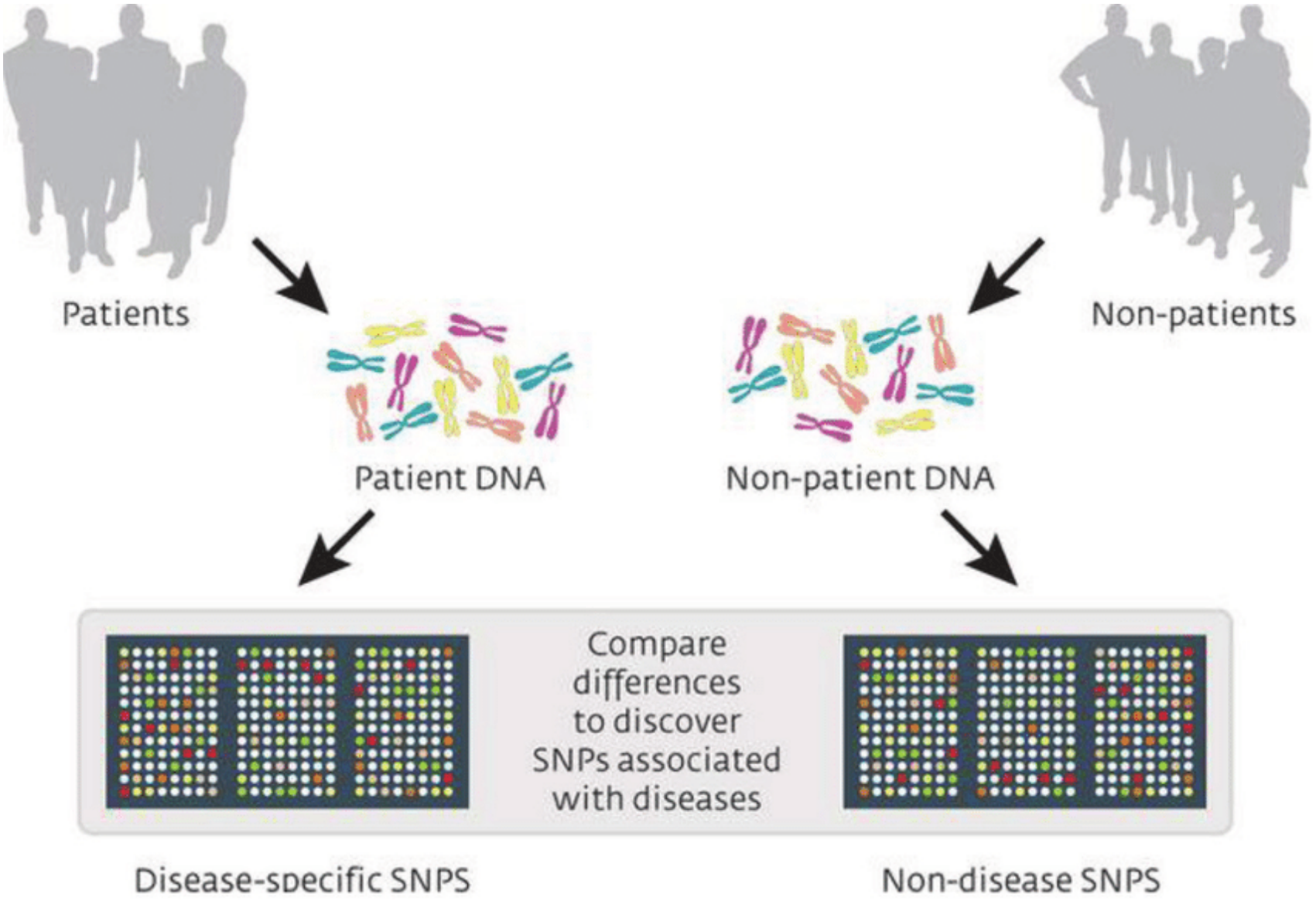
Doug Speed
Quantitative Genetics and Genomics
Aarhus University

Methods for Analysing GWAS Data including prediction of complex traits

Slides at www.ldak.org/lgn

Genome-wide Association Studies



Genome-wide Association Studies

Phenotype	SNPs										
1	2	0	1	1	2	1	0	0	0	...	1
1	1	1	0	1	2	0	0	2	2	...	1
1	1	1	0	1	2	1	0	2	2	...	1
1	0	1	1	1	2	0	0	0	0	...	0
0	2	0	0	2	0	1	0	1	2	...	1
0	2	2	1	2	2	0	0	0	0	...	0
0	0	0	0	1	0	0	0	1	2	...	0
0	1	2	1	1	0	1	0	0	2	...	1
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	...	$X_{500,000}$

There are also covariates, but for simplicity, I ignore them

GWAS Methods

GWAS methods have three main aims:

1 - Detect causal variants

E.g. find risk loci, conditional analysis, fine mapping

2 - Predict traits

E.g. personalized medicine and genomic selection

3 - Understand genetic architecture

E.g. SNP heritability, number of causal loci, heritabilty enrichments

Overview of Talk

Most GWAS methods make assumptions regarding genetic architecture

These include specifying something I call the heritability model

Most GWAS methods use a poor heritability model

By changing the heritability model, we can dramatically improve the performance of most GWAS methods

In particular, I will demonstrate this for prediction methods

Overview of Talk

Most GWAS methods make assumptions regarding genetic architecture

These include specifying something I call the heritability model

Most GWAS methods use a poor heritability model

By changing the heritability model, we can dramatically improve the performance of most GWAS methods

In particular, I will demonstrate this for prediction methods

Heritability Models

Suppose we have GWAS data for n individuals and m SNPs

- Vector Y contains the phenotypes
- Vector X_j contains the genotypes for SNP j
- Y and X_j are standardized (mean zero, variance one)

Assume the linear model $Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

- β_j is the effect size for SNP j
- The heritability contributed by SNP j is $h_j^2 = \beta_j^2$

The heritability model describes how $\mathbb{E}[h_j^2]$ varies across the genome

Heritability Models

Suppose we have GWAS data for n individuals and m SNPs

- Vector Y contains the phenotypes
- Vector X_j contains the genotypes for SNP j
- Y and X_j are standardized (mean zero, variance one)

Assume the linear model $Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

- β_j is the effect size for SNP j
- The heritability contributed by SNP j is $h_j^2 = \beta_j^2$

The heritability model describes how $\mathbb{E}[h_j^2]$ varies across the genome

Heritability Models

Most analyses in human statistical genetics assume $\mathbb{E}[h_j^2]$ is constant

All SNPs are expected to contribute equally

I call this the GCTA Model

GCTA is a method for estimating SNP heritability (h^2_{SNP})

It starts with the linear model $Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

then assumes $\beta_j \sim \mathbb{N}(0, h^2_{SNP}/m)$

With these assumptions, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = h^2_{SNP}/m$

Therefore GCTA assumes that $\mathbb{E}[h_j^2]$ is constant

Heritability Models

Most analyses in human statistical genetics assume $\mathbb{E}[h_j^2]$ is constant

All SNPs are expected to contribute equally

I call this the GCTA Model

GCTA is a method for estimating SNP heritability (h^2_{SNP})

It starts with the linear model $Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

then assumes $\beta_j \sim \mathbb{N}(0, h^2_{SNP}/m)$

With these assumptions, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = h^2_{SNP}/m$

Therefore GCTA assumes that $\mathbb{E}[h_j^2]$ is constant

Heritability Models

Most analyses in human statistical genetics assume $\mathbb{E}[h_j^2]$ is constant

All SNPs are expected to contribute equally

I call this the GCTA Model

The same model is used by the heritability softwares LDSC and MTG2

and by the association softwares Fast-LMM and Bolt-LMM

and by the prediction softwares BayesR, LDpred and lassosum

and by the genetic architecture softwares GENESIS and AI-Mixer

and by my first software, Sparse Partitioning

and is recommended by Tibshirani for lasso and ridge regression

etc, etc.

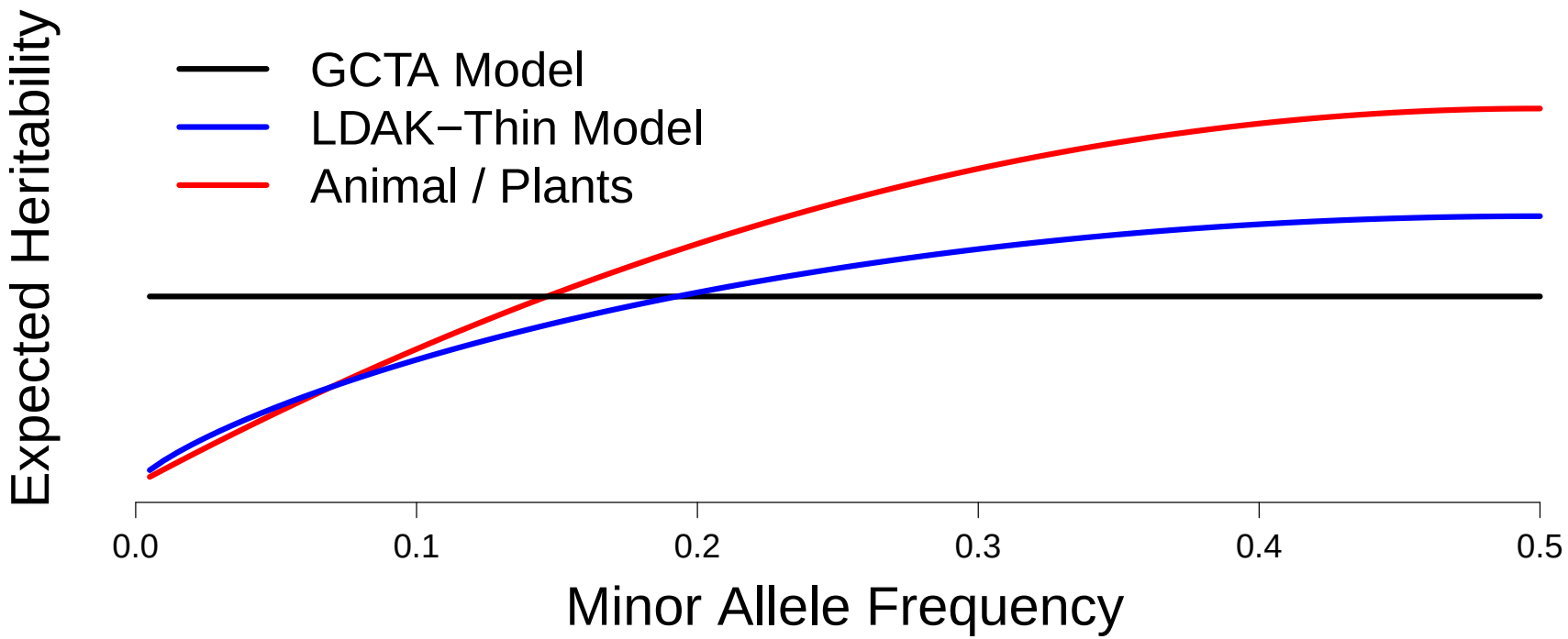
Alternative Models

Can consider models of the form $\mathbb{E}[h_j^2] \propto (p_j(1-p_j))^{1+\alpha}$, where p_j is MAF

GCTA Model obtained by setting $\alpha = -1$

In animal and plant genetics, $\alpha = 0$ is common

I recommend setting $\alpha = -0.25$



Alternative Models

LDAK Model (one parameter): $\mathbb{E}[h_j^2] \propto w_j \times (p_j(1-p_j))^{0.75}$

w_j depends on local levels of linkage disequilibrium

Contribution of a SNP depends on MAF and LD

Baseline LD Model (75 parameters): $\mathbb{E}[h_j^2] = \tau_1 a_{j,1} + \ldots + \tau_{75} a_{j,75}$

Contribution of a SNP depends on MAF, LD and function

e.g., exonic or conserved SNPs contribute more than average

BLD-LDAK Model (66 parameters) - better version of Baseline LD Model

Alternative Models

LDAK Model (one parameter): $\mathbb{E}[h_j^2] \propto w_j \times (p_j(1-p_j))^{0.75}$

w_j depends on local levels of linkage disequilibrium

Contribution of a SNP depends on MAF and LD

Baseline LD Model (75 parameters): $\mathbb{E}[h_j^2] = \tau_1 a_{j,1} + \ldots + \tau_{75} a_{j,75}$

Contribution of a SNP depends on MAF, LD and function

e.g., exonic or conserved SNPs contribute more than average

BLD-LDAK Model (66 parameters) - better version of Baseline LD Model

Alternative Models

LDAK Model (one parameter): $\mathbb{E}[h_j^2] \propto w_j \times (p_j(1-p_j))^{0.75}$

w_j depends on local levels of linkage disequilibrium

Contribution of a SNP depends on MAF and LD

Baseline LD Model (75 parameters): $\mathbb{E}[h_j^2] = \tau_1 a_{j,1} + \ldots + \tau_{75} a_{j,75}$

Contribution of a SNP depends on MAF, LD and function

e.g., exonic or conserved SNPs contribute more than average

BLD-LDAK Model (66 parameters) - better version of Baseline LD Model

Comparing Models

nature genetics

Reevaluation of SNP heritability in complex human traits

Doug Speed[1], Na Cai[2,3], the UCLEB Consortium[4], Michael R Johnson[5], Sergey Nejentsev[6] & David J Balding[1,7]

ANALYSIS
https://doi.org/10.1038/s41588-020-0600-y
nature genetics

Evaluating and improving heritability models using summary statistics

Doug Speed[1,2,3] ✉, John Holmes[4] and David J. Balding[3,4]

Compared models based on how well they fit data for real complex traits

Comparing Models

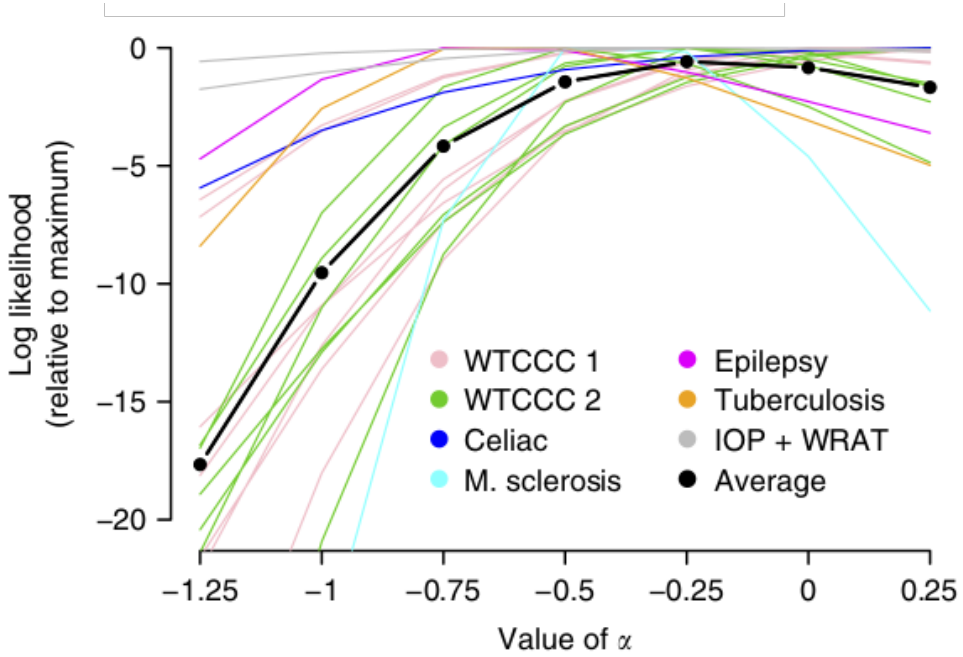
E.g., examined relationship between $\mathbb{E}[h_j^2]$ and MAF by considering models of the form $\mathbb{E}[h_j^2] \propto (p_j(1-p_j))^{1+\alpha}$

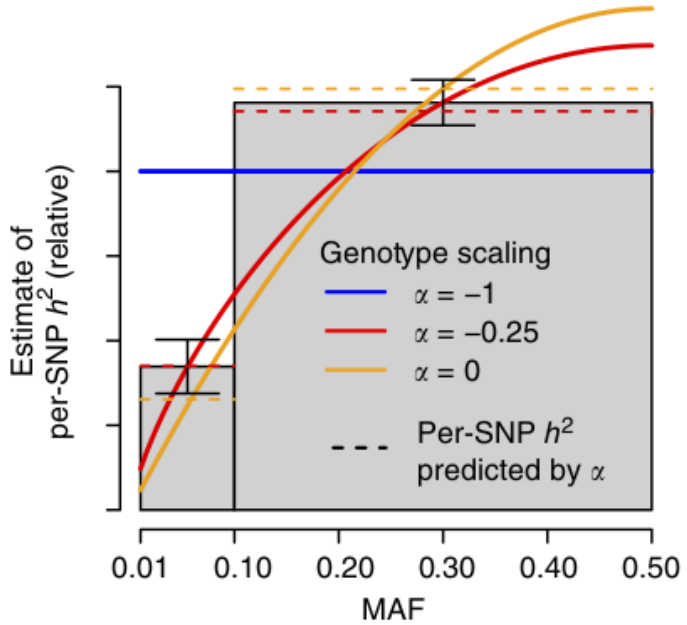
GCTA Model: $\alpha = -1$ Animal and plant genetics: $\alpha = 0$

Comparing Models

E.g., examined relationship between $\mathbb{E}[h_j^2]$ and MAF by considering models of the form $\mathbb{E}[h_j^2] \propto (p_j(1-p_j))^{1+\alpha}$

GCTA Model: $\alpha = -1$ Animal and plant genetics: $\alpha = 0$





Found $\alpha = -0.25$ provides a better fit to the data

The Impact of the Heritability Model

I originally examined its role in SNP-based heritability analyses

Using a poor heritability model resulted in:

Substantial bias when estimating SNP heritability

On average estimates were 40% lower when assuming GCTA Model

Substantial bias when estimating heritability enrichments

Estimates were up to five times higher when assuming GCTA Model

Limited bias when estimating genetic correlations

Although estimates under the GCTA Model tended to be less precise

The Impact of the Heritability Model

I originally examined its role in SNP-based heritability analyses

Using a poor heritability model resulted in:

Substantial bias when estimating SNP heritability

On average estimates were 40% lower when assuming GCTA Model

Substantial bias when estimating heritability enrichments

Estimates were up to five times higher when assuming GCTA Model

Limited bias when estimating genetic correlations

Although estimates under the GCTA Model tended to be less precise

The Impact of the Heritability Model

I originally examined its role in SNP-based heritability analyses

Using a poor heritability model resulted in:

Substantial bias when estimating SNP heritability

On average estimates were 40% lower when assuming GCTA Model

Substantial bias when estimating heritability enrichments

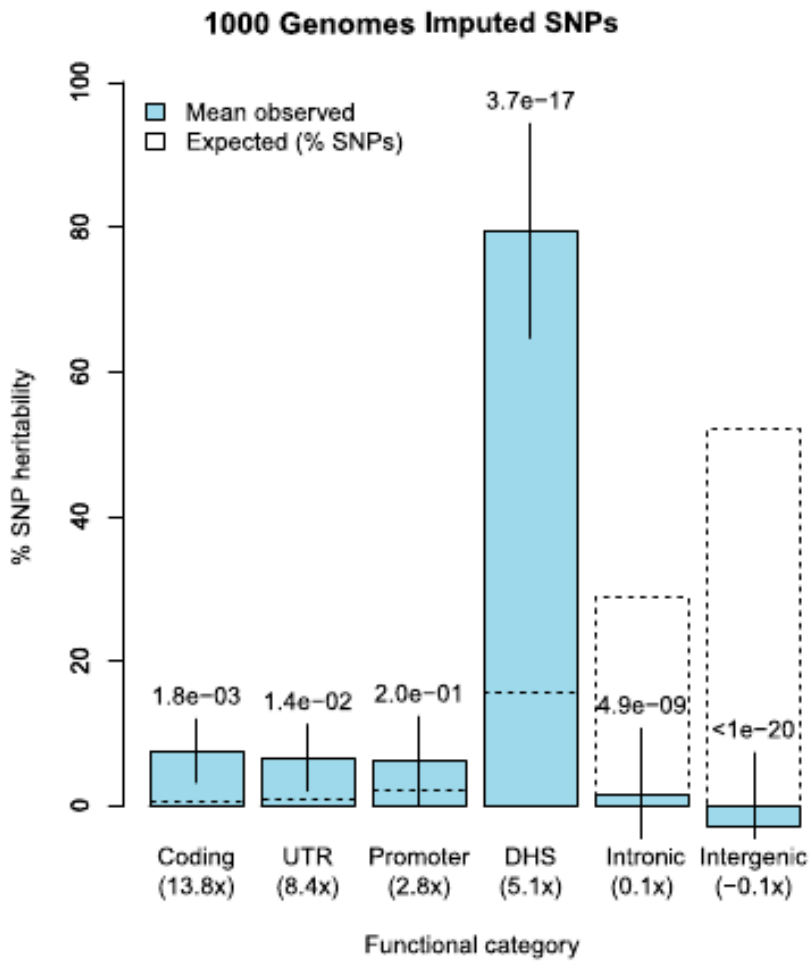
Estimates were up to five times higher when assuming GCTA Model

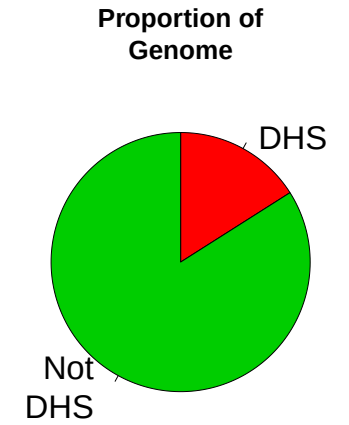
Limited bias when estimating genetic correlations

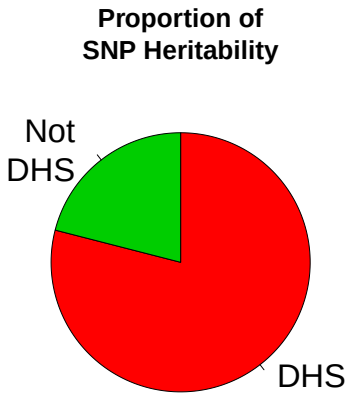
Although estimates under the GCTA Model tended to be less precise

Enrichment of DNase I Hypersensitivity Sites (DHS)

Under the GCTA Model, DHS estimated to explain 79% of h^2_{SNP}



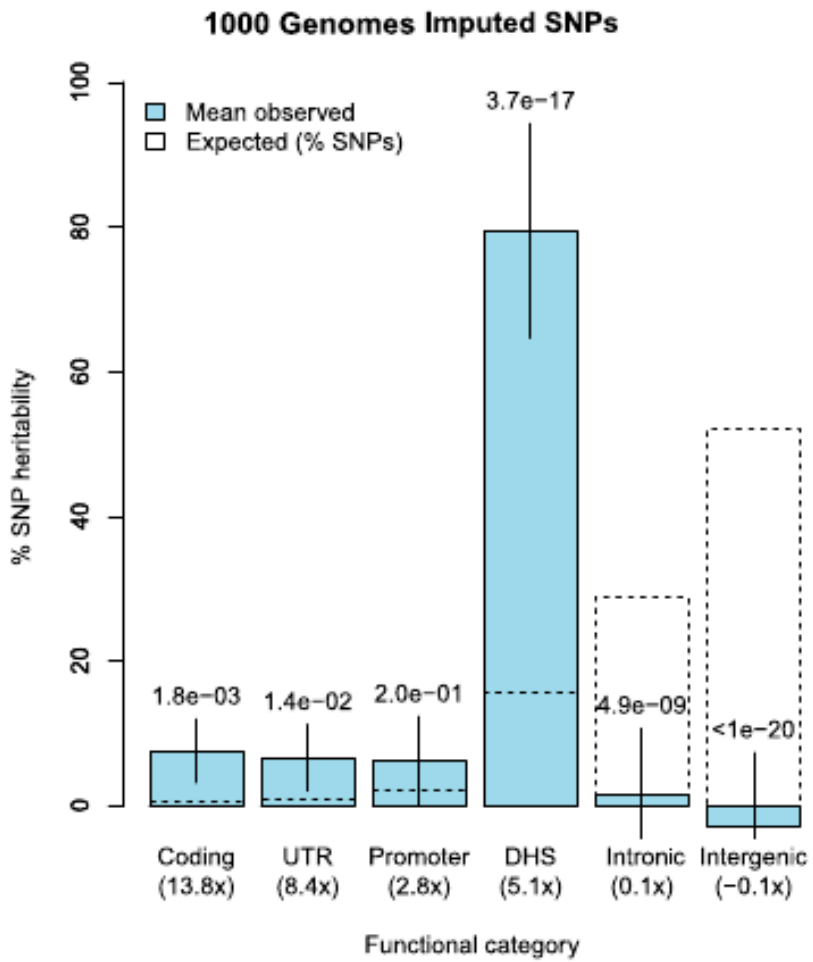


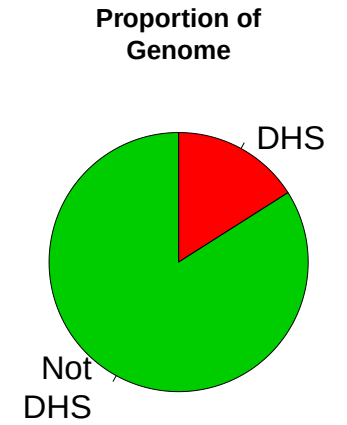


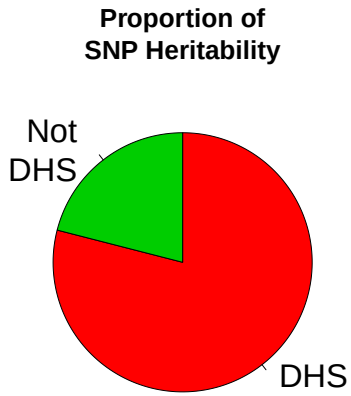
Under the LDAK Model, DHS estimated to explain 24% of h^2_{SNP}

Enrichment of DNase I Hypersensitivity Sites (DHS)

Under the GCTA Model, DHS estimated to explain 79% of h^2_{SNP}







Under the LDAK Model, DHS estimated to explain 24% of h^2_{SNP}

Prediction of Complex Traits

My recent work examines the impact of the heritability model when creating polygenic risk scores (PRS)

PRS take the form $P = X_1\beta_1 + X_2\beta_2 + \ldots + X_m\beta_m$

i.e., a linear SNP-based prediction model

The aim of a prediction tool is to find estimates of β_j that produce the most accurate PRS

Eight Existing PRS Tools

I examined eight of the most commonly-used PRS tools

Individual-level data:

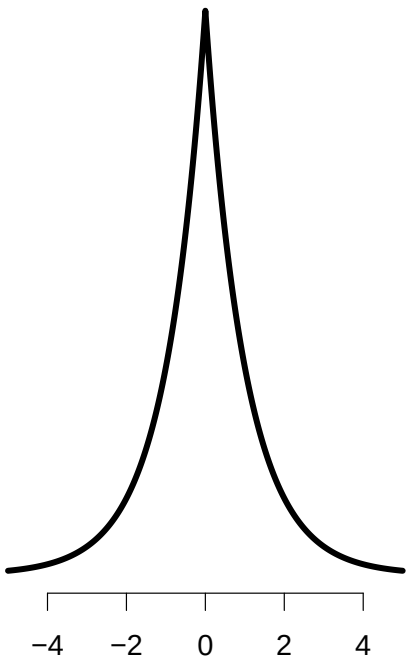
Lasso, Ridge Regression, Bolt-LMM and BayesR

Summary statistics:

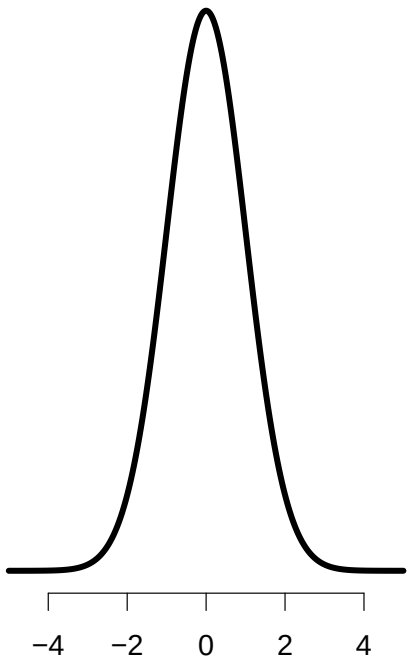
lassosum, sBLUP, LDpred and SBayesR

Eight Existing PRS Tools

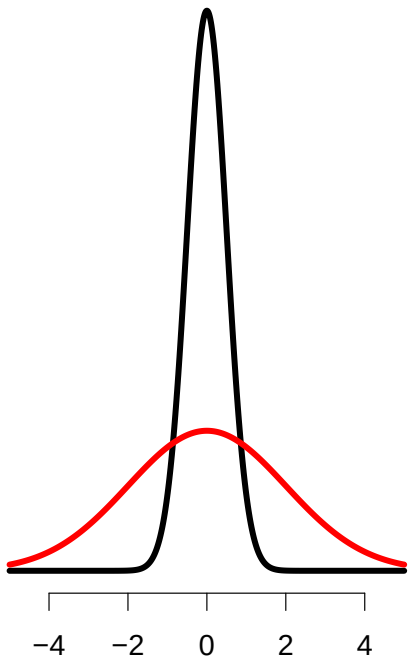
These tools use a variety of prior distribution forms for effect sizes



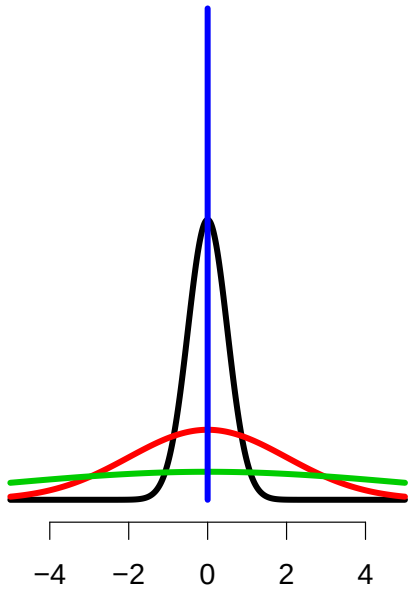
Lasso / lassosum



Ridge / SBLUP



Bolt / LDpred

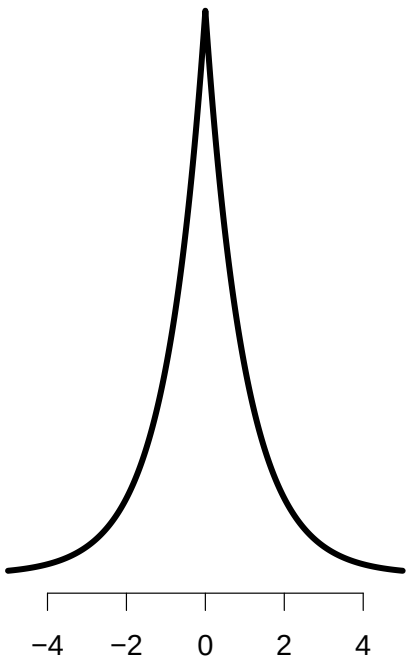


BayesR / SBayesR

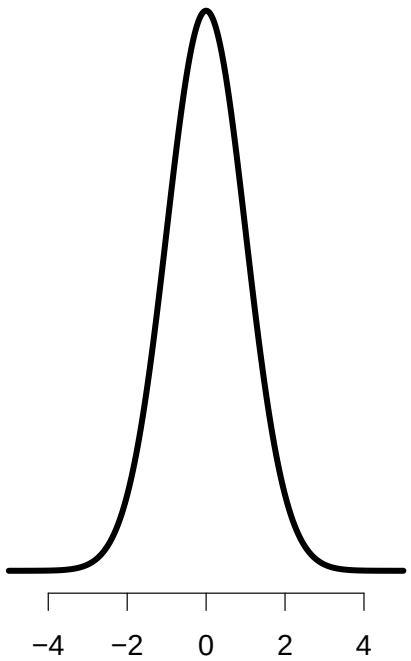
But what heritability model do they assume?

Eight Existing PRS Tools

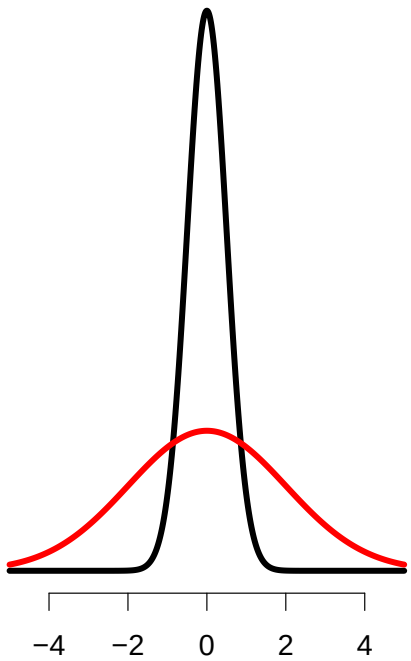
These tools use a variety of prior distribution forms for effect sizes



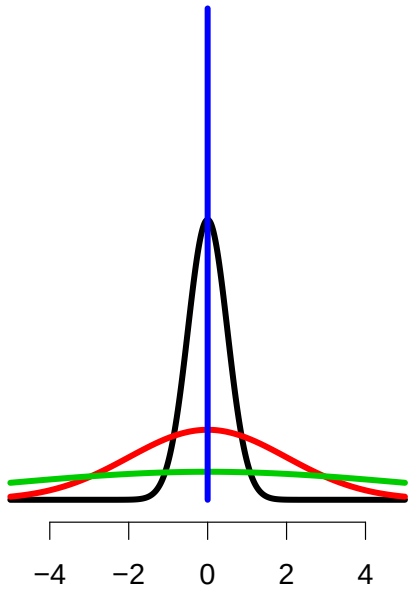
Lasso / lassosum



Ridge / SBLUP



Bolt / LDpred



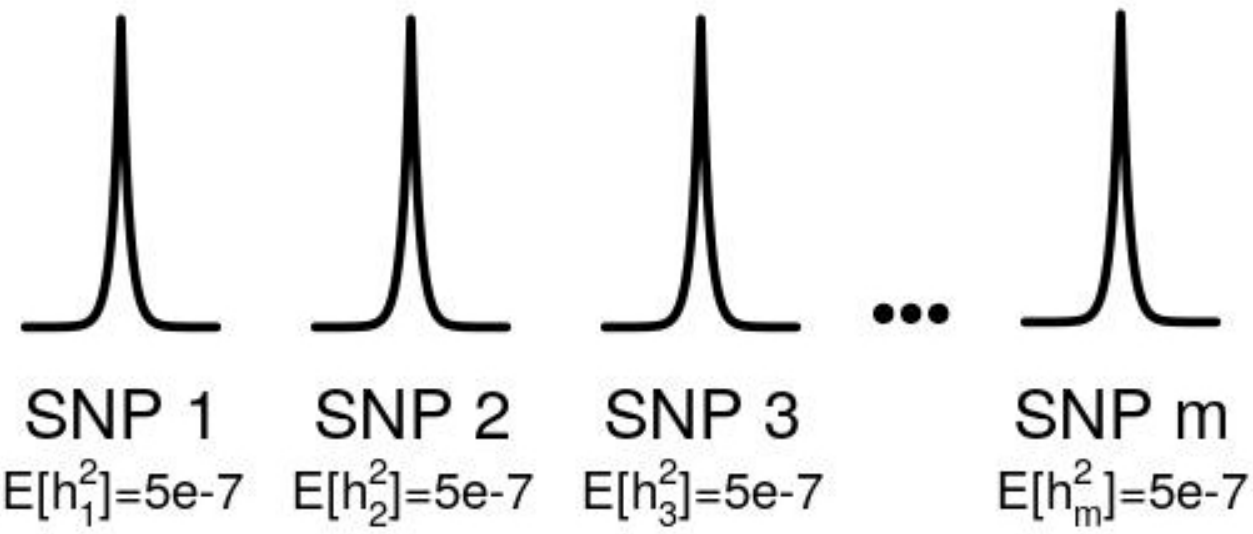
BayesR / SBayesR

But what heritability model do they assume?

Eight Existing PRS Tools

Given the prior form, each tool uses the same parameters for each SNP

E.g., the existing Lasso uses a Laplace distribution with constant rate



If a tool uses the same prior distribution for each β_j

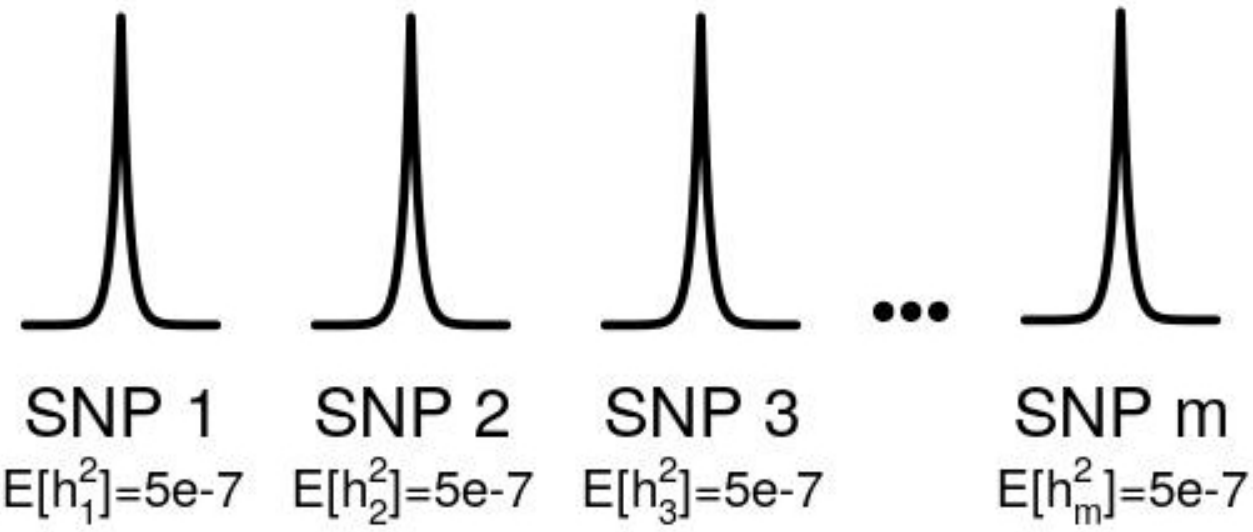
this implies that $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2]$ is constant

Therefore existing tools assume the GCTA Model

Eight Existing PRS Tools

Given the prior form, each tool uses the same parameters for each SNP

E.g., the existing Lasso uses a Laplace distribution with constant rate



If a tool uses the same prior distribution for each β_j

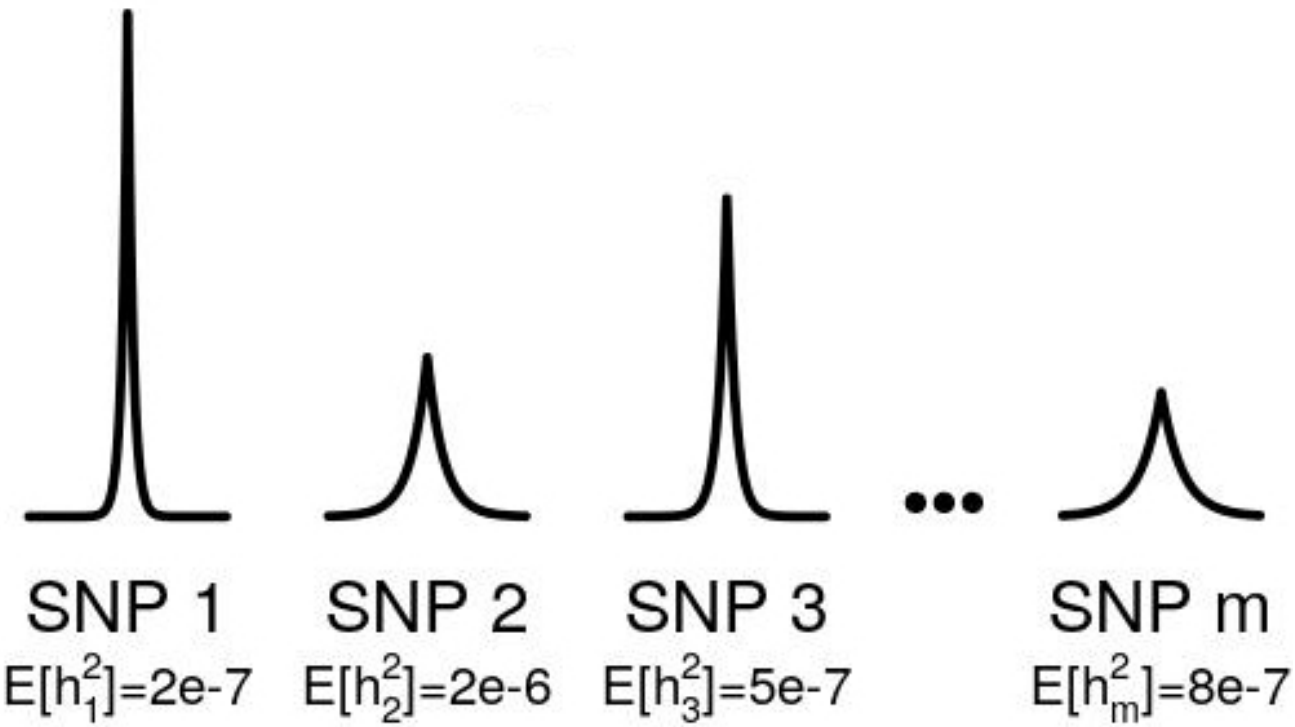
this implies that $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2]$ is constant

Therefore existing tools assume the GCTA Model

Eight New Prediction Tools

I created new versions that allow prior parameters to vary across SNPs

E.g., my Lasso uses a Laplace distribution with SNP-specific rate



Therefore my new tools accommodate different heritability models

Testing the New Tools

Use data from UK Biobank

- On average 280 k individuals and 600 k SNPs
- Have individual-level data for 14 phenotypes
- Have summary statistics for 225 phenotypes

Create prediction models using each of my new tools assuming

- GCTA Model — $\mathbb{E}[h_j^2]$ constant
- LDAK-Thin Model — $\mathbb{E}[h_j^2]$ depends on MAF and LD
- BLD-LDAK Model — $\mathbb{E}[h_j^2]$ depends on MAF, LD and function

Measure accuracy of predicted phenotypes via R^2 (squared correlation)

Testing the New Tools

Use data from UK Biobank

- On average 280 k individuals and 600 k SNPs
- Have individual-level data for 14 phenotypes
- Have summary statistics for 225 phenotypes

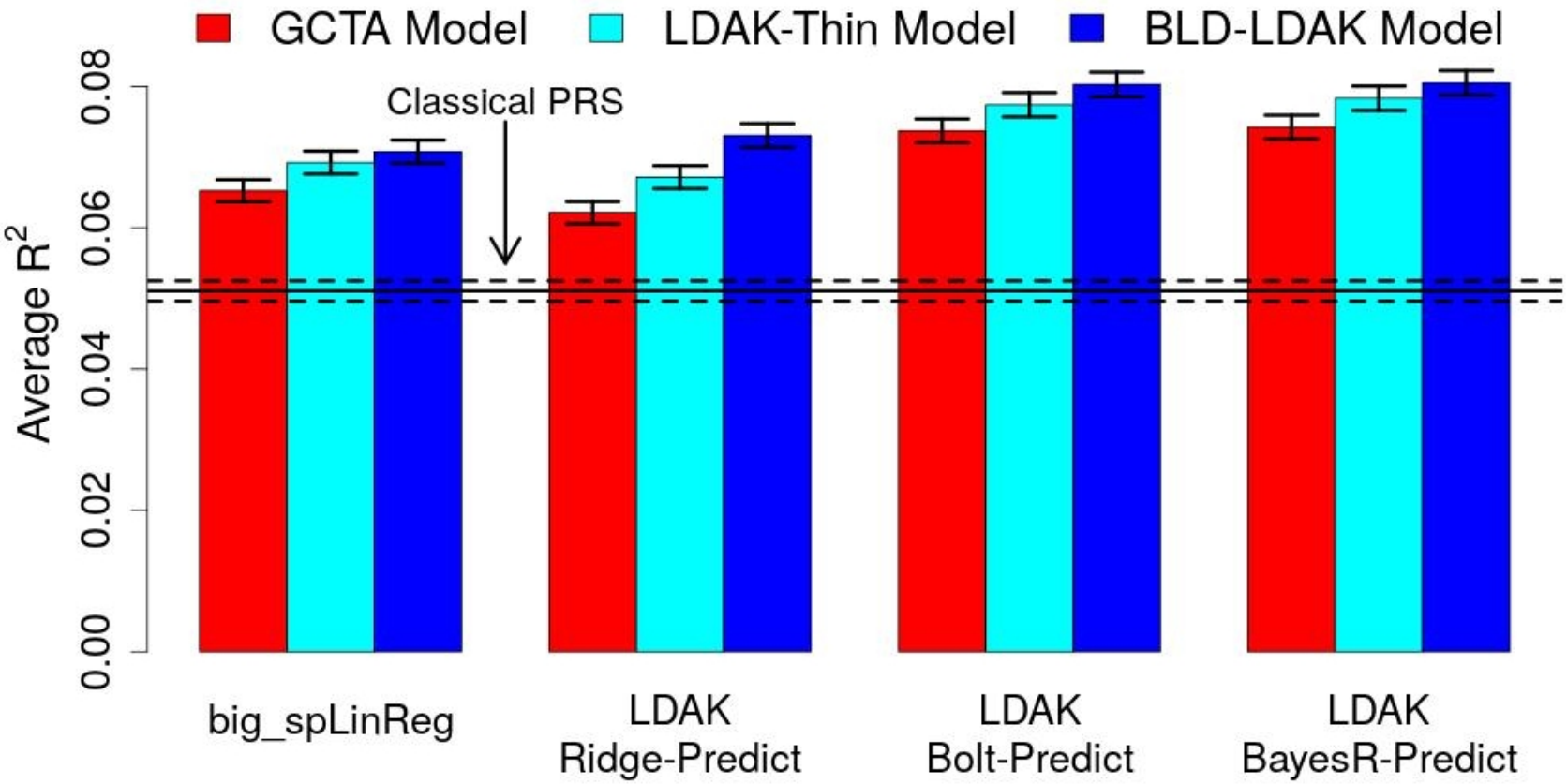
Create prediction models using each of my new tools assuming

- GCTA Model — $\mathbb{E}[h_j^2]$ constant
- LDAK-Thin Model — $\mathbb{E}[h_j^2]$ depends on MAF and LD
- BLD-LDAK Model — $\mathbb{E}[h_j^2]$ depends on MAF, LD and function

Measure accuracy of predicted phenotypes via R^2 (squared correlation)

Performance of New Individual-level Data Tools

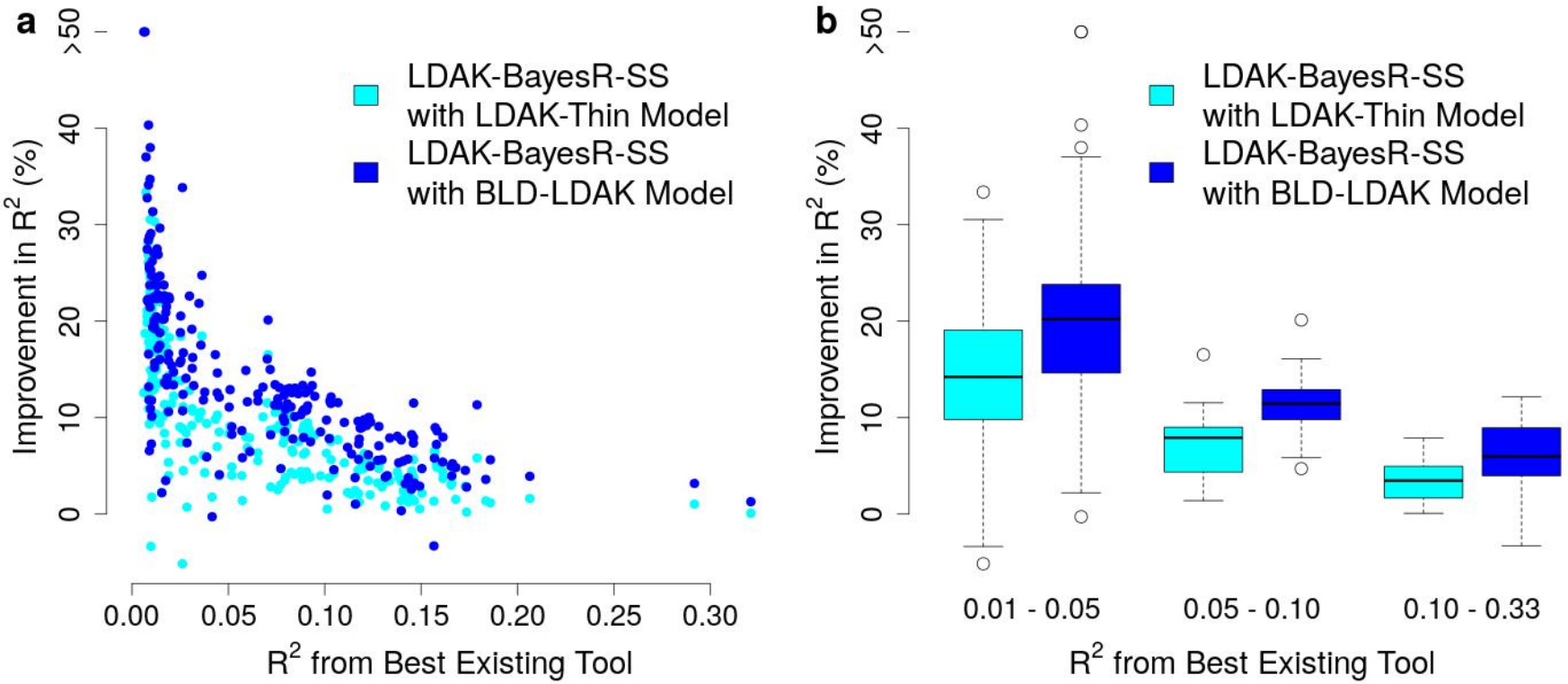
Using a better heritability model ALWAYS improves prediction



Improved accuracy for all four tools and for all 14 phenotypes

Performance of New Summary Statistic Tools

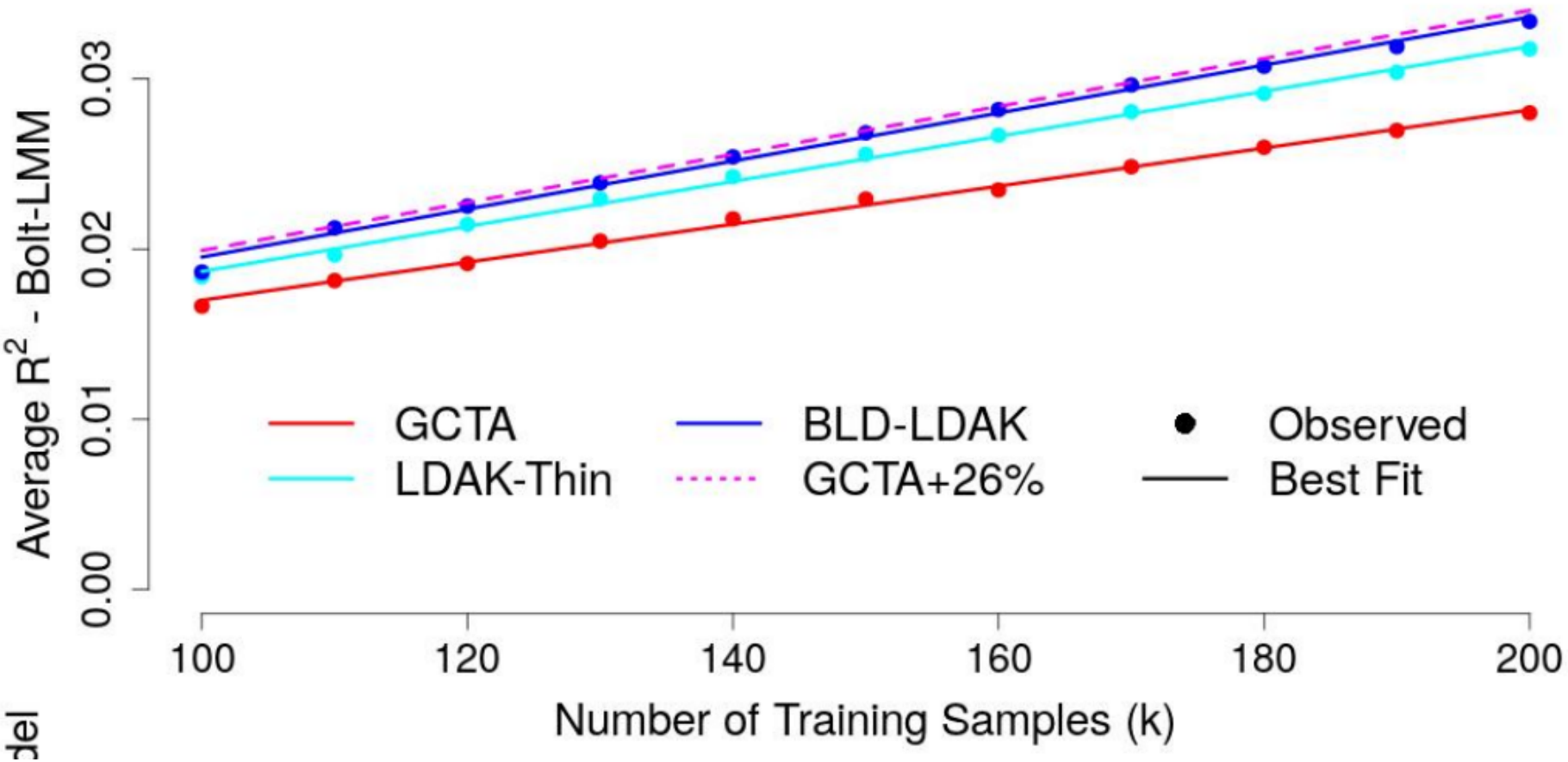
Using a better heritability model almost always improves prediction



Improved accuracy for either 223 or 225 phenotypes (depending on tool)

The Improvement Tends to be Large

Average increase in R^2 is 14% (SD 1%)



This is equivalent to increasing sample size by a quarter

The Improvement Tends to be Large

Classical PRS
Single-SNP model

Improvements in Accuracy (R^2)

The Improvement Tends to be Large

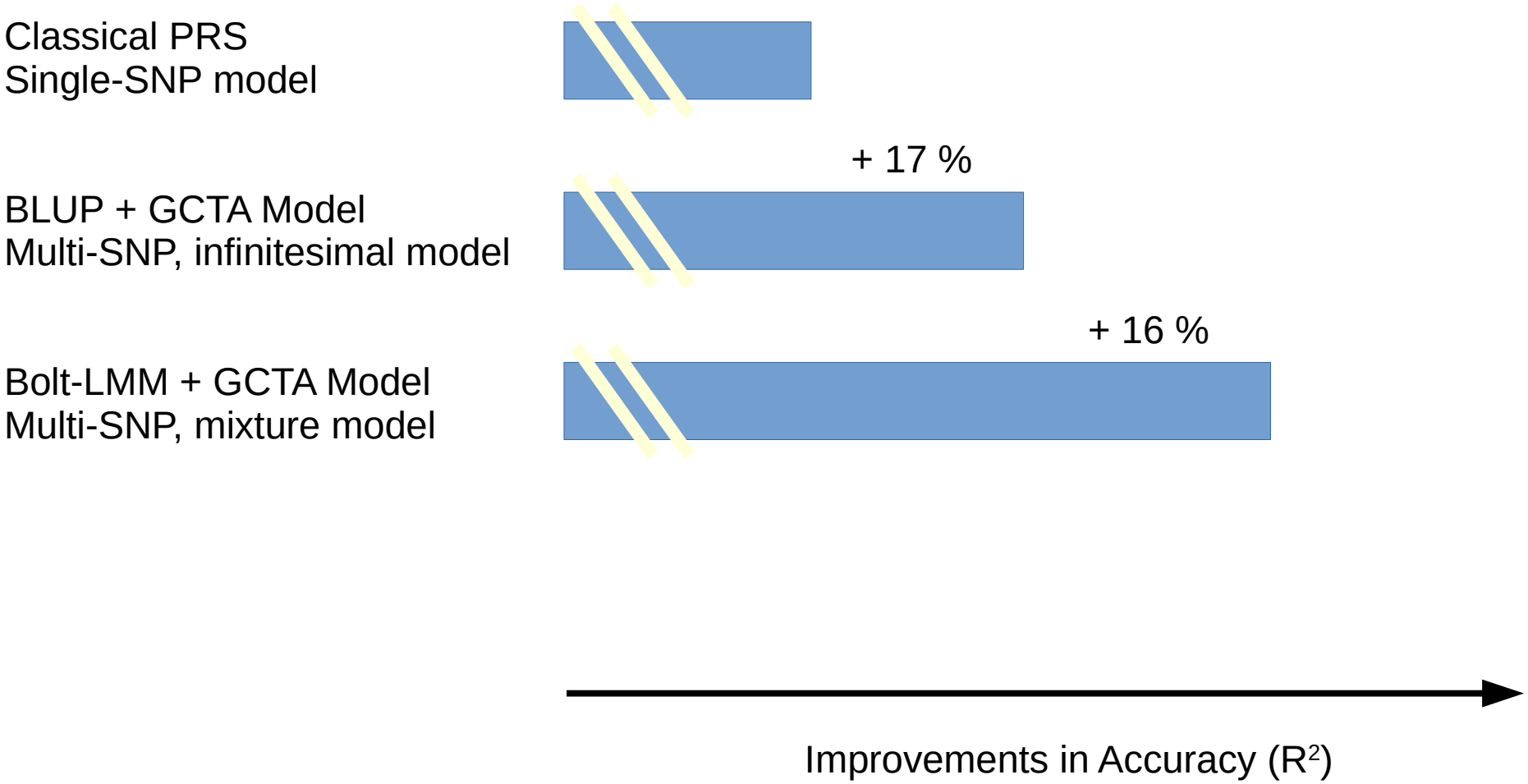
Classical PRS
Single-SNP model

+ 17 %

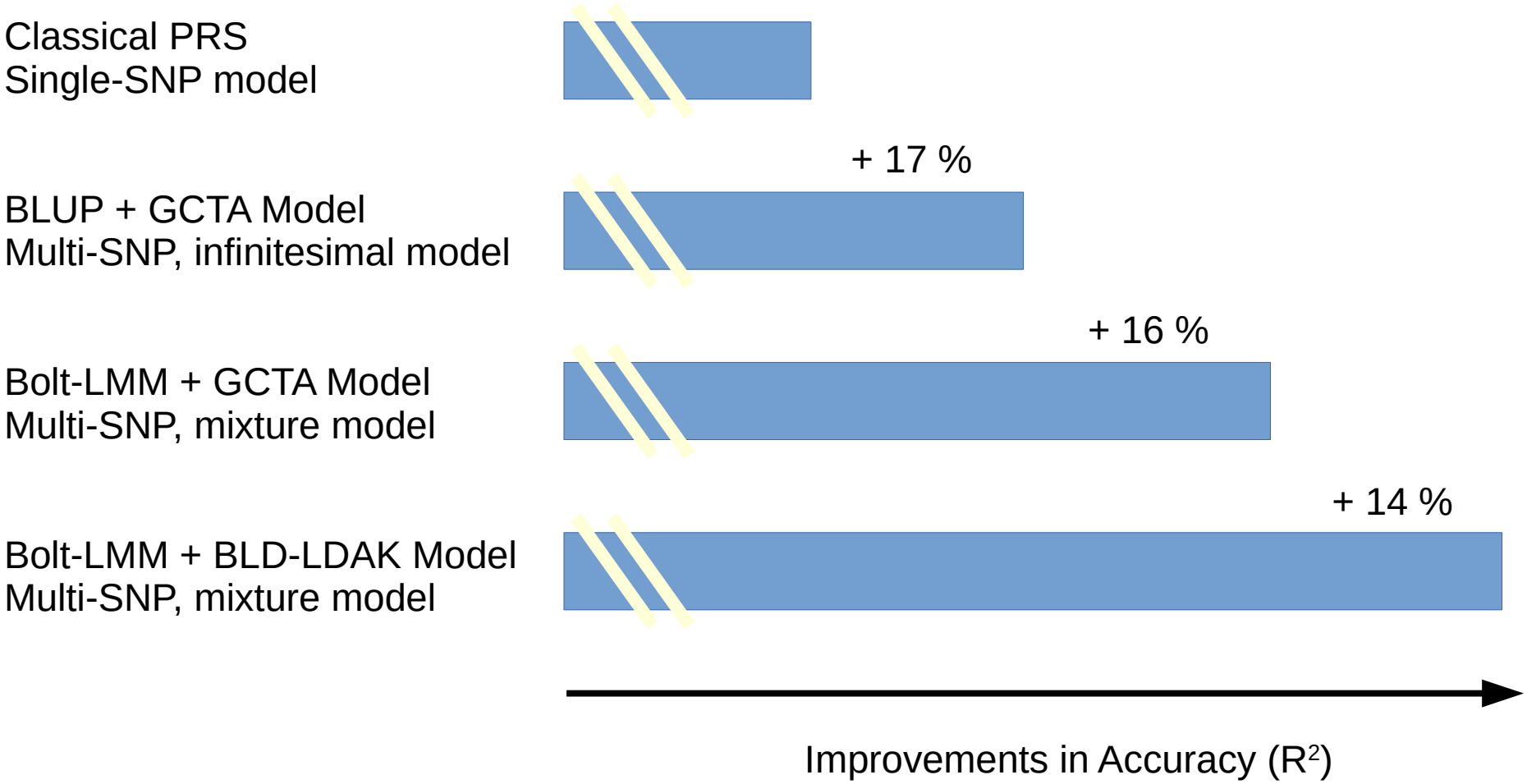
BLUP + GCTA Model
Multi-SNP, infinitesimal model

Improvements in Accuracy (R^2)

The Improvement Tends to be Large



The Improvement Tends to be Large



Computational demands

Individual-level data tools require 20 Gb memory and 3-50 hours CPU

(this is with 200 k individuals and 600 k SNPs)

Use Variational Bayes instead of MCMC

Recommend LDAK-Bolt-Predict - takes 20 hours

Tool can be run in parallel

Provides regular updates and can be restarted if failed to complete

Summary statistic tools require <5 Gb memory and less than 2 hours CPU

Recommend LDAK-BayesR-SS (also provide parallel version)

Computational demands

Individual-level data tools require 20 Gb memory and 3-50 hours CPU

(this is with 200 k individuals and 600 k SNPs)

Use Variational Bayes instead of MCMC

Recommend LDAK-Bolt-Predict - takes 20 hours

Tool can be run in parallel

Provides regular updates and can be restarted if failed to complete

Summary statistic tools require <5 Gb memory and less than 2 hours CPU

Recommend LDAK-BayesR-SS (also provide parallel version)

Computational demands

Individual-level data tools require 20 Gb memory and 3-50 hours CPU

(this is with 200 k individuals and 600 k SNPs)

Use Variational Bayes instead of MCMC

Recommend LDAK-Bolt-Predict - takes 20 hours

Tool can be run in parallel

Provides regular updates and can be restarted if failed to complete

Summary statistic tools require <5 Gb memory and less than 2 hours CPU

Recommend LDAK-BayesR-SS (also provide parallel version)

Read the Paper to Find out More

ARTICLE

https://doi.org/10.1038/s41467-021-24485-y OPEN

Check for updates

Improved genetic prediction of complex traits from individual-level data or summary statistics

Qianqian Zhang[1], Florian Privé[2], Bjarni Vilhjálmsson[1,2] & Doug Speed[1,3,4]✉

The Heritability Model

The heritability model is important

- It is important for SNP-based heritability analysis
- It is very important for prediction
- I guess it is important for other types of methods

Yet people continue to almost exclusively use the GCTA Model

Why is this?

The Heritability Model

The heritability model is important

- It is important for SNP-based heritability analysis
- It is very important for prediction
- I guess it is important for other types of methods

Yet people continue to almost exclusively use the GCTA Model

Why is this?

Ignorance

People have not heard of the heritability model

People who create methods do not realise they are using the GCTA Model

People who use methods do not realise they are using the GCTA Model

Whenever you place a prior distribution on effect sizes

you are assuming a heritability model

Whenever you standardize predictors then use the same prior for all effects

you are assuming the GCTA Model

Ignorance

People have not heard of the heritability model

People who create methods do not realise they are using the GCTA Model

People who use methods do not realise they are using the GCTA Model

Whenever you place a prior distribution on effect sizes

you are assuming a heritability model

Whenever you standardize predictors then use the same prior for all effects

you are assuming the GCTA Model

Ignorance

REGENIE, efficient mixed-model assiciation analysis, Nature Genetics 2021

Technical Report | Published: 20 May 2021

Computationally efficient whole-genome regression for quantitative and binary traits

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell & Jonathan Marchini

Nature Genetics **53**, 1097–1103 (2021) | Cite this article

6694 Accesses | **4** Citations | **35** Altmetric | Metrics

Please note, REGNIE is just one example

Ignorance

Extract from methods section:

covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

Ignorance

Extract from methods section:

covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

Linear Model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

Ignorance

Extract from methods section:

covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

Linear Model

Standardized SNPs

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

Ignorance

Extract from methods section:

covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

Linear Model

Standardized SNPs

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

All effect sizes given the same prior distribution

Ignorance

Extract from methods section:

covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

Linear Model

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

Standardized SNPs

All effect sizes given the same prior distribution

It is not mentioned, however REGENIE assumes the GCTA Model

Confusion

The heritability model is distinct from the prior distribution form

BLUP assumes $\beta_j \sim \mathbb{N}(0, \sigma_g^2)$

LDpred assumes $\beta_j \sim p\ \mathbb{N}(0, \sigma_g^2) + (1-p)\delta_{\{0\}}$

LDpred differs from BLUP with respect to polygeneity

- BLUP assumes all SNPs are causal
- LDpred assumes only a fraction p of SNPs are causal

But the two methods use the same the heritability model

- In BLUP, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = \sigma_g^2$
- In LDpred, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = p\sigma_g^2$

Confusion

The heritability model is distinct from the prior distribution form

BLUP assumes $\beta_j \sim \mathbb{N}(0, \sigma_g^2)$

LDpred assumes $\beta_j \sim p\ \mathbb{N}(0, \sigma_g^2) + (1-p)\delta_{\{0\}}$

LDpred differs from BLUP with respect to polygeneity

- BLUP assumes all SNPs are causal
- LDpred assumes only a fraction p of SNPs are causal

But the two methods use the same the heritability model

- In BLUP, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = \sigma_g^2$
- In LDpred, $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = p\sigma_g^2$

Reluctance to Change

Changing the heritability model appears to be very complicated

The GCTA Model assumes $\mathbb{E}[h_j^2] = \sigma_g^2$

The LDAK-Thin Model assumes $\mathbb{E}[h_j^2] = (p_j(1-p_j))^{0.75}\ \sigma_g^2$

The BLD-LDAK Model assumes $\mathbb{E}[h_j^2] = (a_{j1}\ \tau_1 + \ldots + a_{j66}\ \tau_{66})\sigma_g^2$

But in practice, each model has the form $\mathbb{E}[h_j^2] = q_j\sigma_g^2$

where the q_j are computed in advance (using SumHer)

For example, BEGENIE assumes $\beta_j \sim \mathbb{N}(0, \sigma_g^2)$

they can either assume $\beta_j \sim \mathbb{N}(0, q_j\sigma_g^2)$

or instead scale predictors so that $\mathrm{Var}(X_j) = q_j$

Reluctance to Change

Changing the heritability model appears to be very complicated

The GCTA Model assumes $\mathbb{E}[h_j^2] = \sigma_g^2$

The LDAK-Thin Model assumes $\mathbb{E}[h_j^2] = (p_j(1-p_j))^{0.75}\ \sigma_g^2$

The BLD-LDAK Model assumes $\mathbb{E}[h_j^2] = (a_{j1}\ \tau_1 + \ldots + a_{j66}\ \tau_{66})\sigma_g^2$

But in practice, each model has the form $\mathbb{E}[h_j^2] = q_j\sigma_g^2$

where the q_j are computed in advance (using SumHer)

For example, BEGENIE assumes $\beta_j \sim \mathbb{N}(0, \sigma_g^2)$

they can either assume $\beta_j \sim \mathbb{N}(0, q_j\sigma_g^2)$

or instead scale predictors so that $\mathrm{Var}(X_j) = q_j$

Reluctance to Change

Changing the heritability model appears to be very complicated

The GCTA Model assumes $\mathbb{E}[h_j^2] = \sigma_g^2$

The LDAK-Thin Model assumes $\mathbb{E}[h_j^2] = (p_j(1-p_j))^{0.75}\ \sigma_g^2$

The BLD-LDAK Model assumes $\mathbb{E}[h_j^2] = (a_{j1}\ \tau_1 + \ldots + a_{j66}\ \tau_{66})\sigma_g^2$

But in practice, each model has the form $\mathbb{E}[h_j^2] = q_j\sigma_g^2$

where the q_j are computed in advance (using SumHer)

For example, BEGENIE assumes $\beta_j \sim \mathbb{N}(0, \sigma_g^2)$

they can either assume $\beta_j \sim \mathbb{N}(0, q_j\sigma_g^2)$

or instead scale predictors so that $\mathrm{Var}(X_j) = q_j$

Summary

Most methods assume the GCTA Model

that all SNPs expected to contribute equal heritability

People should instead use a more realistic heritability model

can lead to substantially better performance

My software LDAK contains many different GWAS methods

Heritability analyses using individual-level data and summary statistics

estimating h^2_{SNP}, heritability enrichments, impact of selection

Constructing PRS using individual-level data and summary statistics

improved versions of lasso, ridge regression, Bolt-LMM and BayesR

Summary

Most methods assume the GCTA Model

that all SNPs expected to contribute equal heritability

People should instead use a more realistic heritability model

can lead to substantially better performance

My software LDAK contains many different GWAS methods

Heritability analyses using individual-level data and summary statistics

estimating h^2_{SNP}, heritability enrichments, impact of selection

Constructing PRS using individual-level data and summary statistics

improved versions of lasso, ridge regression, Bolt-LMM and BayesR

Final Slide

Improved genetic prediction of complex traits from individual-level data or summary statistics (Nature Communications 2021)

LDAK software is available at www.ldak.org

Funding from:

Independent Reseach Fund Denmark

EU Horizon 2020

Aarhus University Research Foundation

Lundbeck Foundation

www.ldak.org