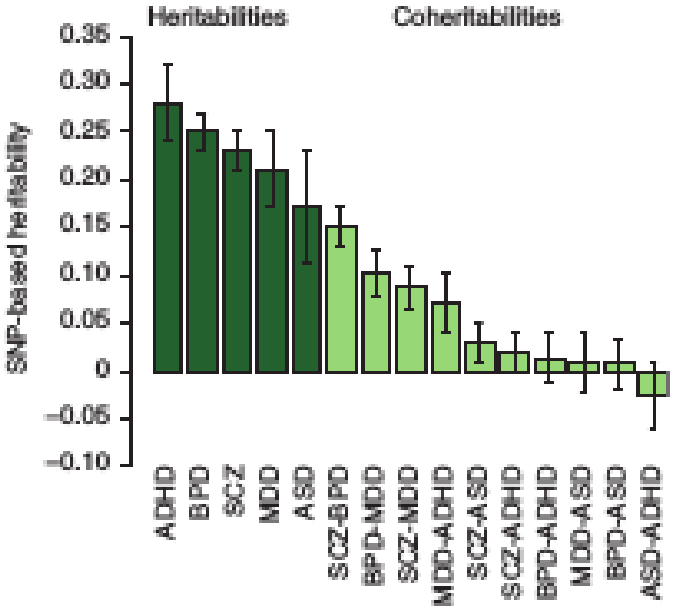
Now interested in the correlation between genetic effects:
$\rho = cor(g_1, g_2)$.

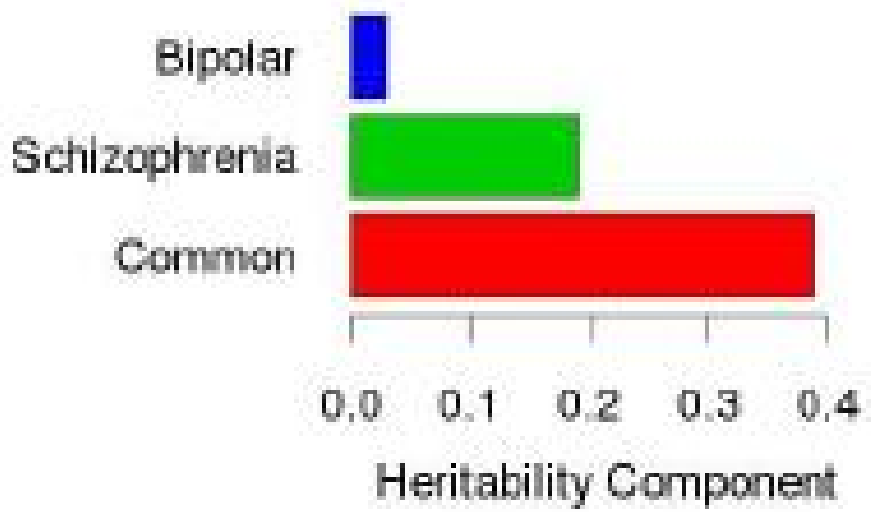
Or equivalently can think of the average correlation between effect sizes:
$\rho = cor(\beta_j, \gamma_j)$

Examining Concordance Between Traits

Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs

Cross-Disorder Group of the Psychiatric Genomics Consortium*





Genome Partitioning

$$
\begin{aligned}
Y = \alpha & \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
& + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
& + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
& + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
& + \ldots + \beta_{500\,000} X_{500\,000} \\
& + e.
\end{aligned}
$$

Assume $\beta_j \sim \mathbb{N}(0, \sigma_g^2/N)$ and $e \sim \mathbb{N}(0, \sigma_e^2)$.

Then $Y \sim \mathbb{N}(\alpha, K\sigma_g^2 + I\sigma_e^2)$, where $K = \frac{XX^T}{N}$

Genome Partitioning

$$
\begin{aligned}
Y = \alpha & \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
& + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
& + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
& + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
& + \ldots + \beta_{500\,000} X_{500\,000} \\
& + e.
\end{aligned}
$$

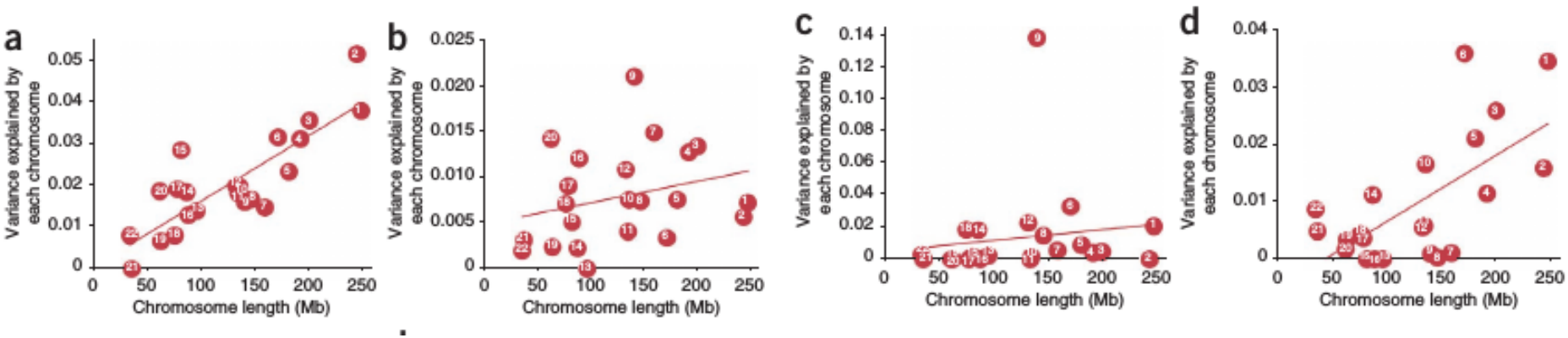
Assume $\beta_j \sim \mathbb{N}(0, \sigma^2_{g1}/N_1)$ and $\beta_k \sim \mathbb{N}(0, \sigma^2_{g2}/N_2)$.

Then $Y \sim \mathbb{N}(\alpha, K_1\sigma^2_{g1} + K_2\sigma^2_{g2} + I\sigma^2_e)$, where $K_1 = \frac{X_1 X_1^T}{N_1}$ and $K_2 = \frac{X_2 X_2^T}{N_2}$

Genome Partitioning

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]



Height BMI vWF QTi

Intensity of Heritability

We define the "intensity of heritability" of a set of SNPs as their heritability divided by their genetic variation.

Can then test for differences in intensity of heritability.

Intensity of Heritability

Are genic SNPs more important than inter-genic SNPs?

		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

Inter-genic defined as all SNPs $>$ 100kbp from a coding region.

Intensity of Heritability

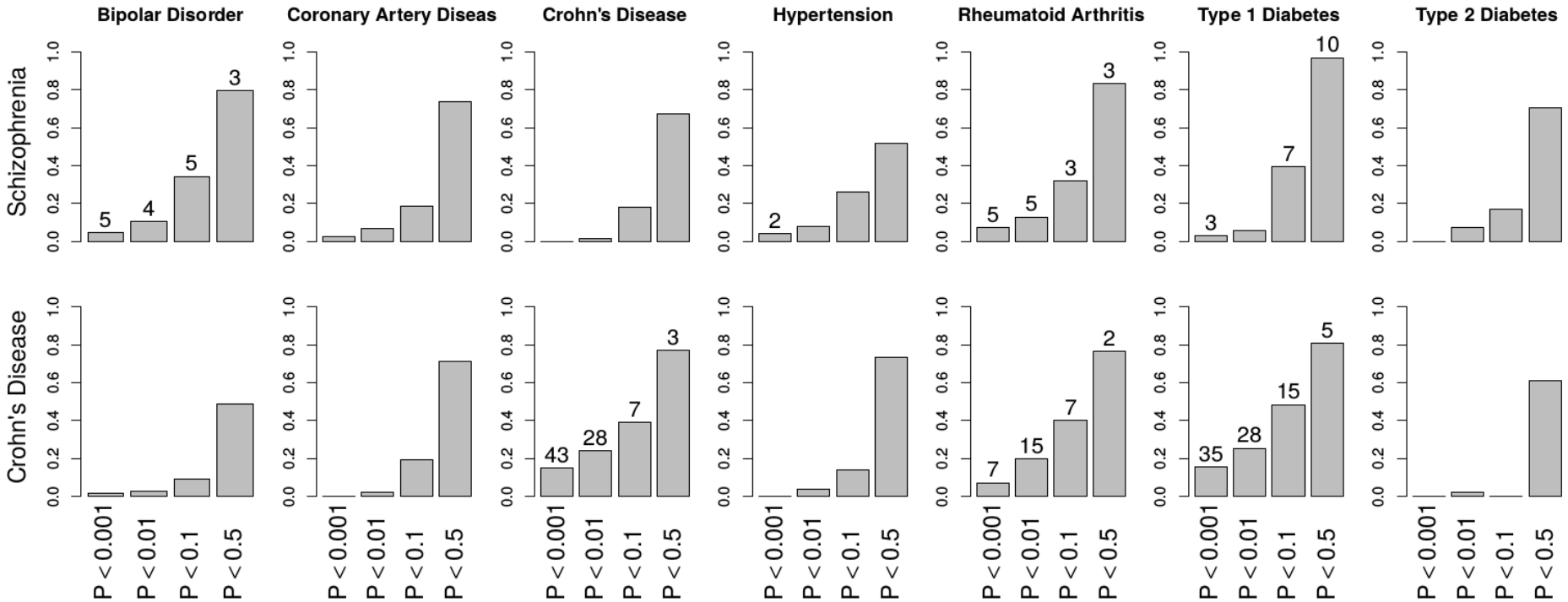
Are genic SNPs more important than inter-genic SNPs?

		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

Can test eQTLs vs non-eQTLs; high-quality SNPs vs low-quality, etc.

Concordance Between Traits

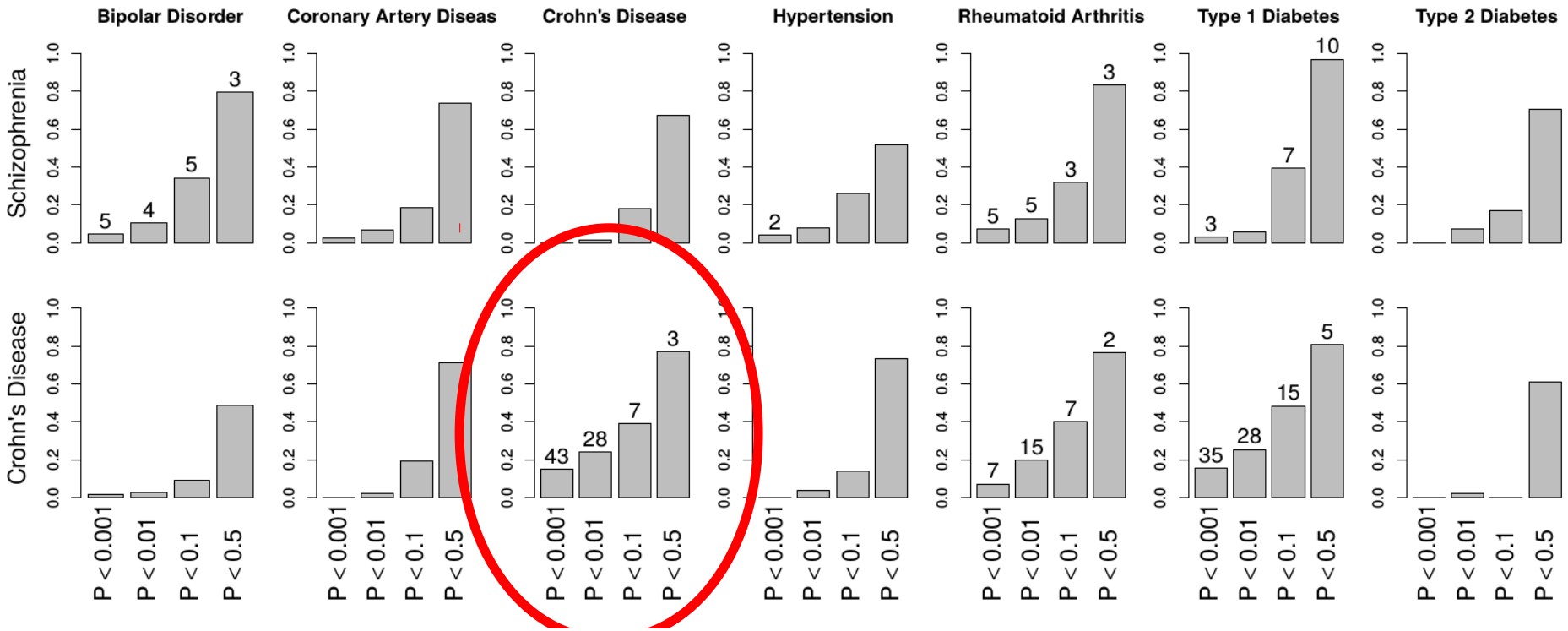
Are SNPs associated with one trait more important for others.



p-values for Schizophrenia and Crohn's obtained from independent studies.

Concordance Between Traits

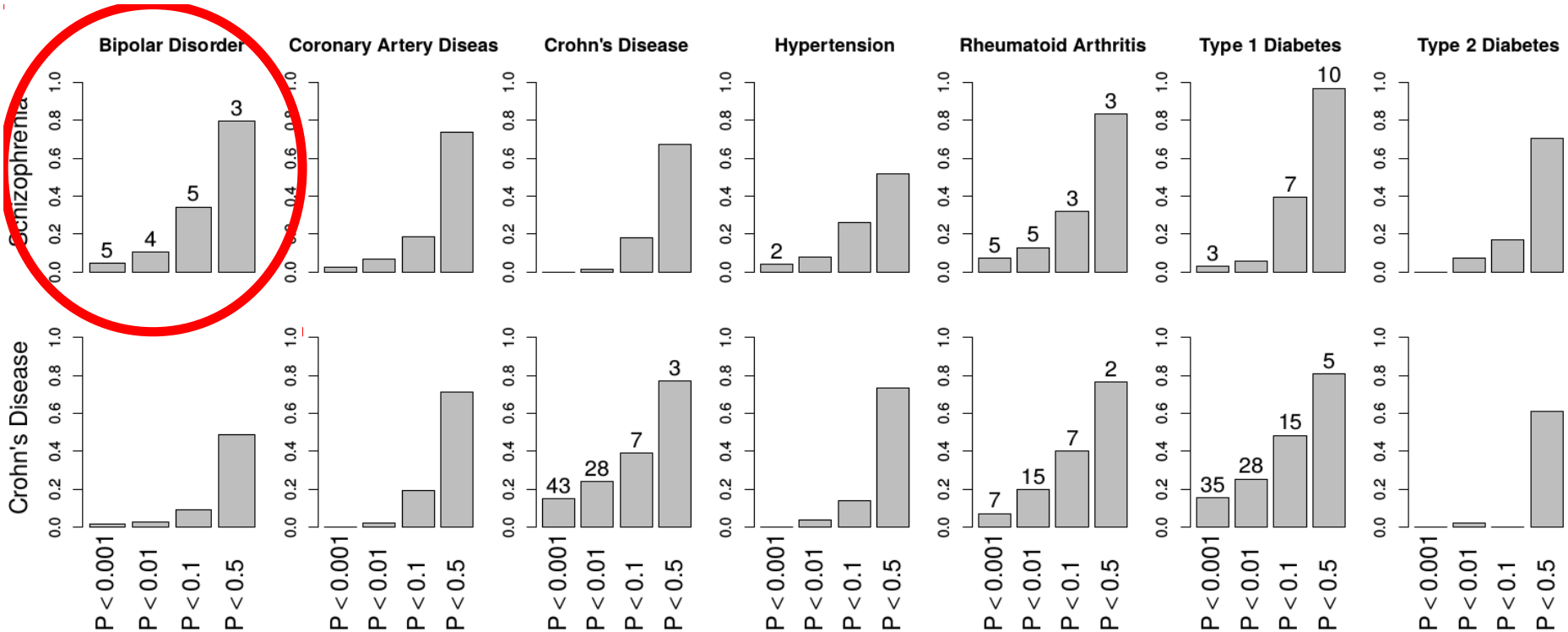
Are SNPs associated with one trait more important for others.



SNPs associated with Crohn's are more important for Crohn's. (Good!)

Concordance Between Traits

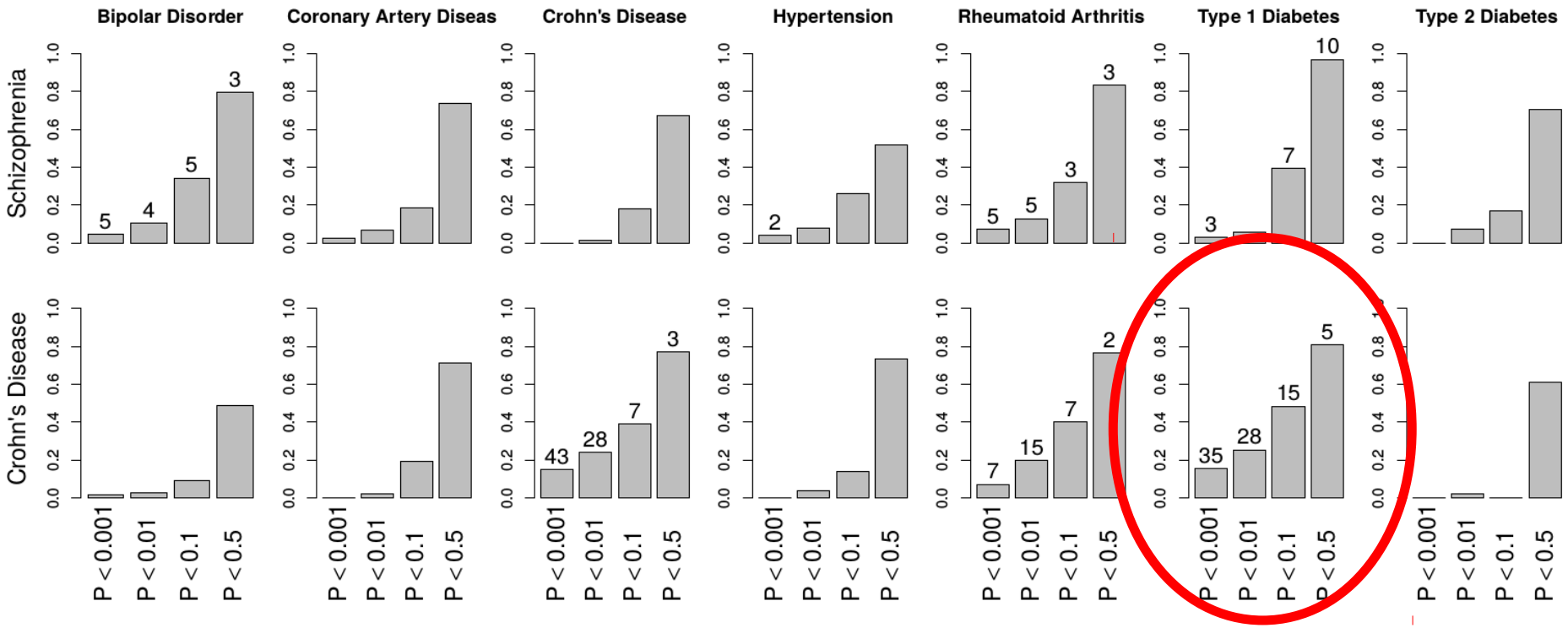
Are SNPs associated with one trait more important for others.



SNPs associated with Schizophrenia are more important for Bipolar.

Concordance Between Traits

Are SNPs associated with one trait more important for others.



Find concordance between Crohn's and Type 1 Diabetes.

Using Intensity of Heritability for Prediction

We are interested in obtaining linear prediction models of the form:

$Y = X_1\beta_1 + X_2\hat{\beta}_2 + \ldots + X_{500\,000}\hat{\beta}_{500\,000}$

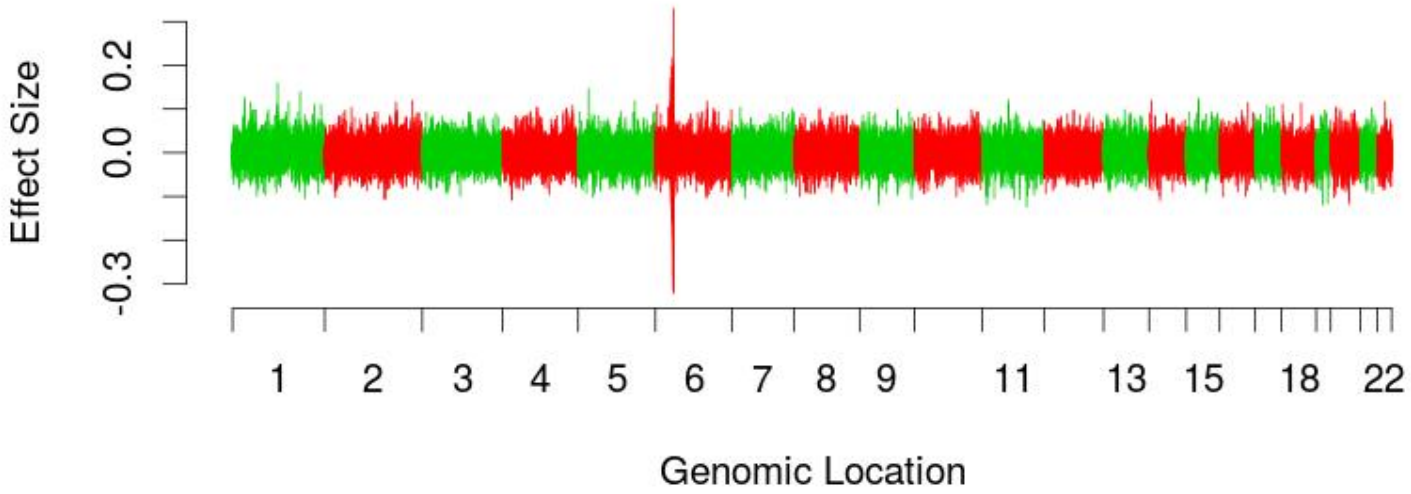
$\hat{\beta}_j$ is the estimated effect size of SNP j (can be zero).

BLUP (Best Linear Unbiased Prediction) assumes the linear mixed model:

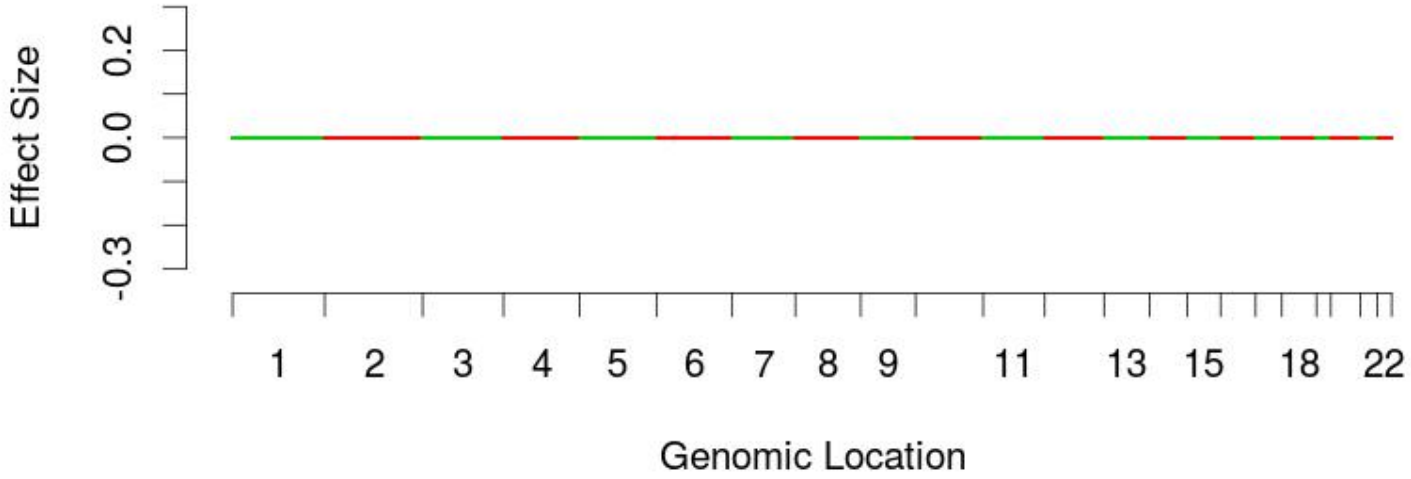
$Y = Z\alpha + g + e \quad \text{with} \quad g \sim \mathbb{N}(0, K\sigma_g^2) \quad \text{and} \quad e \sim \mathbb{N}(0, I\sigma_e^2).$

When $K = \frac{XX^T}{N}$, this is equivalent to assuming $\beta_j \sim \mathbb{N}(0, \sigma_g^2/N)$.

BLUP Excessively Shrinks



Single-SNP Analysis



BLUP

MultiBLUP

MultiBLUP generalizes BLUP:

$$Y = Z\alpha + g_1 + g_2 + \ldots + g_M + e \quad \text{with} \quad g_m \sim \mathbb{N}(0, K_m\sigma^2_{gm}).$$

If $K_1 = \frac{X_1 X_1^T}{N_1}, K_2 = \frac{X_2 X_2^T}{N_2} \ldots, K_M = \frac{X_M X_M^T}{N_M}$

then this is equivalent to a random effects regression model where SNPs in X_1 allowed to have a different effect size prior to those in X_2, etc.

MultiBLUP improves prediction when random effects correspond to distinct intensities of heritability.

MHC / Non-MHC MultiBLUP

Trait	BLUP	MHC/non-MHC MultiBLUP
Bipolar Disorder	0.27	0.27
Coronary Artery Disease	0.13	0.13
Crohn's Disease	0.32	0.29
Hypertension	0.15	0.14
Rheumatoid Arthritis	0.21	**0.35**
Type 1 Diabetes	0.25	**0.56**
Type 2 Diabetes	0.16	0.16
Mean	0.21	0.27

Dividing into MHC and Non-MHC advantageous for RA and TID (but same or worse for others).

Gene-Based Association Testing

To estimate the total contribution of all SNPs, we use the model:

$$
\begin{aligned}
Y = \alpha & \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
& + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
& + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
& + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
& + \ldots + \beta_{500\,000} X_{500\,000} \\
& + e.
\end{aligned}
$$

Gene-Based Association Testing

To test a set of SNPs S, can reduce it to:

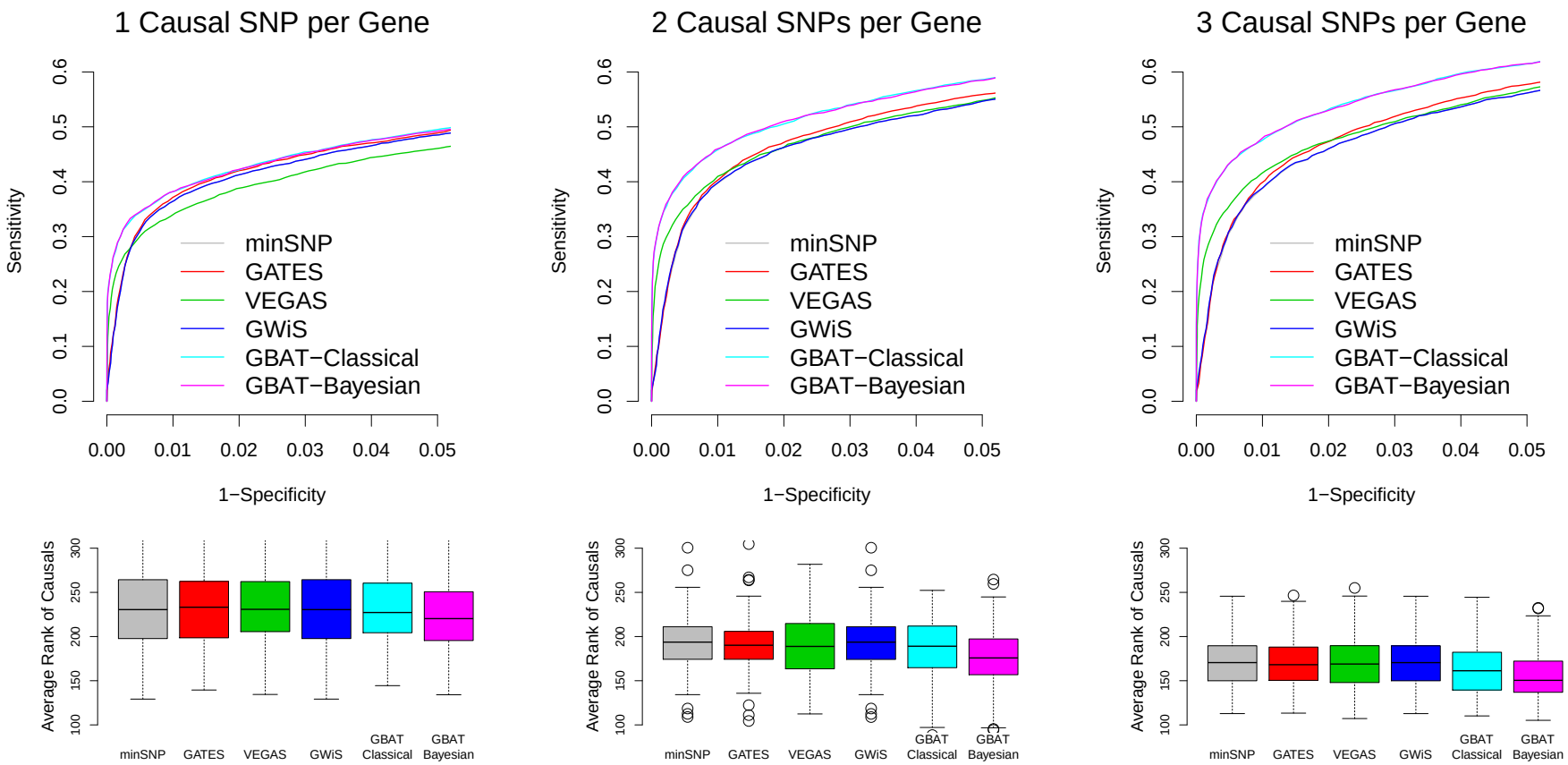
$$
\begin{aligned}
Y = \alpha \\
&+ \cancel{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7} \\
&+ \cancel{\beta_8 X_8 + \beta_9 X_9 +} \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
&+ \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} \cancel{+ \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21}} \\
&+ \cancel{\beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28}} \\
&+ \cancel{\ldots + \beta_{500\,000} X_{500\,000}} \\
&+ e.
\end{aligned}
$$

i.e., $Y = Z\alpha + \sum_{j \in S} X_j \beta_j + e$ with $\beta_j \sim \mathbb{N}(0, \sigma_S^2 / N_S)$.

Perform a likelihood ratio test for $\mathbb{P}(\sigma_S^2 > 0)$.

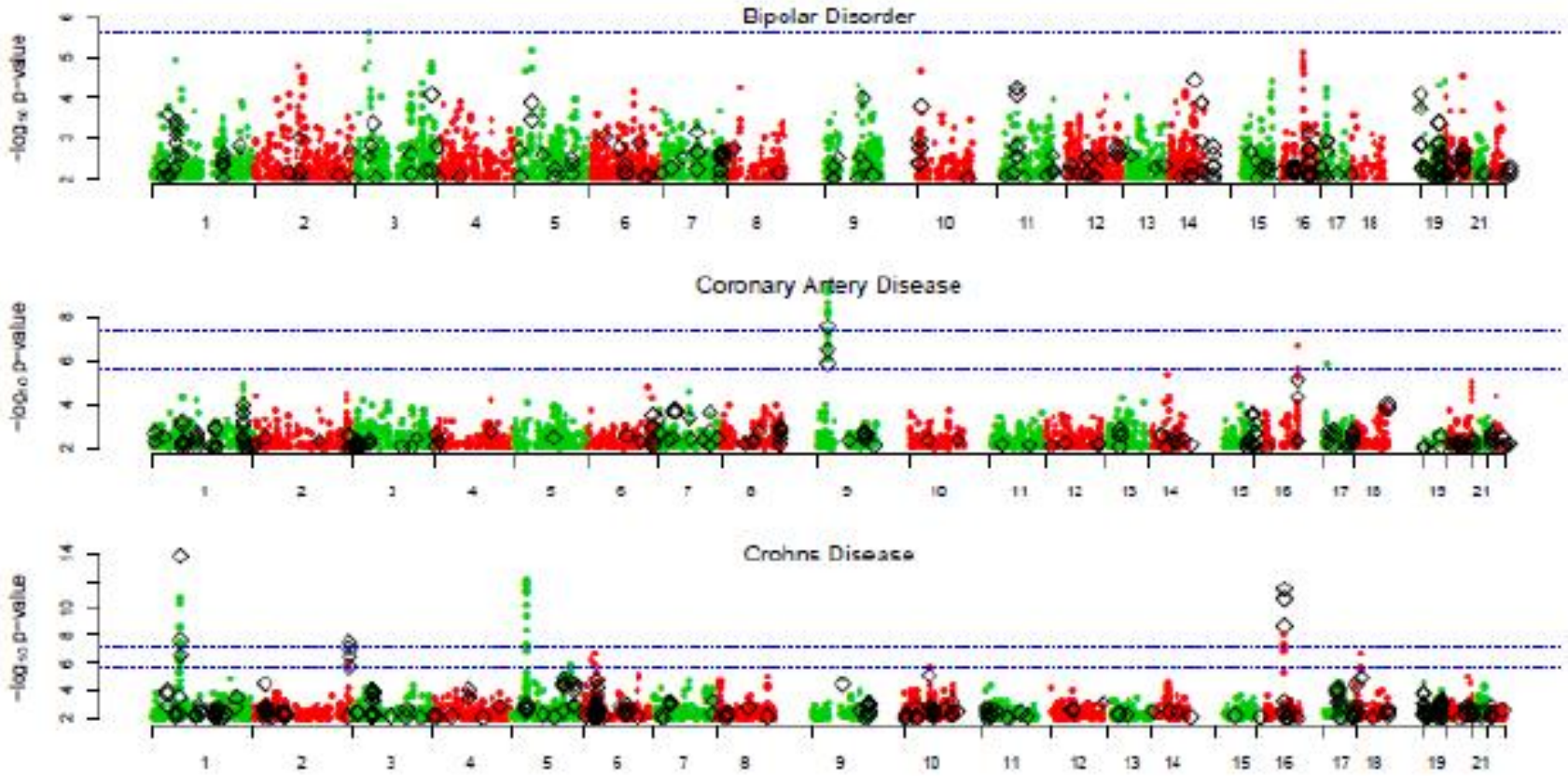
Simulation Study

Generate phenotypes where 50/1000 genes contribute heritability.



GBAT most powerful (and fastest).

GBAT



GBAT provides a complement to single-SNP tests of association.
p-values are well-suited for pathway analysis.

Gene-Based Association Testing

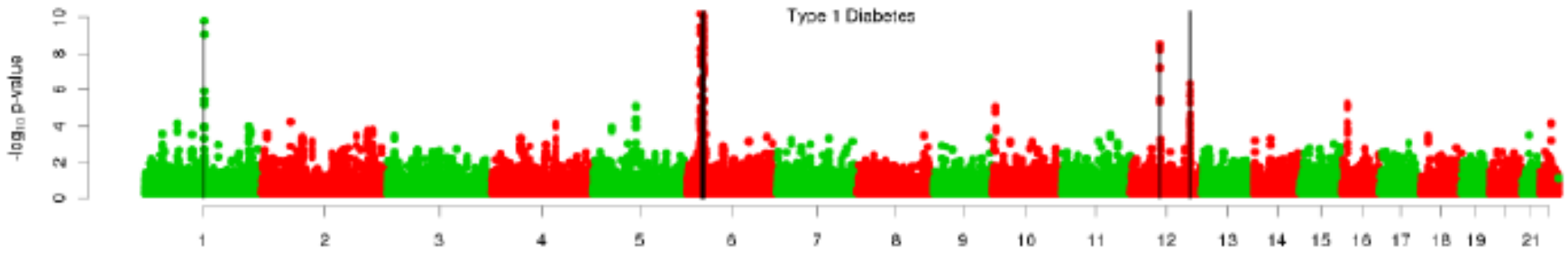
Effectively, you are testing for increased intensity of heritability on a very local level.

Therefore, can incorporate this gene-based (or regional) testing in MultiBLUP.

Adaptive MultiBLUP

Step 1: Divide genome into (say) 75kbp overlapping chunks.

Step 2: Test each chunk for association (using GBAT).

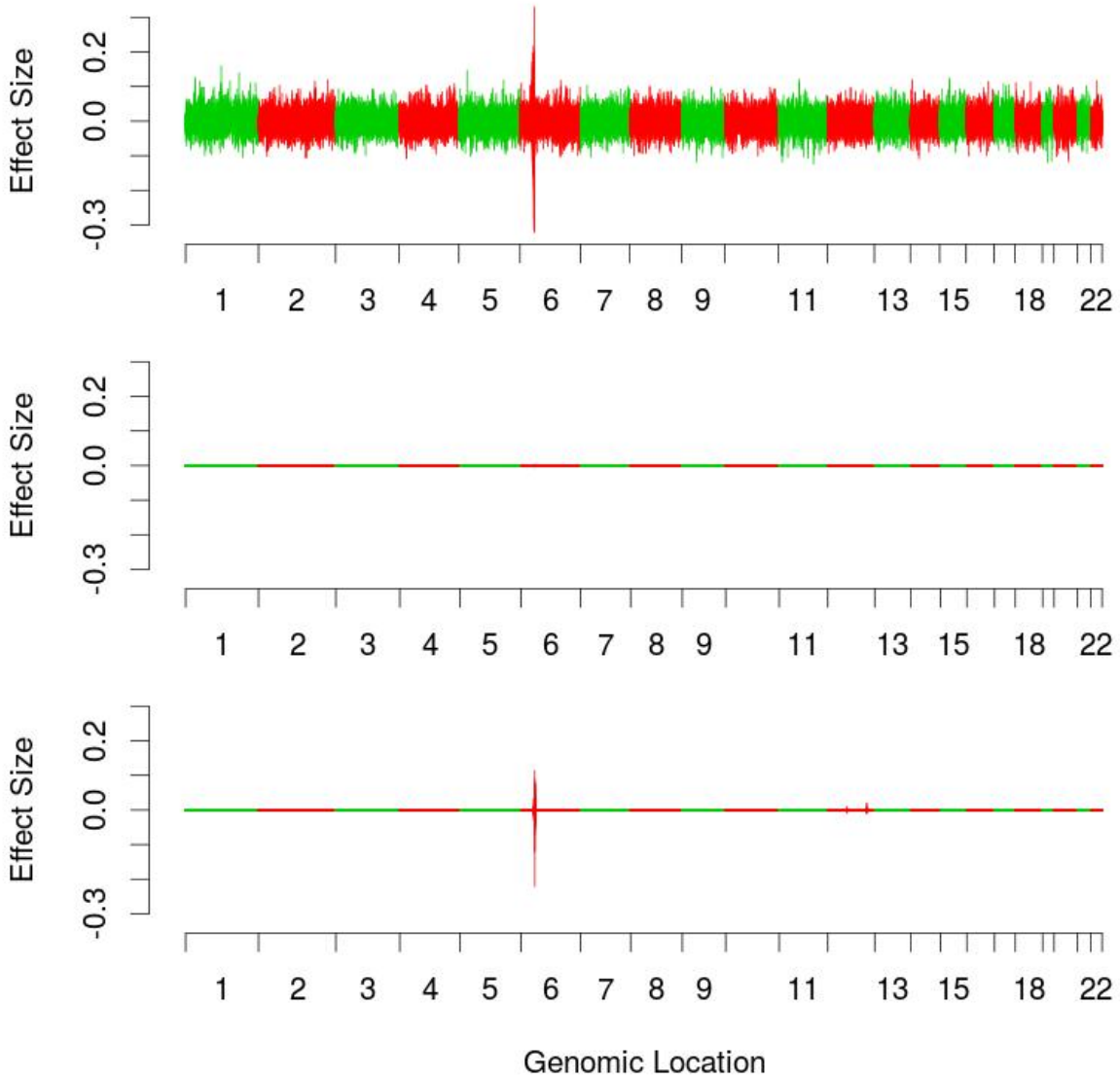


Step 3: Identify all significant chunks (say $P < 10^{-5}$).
(Merge these chunks with neighbouring chunks with $P < 0.01$.)

E.g., for Type 1 Diabetes, obtain 4 local regions.

Step 4: Run MultiBLUP with five random effects.

Adaptive MultiBLUP Offers Flexible Shrinkage



Genetic Profile Risk Scores

BLUP

Adaptive MultiBLUP

Performance of Adaptive MultiBLUP

	Current methods				MultiBLUP	
Trait*	BLUP	Risk Score	Stepwise Regression	BSLMM	MHC/non-MHC	Adaptive
BD	**0.27**	0.25 (1)	0.02	0.27	0.27	0.27
CAD	0.13	0.12 (1)	0.08	0.15	0.13	**0.16**
CD	0.32	0.28 (1)	0.18	0.34	0.29	**0.36**
HT	0.15	0.14 (1)	0.00	0.14	0.14	**0.17**
RA	0.21	0.28 (3)	0.32	0.33	0.35	**0.37**
T1D	0.25	0.34 (5)	0.54	0.57	0.56	**0.59**
T2D	0.16	0.14 (1)	0.10	0.17	0.16	**0.18**
Mean	0.21	0.22	0.18	0.28	0.27	**0.30**

Other methods offer performance similar to Adaptive MultiBLUP: BSLMM (Bayesian Sparse Linear Mixed Models), BayesR, SparSNP.

But I believe Adaptive MultiBLUP most computationally efficient: Can be applied to upwards of 50 000 samples.

Different Kinship Matrices

The "Yang" kinship standardizes each SNP: $K_{ik} = \sum_j^N \frac{(S_{ij} - \bar{S}_j)((S_{kj} - \bar{S}_j)}{Var(S_j)}$

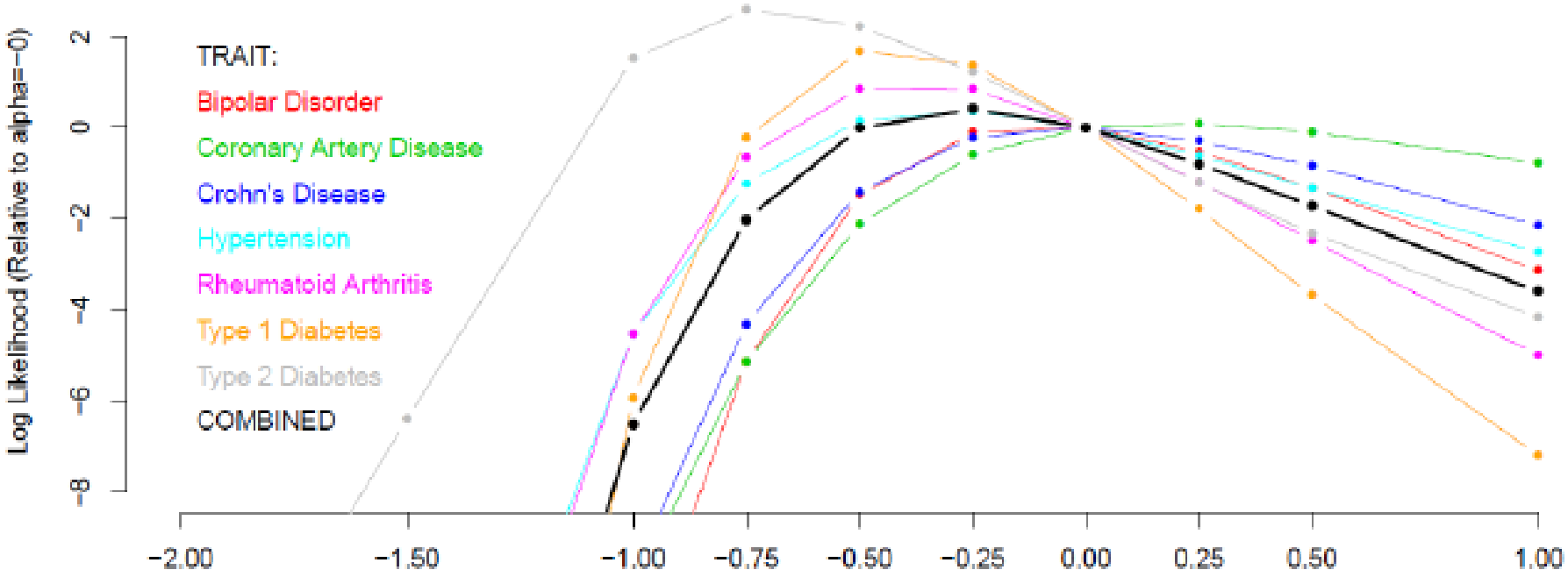
This corresponds to assumption each SNP contributes equal heritability.

Animal/plant geneticists generally use: $K_{ik} = \frac{\sum_j^N (S_{ij} - \bar{S}_j)((S_{kj} - \bar{S}_j)}{Constant}$

This corresponds to assumption each SNP has equal effect magnitude.

Different Kinship Matrices

Can interrogate genetic architecture by trying alternative scalings.



Rare Variants Bigger Effects $\longleftrightarrow$ Common Variants Bigger Effects

Evidence that rarer variants tend to have (slightly) larger effects.

Null Distribution for Likelihood Ratio Test

Likelihood ratio test statistic: $T = 2 \times (l_1 - l_0)$.

l_1 & l_0 are maximum log likelihoods under null & alternative models.

Standard MLE theory assumes $T \sim \chi^2(1)$.

Null Distribution for Likelihood Ratio Test

Likelihood ratio test statistic: $T = 2 \times (l_1 - l_0)$.

l_1 & l_0 are maximum log likelihoods under null & alternative models.

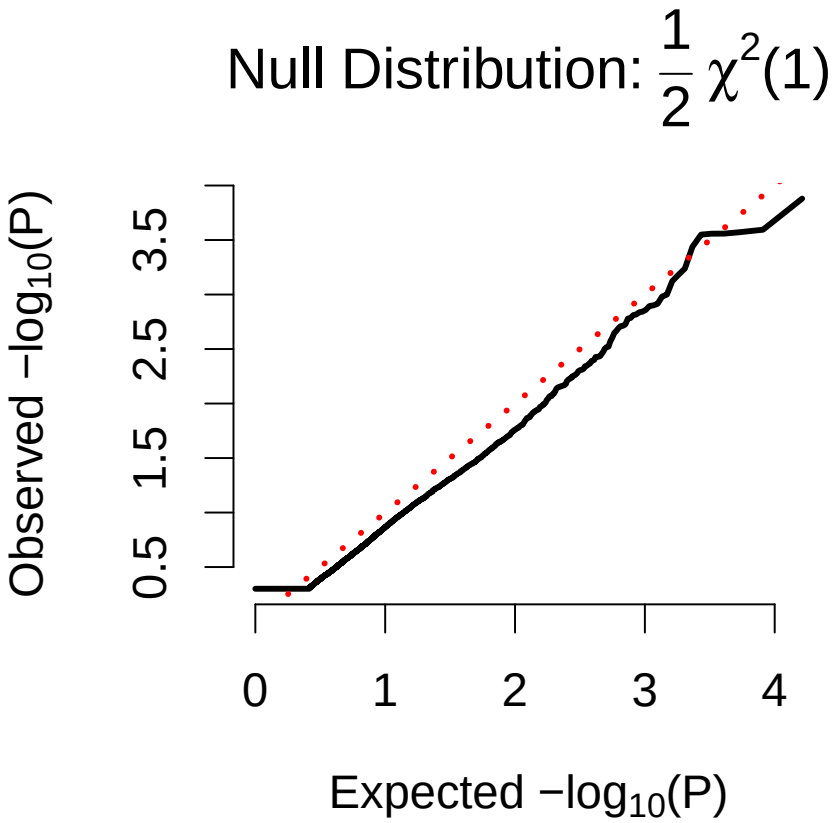
Standard MLE theory assumes $T \sim \chi^2(1)$.

However, because testing $\sigma_g^2 > 0$ (one sided / on boundary?)

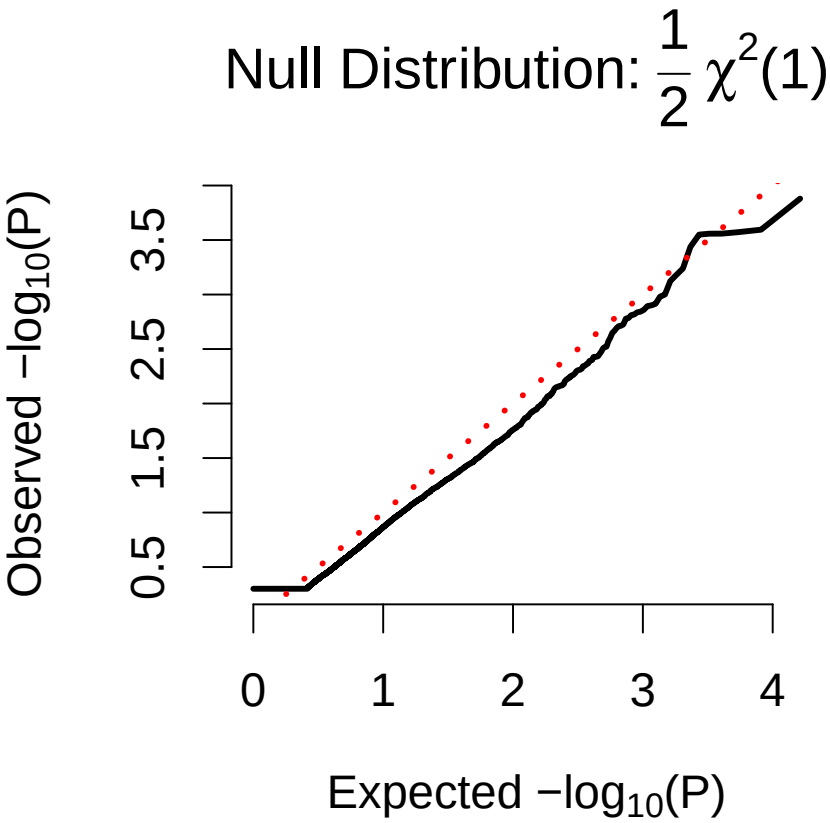
MLE theory (?) assumes $T \sim \frac{1}{2}\delta_{\{0\}} + \frac{1}{2}\chi^2(1)$.

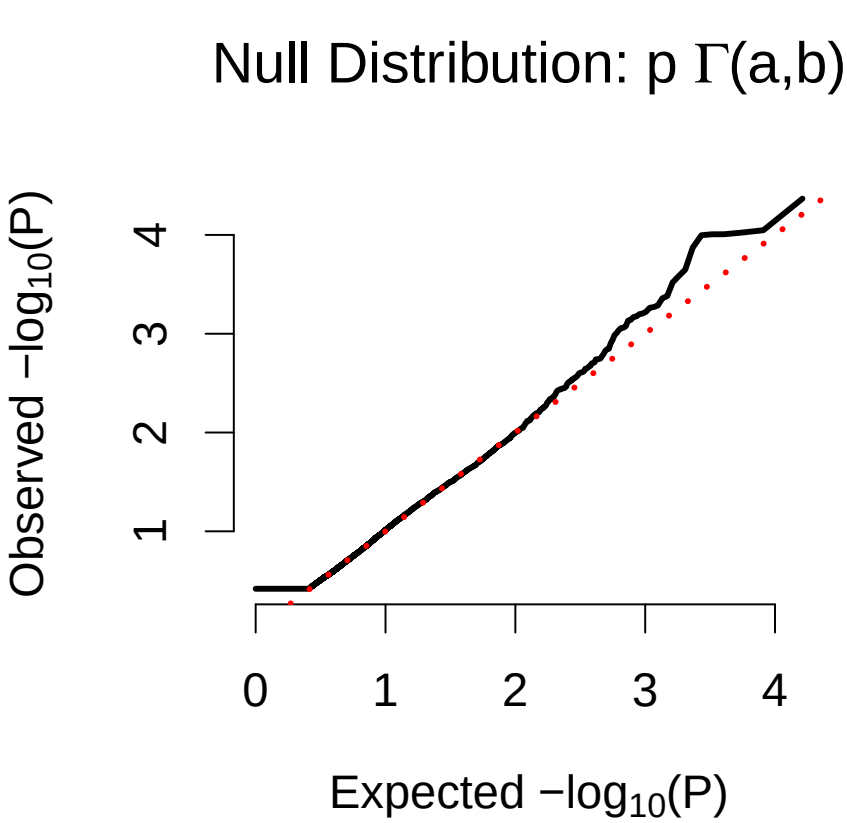
This leads to conservative p-values (because a random effect?).

Null Distribution for Likelihood Ratio Test



Null Distribution for Likelihood Ratio Test





Estimate p, a, and b from permutation - important for gene-based tests.

Bayesian REML

The basic model:

$Y = Z\alpha + g + e \quad$ with $\quad g \sim \mathbb{N}(0, K\sigma_g^2) \quad$ and $\quad e \sim \mathbb{N}(0, I\sigma_e^2)$

In LDAK, the user can specify a beta prior for $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$.

For the general model uses a Dirichlet prior for the M heritabilities.

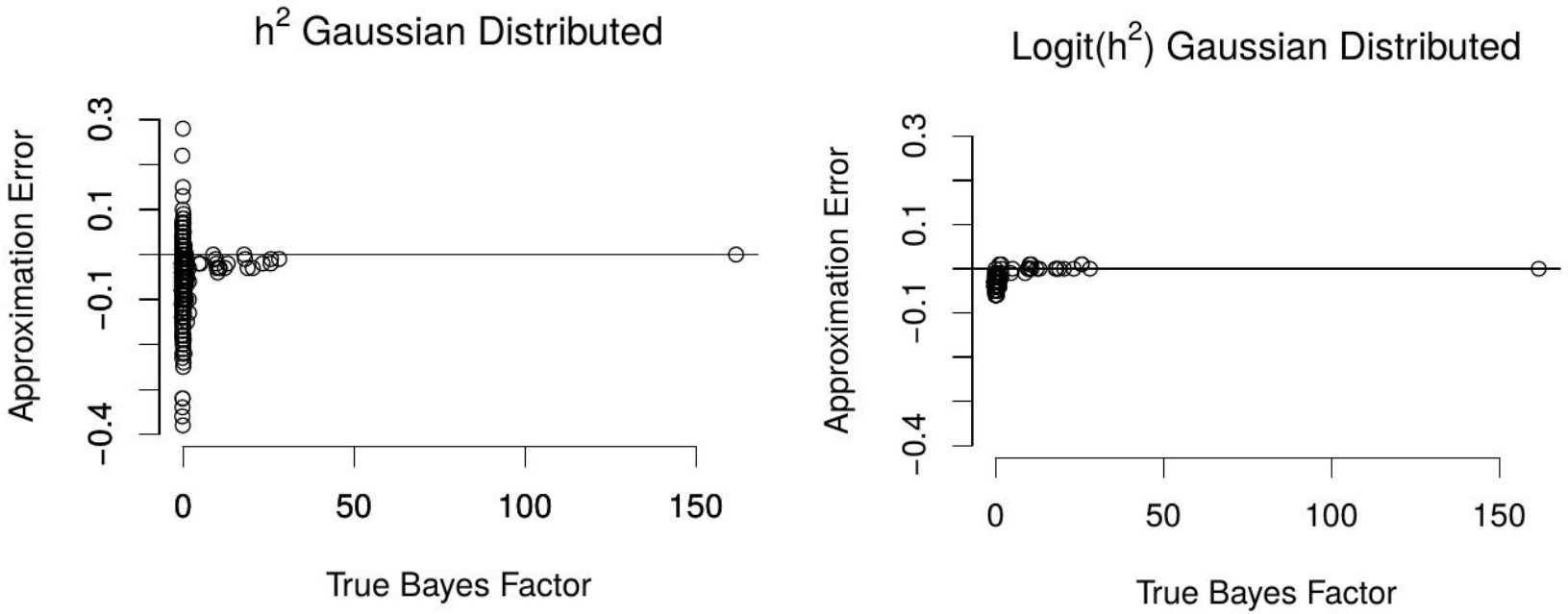
LDAK estimates the posterior distribution of h^2 and reports Bayes Factors.

Particularly useful for gene-based tests.

Improved Estimation of h^2

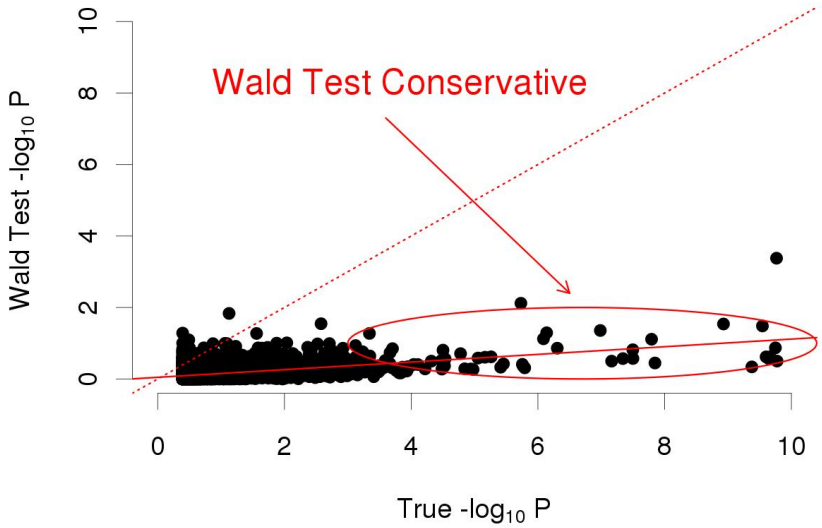
Standard REML estimates assumes a normal distribution for h^2. Figure too large - but basically shows breakdown for small h^2. LDAK assumes a normal distribution for $logit(h^2) = h^2/(1 - h^2)$.

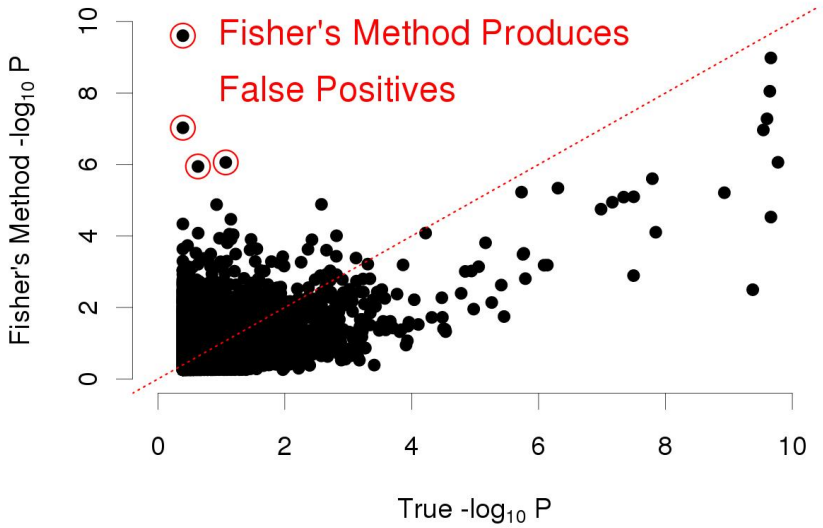
Leads to More Precise Inferences



Can compute Bayes Factors more accurately.

Heritability Meta Analysis

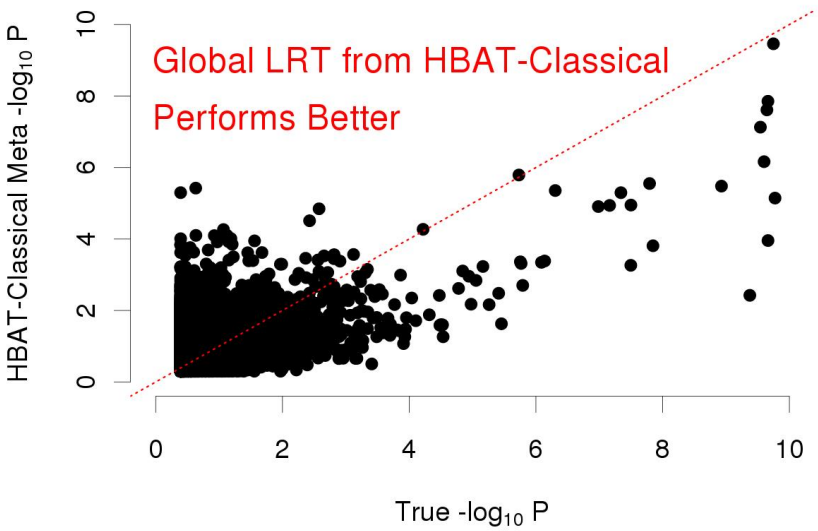


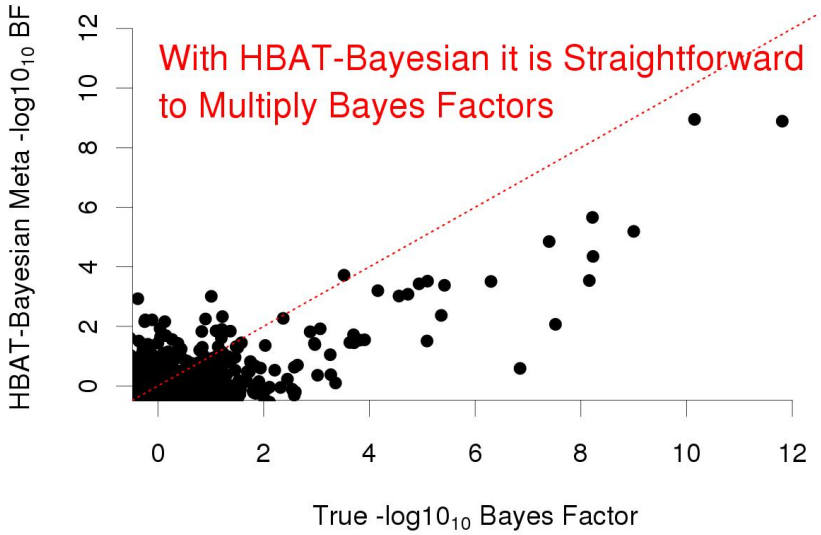


Previous approaches obtain a global estimate of $\mathbb{P}(h_S^2)$ by inverse-variance weighting of individual cohort estimates of h_S^2.

Then either use a Wald Test, or combine individual cohort p-values using Fisher's Method.

Heritability Meta Analysis





By contrast, valid to use inverse-variance weighting on logit scale.
Can derive an approximate global likelihood ratio test statistic for the fixed effect model.

But we prefer combining Bayes Factors ($\sim$ random effect model).

Using Eigen-Decompositions

Estimate variance components using Average Information REML.
(Similar to Newton-Raphson, but approximates the second derivative.)

For basic model, each iteration requires inversion of $V = K\sigma_g^2 + I\sigma_e^2$.

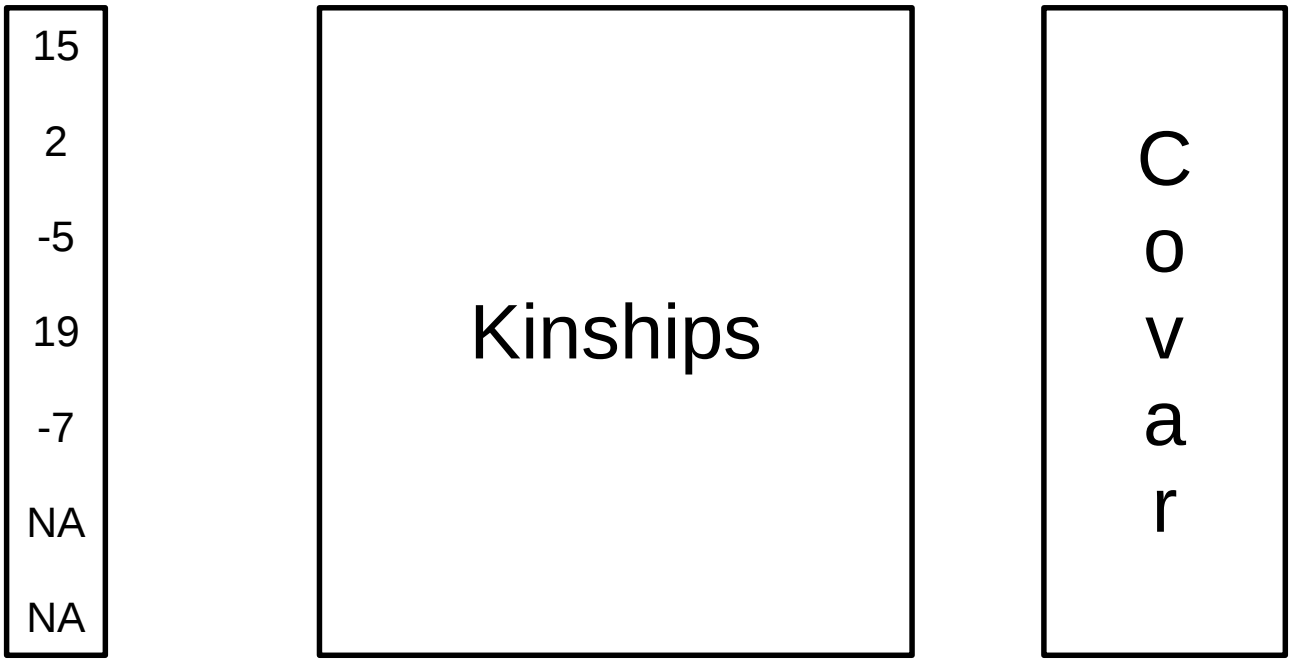
This becomes trivial given the eigen-decomposition $K = UEU^T$:
$V = U(E\sigma_g^2 + I\sigma_e^2)U^T$ and $V^{-1} = U(E\sigma_g^2 + I\sigma_e^2)^{-1}U^T$

Alternatively, can use Woodbury Matrix Identity:
$V^{-1} = (\frac{XX^T}{N}\sigma_g^2 + I\sigma_e^2)^{-1} = \frac{1}{\sigma_e^2}I - \frac{X}{\sigma_e^2}(\frac{X^TX}{\sigma_e^2} + \frac{N}{\sigma_g^2})^{-1}\frac{X^T}{\sigma_e^2}$.

When K low-rank ($N < n$) this leads to huge speed ups.

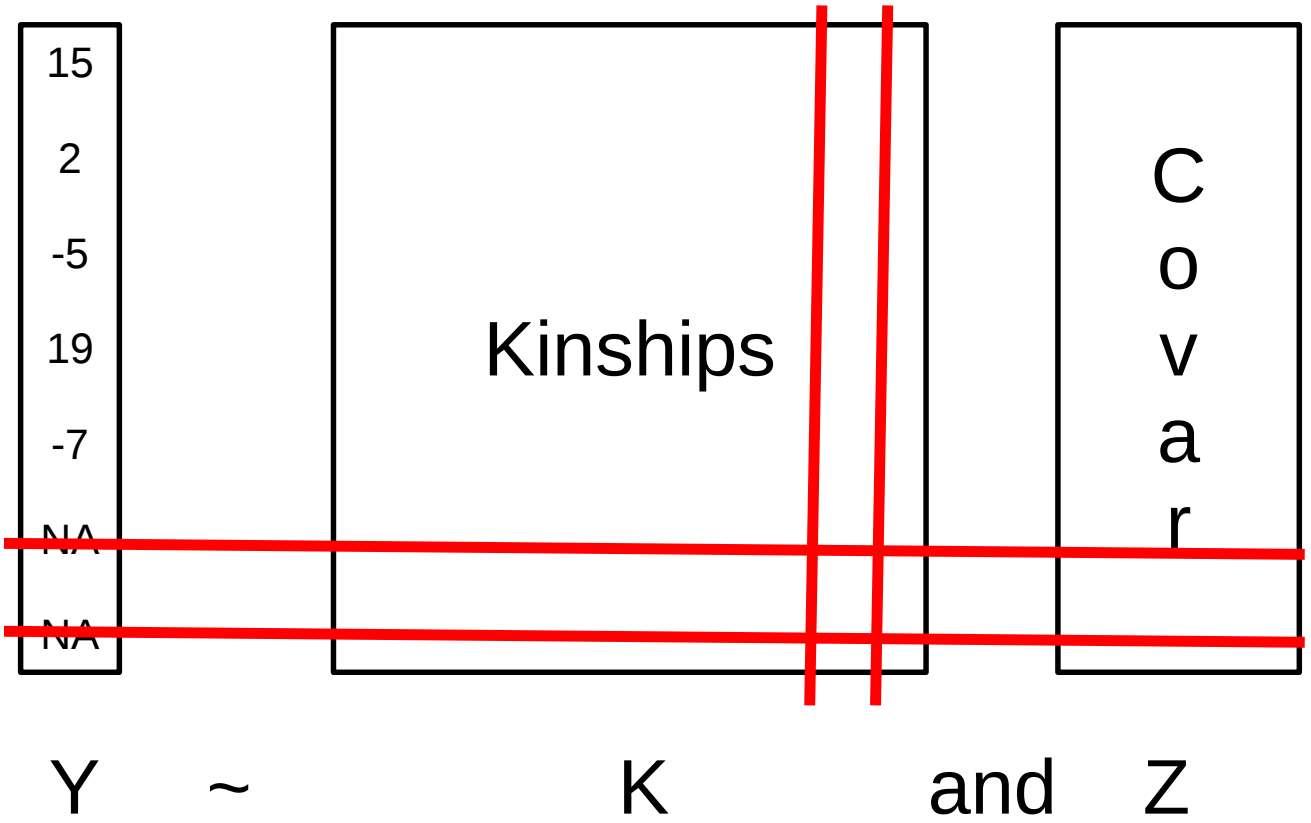
Missing phenotypic values

Suppose wish to estimate h^2 but fraction $1 - p$ phenotypes missing:



Missing phenotypic values

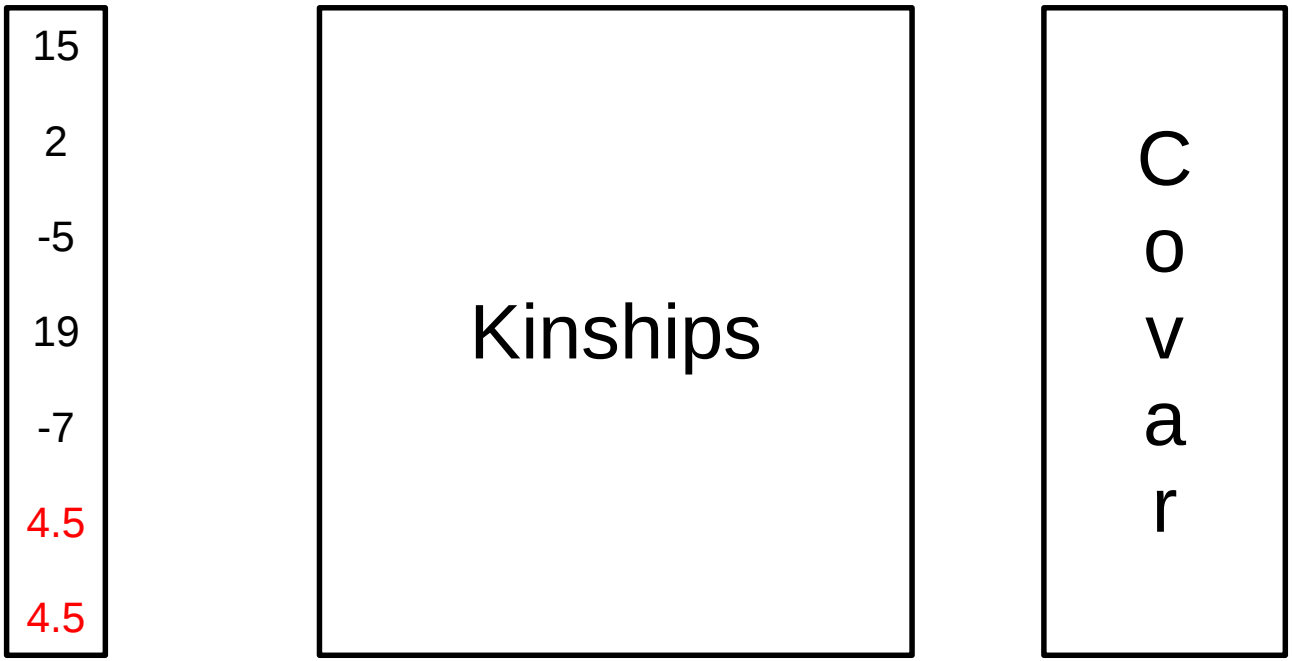
Suppose wish to estimate h^2 but fraction $1 - p$ phenotypes missing:



Would typically just exclude affected individuals.

Missing phenotypic values

Alternative: replace missing with the mean and analyse all:



Resulting estimate will have expected value ph^2.

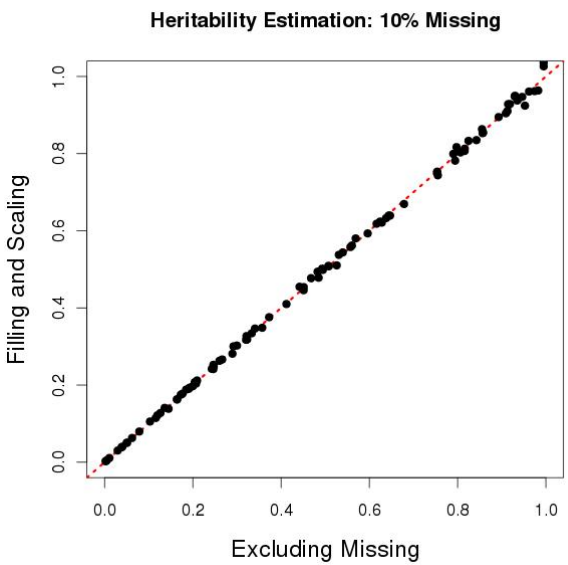
The Dentist Method

Therefore, we propose you first "fill" missing values, then "scale" the resulting estimate of h^2 by $\frac{1}{p}$.

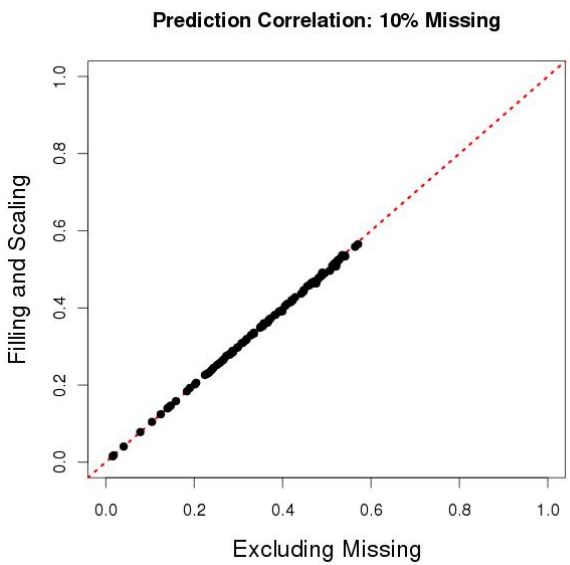
Can use the same trick for BLUP or mixed model association analysis.

The Dentist Method

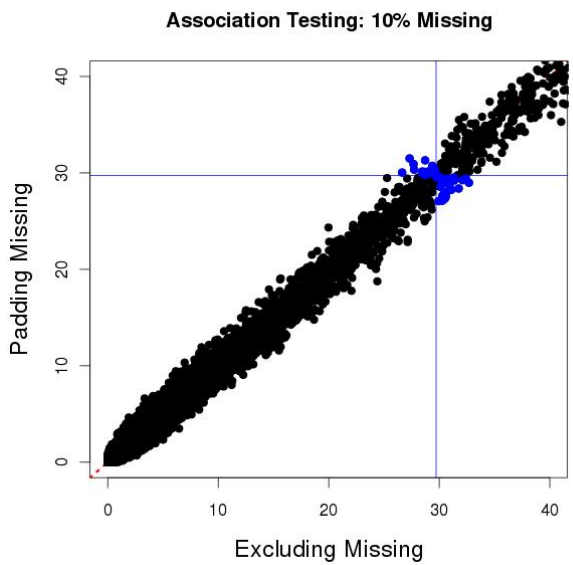
Simulated data with 10% of phenotypic values missing.



Heritability Analysis



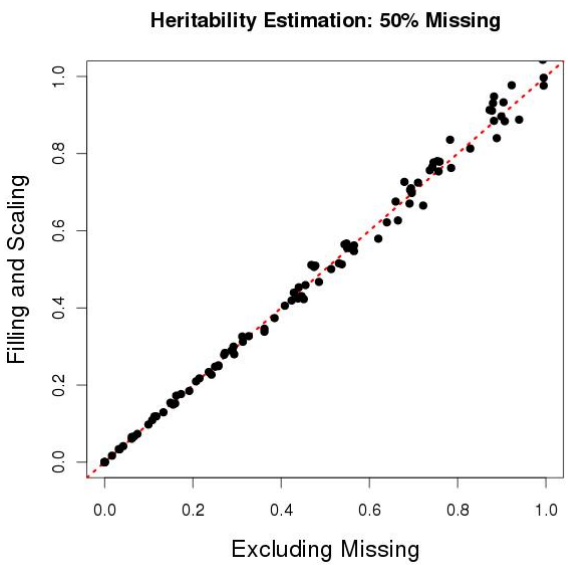
BLUP Performance



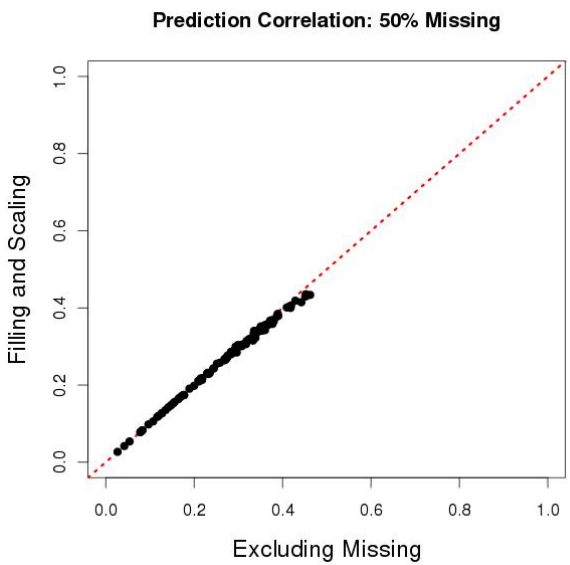
MM Association Testing

The Dentist Method

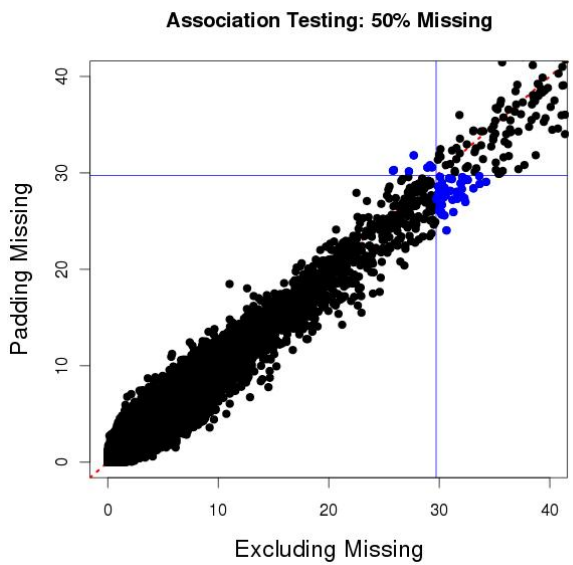
Simulated data with 50% of phenotypic values missing.



Heritability Analysis



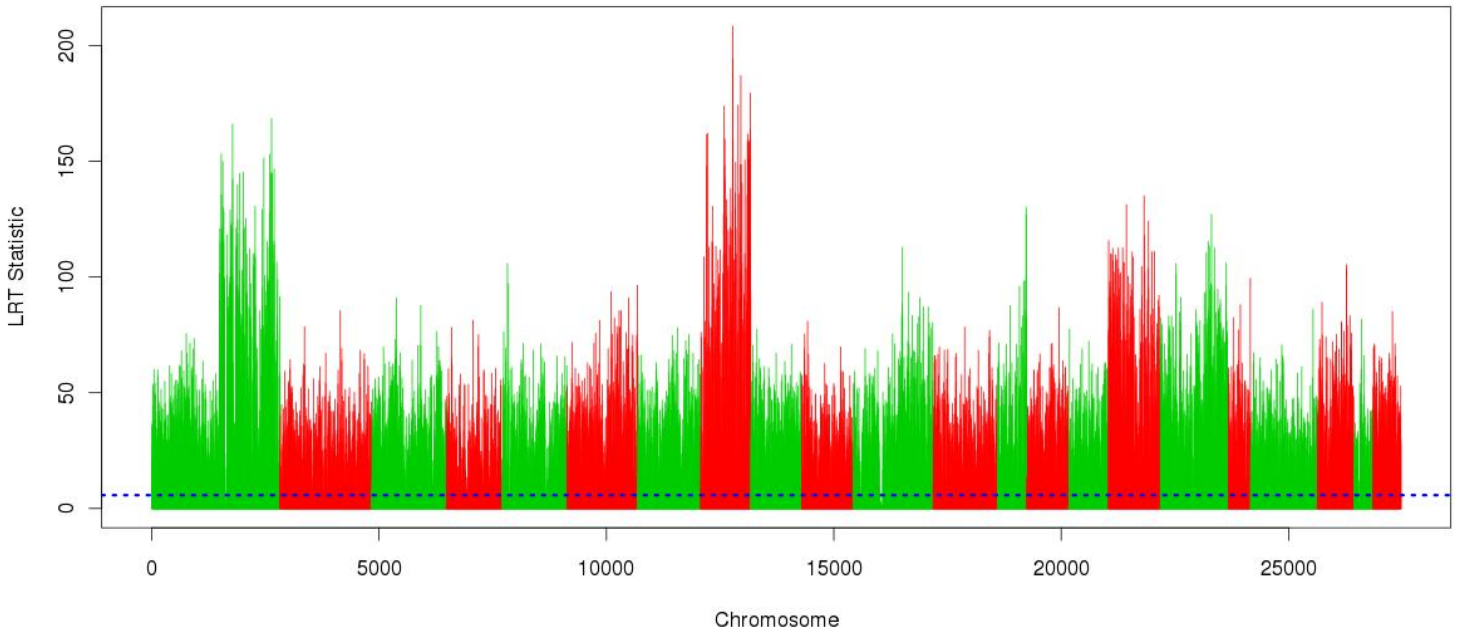
BLUP Performance



MM Association Testing

Why Use the Dentist Method?

When analysing high-dimensional data, filling in missing phenotypic values allows us to use the same decomposition for each phenotype.



Appllied to breast cancer data; $n = 1674$, 27k gene expressions.
Less than 1% missing values, but these affect $\sim 80\%$ of phenotypes.
Filling reduces analysis from weeks to hours.

Outstanding Issues

Have efficient algorithm for inverting V when:

$V = K\sigma_g^2 + I\sigma_e^2$ for K full-rank.

$V = K\sigma_g^2 + I\sigma_e^2$ for K low-rank.

$V = K_1\sigma_{g1}^2 + K_2\sigma_{g2}^2 + I\sigma_e^2$ for K_1 full-rank, K_2 low-rank.

How about for $V = K_1\sigma_{g1}^2 + K_2\sigma_{g2}^2 + I\sigma_e^2$ when K_1 and K_2 full-rank?

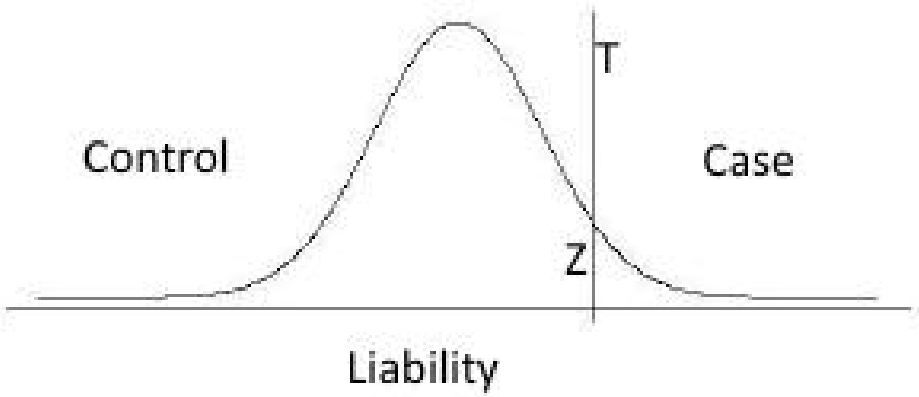
Some applications approximate K_2 with a low-rank matrix.
OK if σ_{g2}^2 a nuisance parameter.

REML for Case/Control Data

Liability model for binary traits:

We only observe whether L above or below a threshold T

T determined by disease prevalence K



Desired Model:

$L = Z\alpha + g + e$ with $g \sim \mathbb{N}(0, K\sigma_g^2)$ and $e \sim \mathbb{N}(0, I\sigma_e^2)$

where $Y_i = 1$ iff $Z_i > T$.

Common practice is to analyse as if a continuous phenotype, then transform to liability scale.

Probit REML

If we knew each individual's liability values, estimating σ_g^2 and σ_e^2 would be straightforward (and trivial given eigen-decomposition of K).

Therefore, envisage MCMC scheme where at each step we first sample L_i, then maximise (or sample from) σ_g^2 and σ_e^2.

Summary

- Estimating SNP heritability
- Intensity of Heritability
- Gene-based Association Testing
- Adaptive MultiBLUP
- Bayesian REML and Heritability Meta-Analysis
- Efficient REML with Multiple Random Effects???
- Probit REML???

Acknowledgments

Work with David Balding.

Money from Medical Research Council.

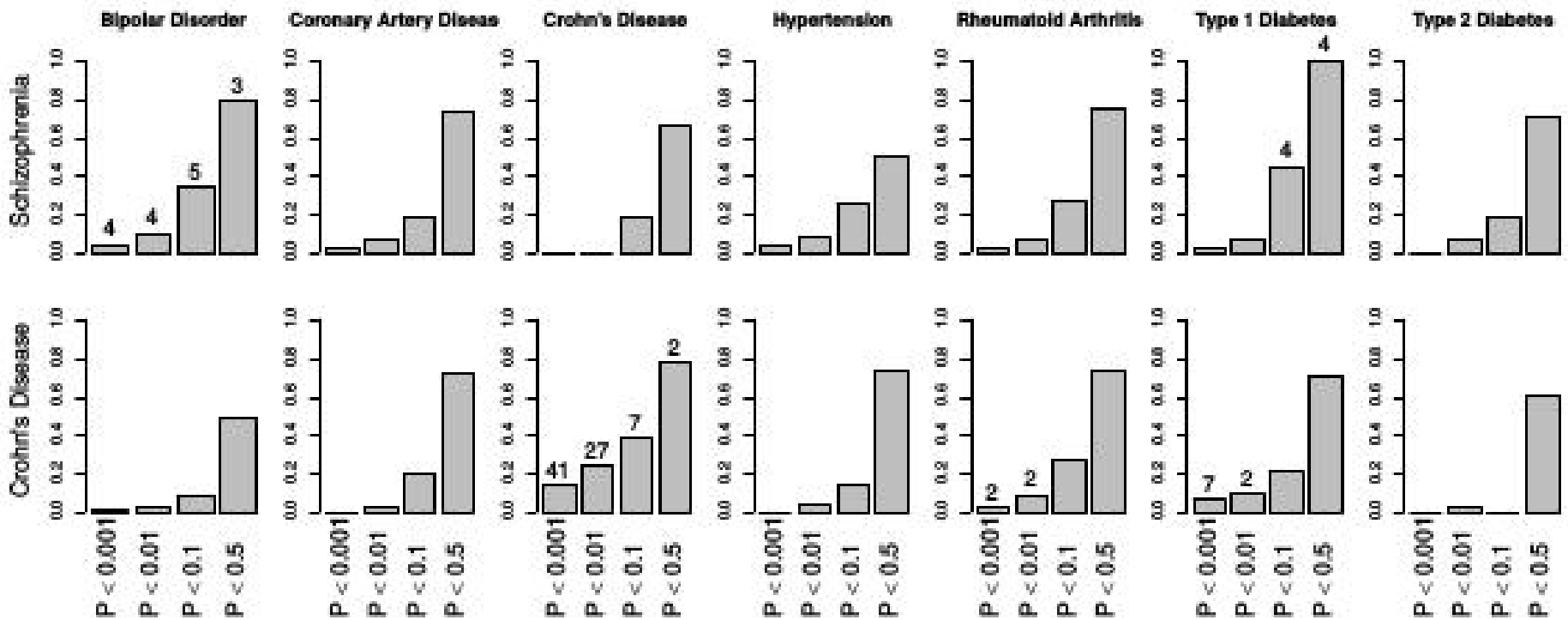
LDAK available at www.ldak.org.

I'm here til end of May

Find me in Old Geology South

Concordance Between Traits

Are SNPs associated with one trait more important for others.



Repeated with MHC region excluded.