Prediction of quantitative traits using marker data

Peter M. Visscher & Michael E. Goddard

peter.visscher@uq.edu.au

Mike.Goddard@depi.vic.gov.au

Key concepts

- Prediction of phenotypic values is limited by heritability
- Accuracy of prediction depends on
  - how well marker effects are estimated (sample size)
  - how well marker effects are correlated with causal variants (LD)
- Estimation of marker effects and prediction in the same data leads to (severe) bias
- Variance explained by a SNP-based predictor is not the same as the variance explained by those SNPs
- Best prediction methods take genetic values as random effects

WORLD WIDE SIRES, LTD.
YOUR FOUNDATION...YOUR FUTURE

7HO10780 UNICORN MILLION ABERLIN-ET *TR *TV
*TL *TY *TD
USA 000066985571
MILLION X GOLDWYN X O MAN
100% Registered Holstein Ancestry

ABERLIN

...orn Million Aberlin-ET

DAM RC-LC Goldwyn ATM

Copyright © 2001 by the Genetics Society of America

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes[†] and M. E. Goddard[†,‡]

Production				Management Traits	
TPI	2206	PTA Ty...	3.32	SCE / Rel.%	7/69
NM$	561	Udder Comp...	3.53	DCE / Rel.%	7/65
PTA Milk (lbs)	836	Feet & Leg Comp...	2.53	SSB / Rel.%	7.4/56
PTA Protein (lbs)	29	Body Composite	2.31	DSB / Rel.%	8.1/57
PTA Protein (%)	0.01	Dairy Composite	1.90	SCS	2.65
PTA Fat (lbs)	38	Type Reliability %	72	Productive Life	4.6
PTA Fat (%)	0.02	Dtrs / Herds	0/0	DPR / Rel.%	0.8/61
Production Reliability %	76	aAa	342	SCR / Rel.%	2.6/89
Dtrs / Herds	0/0				

"Genomic selection" =
'precision medicine' for cows

Take-home from animal breeding

(1) Don’t need genome-wide significant effects

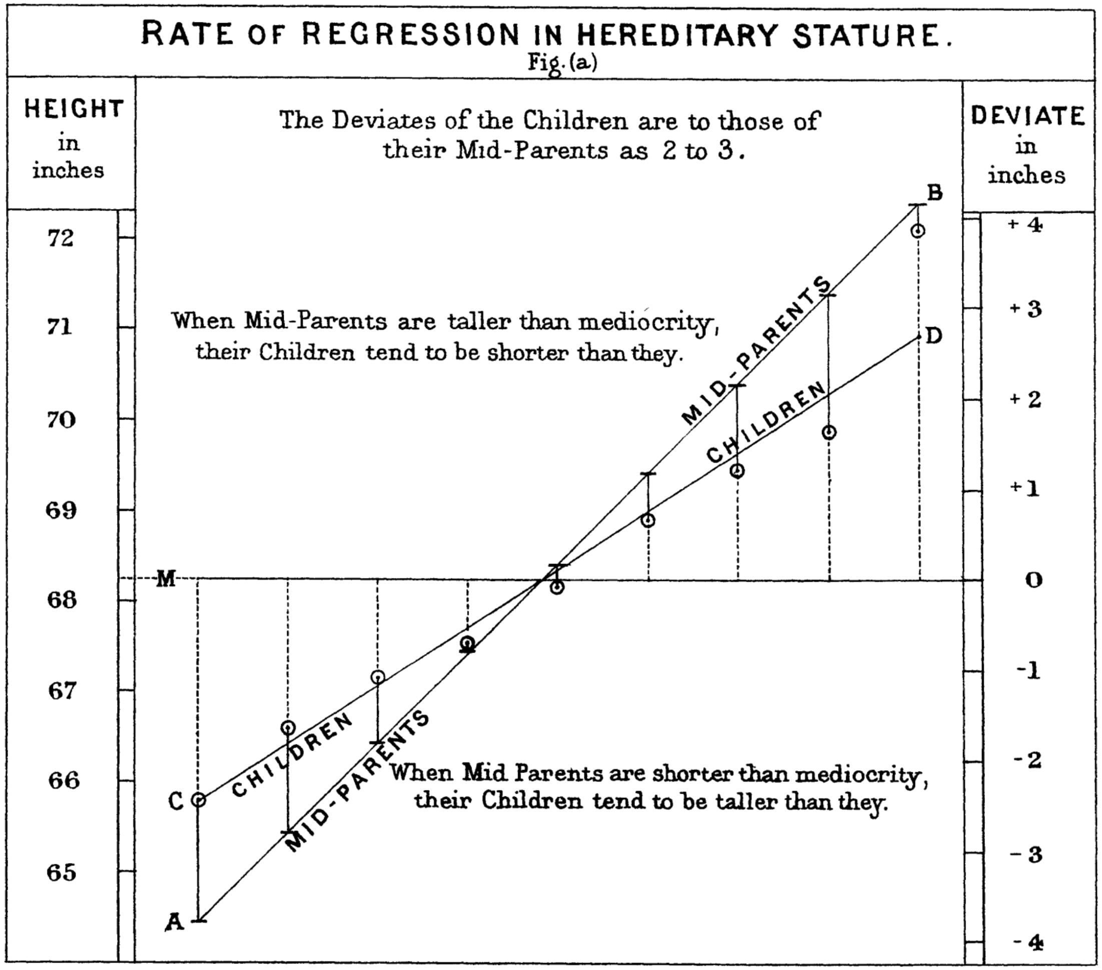
(2) Don’t need to know causal variants

(3) Don’t need to know function

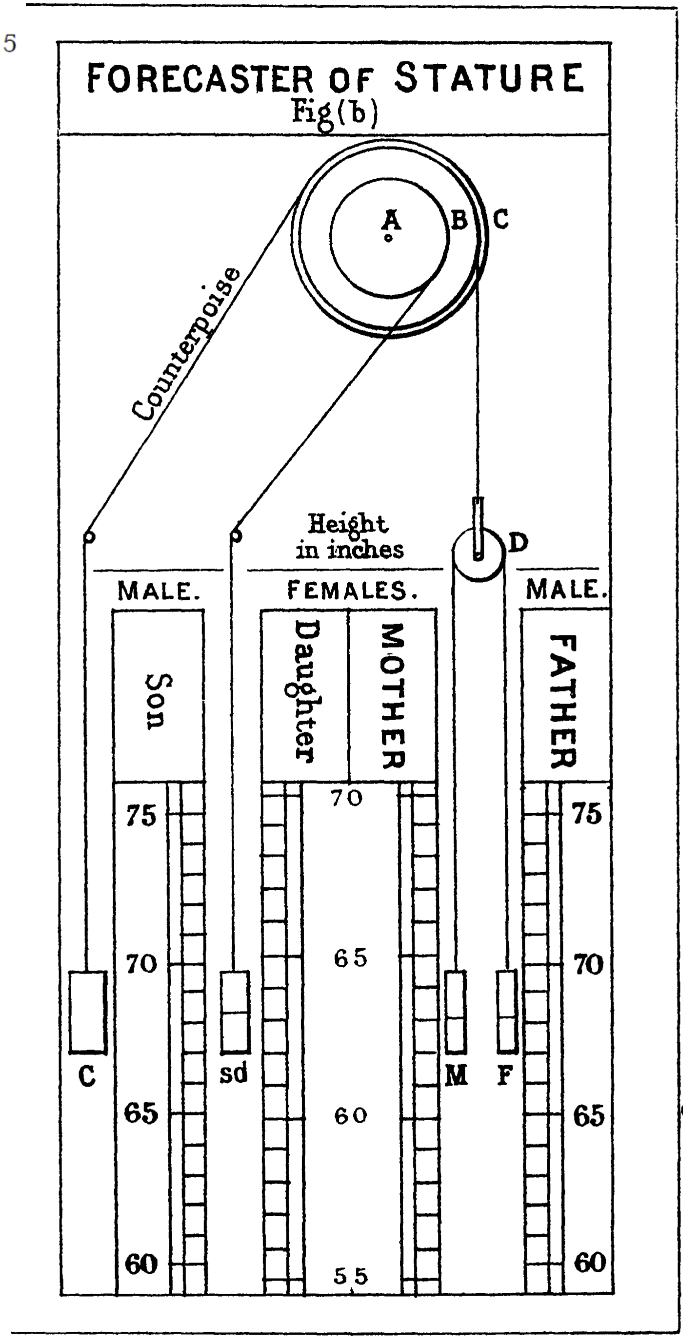
(4) Use all phenotypic and SNP data simultaneously

Regression Towards Mediocrity in Hereditary Stature.
Author(s): Francis Galton
Source: *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15 (1886), pp. 246-263

RATE OF REGRESSION IN HEREDITARY STATURE.
Fig.(a)



FORECASTER OF STATURE
Fig(b)



Journ. Anthropolog. Inst., Vol.XV, Pl.IX

A quantitative genetics model

y = fixed effects + G + E

G = A + D + I

Possible predictions:

- Predict y from fixed effects and G
- Predict G from A
- Predict y from A using pedigree (IBD)
- **Predict y from A using markers**

Fixed effects models

- Linear regression using GWAS data
- Widely used in human genetics (‘profile scoring’, ‘polygenic risk scores’)
- Properties and pitfall
  - chance association can lead to bias
  - relationship between prediction accuracy and heritability
  - PLINK implementation

Prediction using linear regression

$y = \beta x + e$

- Usually, β and x are considered ‘fixed’
- For SNPs, x is random with variance 2p(1-p) assuming HWE
- Later we will consider the case where β is random

Chance association

m markers, sample size *N*

All $\beta = 0$

Multiple linear regression of *y* on *m* markers

$E(R^2) = m/N$ {strictly $m/(N-1)$}

→ Variation "explained" by chance

[Wishart, 1931]

Selection bias

- Select *m* ‘best’ markers out of *M* in total
- ‘Prediction’ in same sample (in-sample prediction)

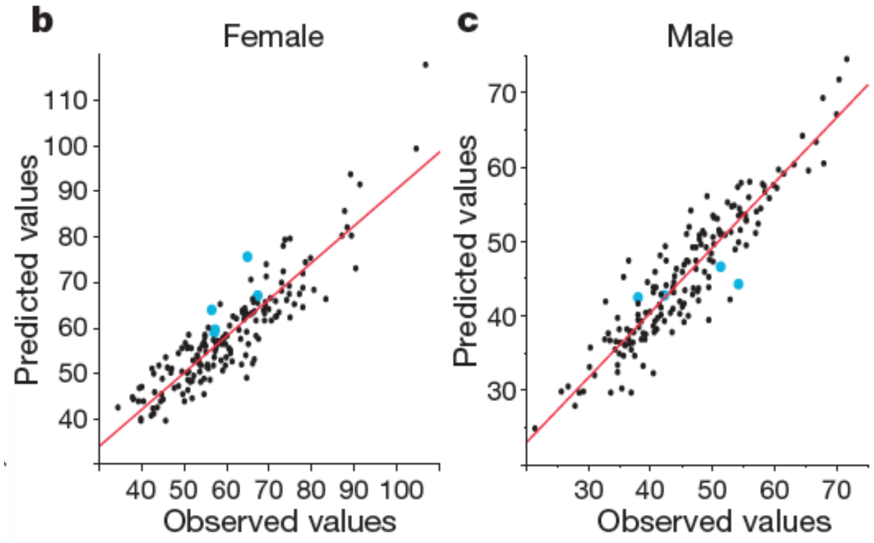
E(R^2) >> m/N

→ Lots of variation explained by chance

ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel



~15 best markers selected from 2.5 million markers

Least squares prediction

$$R_m^2 = \mathrm{var}(a) / \mathrm{var}(y) = h^2$$

$$E(\hat{R}^2_{y,\hat{y}}) \approx h^2 / [1 + m / \{Nh^2\}]$$

Even if we knew all *m* causal variants but needed to estimate their effect sizes then the variance explained by the predictor is less than the variance explained by the causal variants in the population.

[Daetwyler et al. 2008, PLoS Genetics; Visscher, Yang, Goddard 2010, Twin Research Human Genetics 2010]

Take-home

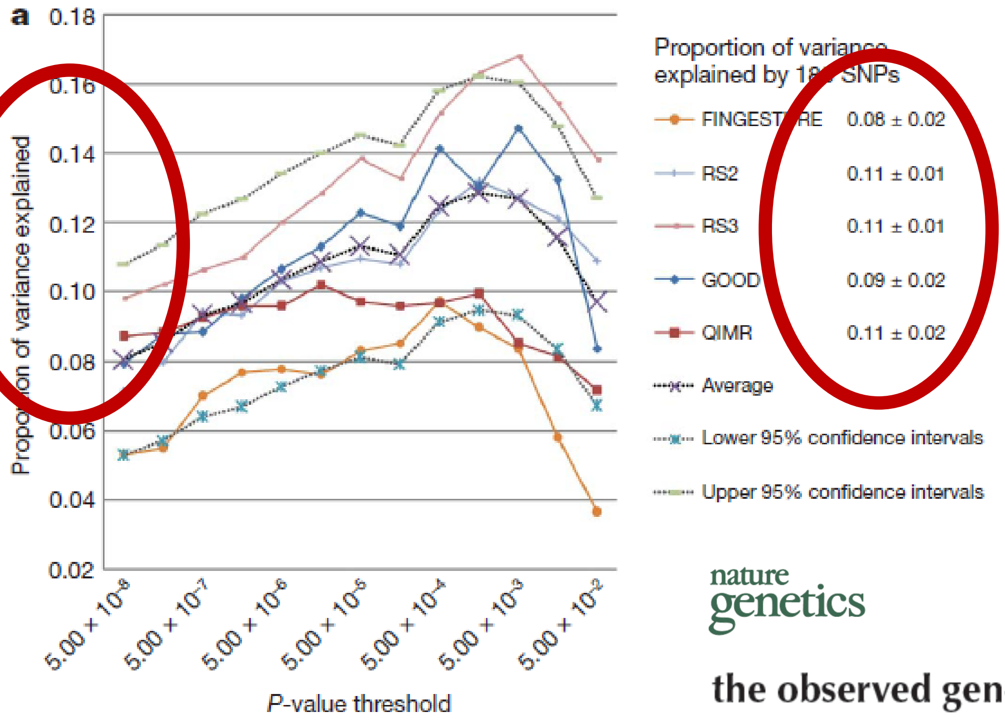
Estimation of variance contributed by (all) loci is not the same as prediction accuracy

unless the effect sizes are estimated without error

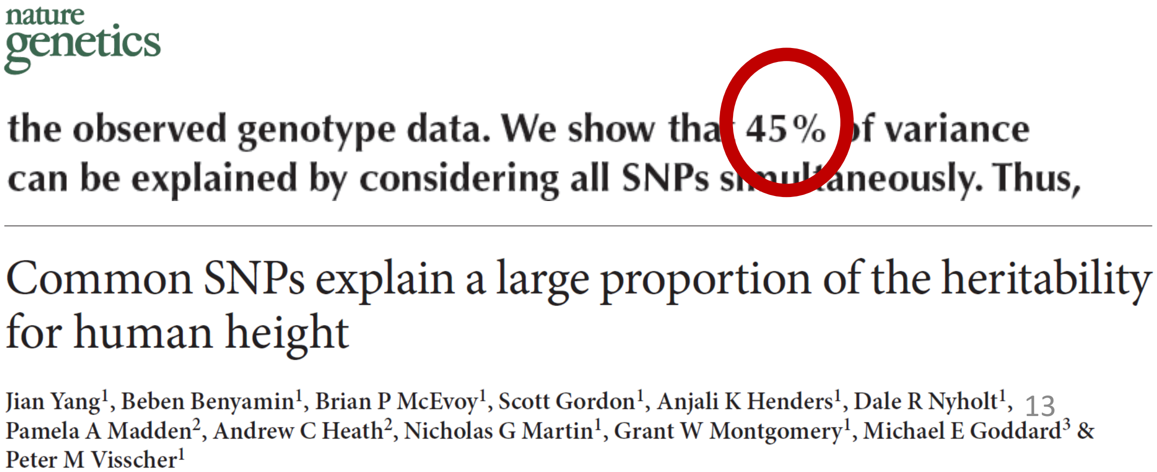
LETTER

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height



SNPs explain 45% of variation
Prediction R^2 ~ 10%



nature genetics

the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus,

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]

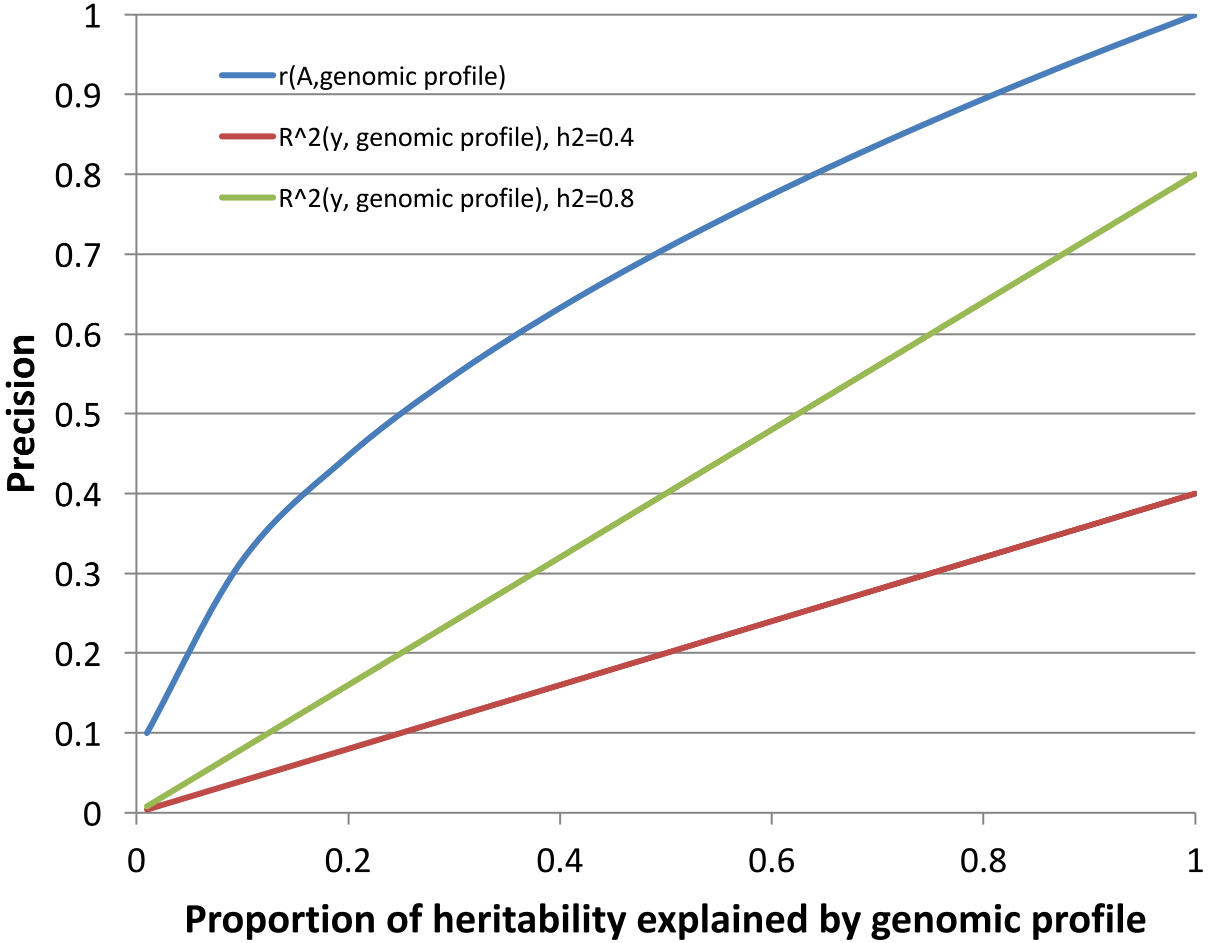
Measures of how well a predictor works

- "Accuracy" (animal breeding)
  - Correlation between true genome-wide genetic value and its predictor
- R^2 from a regression of outcome on predictor (human genetics)
- Area-under-curve from ROC analyses (disease classification)

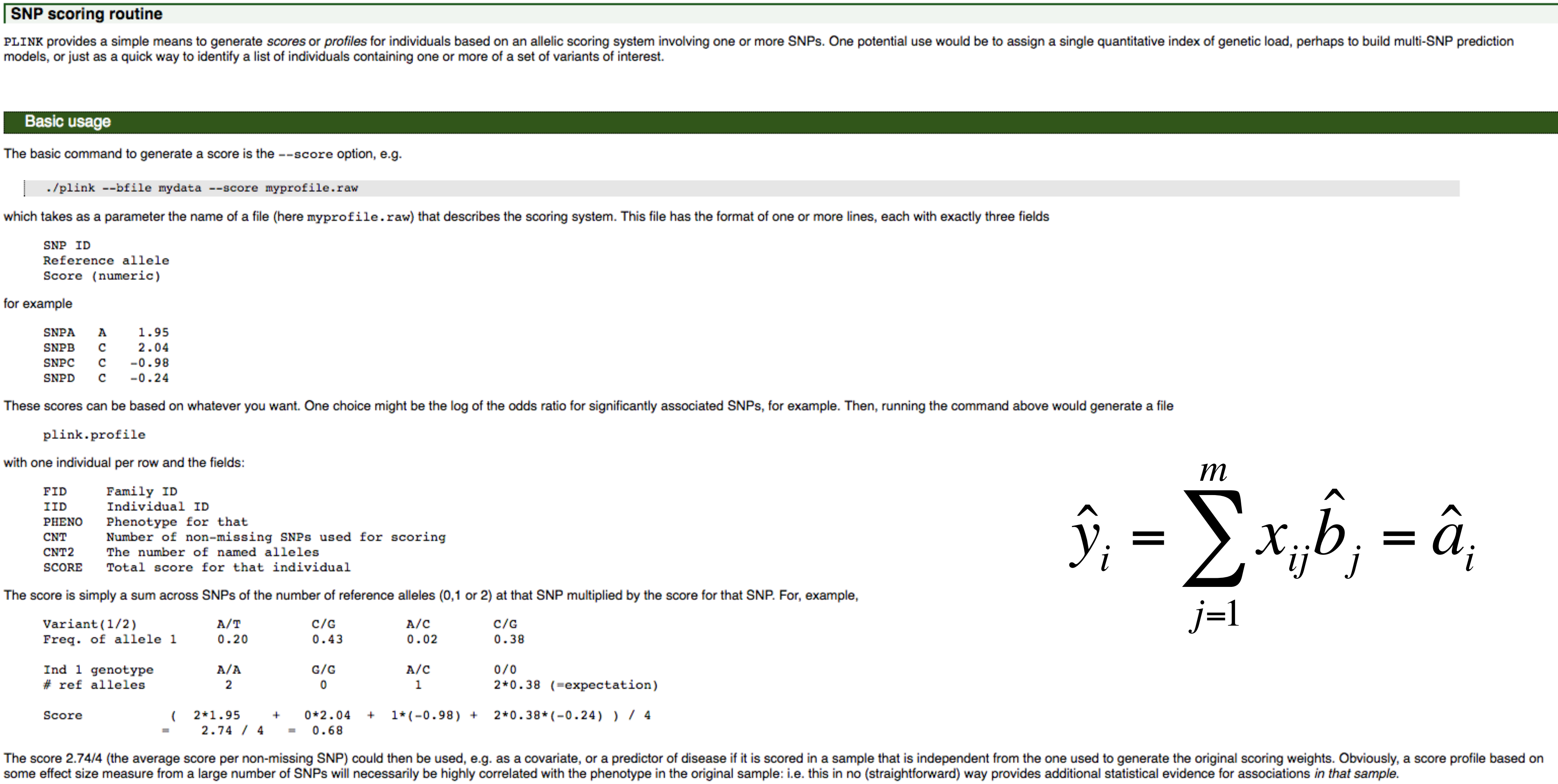
Limits of prediction

- A perfect predictor of A can be a lousy predictor of a phenotype
- The regression R^2 has a maximum that depends on heritability

Predictions from known variants



PLINK profile scoring



SNP scoring routine

PLINK provides a simple means to generate *scores* or *profiles* for individuals based on an allelic scoring system involving one or more SNPs. One potential use would be to assign a single quantitative index of genetic load, perhaps to build multi-SNP prediction models, or just as a quick way to identify a list of individuals containing one or more of a set of variants of interest.

Basic usage

The basic command to generate a score is the --score option, e.g.

```
./plink --bfile mydata --score myprofile.raw
```

which takes as a parameter the name of a file (here myprofile.raw) that describes the scoring system. This file has the format of one or more lines, each with exactly three fields

```
SNP ID
Reference allele
Score (numeric)
```

for example

```
SNPA   A    1.95
SNPB   C    2.04
SNPC   C   -0.98
SNPD   C   -0.24
```

These scores can be based on whatever you want. One choice might be the log of the odds ratio for significantly associated SNPs, for example. Then, running the command above would generate a file

```
plink.profile
```

with one individual per row and the fields:

```
FID     Family ID
IID     Individual ID
PHENO   Phenotype for that
CNT     Number of non-missing SNPs used for scoring
CNT2    The number of named alleles
SCORE   Total score for that individual
```

$$\hat{y}_i = \sum_{j=1}^{m} x_{ij}\hat{b}_j = \hat{a}_i$$

The score is simply a sum across SNPs of the number of reference alleles (0,1 or 2) at that SNP multiplied by the score for that SNP. For, example,

```
Variant(1/2)        A/T        C/G        A/C        C/G
Freq. of allele 1   0.20       0.43       0.02       0.38

Ind 1 genotype      A/A        G/G        A/C        0/0
# ref alleles        2          0          1         2*0.38 (=expectation)

Score             (  2*1.95    +    0*2.04  +  1*(-0.98) +  2*0.38*(-0.24) ) / 4
                =    2.74 / 4    =  0.68
```

The score 2.74/4 (the average score per non-missing SNP) could then be used, e.g. as a covariate, or a predictor of disease if it is scored in a sample that is independent from the one used to generate the original scoring weights. Obviously, a score profile based on some effect size measure from a large number of SNPs will necessarily be highly correlated with the phenotype in the original sample: i.e. this in no (straightforward) way provides additional statistical evidence for associations *in that sample*.

In class demo

- 180 height variants from Lango-Allen et al. 2010
  - Estimation of b from data (N ~ 4000)
  - Using b from Lango-Allen paper
- Taking the top 180 SNPs from GWAS

Random effect models

Prediction of genetic value using better predictors

Model with additive inheritance

$y = g + e$

$V(g) = G\sigma_g^2$, $V(e) = I\sigma_e^2$, $V(y) = V = G\sigma_g^2 + I\sigma_e^2$,

Aim is to predict g for individuals

Eg to predict future risk of a disease

Prediction of genetic value

$y = g + e$

$V(g) = G\sigma_g^2$, $V(e) = I\sigma_e^2$, $V(y) = V = G\sigma_g^2 + I\sigma_e^2$,

Best prediction is

g-hat = $E(g \mid y)$

If y and g are bivariate normal

$E(g \mid y\ ) = b'y = \sigma_g^2\ GV^{-1}\ y$

Prediction of genetic value

Eg Unrelated individuals

$V(g) = Ih^2$, $V(e) = I(1-h^2)$, $V(y) = I$,

Best prediction is

g-hat = $E(g \mid y) = b'y = \sigma_g{}^2\ GV^{-1}\ y = h^2\ y$

Prediction of genetic value

$y = g + e$, $g = Zu$

$V(u) = I\sigma_u^2$, $V(Zu) = ZZ'\sigma_u^2$,

Best prediction is

u-hat = $E(u \mid y)$

If y and u are multivariate normal

$E(u \mid y\ ) = b'y = \sigma_u^2 Z'V^{-1}\ y$

Prediction of genetic value

y = g + e, g = Zu

V(u) = $I\sigma_u^2$, V(Zu) = $ZZ'\sigma_u^2$,

u-hat = E(u | y) = b'y = $\sigma_u^2 Z'V^{-1}$ y

g-hat = Z u-hat = $\sigma_u^2 ZZ'V^{-1}$ y = $\sigma_g^2 GV^{-1}$ y

Prediction of genetic value

$y = g + e$, $g = Zu$

If y and u are multivariate normal

$E(u \mid y) = b'y = \sigma_u^2 Z'V^{-1} y$

The SNP effects are unlikely to be normally distributed with equal variance

Prediction of genetic value

Best prediction

u-hat = $E(u \mid y)$

$= \int u \, P(u \mid y) \, du$

Bayes theorem

$$P(u \mid y) = \underset{\overset{\uparrow}{\text{Likelihood}}}{P(y \mid u)} \; \underset{\overset{\uparrow}{\text{prior}}}{P(u)} \; / \; P(\text{data})$$

Prediction of genetic value

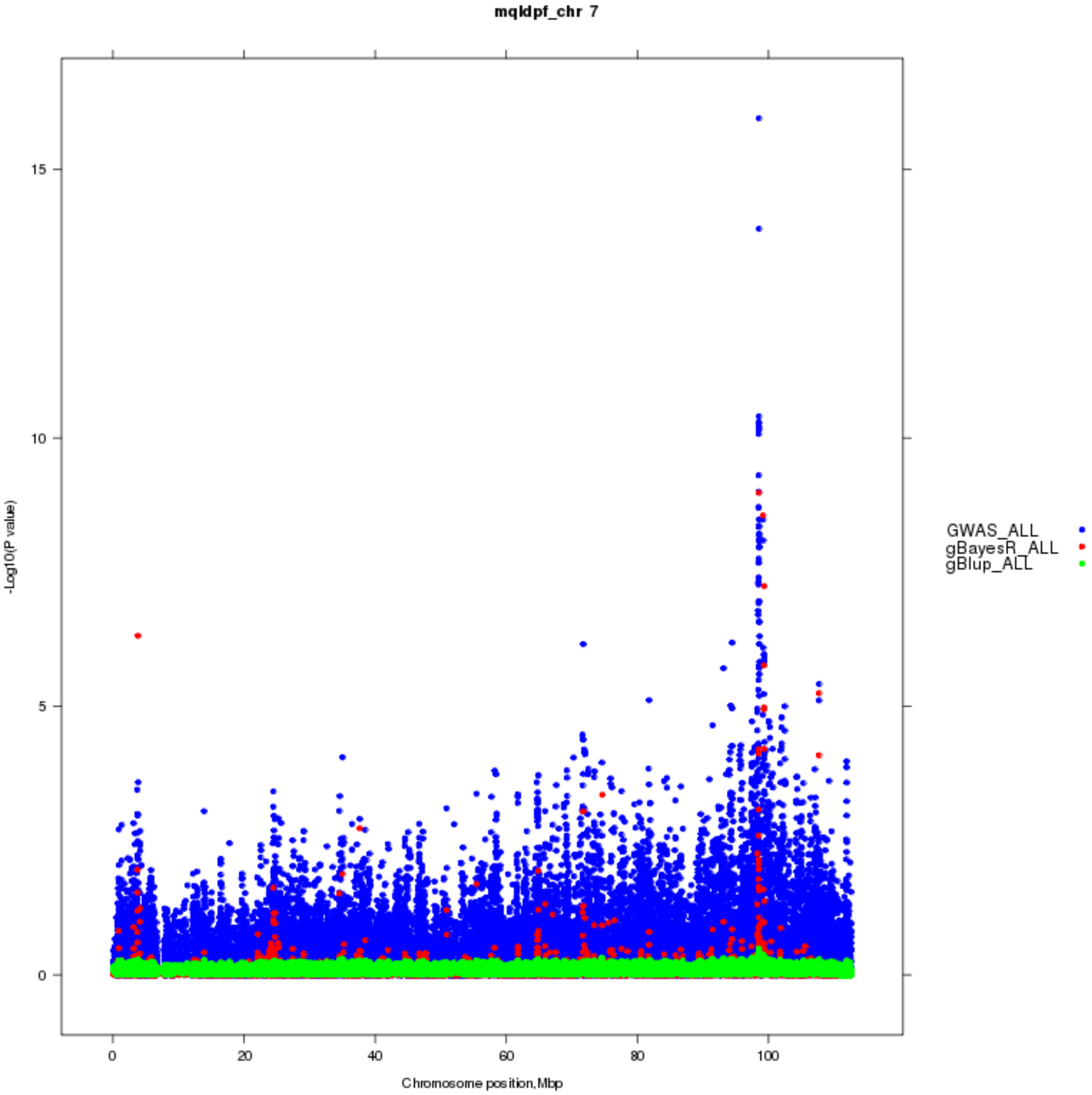
Bayesian estimation

$E(u \mid y) = \int u \, P(y \mid u) \, P(u) / P(y) \, du$

Distribution of SNP effects

Normal → BLUP
t-distribution → Bayes A
Mixture → Bayes B (Meuwissen et al 2001)

Mixture of N → Bayes R (Erbe et al 2012)
$u \sim N(0,\sigma_i^2)$ with probability π_i
$\sigma_i^2 = \{0, 0.0001, 0.001, 0.01\} \, \sigma_g^2$
Accuracy is greatest if assumed distribution matches real distribution.


mqldpf_chr 7
-Log10(P value)
0
5
10
15
0
20
40
60
80
100
Chromosome position,Mbp
GWAS_ALL ●
gBayesR_ALL ●
gBlup_ALL ●

Prediction of genetic value

Other methods of prediction

Estimate effect of each SNP one at a time and add
g-hat = Z u-hat
u-hat estimated from single SNP regression

Biased E(g | g-hat) ≠ g-hat
Less accurate because ignores LD between SNPs
and treats u as fixed effects

Prediction of genetic value

Real data

- 4500 bulls and 12000 cows (Holstein and Jersey)
- 600,000 SNPs genotyped
- Train using bulls born < 2005
- Test using bulls born >= 2005

Correlation of EBV and daughter average

	Protein	Stature	Milk	Fat%
BLUP	0.66	0.52	0.65	0.72
Bayes R	0.66	0.54	0.68	0.82

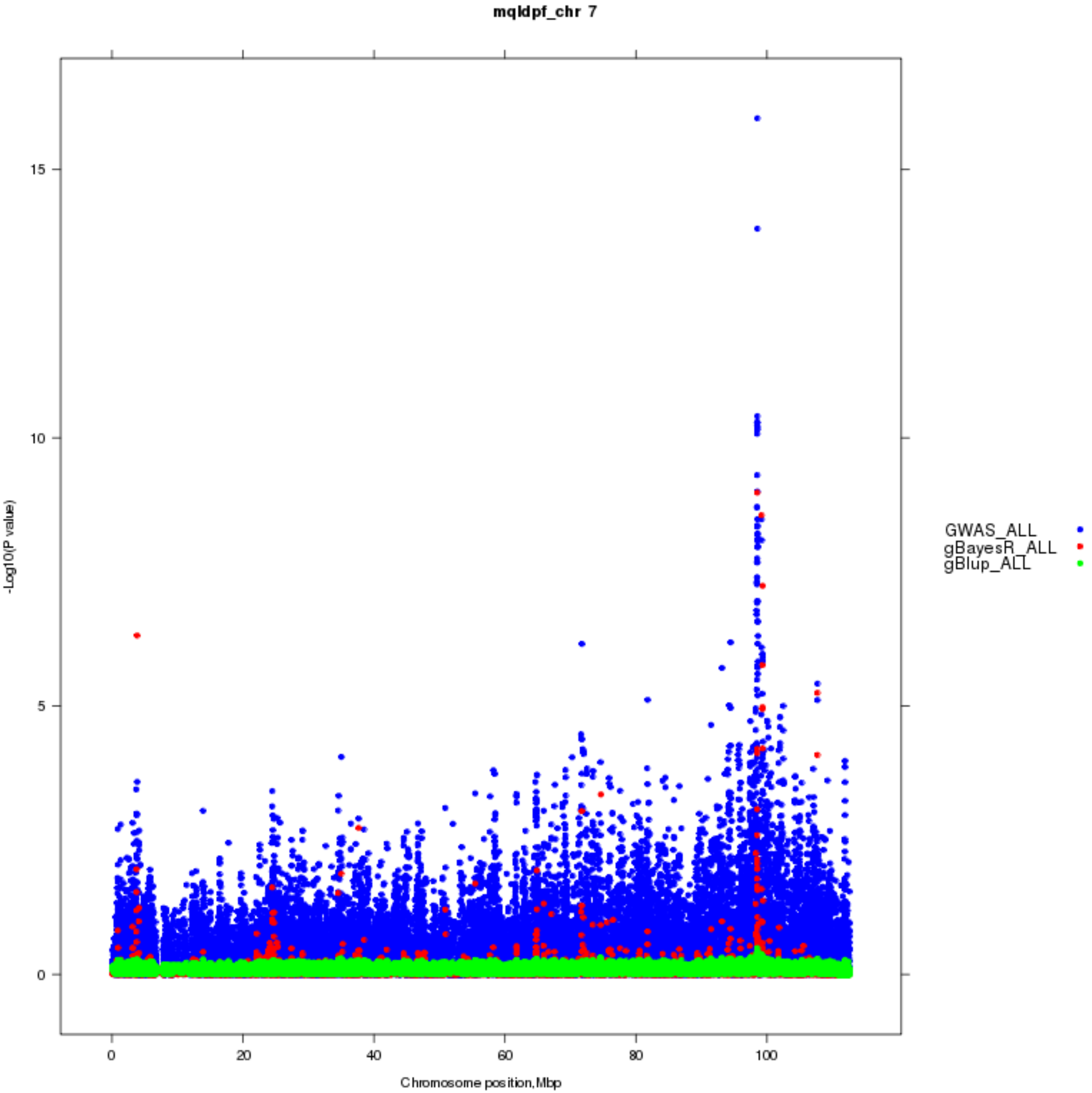
Genetic architecture

Proportion of SNPs from distribution with variance

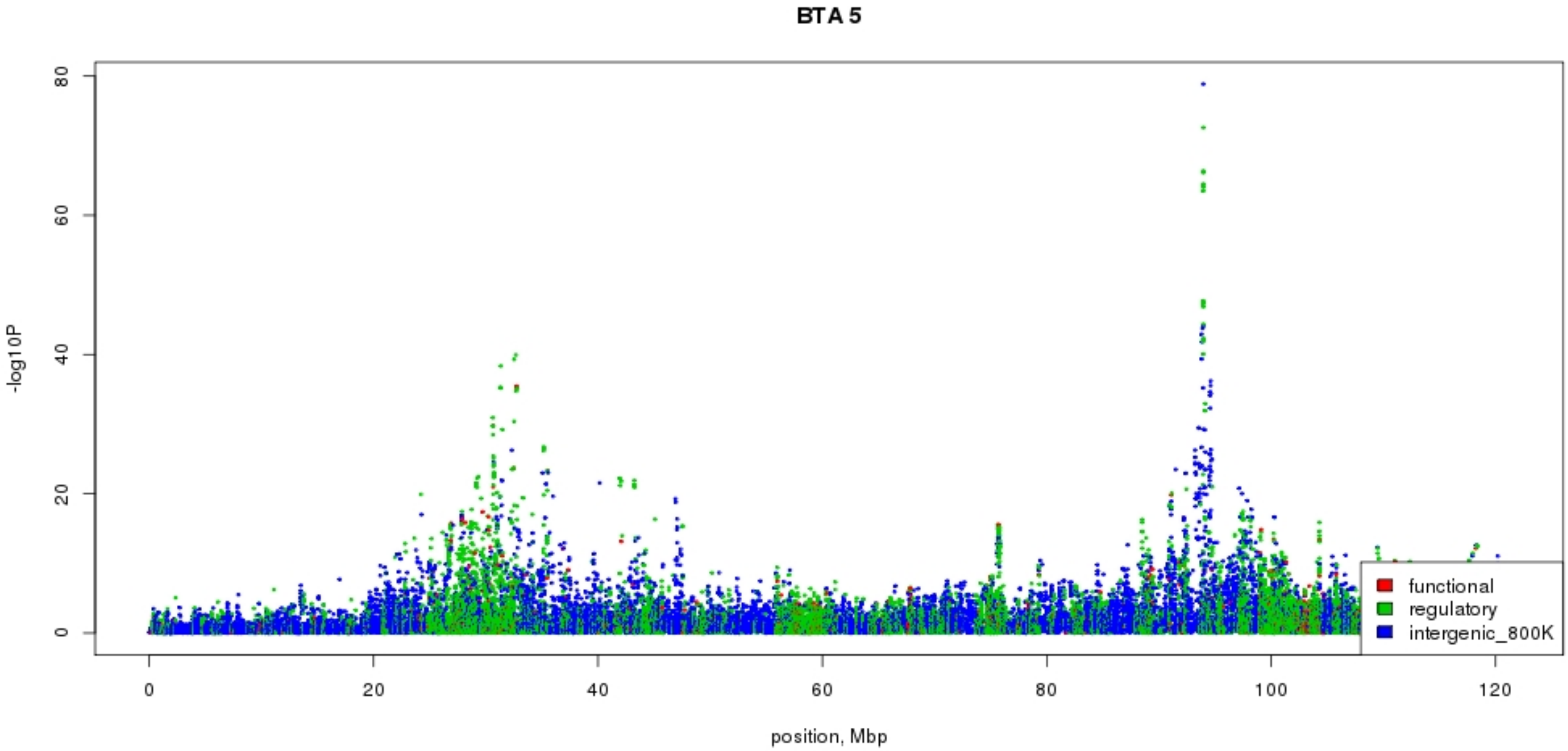
Trait	0.01%	0.1%	1%	polygenic (%)
RFI	7498	296	6	11
LDPF	1419	254	36	27
Mean	4029	271	19	25

Integration of prediction and mapping of causal variants

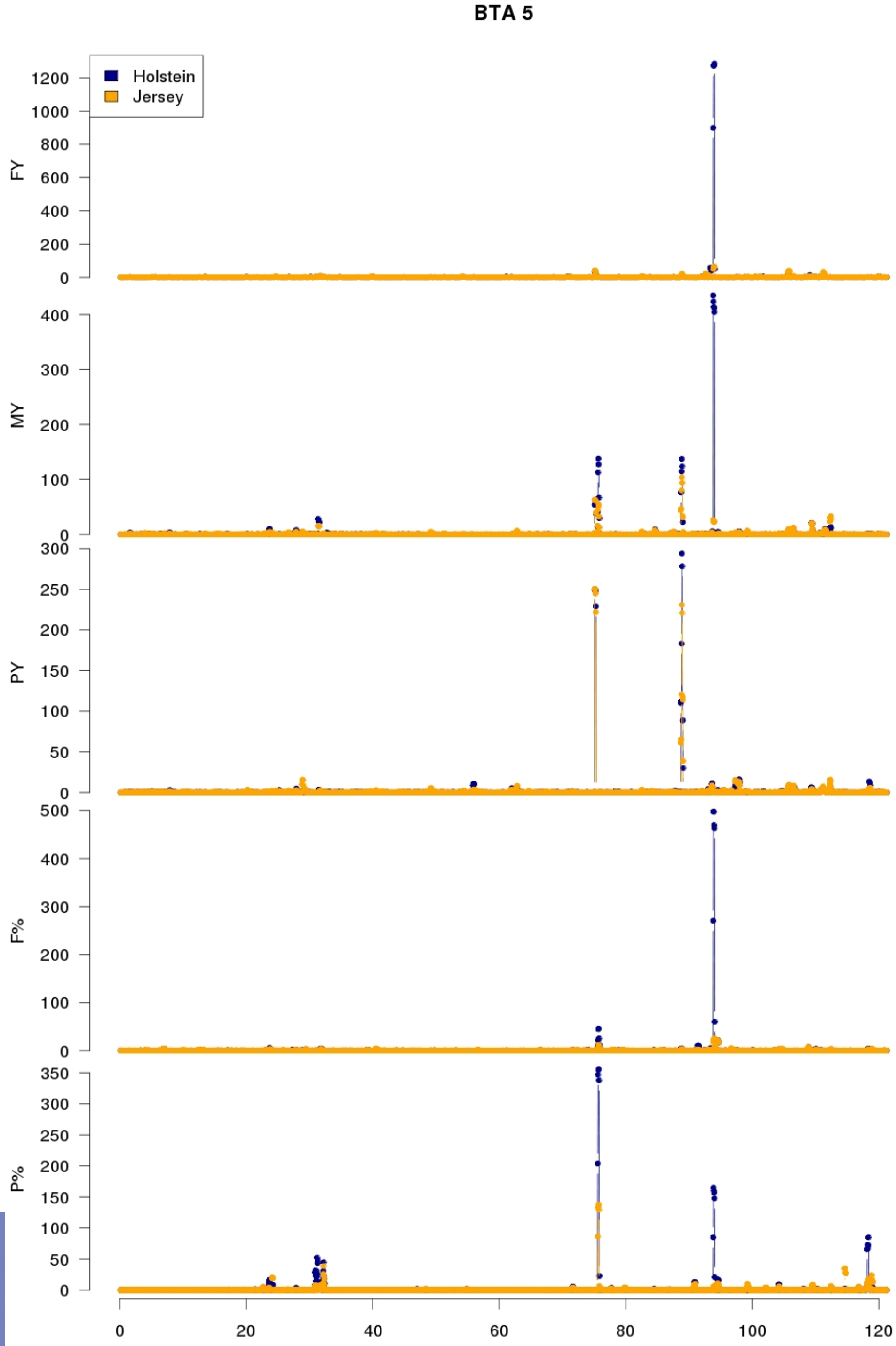
Same Bayesian models as used for prediction can be used for mapping causal variants of complex traits


mqldpf_chr 7
-Log10(P value)
0
5
10
15
0
20
40
60
80
100
Chromosome position,Mbp
GWAS_ALL
gBayesR_ALL
gBlup_ALL

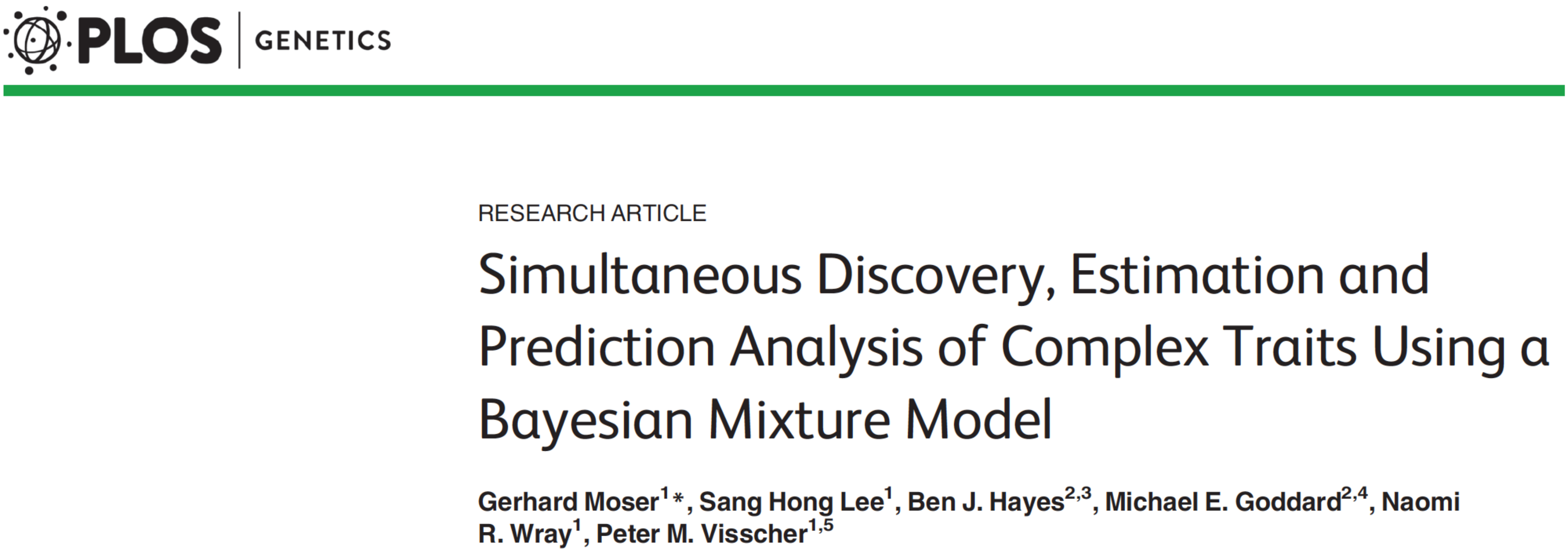
Mapping QTL – Milk on BTA5



Mapping QTL – Milk on BTA5



Application to human disease data (WTCCC)



PLOS | GENETICS

RESEARCH ARTICLE

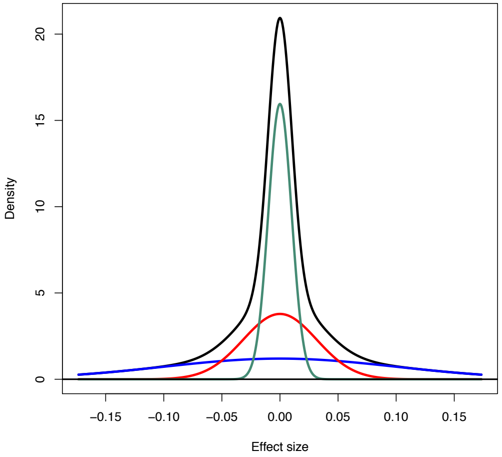
Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

Model

- Assumes true SNP effects are derived from a series of normal distributions
- Prior assumptions
  - Effects size of SNP *k*

$$\sigma_k^2 = \begin{cases} \pi_1 \times N\left(0, 0 \times \sigma_g^2\right) \\ \pi_2 \times N\left(0, 10^{-4} \times \sigma_g^2\right) \\ \pi_3 \times N\left(0, 10^{-3} \times \sigma_g^2\right) \\ \pi_4 \times N\left(0, 10^{-2} \times \sigma_g^2\right) \end{cases}$$



  - Mixing proportion, **π**
    - *Dirichlet* distribution, $p(\pi_1, \dots, \pi_4) \sim D(\delta, \dots, \delta), with\ \delta = 1$

  - Genetic variance
    - hyper-parameter estimated from data, $\sigma_g^2 \sim \chi^{-2}(v_0, S_0^2)$

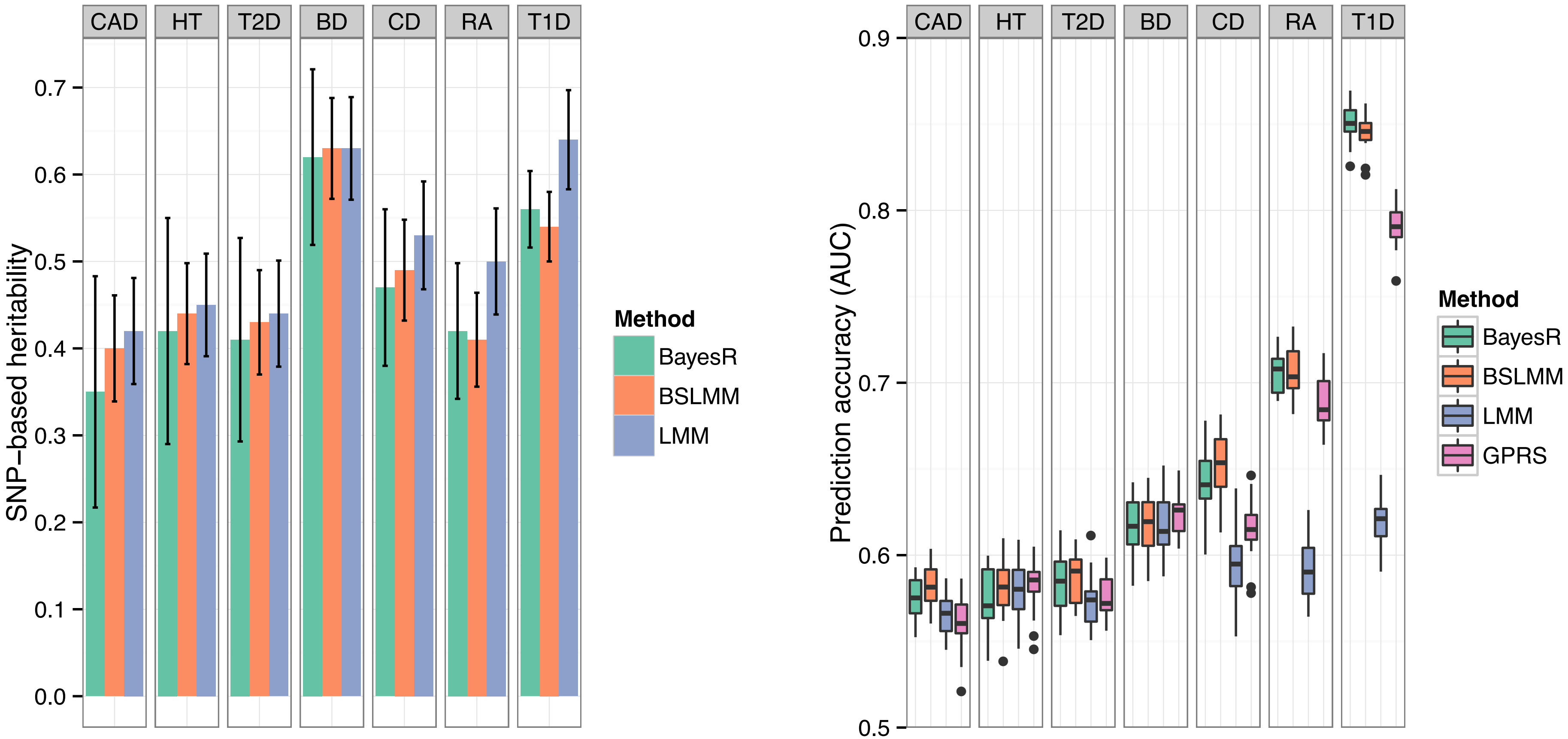
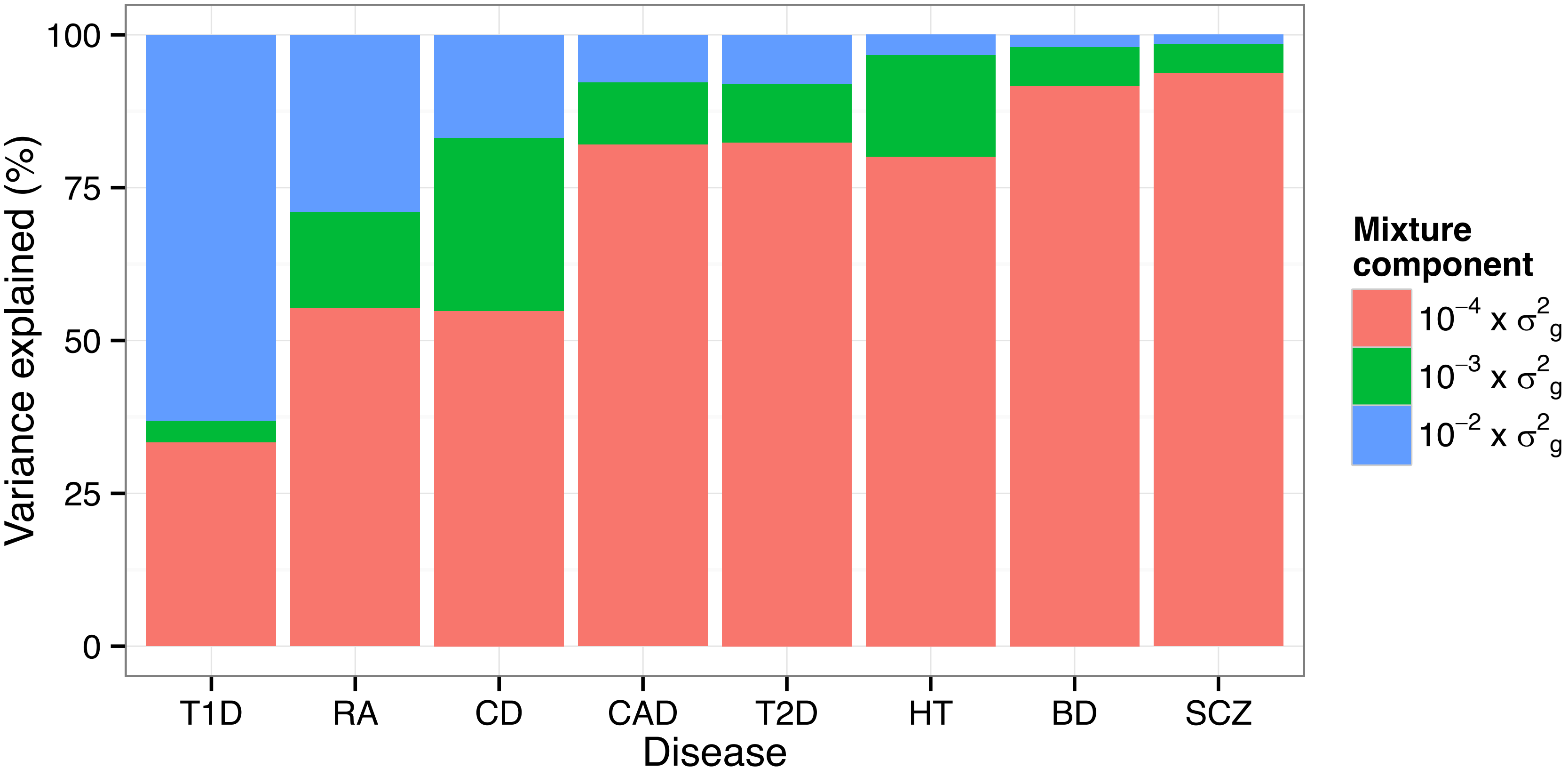


Figure 4. Comparison of performance of BayesR, BSLMM, LMM and GPRS in WTCCC data. (A) Estimates of SNP-based heritability on the observed scale. Antennas are standard deviations of posterior samples for BayesR and BSLMM or standard errors for LMM. GPRS does not provide estimates of heritability. (B) Distribution of the area under the curve (AUC). The single boxplots display the variation in estimates among 20 replicates. In each replicate, the data set was randomly split into a training sample containing 80% of individuals and a validation sample containing the remaining 20%.

Expected proportion of total SNP variance explained by each mixture

(Number of SNPs in class × variance assigned to SNP) / sum of marker variance



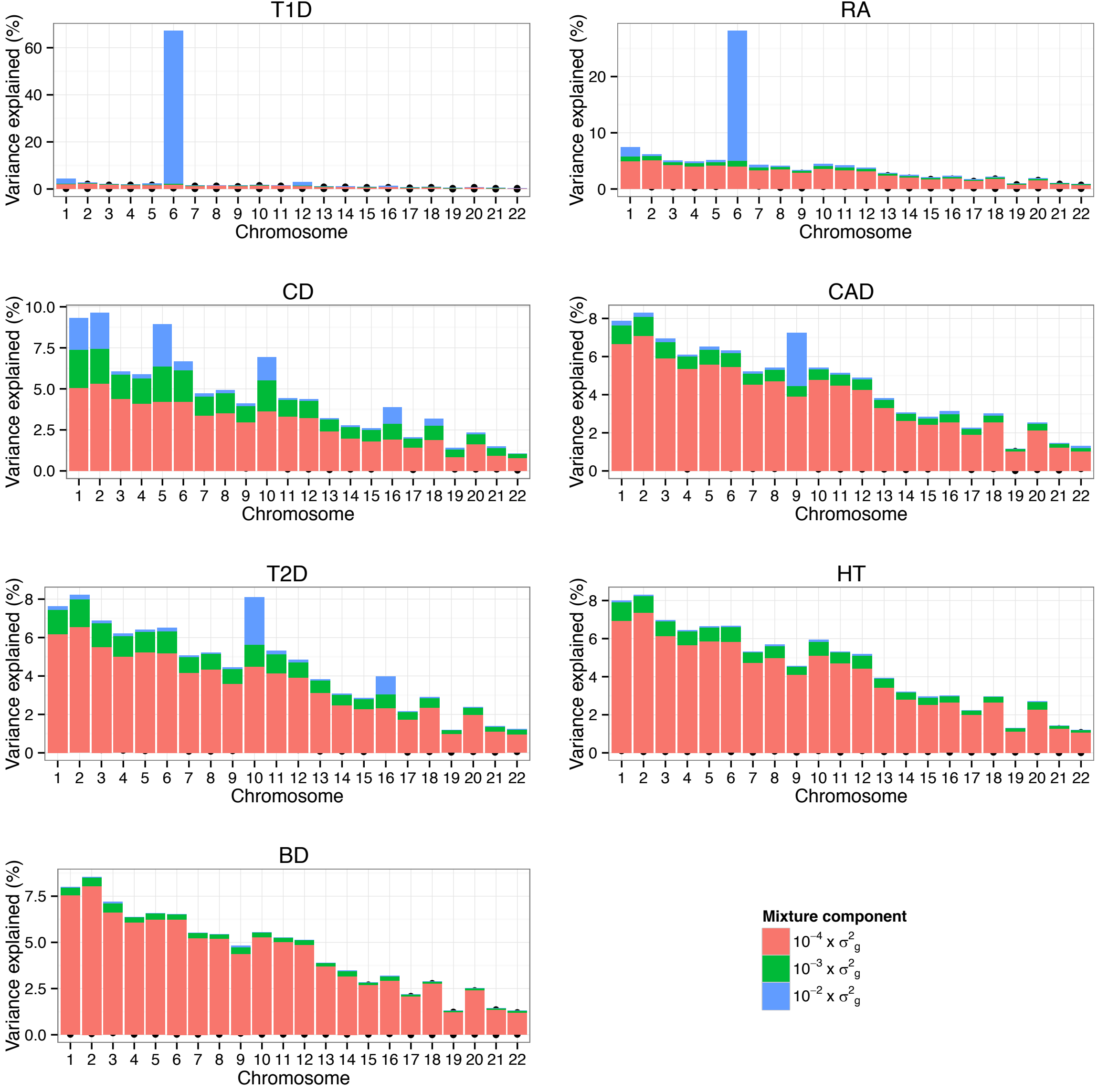
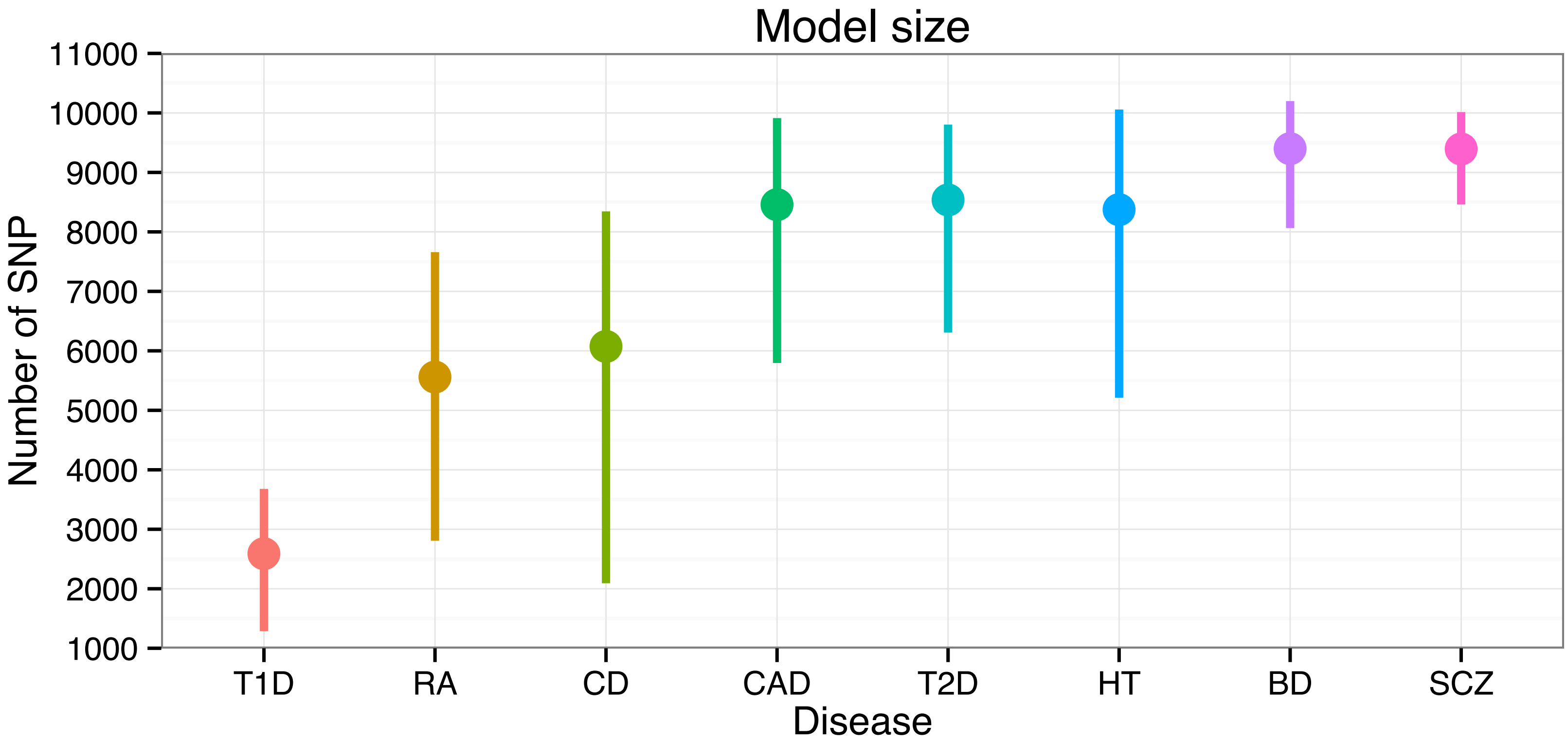


Figure 6. Proportion of genetic variance on each chromosome explained by SNPs with different effect sizes underlying seven traits in WTCCC. Proportion of additive genetic variation contributed by individual chromosomes and the proportion of variance on each chromosome explained by SNPs with different effect sizes. For each chromosome we calculated the proportion of variance in each mixture component as the sum of the square of the sampled effect sizes of the SNPs allocated to each component divided by the sum of the total variance explained by SNPs. The colored bars partition the genetic variance in contributions from each mixture class.

Posterior mean of number of SNPs estimated by BayesR

- Posterior mean and 95% posterior credible interval
- WTCC1+SCZ swedish



Prediction of genetic value Summary

Best prediction is g-hat = E(g | y)

Genetic values treated as random effects

Eg $g \sim N(0, G\sigma_g^2)$

Equivalent model to predict SNP effects u

E(u | y) depends on prior distribution of u

→ Bayesian models

g-hat = Z u-hat gives higher accuracy than assuming

$g \sim N(0, G\sigma_g^2)$

Bayesian models integrate prediction and mapping of causal variants

Key concepts

- Prediction of phenotypic values is limited by heritability
- Accuracy of prediction depends on
  - how well marker effects are estimated (sample size)
  - how well marker effects are correlated with causal variants (LD)
- Estimation of marker effects and prediction in the same data leads to (severe) bias
- Variance explained by a SNP-based predictor is not the same as the variance explained by those SNPs
- Best prediction methods take genetic values as random effects

Supplementary derivations

Theory (additive model)
m unlinked causal variants

$$y_i = \sum_{j=1}^{m} x_{ij} b_j + e_i = a_i + e_i$$

$$\mathrm{var}(y) = \sum_{j=1}^{m} \mathrm{var}(x_j) b_j^2 + \mathrm{var}(e) = \mathrm{var}(a) + \mathrm{var}(e)$$

$$\mathrm{cov}(y_i, y_k) = \sum_{j=1}^{m} \mathrm{cov}(x_{ij}, x_{kj}) b_j^2 + \mathrm{cov}(e_i, e_k)$$

$$= \mathrm{cov}(a_i, a_k) + \mathrm{cov}(e_i, e_k)$$

$$= \mathrm{cov}(a_i, a_k) \text{ if } \mathrm{cov}(e_i, e_k) = 0$$

Prediction

$$\hat{y}_i = \sum_{j=1}^{m} x_{ij}\hat{b}_j = \hat{a}_i$$

$$\mathrm{var}(\hat{y}) = \sum_{j=1}^{m} \mathrm{var}(x_j)\hat{b}_j^2 = \mathrm{var}(\hat{a})$$

$$\mathrm{cov}(\hat{y}_i, \hat{y}_k) = \sum_{j=1}^{m} \mathrm{cov}(x_{ij}, x_{kj})\hat{b}_j^2 = \mathrm{cov}(\hat{a}_i, \hat{a}_k)$$

- theory -

$$\operatorname{cov}(\hat{y}_i, y_i) = \operatorname{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij}b_j + e_i\}$$

$$= \sum_{j=1}^{m} \operatorname{var}(x_{ij})\hat{b}_j b_j + \sum_{j=1}^{m} x_{ij} \operatorname{cov}(\hat{b}_j, e_i)$$

If *b* estimated from the same data in which prediction is made, then the second term is non-zero

Effect of errors in estimating SNP effects (least squares; single SNP)

$$y_i = x_i b + e_i$$

$$\hat{b} = b + \varepsilon$$

$$E(\hat{b}) = b$$

$$\mathrm{var}(\hat{b}) = \mathrm{var}(\varepsilon) = \sigma_e^2 / \Sigma x^2 \approx \mathrm{var}(y) / \{N\,\mathrm{var}(x)\}$$

$$\mathrm{var}(x) = 2p(1-p) \text{ under HWE}$$

$$\text{Define } \mathrm{R}_{SNP}^2 = \mathrm{var}(x) b^2 / \mathrm{var}(y)$$

= contribution of single SNP to heritability

- effects of errors -

$$\hat{R}^2_{y,\hat{y}} = \operatorname{cov}(y,\hat{y})^2 / \{\operatorname{var}(y)\operatorname{var}(\hat{y})\}$$

$$E[\operatorname{cov}(y,\hat{y})] = E[\operatorname{cov}(xb, x\hat{b})] = \operatorname{var}(x_i)E(\hat{b})b$$

$$= \operatorname{var}(x)b^2$$

$$E[\operatorname{var}(\hat{y})] = E[\operatorname{var}(x\hat{b})] = \operatorname{var}(x)E[\hat{b}^2]$$

$$= \operatorname{var}(x)[b^2 + \operatorname{var}(\hat{b})] \approx \operatorname{var}(x)b^2 + \operatorname{var}(x)\operatorname{var}(y)/[N\operatorname{var}(x)]$$

$$= \operatorname{var}(x)b^2 + \operatorname{var}(y)/N$$

$$E(\hat{R}^2_{y,\hat{y}}) \approx R^2_{SNP} / [1 + 1/\{NR^2_{SNP}\}]$$