LDpred2: better, faster, stronger

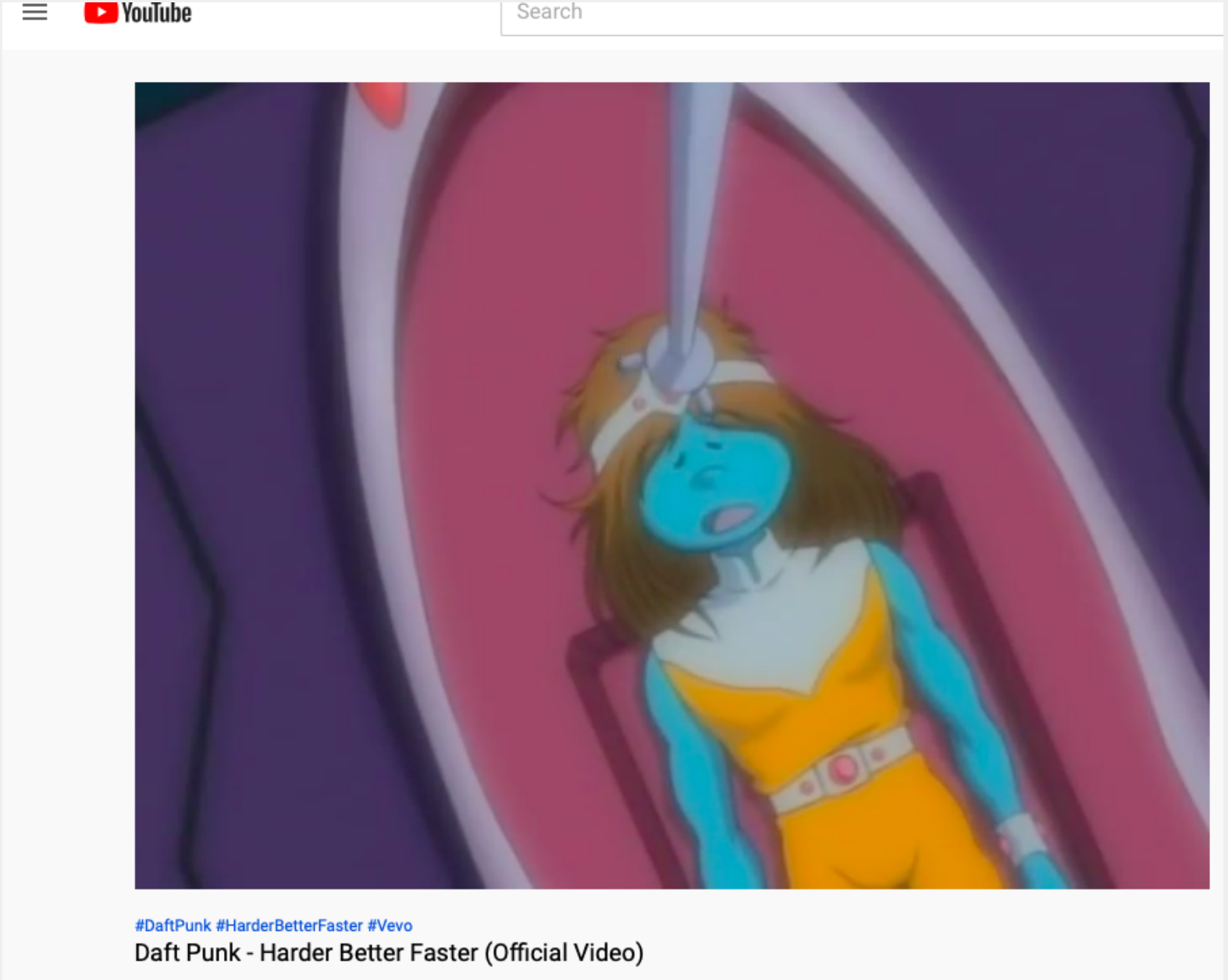
Privé, Arbel, Vilhjálmsson;BiorXiv 2020

Discussion led by: Bárbara Bitarello

02 - 07 - 2020

Background: LDpred

Important background



What does LDpred do?

Published in 2015: $\sim$ 435 citations (less than PRSice from same year)

- matrix of correlation between genetic variants (LD matrix), summary statistics from GWAS (β, $p - value$), genotype and phenotype files from test and validation sets

Infinitesimal:

- All markers are causal
- Effect sizes drawn from Gaussian
- Computationally efficient
- Not very plausible

Non-infinitesimal

- Assumes p of variants are causal - more plausible
- Analytical solution hard - approximate MCMC Gibbs sampler (not efficient nor robust)

What does LDpred do?

But *actually* also requires:

- big LD reference panel, correct model specifications - not trivial
- Steps:
  - Coordinating summary stats, LD reference genotypes, validation or test genotypes
  - Estimating weights for variants - which requires additional parameters.
  - Calculating PRS
  - User needs to calculate partial-R^2 on their own (e.g. in R)

LDpred model - Overview

LDpred uses a Bayesian framework to assing effect sizes from provided summary statistics and LD information

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \tag{1}$$

LDpred model - Overview

Unlinked markers and non-infinitesimal architecture
Effects are drawn from a mixture distribution:

$$\beta_j \sim \begin{cases} N(0, \frac{h^2}{Mp}), & \text{with probability } p. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

-LDpred1: h^2 estimated with constrained LD score regression (fixed intercpept=1)
-Gibbs sampler algorithm:

Gibbs sampler algorithm: main steps

1. residualized effect sizes for each variant j : $\tilde{\beta}_j$
2. probability that variant j is causal: $\bar{p}_j$
3. β_j is sampled according to:

$$\beta_j|\tilde{\beta}_j \sim \begin{cases} N(\frac{1}{1+\frac{Mp}{nh^2}}\tilde{\beta}_j, \frac{1}{1+\frac{Mp}{nh^2}}\frac{1}{n}), & \text{with probability } p. \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

4. posterior mean of $\beta_j|\tilde{\beta}_j$: ω_j

Gibbs sampler algorithm: main steps

Algorithm 1 LDpred, with hyper-parameters p and h^2, LD matrix $\boldsymbol{R}$ and summary statistics $\hat{\gamma}$, se$(\hat{\gamma})$ a

$\hat{\boldsymbol{\beta}} \leftarrow \dfrac{\hat{\gamma}}{\text{se}(\hat{\gamma}) \cdot \sqrt{n}}$ ▷ Initialization of scaled marginal effects (see previous se

$\boldsymbol{\Omega} \leftarrow \mathbf{0}$ ▷ Initialization of posterior

for $k = 1, \ldots, N_{\text{burn-in}} + N_{\text{iter}}$ **do** ▷ Gibbs iter

  for each variant j **do** ▷ All va

    Compute $\tilde{\beta}_j$ according to (3)

    Compute $\bar{p}_j$ according to (4)

    Sample β_j according to (5)

    Compute ω_j according to (6)

  end for

  if $k > N_{\text{burn-in}}$ **then**

    $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} + \boldsymbol{\omega}$

  end if

end for

$\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega}/N_{\text{iter}}$ ▷ Average of all $\boldsymbol{\omega}$ after b

Return $\boldsymbol{\Omega} \cdot \text{se}(\hat{\gamma}) \cdot \sqrt{n}$ ▷ Return posterior means, scaled back (see previous se

LDpred: pros and cons overview

PROS:

- elegant modelling of genetic architecture
- assigns weights to variants instead of arbitrary P+T
- also offers P+T in the same framework
- mostly runs PLINK in the background, and Python scripts

CONS:

- Errors messages are cryptic
- Slow
- Gibbs sampler extremely sensitive to model parameters
- particularly bad for long-range LD regions (e.g HLA)
- MCMC setup might or not improve things and makes it it *much slower*
- No manual available.

New method: LDpred2

LDpred2: what's new?

- Runs in bigsnpr package in R.
- LDpred-auto: learns parameters from the data. Stronger
- More accurate PRS: simulation and real data benchmarking
- Compares favorably to LDpred 1 and other methods [sort of]
- parallelization in C++ - FASTER
- has tutorial!! - Better `https://privefl.github.io/bigsnpr/articles/LDpred2.html`

Simulations: methods

Binary phenotypes; each set 10X (average AUC is reported)

- UKBB data
- unrelated individuals - 360K
  - 10,000 for validation, LD reference
  - 300,000 for GWAS
  - $\sim$ 52,000 as test set
- HapMap3 variants - 1.1 Million
- $h^2 = 0.4$ or $h^2 = 0.3$, prevalence 15%
- $M = \{300, 3000, 30000, 300000\}$
- Variance of genetic liability=h^2
- HLA region
- Implemented in bigsnpr

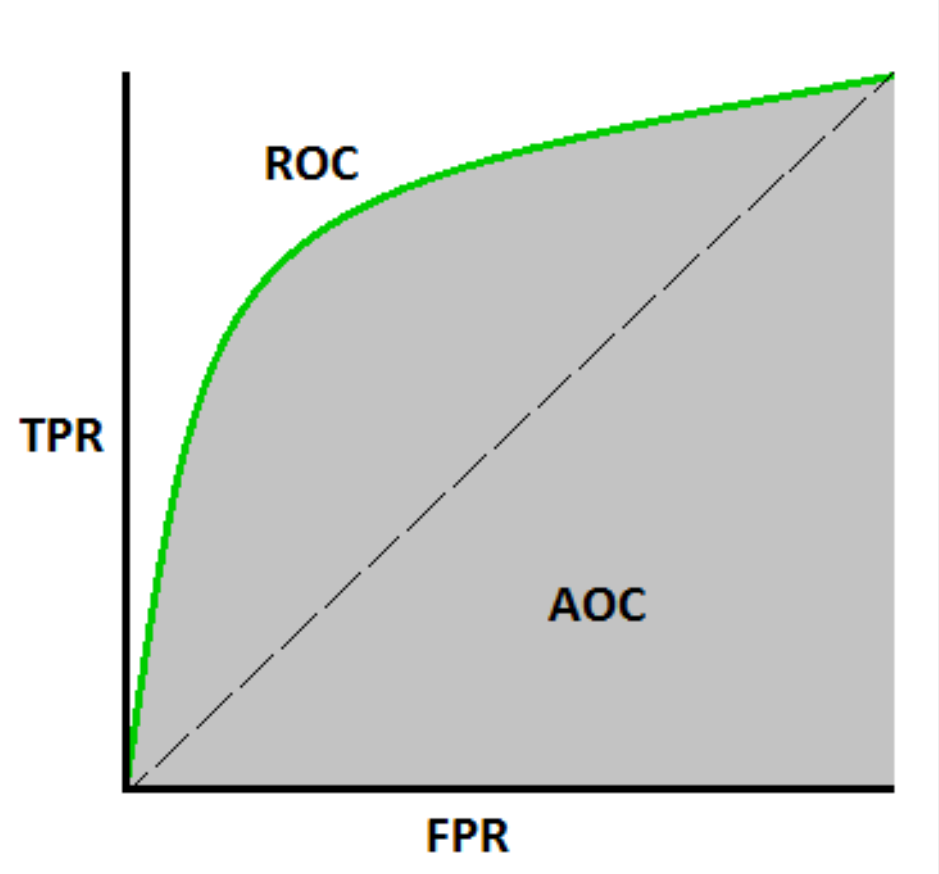
Real data: methods

- Unrelated individuals - 360K
- All case-control phenotypes
- 10,000 for validation, LD reference
- $\sim$ 352,000 as test set
- Compare LDpred1, LDpred2, C+T, SCT, lassosum, PRS-CS
- summary statistics:

Trait	GWAS citation	GWAS sample size	GWAS #variants
Breast cancer (BRCA)	Michailidou *et al.* (2017)	137,045 / 119,078	11,792,542
Rheumatoid arthritis (RA)	Okada *et al.* (2014)	29,880 / 73,758	9,739,303
Type 1 diabetes (T1D)	Censin *et al.* (2017)	5913 / 8828	8,996,866
Type 2 diabetes (T2D)	Scott *et al.* (2017)	26,676 / 132,532	12,056,346
Prostate cancer (PRCA)	Schumacher *et al.* (2018)	79,148 / 61,106	20,370,946
Depression (MDD)	Wray *et al.* (2018)	59,851 / 113,154	13,554,550
Coronary artery disease (CAD)	Nikpay *et al.* (2015)	60,801 / 123,504	9,455,778
Asthma	Demenais *et al.* (2018)	19,954 / 107,715	2,001,280

Table 1: Summary of external GWAS summary statistics used. The GWAS sample size is the number of cases / controls in the GWAS.

Methods: performance comparisons



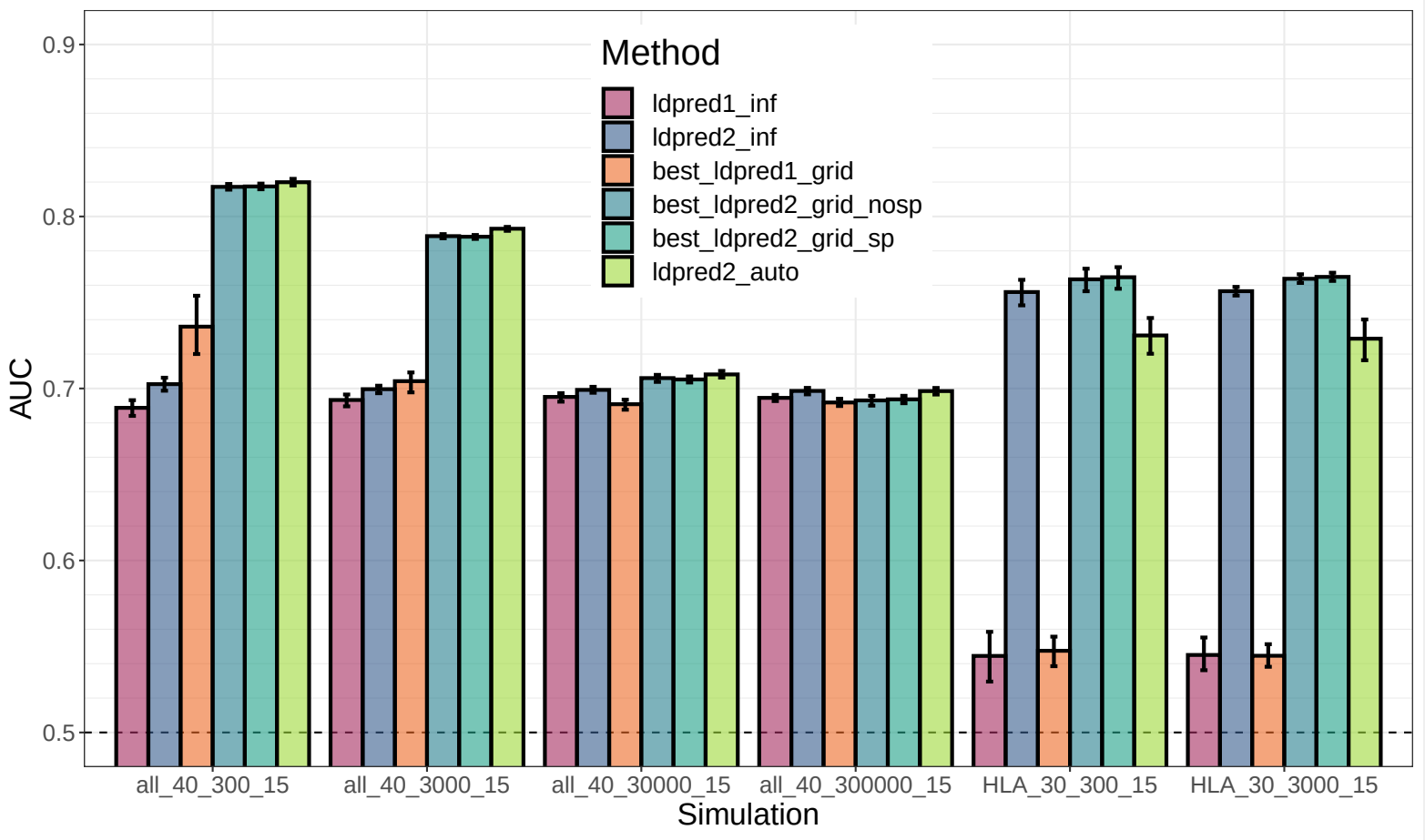
$$TPR = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

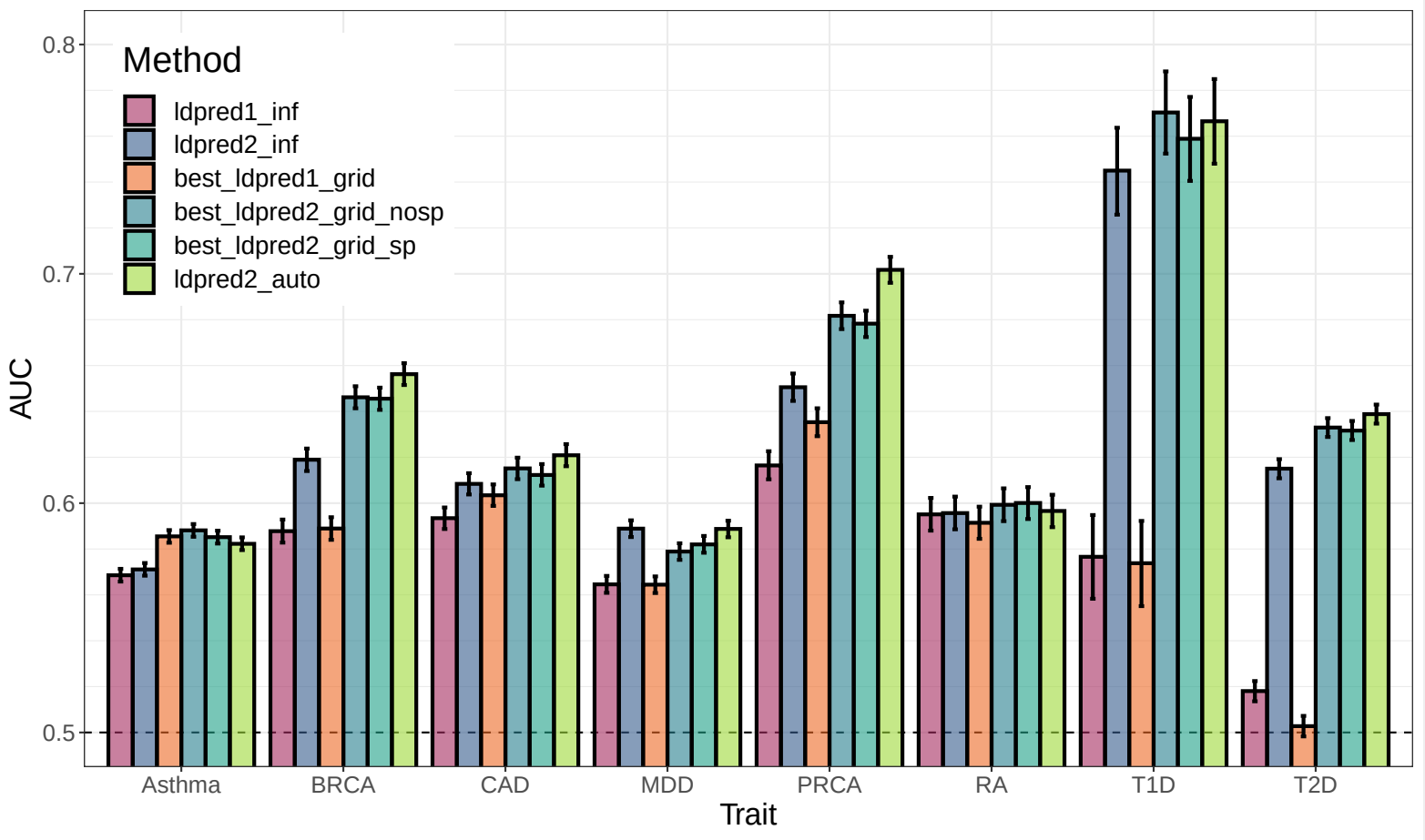
$$FPR = 1 - Specifcity$$

Image: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

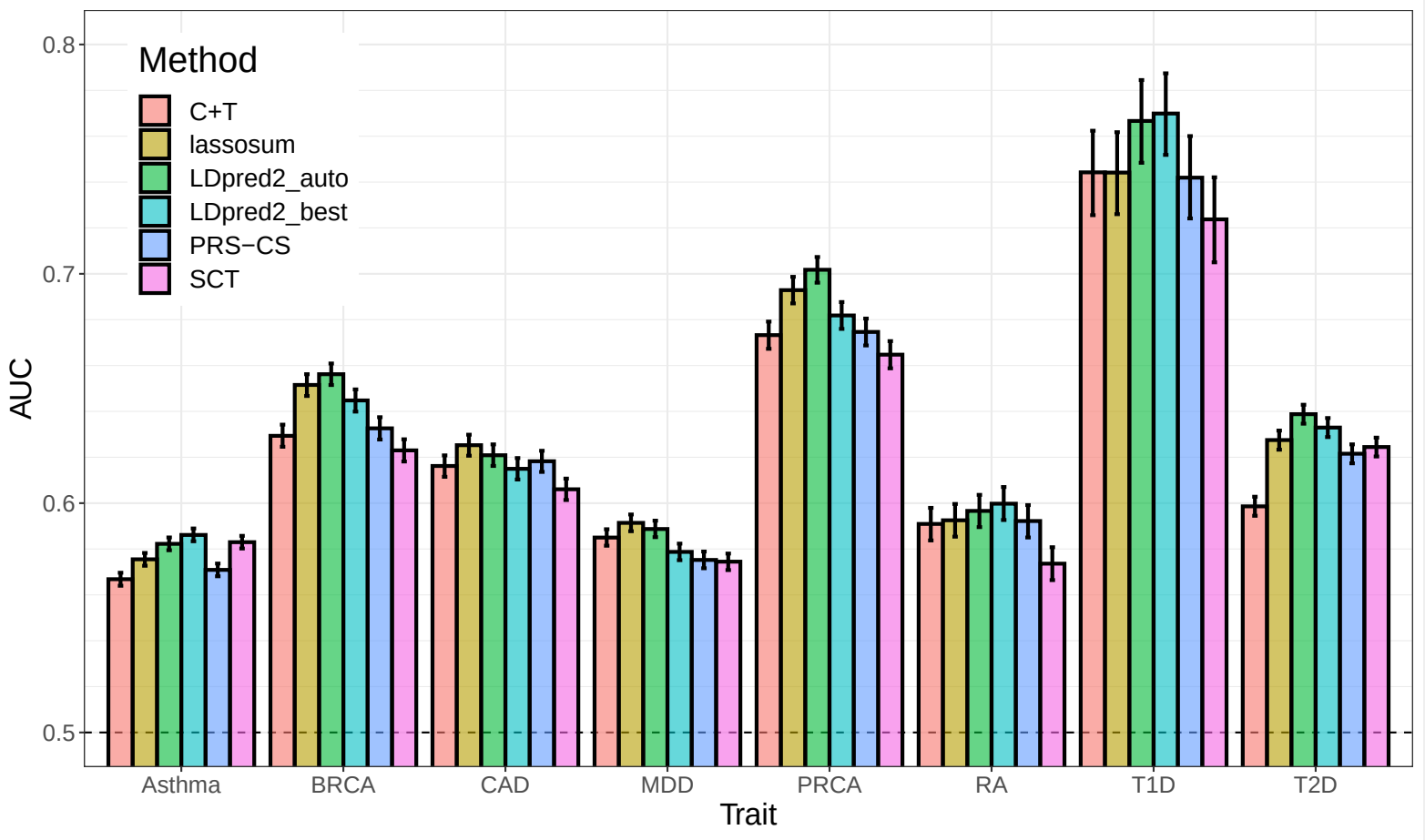
Simulations: Results



Real Data: Results



Real Data: Results



Conclusions

- Strengths: long-range LD and less polygenic traits, does not require validation step
- solves gibbs sampler inconsistencies
- higher prediction accuracy then LDpred1
- Use HapMap3 variants

- Not really better than lassosum?
- Still kinda slow

QC for LDpred2-auto

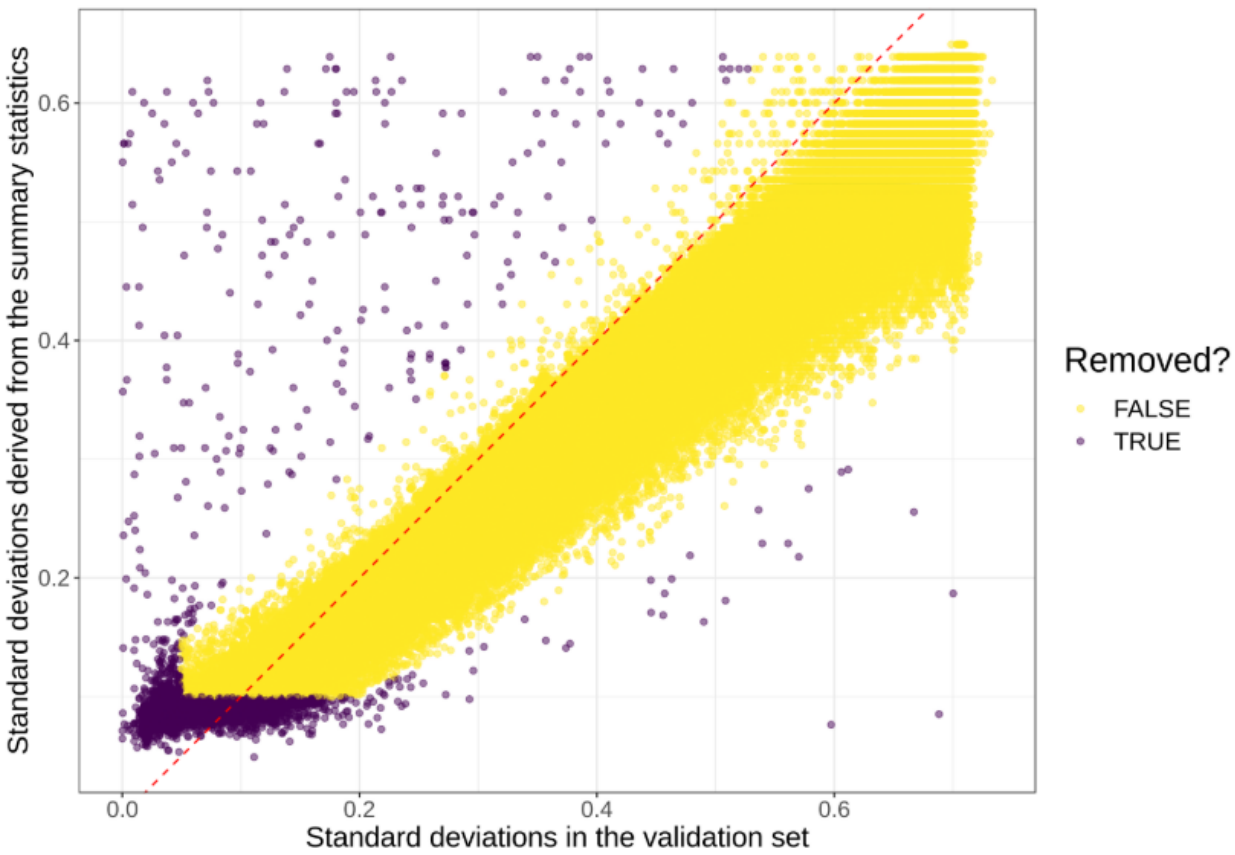


Figure S2: Standard deviations derived from summary statistics of breast cancer based on equation (S1) versus the standard deviations of genotypes of individuals in the validation set. Coloring shows the quality control applied in this paper.