

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Doug Speed

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 6. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Permission request regarding CAGI 6 recordings and hybrid presentations

I understand that my presentation as associated with CAGI 6 will be recorded and broadcasted via the internet through CAGI's designated technology provider.

As the copyright holder to this presentation entitled [*MegaPRS Prediction of Complex Traits*], I hereby agree that the copyright for the online broadcast and recording of my presentation be released to CAGI and my recording be uploaded on to the CAGI website.

Doug Speed

MegaPRS Prediction of Complex Traits

Aarhus University, Denmark

Polygenic Risk Scores

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

where X_j contains genotypes for SNP j and β_j is its effect size

There are many tools for constructing PRS

each tries to obtain the most accurate estimates of β_j

The best tools estimate β_j jointly

these require prior distributions for β_j (because $m \geq n$)

For example, a simple solution is to assume $\beta_j \sim N(0, \sigma^2)$

this prior is used by BLUP, ridge regression, etc

Advantage of MegaPRS 1

Most PRS tools assume that all SNPs are expected to contribute equally
i.e., that $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2]$ is constant across the genome

E.g., BLUP assumes $\beta_j \sim N(0, \sigma^2)$
and therefore $\mathbb{E}[h_j^2] = \mathbb{E}[\beta_j^2] = \sigma^2$

MegaPRS relaxes this assumption
it allows the user to specify $\mathbb{E}[h_j^2]$ (the “heritability model”)

E.g., the MegaPRS version of BLUP assumes $\beta_j \sim N(0, q_j \sigma^2)$
where the q_j are set according to the chosen heritability model

Advantage of MegaPRS 2

MegaPRS allows for a wide variety of prior forms

MegaPRS BLUP uses $\beta_j \sim N(0, q_j\sigma^2)$

MegaPRS Lasso assumes $\beta_j \sim DE(0, \lambda/q_j)$

MegaPRS Bolt-LMM assumes $\beta_j \sim pN(0, q_j\sigma_1^2) + (1 - p)N(0, q_j\sigma_2^2)$

MegaPRS BayesR uses a mixture of three normals and a point mass

Thus MegaPRS includes generalized versions of most existing PRS tools
e.g., SBLUP, lassosum, LDpred and SBayesR

Advantages of MegaPRS 3, 4 & 5

MegaPRS does not require a validation dataset

i.e., can construct a PRS from just one set of summary statistics
does this using “pseudo summary statistics”

MegaPRS is computationally efficient

estimates β_j using Variational Bayes (not MCMC)
takes one hour to construct PRS for $m \approx 1\text{ M}$ SNPs

MegaPRS is robust

not necessary to remove high-LD regions
no issues with convergence

Performance on UK Biobank Data

I tested MegaPRS using data from UK Biobank
used summary statistics for 225 phenotypes
600 k SNPs, average sample size 280 k individuals

Considered three heritability models

$\mathbb{E}[h_j^2]$ is constant (assumption of existing tools)

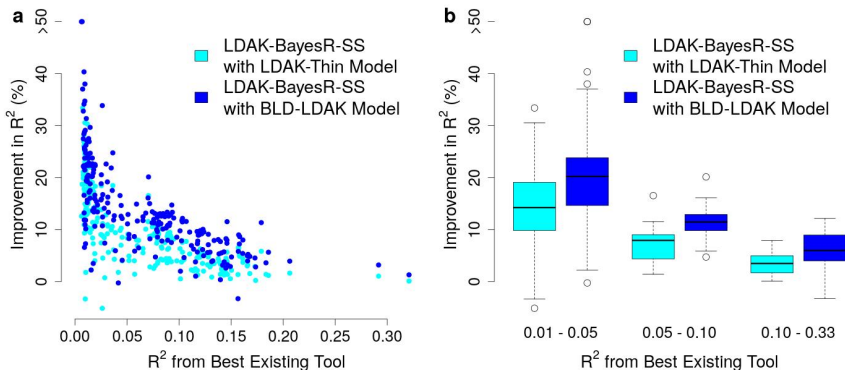
LDAK-Thin Model ($\mathbb{E}[h_j^2]$ depends on MAF and LD)

BLD-LDAK Model ($\mathbb{E}[h_j^2]$ depends on MAF, LD and function)

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

Performance on UK Biobank Data

MegaPRS out-performed existing tools for 223 out of 225 phenotypes



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

equivalent to increasing sample size by a quarter

MegaPRS Settings for CAGI Challenge

Assumed the BLD-LDAK heritability model

$\mathbb{E}[h_j^2]$ depends on MAF, LD and functional annotations

Used the BayesR prior distribution

$$\beta_j \sim \pi_1 N(0, q_j \sigma^2) + \pi_2 N(0, q_j \sigma^2 / 10) + \pi_3 N(0, q_j \sigma^2 / 100) + \pi_4 \delta\{0\}$$

MegaPRS requires an ancestrally-matched reference panel

I used 2000 unrelated, white British individuals from UK Biobank

Final Slide

MegaPRS is available within LDAK (www.ldak.org)

For full details, see “Improved genetic prediction of complex traits from individual-level data or summary statistics” (Nature Comms 2021)

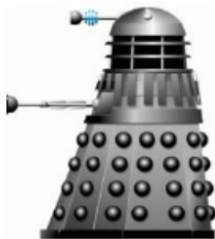
Funding from:

Independent Research Fund Denmark

EU Horizon 2020

Aarhus University Research Foundation

Lundbeck Foundation



www.ldak.org