Progress in complex phenotype prediction: the CAGI6 Polygenic Risk Score challenge

Sung Chun, PhD

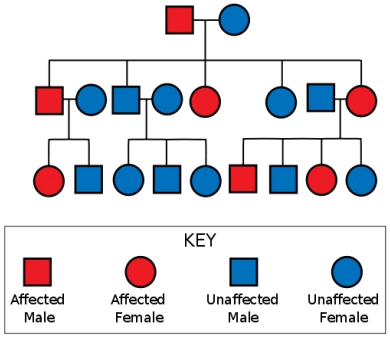
Scientist

Boston Children's Hospital, Harvard Medical School

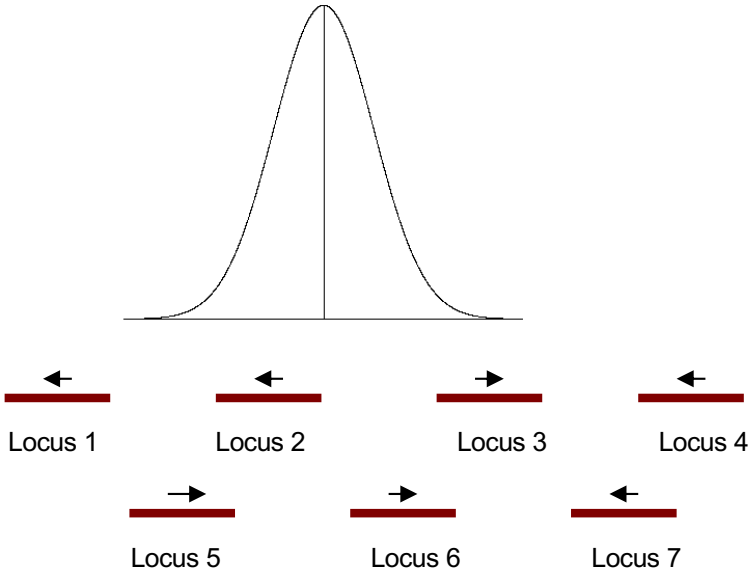
October 25, 2022

Complex phenotypes

Inheritance at each locus is Mendelian. Loci are independent

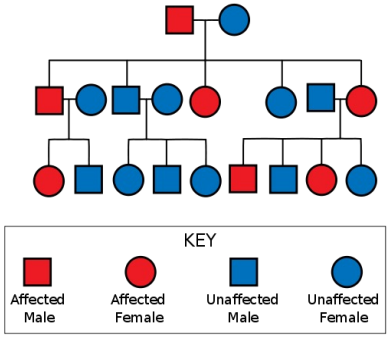


Phenotype is additive over locus effects -> normal distribution

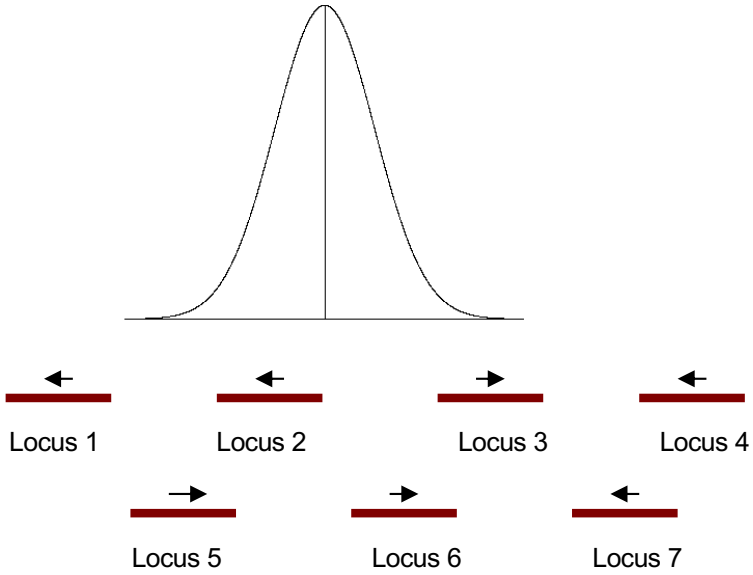


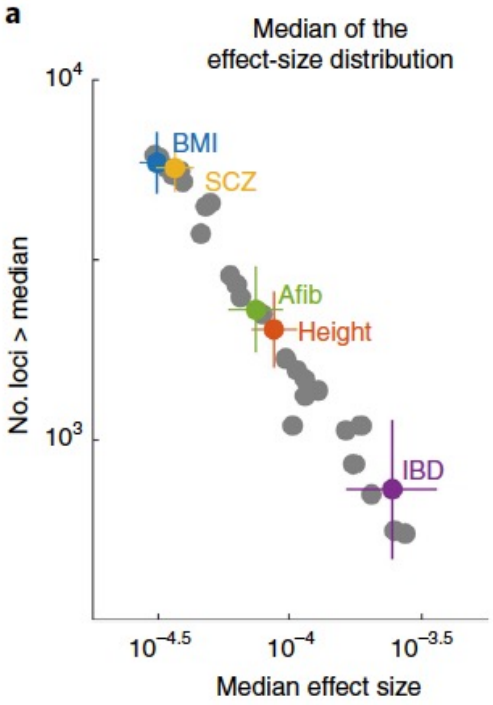
Complex phenotypes

Inheritance at each locus is Mendelian. Loci are independent



Phenotype is additive over locus effects -> normal distribution

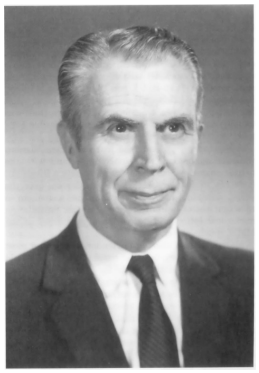




O'Connor 2021

Complex phenotype prediction

BLUP: Best Linear Unbiased Predictor
(Random Effect Model)

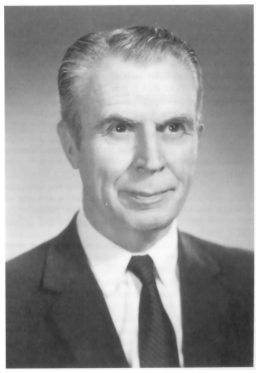


Charles Handerson
(1911 – 1989)

- Infinitesimal model
- All SNPs are causal
- Genetic effects are random
- Predict the expected genetic effect

Complex phenotype prediction

BLUP: Best Linear Unbiased Predictor
(Random Effect Model)



Charles Handerson
(1911 – 1989)

- Infinitesimal model
- All SNPs are causal
- Genetic effects are random
- Predict the expected genetic effect

GENOME RESEARCH

Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray, Michael E. Goddard and Peter M. Visscher

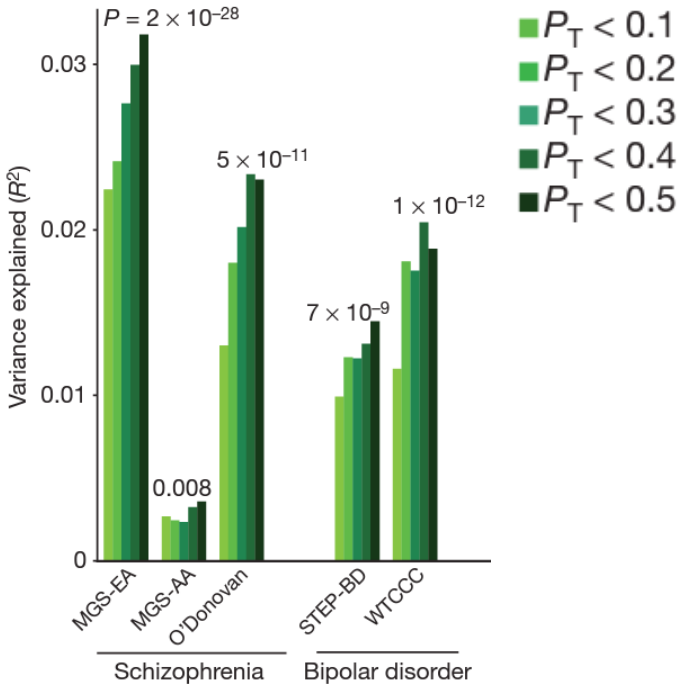
Genome Res. 2007 17: 1520-1528; originally published online Sep 4, 2007;
Access the most recent version at doi:10.1101/gr.6665407

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Phenotype prediction by thresholding on p-values of GWAS association statistics



Pruning and Thresholding

Fixed Effect Model

$$y_i = \sum_j \beta_j \times x_{ij}$$

Genetic risk of individual i

Genotype of SNP j and individual i

Effect size of SNP j

LETTERS

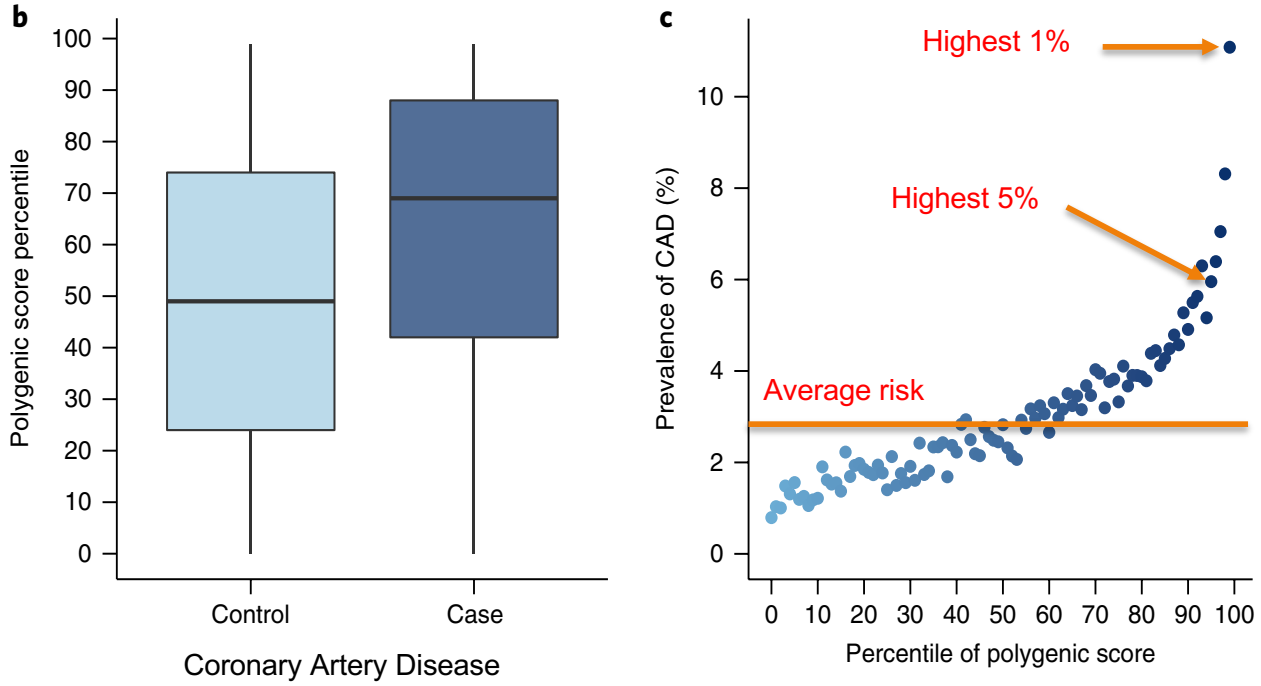
Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Polygenic Risk Scores (PRS)

- Summary statistics are easily available
- Many methods require a separate small individual level dataset to tune parameters
- Methods: BSLMM, LDPred, lassosum, PRScs, BayesR, MegaPRS, AnnoPred, NPS, etc

Extreme tails of PRS are highly predictive



Khera et al. 2018

CAGI6 Polygenic Risk Prediction challenge

- Evaluates the accuracy of statistical methods
- New area for CAGI
- Massive protected patient data for training/validation cohort
- Receive software submission
- Synthetic data to augment real data challenge

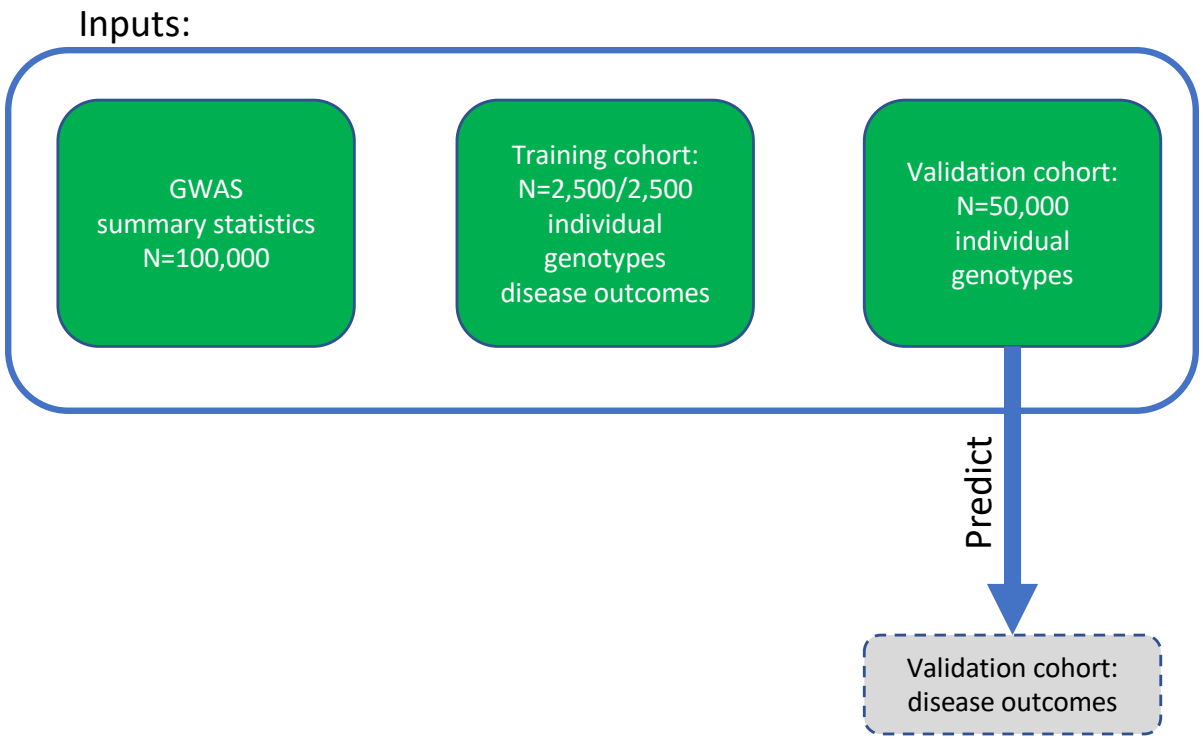
Phenotypes: Real and simulated traits

- Four Real phenotypes:
  - Breast Cancer
  - Early-onset Coronary Artery Disease
  - Inflammatory Bowel Disease (IBD)
  - Type 2 Diabetes

- 30 Simulated phenotypes (optional challenge)

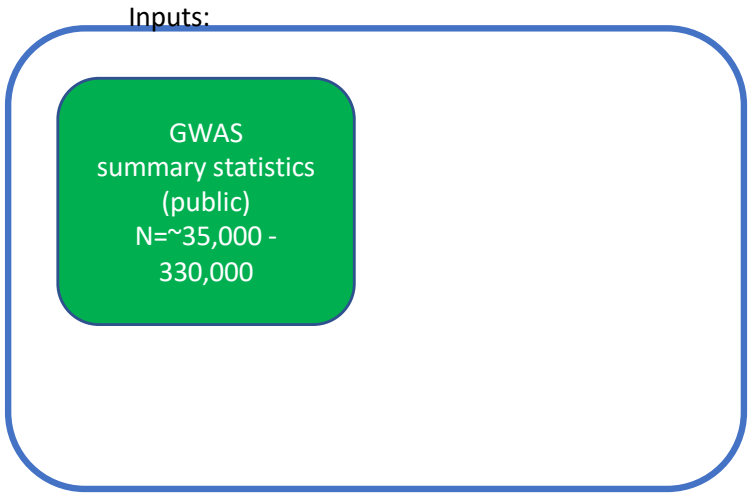
Source: Chun and Imakaev et al. AJHG 2020

Datasets: Simulated phenotypes

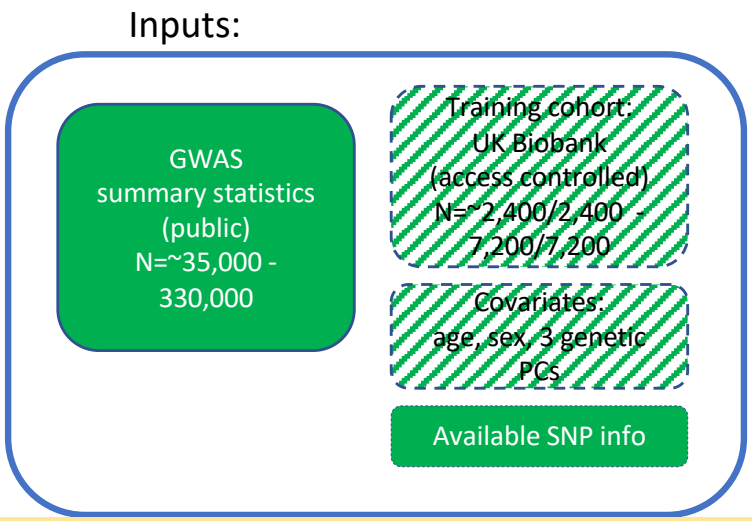


Source: Chun and Imakaev et al. AJHG 2020

Datasets: Real phenotypes



Datasets: Real phenotypes

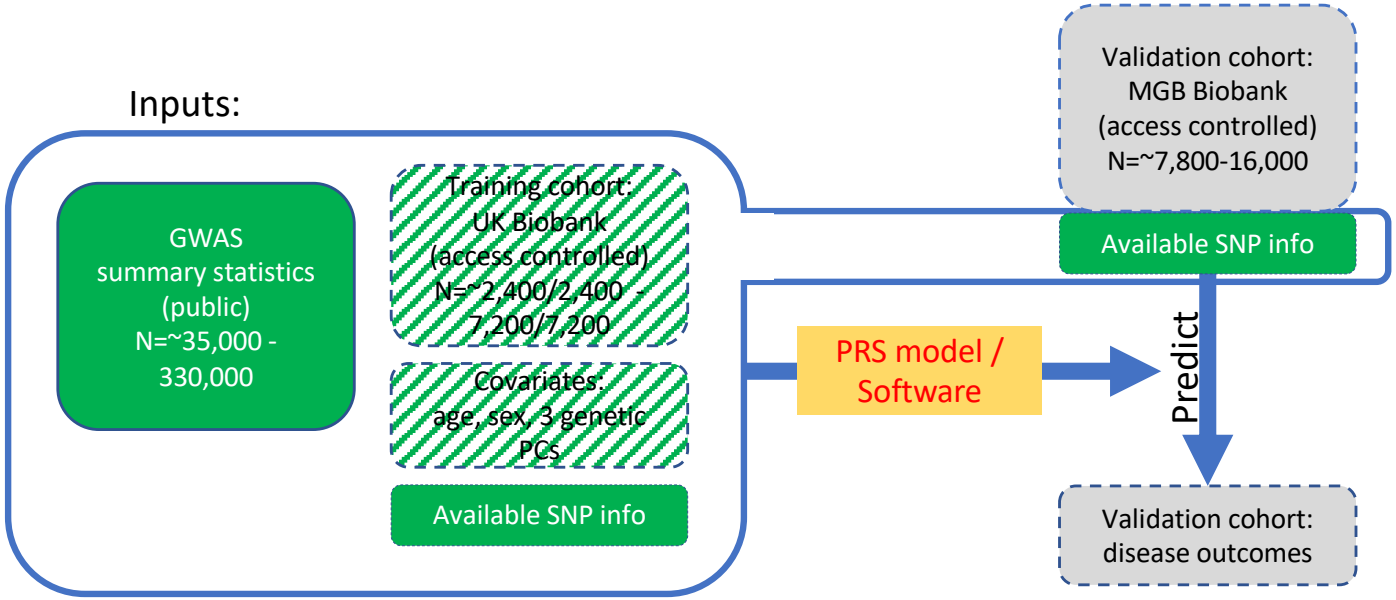


Training cohort from UK Biobank (N)

	Case	Control
Breast cancer	3,956	3,956
IBD	2,483	2,483
CAD	2,000	2,000
Type 2 Diabetes	7,298	7,298

Training on UK Biobank data was handled centrally via software submission. UK Biobank data are not directly shared with participants.

Datasets: Real phenotypes



Training on UK Biobank data was handled centrally via software submission.
UK Biobank data are not directly shared with participants.

Validation Cohort: MGB Biobank

Journal of *Personalized Medicine*

MDPI

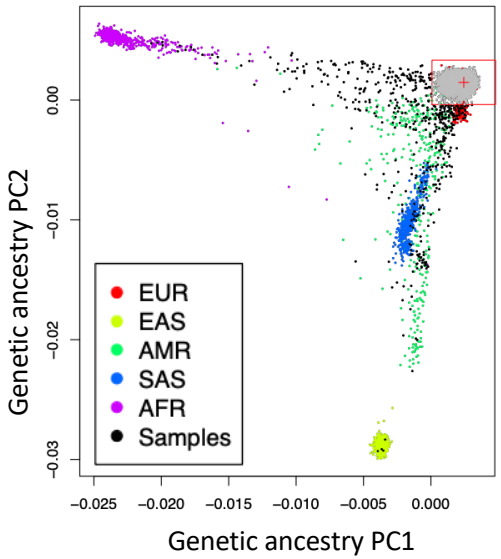
Article

Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations

Elizabeth W. Karlson [1,*], Natalie T. Boutin [2], Alison G. Hoffnagle [3] and Nicole L. Allen [1]

Sample sizes:

- 16,839 Whites
- 1,100 African Americans
- 403 Asians



Overview of submissions

Team 1	• Applied an AI technique • Incorporate covariates directly into the model	Software submission (Pre-trained)
Team 2	• Applied MegaPRS	Per-SNP effect sizes
Team 3	• Aggressively using variant annotations • Based on LD Pruning + Thresholding technique	Software submission (Training requested)
Team 4	• A sparse lasso regression technique	Software submission (Pre-trained)
Team 5	• Multiple published PRS methods	Per-SNP effect sizes

Evaluation metrics

Logistic regression of:

$$y \sim age + sex + PC1 + PC2 + PC3 + \cdots + PC10 + PRS$$

- AUC of the full model including PRS and covariates

Primary metric

- $R^2_{nagelkerke}$: the likelihood of data under the full model compared to the likelihood under the null model

- 5% Tail Odds Ratio (OR): The odd at top 5% of the score relative to the odd of the rest of distribution (covariates included)

$$\frac{\#\{y = 1 \wedge \hat{y} > q_{0.95}\}}{\#\{y = 0 \wedge \hat{y} > q_{0.95}\}} / \frac{\#\{y = 1 \wedge \hat{y} < q_{0.95}\}}{\#\{y = 0 \wedge \hat{y} < q_{0.95}\}}$$

Methods for baseline comparison

- Negative control:
  - Logistic regression with only covariates (no PRS)

$$y \sim age + sex + PC1 + PC2 + PC3 + \cdots + PC10$$

- 4 widely used or current state-of-the-art methods:
  - LD Pruning + Thresholding (P+T)
  - LDPred v1
  - PRS-cs
  - NPS

Prediction accuracy ($R^2_{nagelkerke}$)

Breast Cancer

Model	$R^2_{nagelkerke}$	Rank
NPS	0.111	-
PRScs	0.109	-
MegaPRS	**0.101**	**1**
Team 2-1	0.100	2
Team 5-C	0.093	3
Team 5-L1	0.091	4
Team 3-4	0.091	5
Team 5-A	0.090	6
P+T	0.090	-
Team 5-T	0.089	7
LDPred 1	0.088	-
Team 4-3	0.087	8
Team 3-3	0.085	9
Team 5-L2i	0.085	10
Team 4-1	0.084	11
Team 3-2	0.082	12
Team 5-L2	0.074	13
Team 4-2	0.073	14
Cov only	**0.073**	**-**
Team 3-5	0.072	15
Team 3-6	0.072	16
Team 3-1	0.072	17
Team 1	n/a	-

IBD

Model	$R^2_{nagelkerke}$	Rank
MegaPRS	**0.173**	**1**
NPS	0.157	-
PRScs	0.157	-
Team 2-1	0.156	2
Team 5-A	0.148	3
Team 5-L1	0.141	4
Team 5-L2	0.141	5
P+T	0.140	-
Team 5-C	0.139	6
LDPred 1	0.134	-
Team 5-T	0.125	7
Team 5-L2i	0.124	8
Team 4-1	0.103	9
Team 4-3	0.103	10
Team 1	0.102	11
Team 4-2	0.099	12
Cov only	**0.099**	**-**
Team 3-5	0.098	13
Team 3-4	0.098	14
Team 3-2	0.098	15
Team 3-1	0.098	16
Team 3-3	0.098	17
Team 3-6	0.098	18

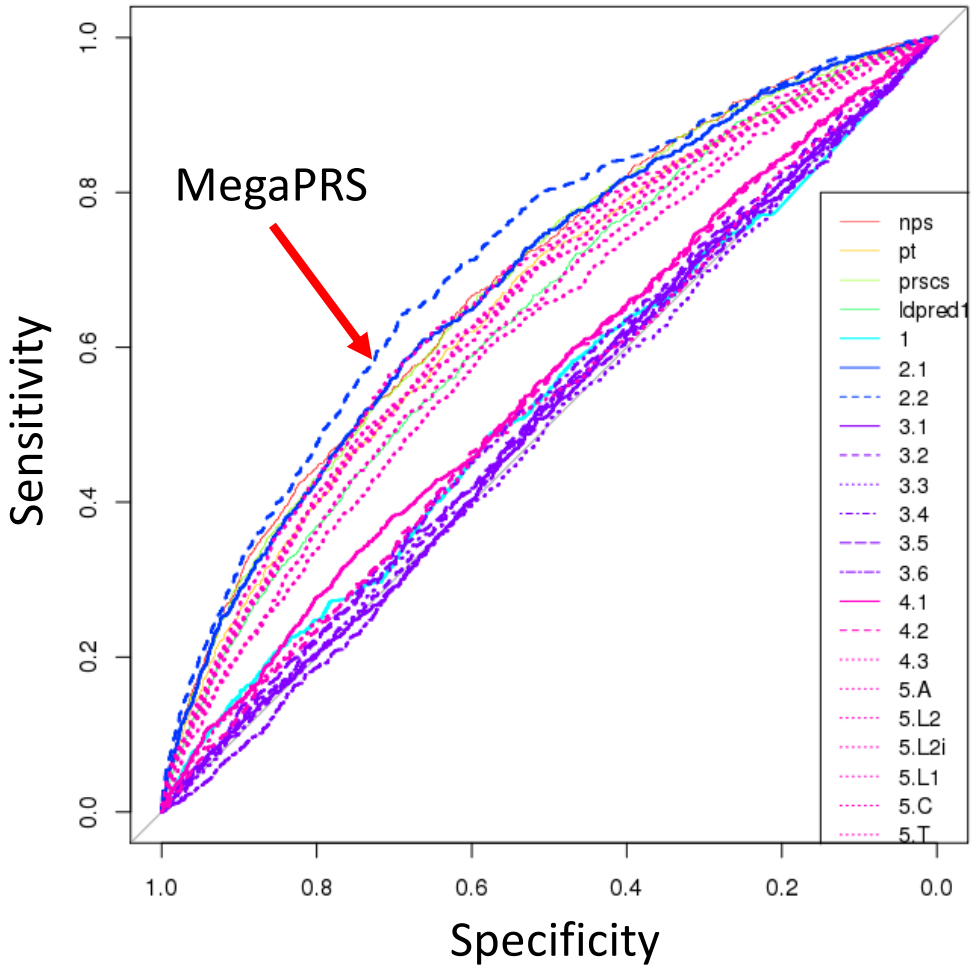
Coronary Artery Disease

Model	$R^2_{nagelkerke}$	Rank
PRScs	0.160	-
NPS	0.160	-
P+T	0.155	-
MegaPRS	**0.155**	**1**
Team 3-4	0.154	2
Team 5-L2	0.153	3
Team 5-C	0.153	4
Team 5-L2i	0.153	5
Team 5-A	0.153	6
LDPred 1	0.151	-
Team 3-3	0.150	7
Team 3-2	0.148	8
Team 5-T	0.147	9
Team 3-1	0.146	10
Team 3-6	0.146	11
Team 3-5	0.146	12
Team 4-1	0.143	13
Team 5-L1	0.143	14
Team 4-2	0.142	15
Team 4-3	0.142	16
Team 1	0.142	17
Cov only	**0.142**	**-**

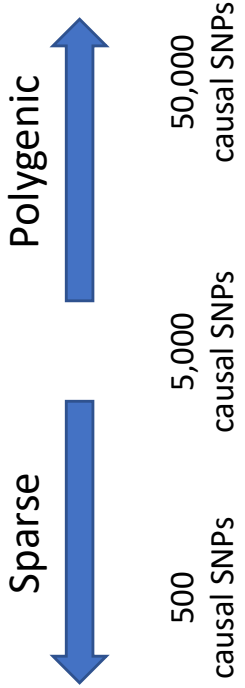
Type 2 Diabetes

Model	$R^2_{nagelkerke}$	Rank
NPS	0.130	-
PRScs	0.130	-
MegaPRS	**0.127**	**1**
Team 5-A	0.118	2
LDPred 1	0.117	-
Team 5-C	0.117	3
Team 4-3	0.112	4
Team 5-L2	0.111	5
Team 5-L2i	0.107	6
P+T	0.107	-
Team 1	0.105	7
Team 5-T	0.099	8
Team 4-1	0.099	9
Team 3-4	0.098	10
Team 3-3	0.095	11
Team 5-L1	0.095	12
Team 3-2	0.095	13
Team 3-1	0.094	14
Team 3-5	0.093	15
Team 3-6	0.093	16
Team 4-2	0.092	17
Cov only	**0.092**	**-**

ROC of IBD PRS models

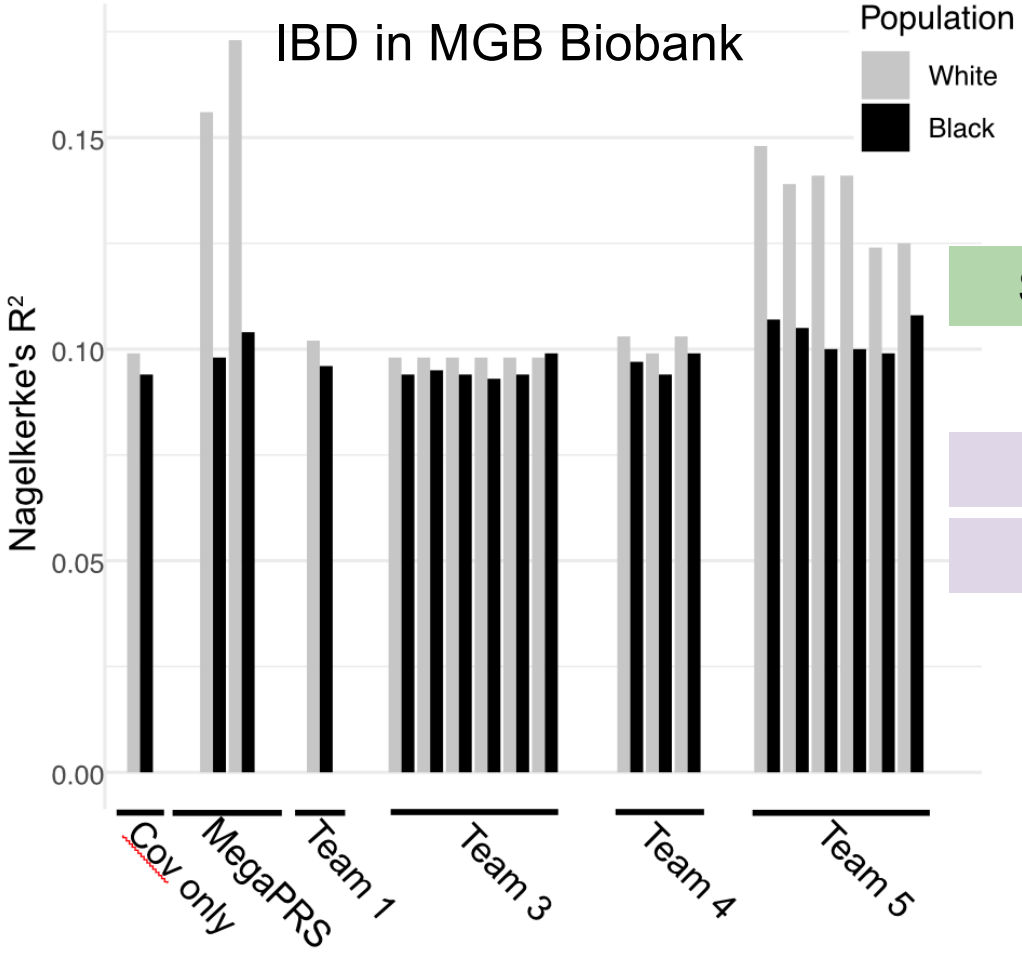


Prediction accuracy in simulated datasets



		Validation	
% causal SNPs	Method	$R^2_{Liability}$	95% CI
1%	P+T	0.074	
	LDPred 1	0.103	
	PRS-CS	0.110	
	NPS	0.123	
	MegaPRS	**0.145***	**[0.143 - 0.147]**
0.1%	P+T	0.199	
	LDPred 1	0.115	
	PRS-CS	0.224	
	NPS	0.256	
	MegaPRS	**0.297***	**[0.292 - 0.302]**
0.01%	P+T	0.307	
	LDPred 1	0.219	
	PRS-CS	0.327	
	NPS	0.463	
	MegaPRS	**0.460**	**[0.453 - 0.467]**

Poor transferability across populations



Sample size in MGB Biobank		
	Case	Control
White	820	16,335
Black	26	1,100

Conclusions

- We would not have an improvement over the existing state-of-the-art methods for any of the phenotypes.
    - Looks like we do!

- Investment in the statistics does not pay off any longer.
    - Looks like it still does!

CAGI6 PRS challenge: Limitations

- Technically difficult challenge – 22 groups signed up initially; only 5 groups made submission
- Bringing experts into CAGI – need to demonstrate a clear benefit of participating
- Bringing new computational biology groups into the field – help getting over the hurdle of training data access

CAGI6 PRS challenge: Limitations

- A machine learning-based prediction model did not perform well – not designed to evaluate the full potential
- Need a mechanism to share training cohorts more easily and make more covariates available
- Need to include more diverse ancestry

Acknowledgements

- Co-assessor/data provider: Shamil Sunyaev (Harvard/BWH)
- Data provider: Nikolaos Patsopoulos (Harvard/BWH)
- Data provider: Benjamin Neale (Broad Institute)
- CAGI organizer: Constantina Bakolitsa (UC Berkeley)
- CAGI organizer: Predrag Radivojac (Northeastern University)

We would like thank all CAGI6 PRS challenge participants.

This research has been conducted using the UK Biobank Resource under Application Number 31063.

We are looking for postdocs.