Disclosure Slide

for Samuel S. Kim

I have nothing to disclose

Improving the informativeness of Mendelian disease-derived pathogenicity scores for common disease using AnnotBoost

Samuel Kim

Alkes Price Group

10.28.2020

Kim et al. bioRxiv 2020 (accepted in principle, *Nat. Commun.*)

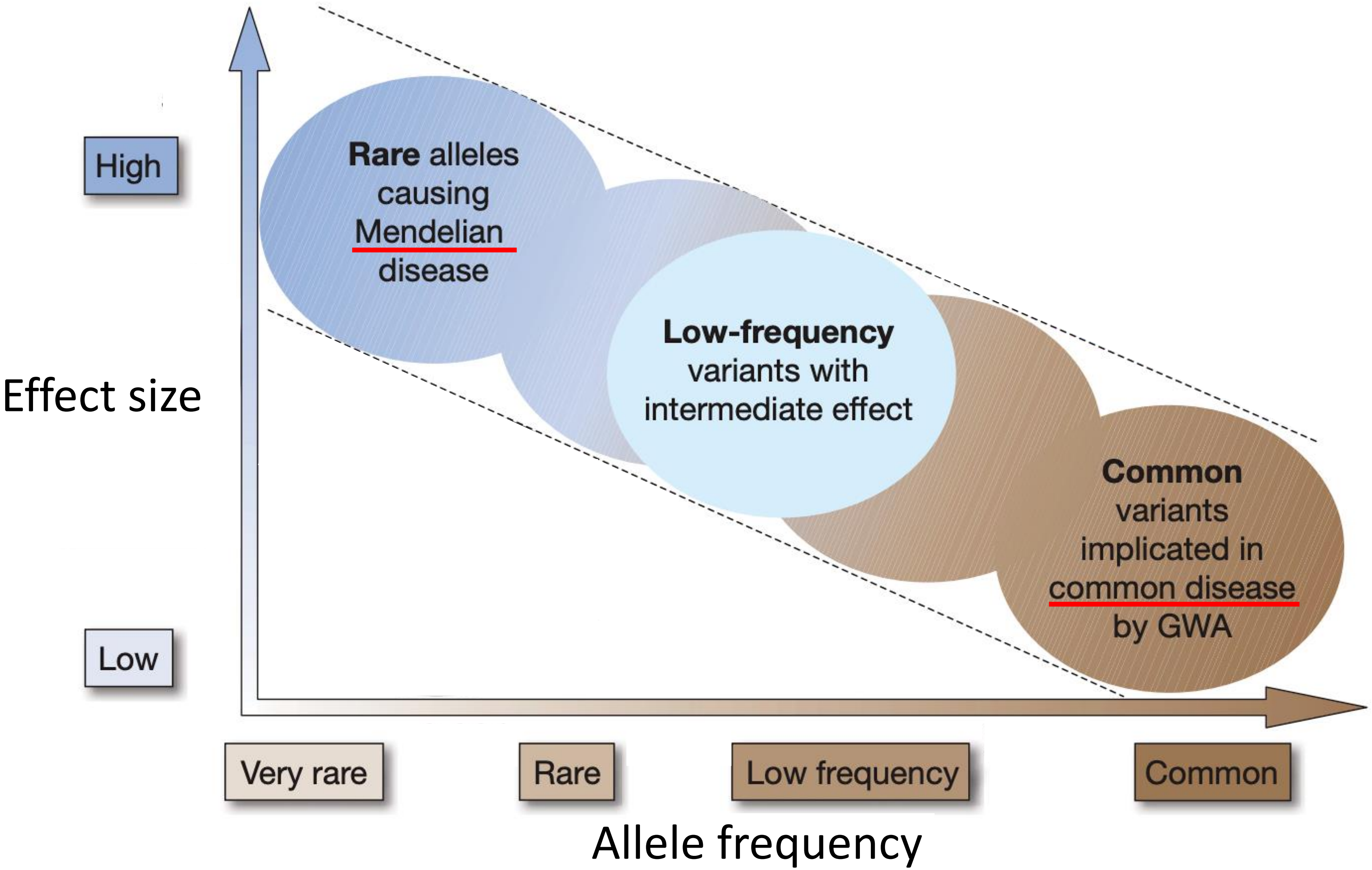
Outline

- Motivation
- Methods: assessing informativeness of existing pathogenicity scores
- Methods: improving the informativeness of existing pathogenicity scores
- Results

Outline

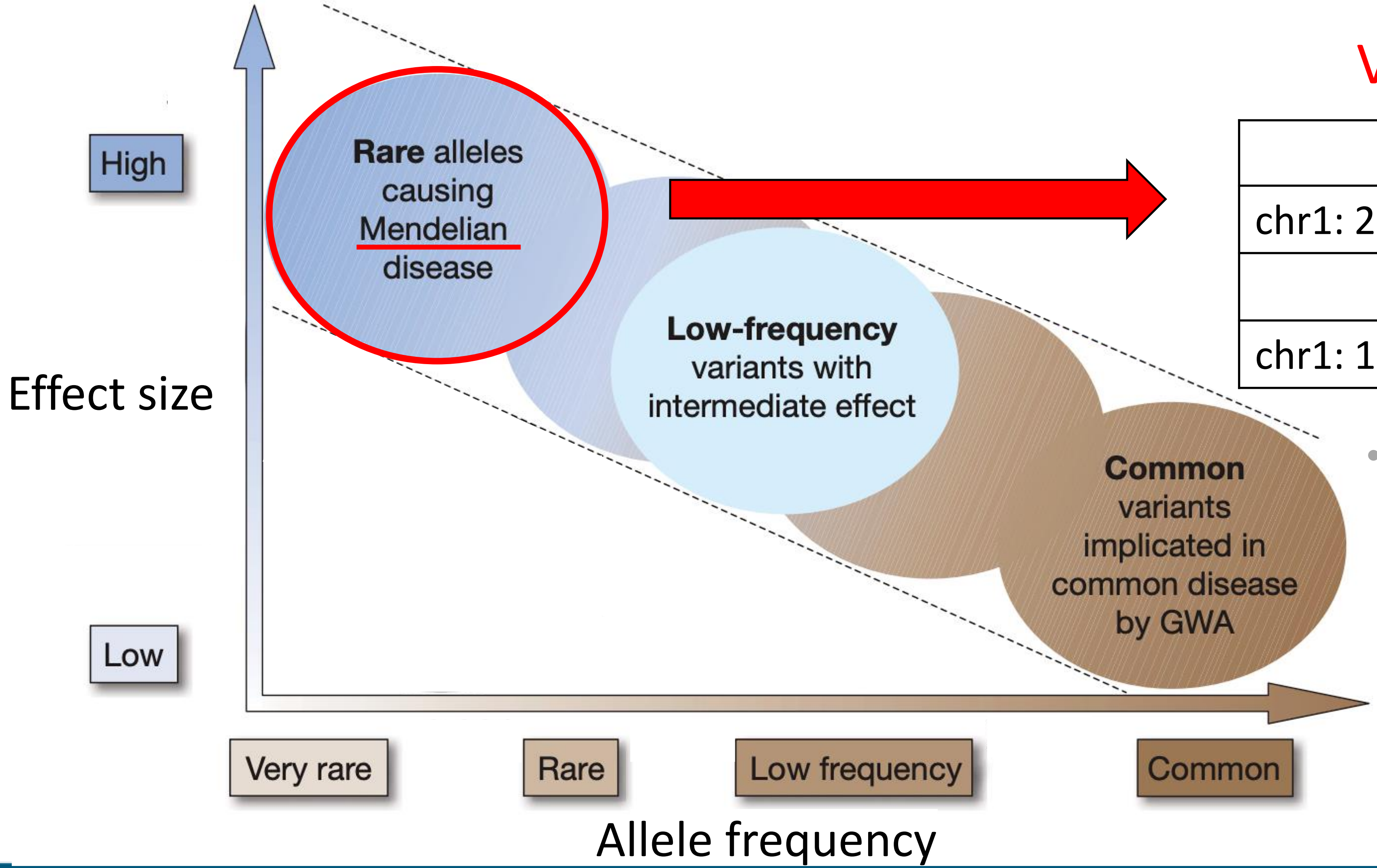
- ✓ Motivation
- Methods: <u>assessing</u> informativeness of existing pathogenicity scores
- Methods: <u>improving</u> the informativeness of existing pathogenicity scores
- Results

Mendelian disease and common disease: the big divide?



(Figure from Manolio et al. Nature 2009; Antonarakis et al. Nat Rev Genet 2010)

Mendelian disease-derived pathogenicity scores prioritize pathogenic, rare variants for gene discovery / diagnosis

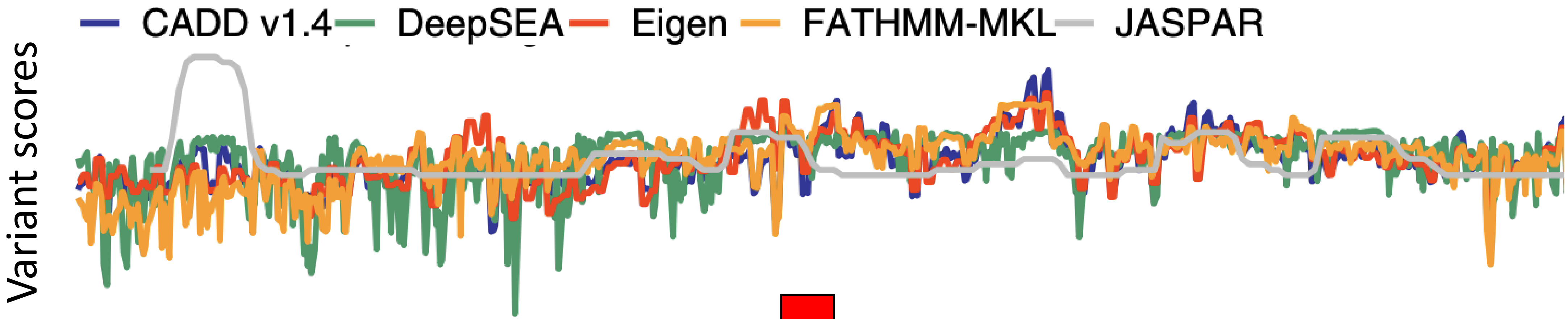


Variant scores

SNVs	CADD
chr1: 225528175 G>T	53
…	…
chr1: 1152488 A>G	4

- Can be genome-wide or missense variants only

What is the contribution of Mendelian disease-derived pathogenicity scores to common diseases?



???

Type 2 Diabetes

Height

Rheumatoid arthritis

Common disease
(low-frequency and common variants)

Schizophrenia

Autism

LDL, HDL

(CADD: Kircher et al. NG 2014, DeepSEA: Zhou et al. Nat Methods 2015
Reviewed in Eilbeck et al. Nat Rev Genet 2017, Figure from Kircher et al. Nat Commun 2019)

ASHG
American Society of Human Genetics

Shared genetic architecture between Mendelian disease and common disease

- Gene overlap between monogenic diseases and complex traits

 - e.g. LDLR: monogenic hypercholesterolemia and cardiovascular diseases

- Significant comorbidities
- Mendelian disease genes are enriched in GWAS closest genes
- *Limitation*: previous analyses were either gene-based or limited to genome-wide significant SNPs

Our goals: pathogenicity score → common disease

1. Assess informativeness of Mendelian disease-derived pathogenicity scores for 41 common diseases and complex traits

Our goals: pathogenicity score → common disease

1. Assess informativeness of Mendelian disease-derived pathogenicity scores for 41 common diseases and complex traits

2. Develop a framework to improve their informativeness for common disease

Outline

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

- Motivation
- Methods: assessing informativeness of existing pathogenicity scores
- Methods: improving the informativeness of existing pathogenicity scores
- Results

Method building Mendelian disease-derived pathogenicity annotations

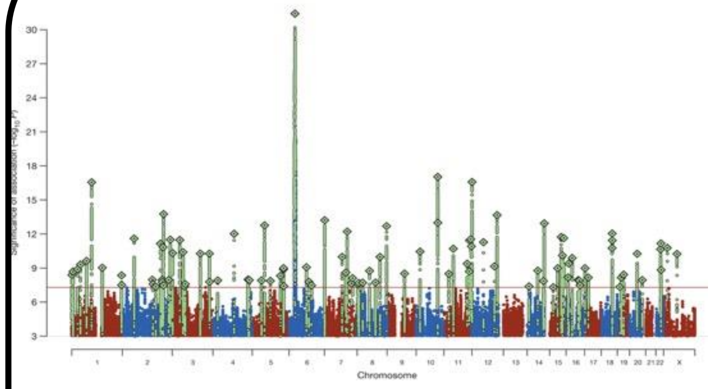
- Pathogenicity scores overwhelmingly predict pathogenic rare SNPs.
- Hypothesis: Mendelian disease variants and common disease variants share <u>similar properties</u>.

To evaluate this hypothesis,

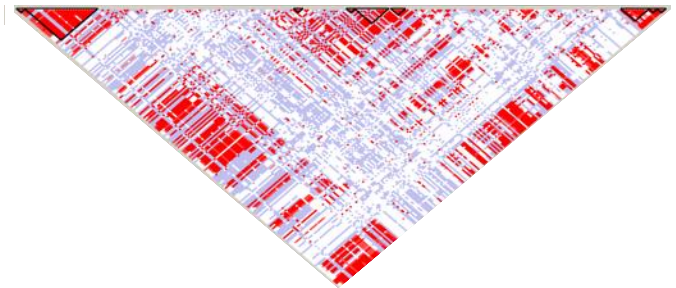
- Given a pathogenicity score, applied S-LDSC on binary annotations to 41 complex traits (avg. N = 320K; 30 from UK Biobank)

To evaluate disease heritability enrichment, used stratified LD score regression (S-LDSC)

Input



Summary statistics



Reference panel (1000G EUR)

Functional annotations

$$E\left[\chi_j^2\right] = N\sum_c \tau_c \ell(j,c) + 1$$

Output

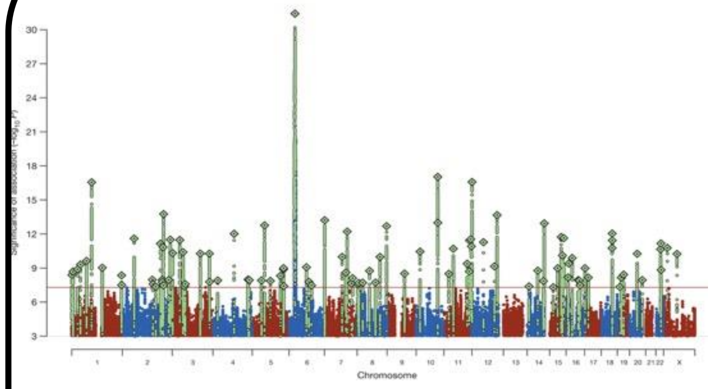
1. Enrichment = Prop. h^2g / Prop. SNPs
2. Standardized effect size (τ*) = $M\tau_c sd(c) / h^2g$

That is, proportionate change in per-SNP heritability associated to a one sd(annotation$_c$) increase, conditional on all other annotations in the model.

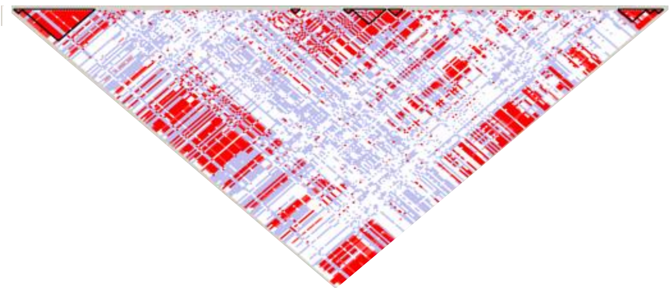
(Finucane et al. Nat. Genet. 2015)

To evaluate disease heritability enrichment, used stratified LD score regression (S-LDSC)

Input



Summary statistics



Reference panel (1000G EUR)

Functional annotations

- Annotations with τ^* = 0: no unique information
- Annotations with significantly positive or negative τ^* are conditionally informative, after considering all other annotations in the model.

(Finucane et al. Nat. Genet. 2015)

Outline

- Motivation
- Methods: assessing informativeness of existing pathogenicity scores
- Methods: improving the informativeness of existing pathogenicity scores
- Results

AnnotBoost: a gradient boosting-based ML framework to impute and denoise existing pathogenicity scores

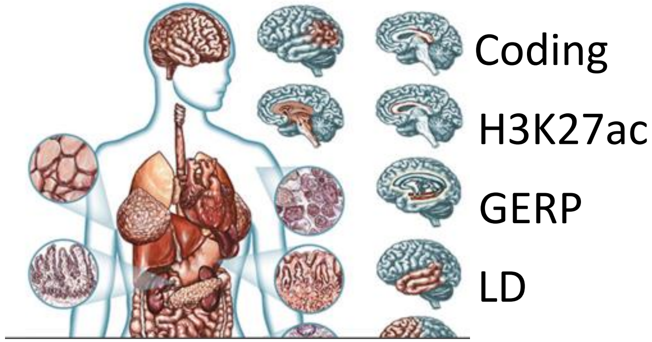
Input

SNPs	Annotation X
rs575272151	1
…	…
rs540538026	1

Pathogenicity score to boost

Training set

Validation set



75 baseline-LD features

*Not phenotype-specific

*Implements XGBoost to take account of nonlinearity

AnnotBoost: a gradient boosting-based ML framework to impute and denoise existing pathogenicity scores

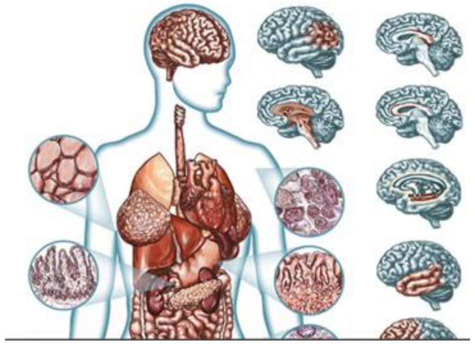
Input

SNPs	Annotation X
rs575272151	1
…	…
rs540538026	1

Training set

Validation set

Pathogenicity score to boost

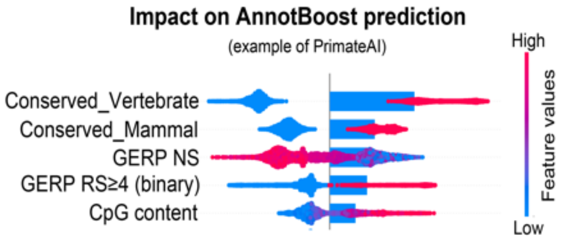


75 baseline-LD features

Output

SNPs	Annotation X⇧
rs575272151	0.80
rs540538026	0.25

Boosted score



Feature importance

Analysis

Heritability enrichment (S-LDSC)

Heritability model fit ($logl_{SS}$)

AnnotBoost model training

Example shown with CADD score.

SNPs	CADD (Kircher et al. NG 2014)
rs184094753	55
rs11588155	0.001
rs28359608	20

AnnotBoost works regardless binary or continuous-valued score.

ASHG
American Society of Human Genetics

AnnotBoost model training

SNPs	CADD
rs184094753	55
...	...
rs28359608	20

Sort →

SNPs	CADD
rs184094753	55
...	...
rs11588155	0.001

→ Top 10%: label ‘1’ (positive SNPs)

→ Bottom 40%: label ‘0’ (control SNPs)

(Without using external disease data)

AnnotBoost achieved robust results with different training ratios.

AnnotBoost model training

SNPs	CADD
rs184094753	55
...	...
rs28359608	20

Sort →

SNPs	CADD
rs184094753	55
...	...
rs11588155	0.001

→ Top 10%: label ‘1’ (positive SNPs)

→ Bottom 40%: label ‘0’ (control SNPs)

AnnotBoost training

Even (resp. odd) chr SNPs	GERP	Coding	H3K27ac		CpG	CADD (binary label)
rs184094753	0	0	0	…	0.3	1
…	1	0	1	…	0.1	…
rs11588155	0	1	0	…	0.5	0

baseline-LD features $[X_{train}]$

Y_{train}

AnnotBoost model training

SNPs	CADD
rs184094753	55
…	…
rs28359608	20

Sort →

SNPs	CADD
rs184094753	55
…	…
rs11588155	0.001

→ Top 10%: label '1' (positive SNPs)

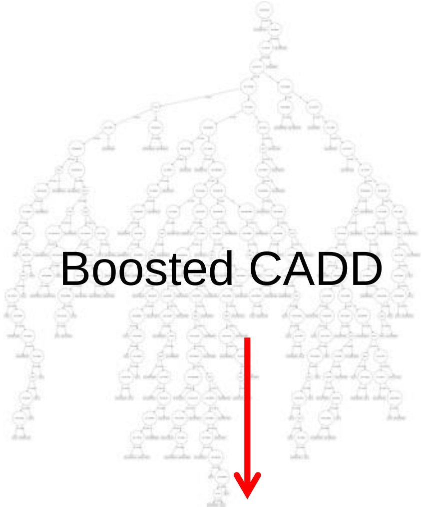
→ Bottom 40%: label '0' (control SNPs)

AnnotBoost training

Even (resp. odd) chr SNPs	GERP	Coding	H3K27ac		CpG	CADD (binary label)
rs184094753	0	0	0	…	0.3	1
…	1	0	1	…	0.1	…
rs11588155	0	1	0	…	0.5	0

baseline-LD features [X_{train}]

Y_{train}



Boosted CADD

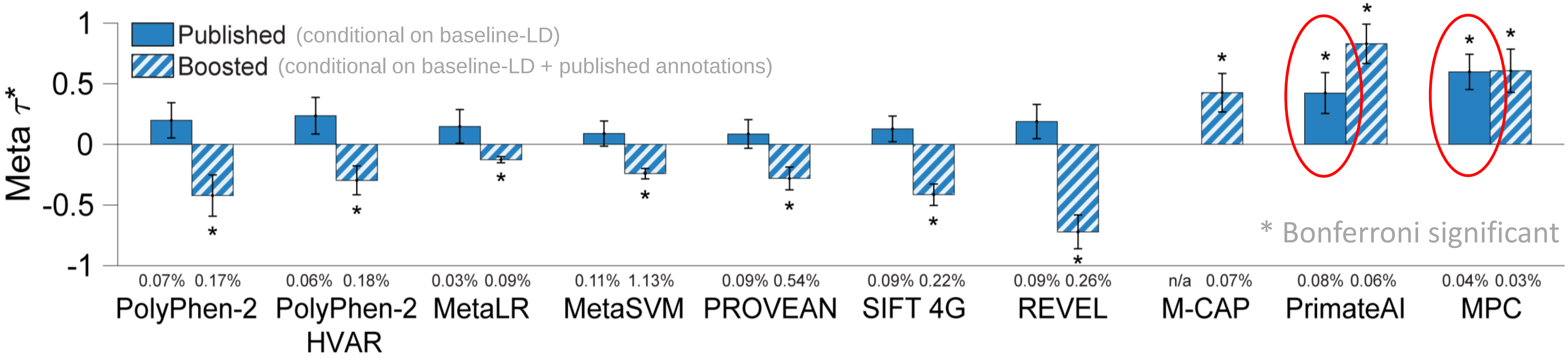
S-LDSC

Outline

- Motivation
- Methods: assessing informativeness of existing pathogenicity scores
- Methods: improving the informativeness of existing pathogenicity scores
- ✓ Results

AnnotBoost improves the Informativeness of Mendeli
derived missense scores

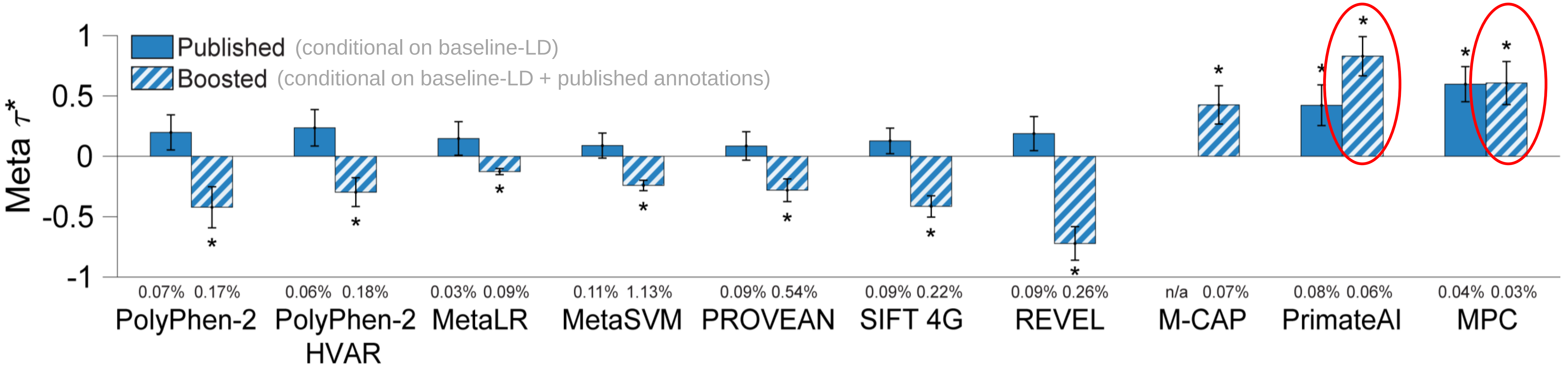
- Two missense scores are conditionally informative (with significant τ*)



- PrimateAI: eliminating common missense variants identified in other primate species
- MPC: identifying regions within genes that are depleted for missense variants in ExAC data

AnnotBoost improves the Informativeness of Mendeli derived missense scores

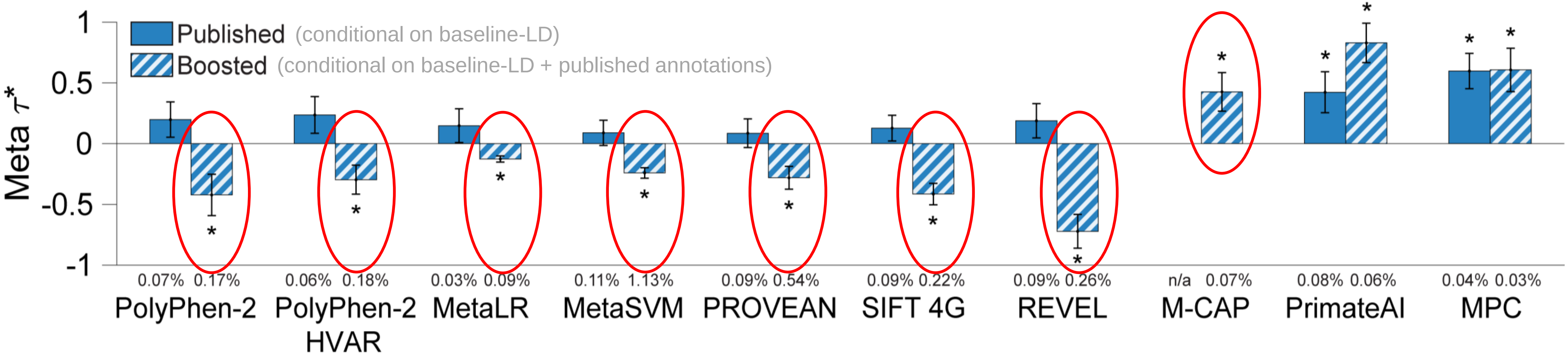
- AnnotBoost generates orthogonal signals from published scores



- PrimateAI: eliminating common missense variants identified in other primate species
- MPC: identifying regions within genes that are depleted for missense variants in ExAC data

(Sundaram et al. Nat Genet 2018, Samocha et al. biorxiv)

AnnotBoost improves the Informativeness of Mendeli
derived missense scores

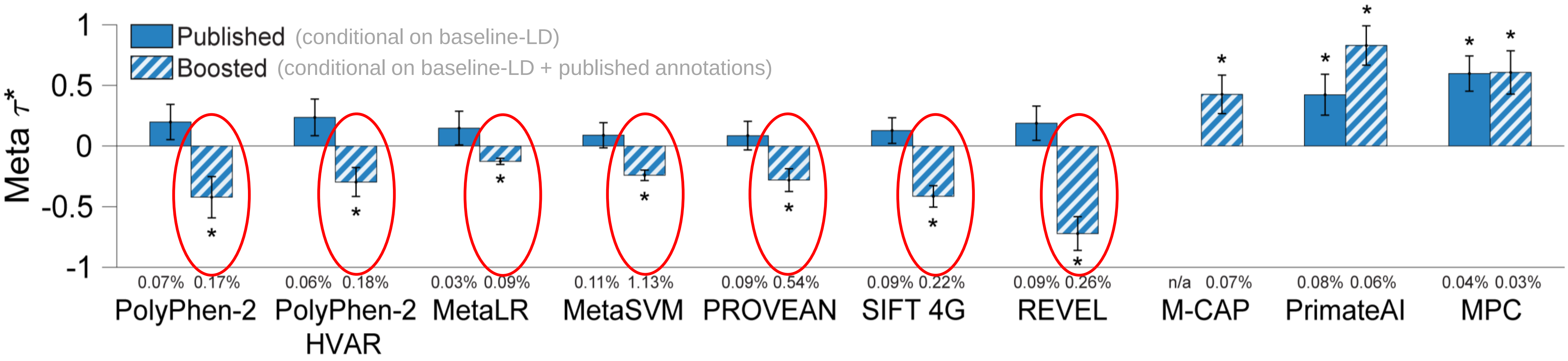


Non-significant (published) → significant (boosted)

Imputed non-coding SNPs (driven by conservation features): >85% signals

- M-CAP: ensemble model trained on HGMD pathogenic vs. ExAC benign variants

AnnotBoost improves the Informativeness of Mendeli derived missense scores



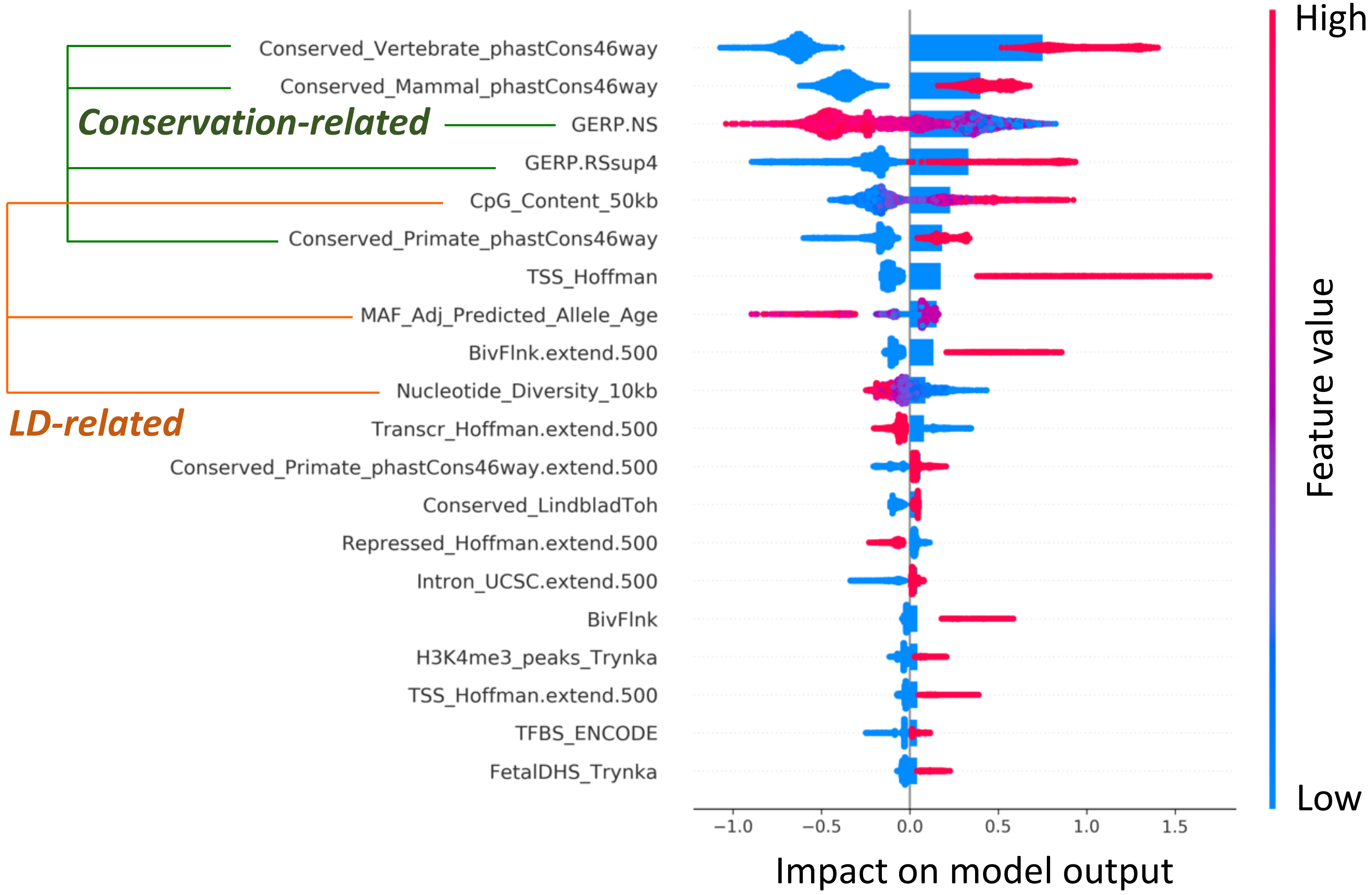
Neg τ^* = Enriched but less enriched than expected

e.g. REVEL: 4.7x enriched (expected enrichment 8.0x)

(Neg τ^*: Hormozdiari et al. Nat. Comm. 2019)

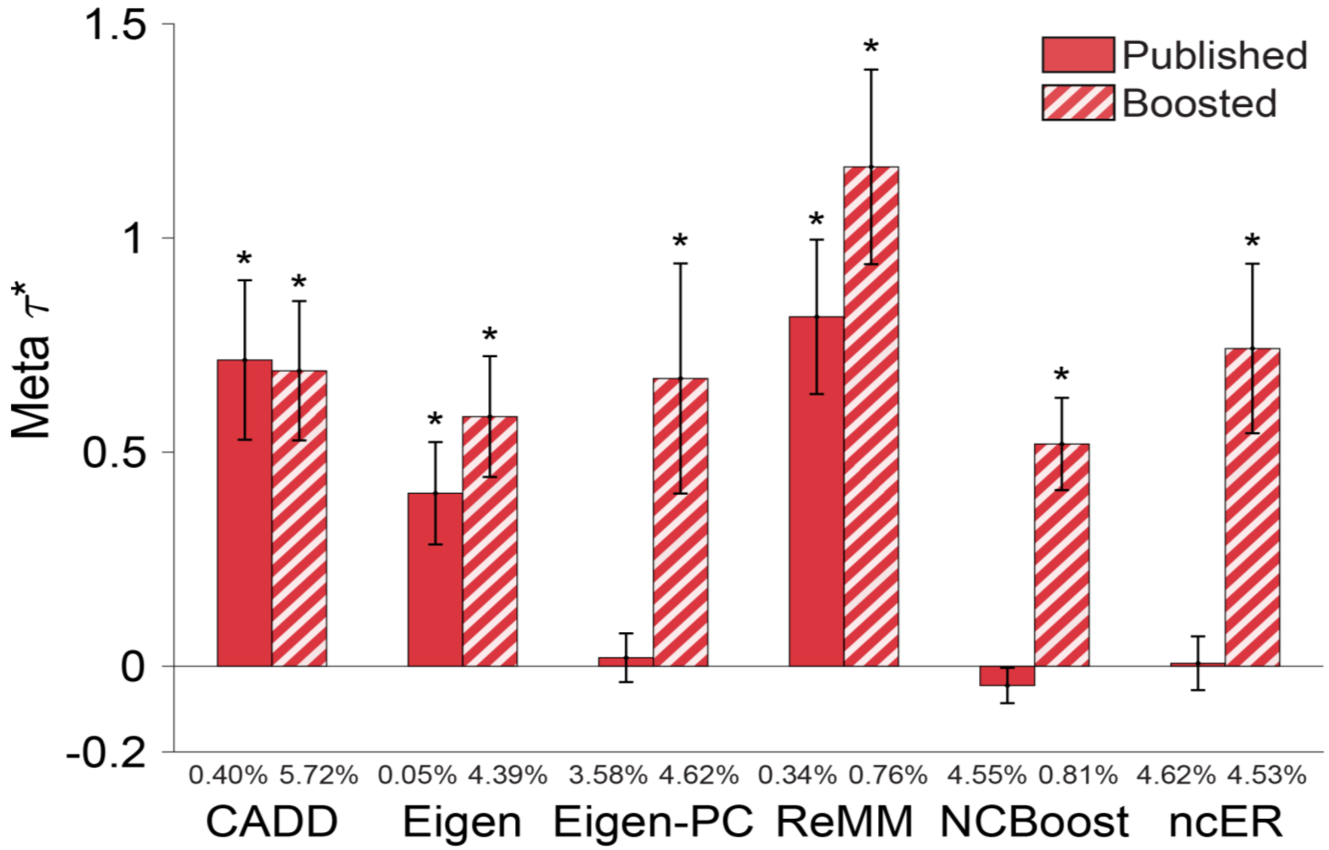
Which genomic features are driving AnnotBoost predictions?

- Improve interpretability; signed impact of features driving PrimateAI⇧:



(SHAP feature importance: Lundberg & Lee. NIPS 2017)

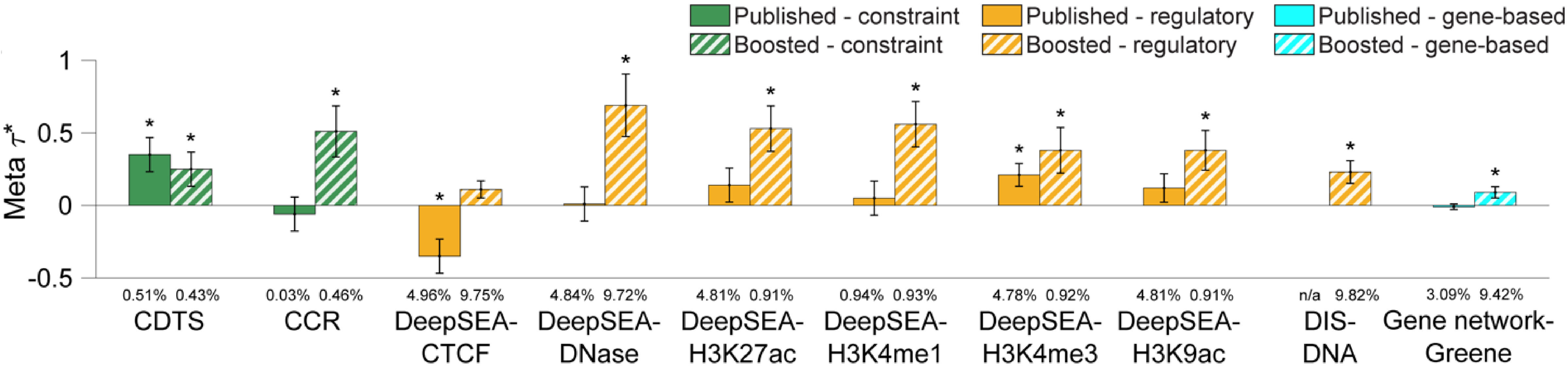
AnnotBoost improves the Informativeness of Mendeli derived genome-wide scores



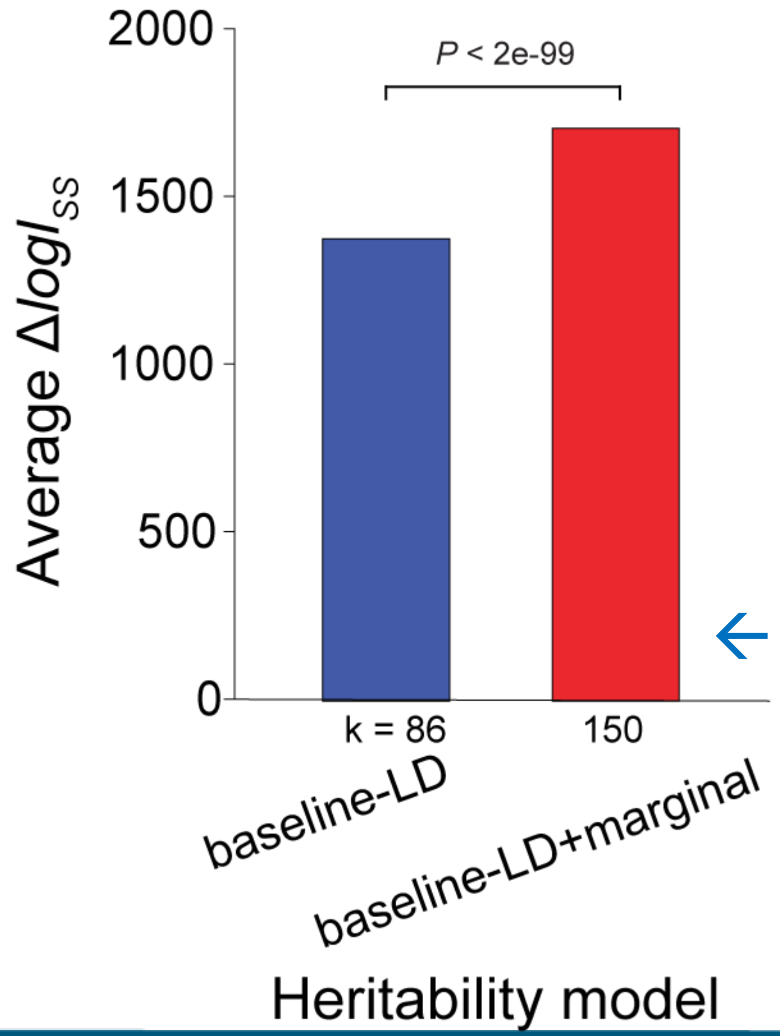
- Eigen, Eigen-PC, NCBoost, ncER: imputed SNPs 17-54% overall signals
- CADD, ReMM: denoised previously scored SNPs

AnnotBoost improves the Informativeness of constraint, epigenetic, gene scores

- Imputed SNPs retained 55% of overall signal, on average



Boosted scores significant improved heritability model fit ($\Delta\log l_{SS}$) by +23.9% in all 30/30 UK Biobank traits

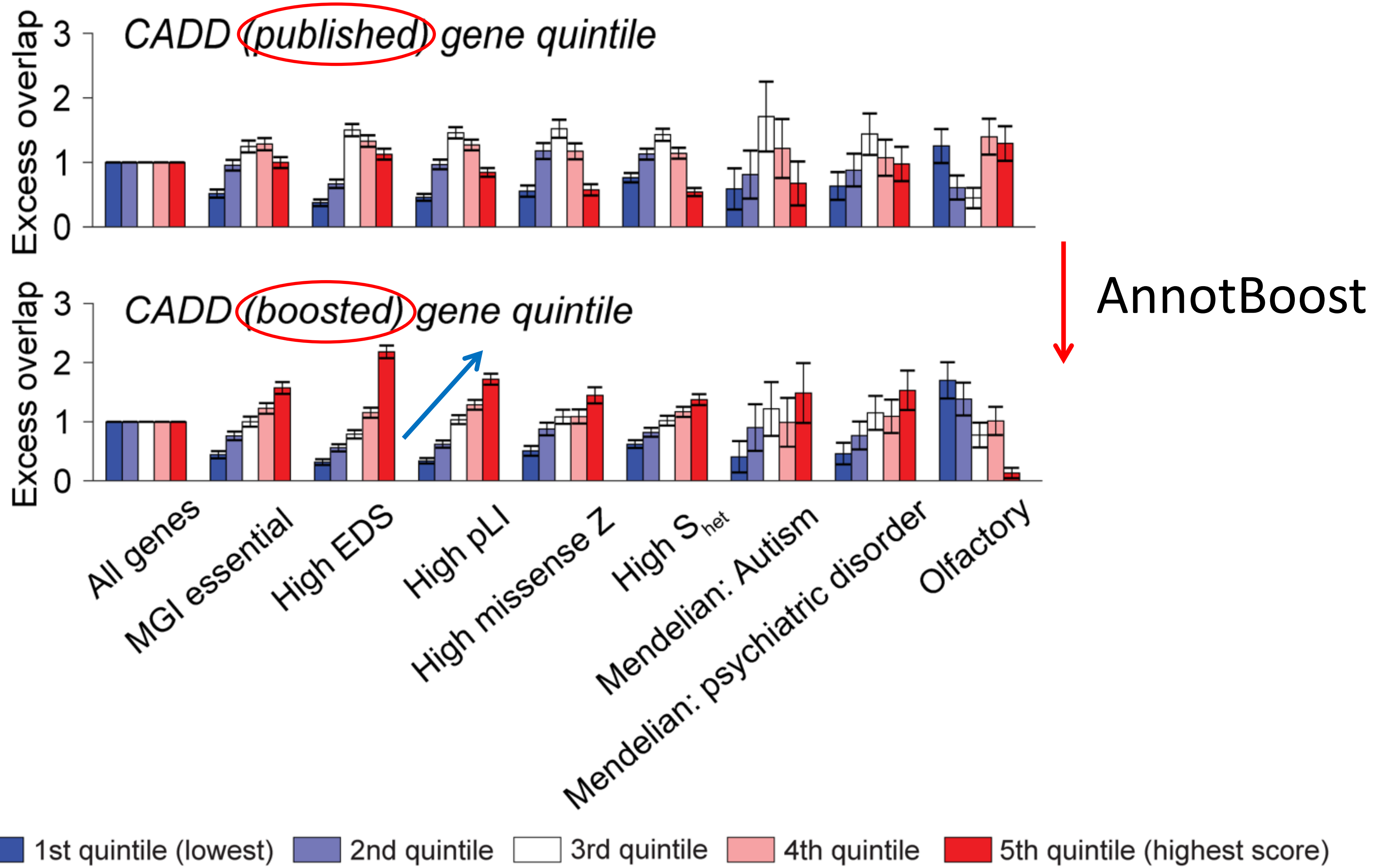


← +64 new annotations

$\Delta\log l_{SS}$ = $\log l_{SS}$(model) - $\log l_{SS}$(MAF+LD model)

- baseline-LD+marginal better predicted disease-associated fine-mapped SNPs by +4.9% to +21.3%.

AnnotBoost can help identify biologically important genes



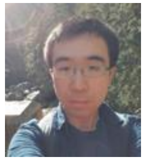
(Lek et al. Nature 2016, Samocha et al. Nat Genet 2014, Cassa et al. Nat Genet 2017)

Conclusions

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

- Developed AnnotBoost to study shared variant properties between Mendelian disease variants and common disease variants.
- Our new annotations significantly improved the heritability model (+23.9%), motivating their inclusion in future fine-mapping studies.
- AnnotBoost can be applied to future pathogenicity scores to improve our understanding of genetic architecture of complex traits and identify biologically important genes.

Acknowledgements

- **Alkes Price** 
- Kushal Dey 
- Omer Weissbrod 
- Carla Márquez-Luna 
- Steven Gazal 
- Bryce van de Geijn 
- Farhad Hormozdiari 
- Huwenbo Shi 
- UK Biobank
- NIH for funding

Price Group @ HSPH

samsungilkim sungil@mit.edu github.com/samskim/annotboost

Thank you!

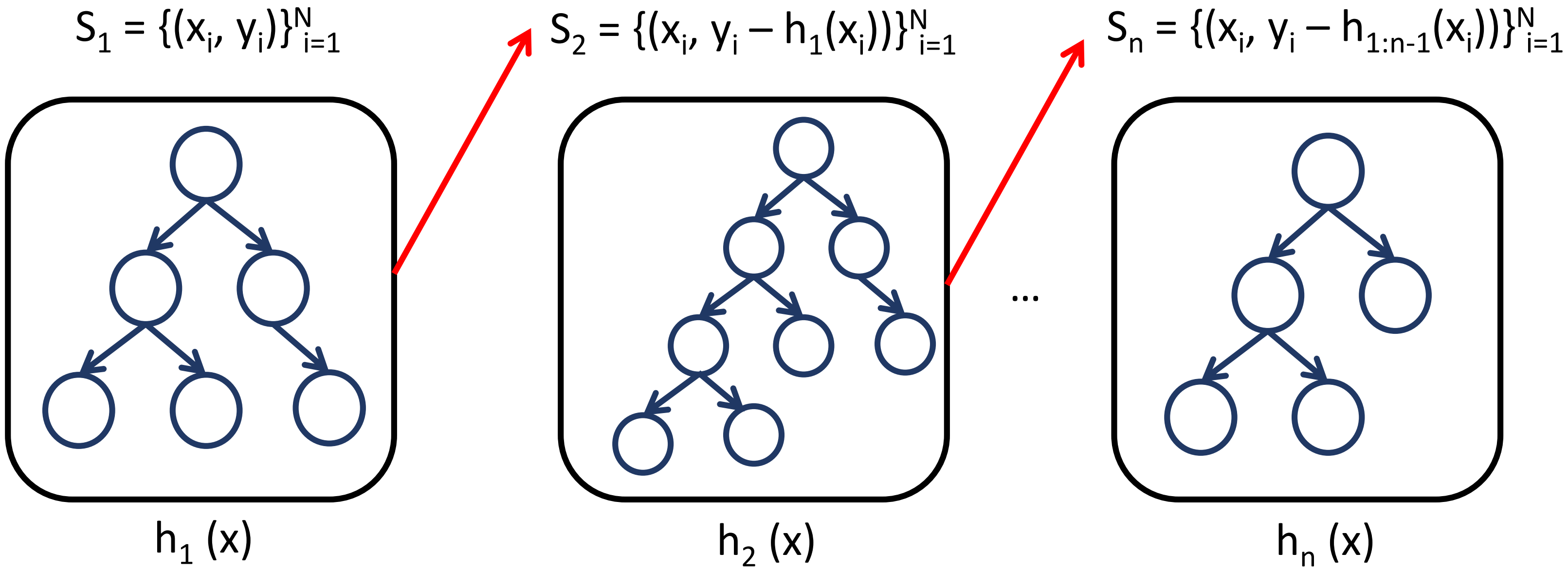
sungil [at] mit.edu
samsungilkim
samuel-s-kim
github.com/samskim/annotboost

Kim SS, Dey KK, Weissbrod O, Marquez-Luna C, Gazal S, Price AL. Improving the informativeness of Mendelian disease-derived pathogenicity scores for common disease. 2020 bioRxiv. (accepted in principle, *Nat. Commun*.)

Supplementary slides

AnnotBoost implements gradient boosting to leverage nonlinearity among features

- S-LDSC takes account of linear interactions in the model.
- Gradient boosting (decision tree-based) accounts for nonlinearity.



Classification model $H = \alpha h_1(x) + \beta h_2(x) + \ldots + \gamma h_n(x)$

where α,β,γ are optimal weights

(Chen & Guestrin, KDD 2016)

ASHG American Society of Human Genetics

Applied AnnotBoost to missense + genome-wide pathogenicity scores

Score	Description	Coverage (% SNPs scored)
PolyPhen-2	Impact of missense variants using protein sequence and structure using HumDiv	0.28%
PolyPhen-2-HVAR	Impact of missense variants using protein sequence and structure using HumVar	0.28%
MetaLR	Deleterious missense mutations using ensemble scoring (logistic regression)	0.32%
MetaSVM	Deleterious missense mutations using ensemble scoring (support vector machine)	0.32%
PROVEAN	Impact of an amino acid change on protein function	0.31%
SIFT 4G	Impact of an amino acid change on protein function	0.31%
REVEL	Pathogenic missense variants using ensemble scoring	0.32%
M-CAP	Pathogenic rare missense variants	0.03%
PrimateAI	Impact of missense variants using deep neural networks	0.26%
MPC	Regional missense constraint	0.10%
MVP	Impact of missense variants using deep neural networks	0.29%
CADD	Predicted deleterious variants using ensemble scoring	100%
Eigen	Putatively causal variants using unsupervised learning	83.79%
Eigen-PC	Putatively causal variants using unsupervised learning using the lead eigenvector	83.79%
ReMM	Pathogenic regulatory variants using ensemble scoring	100%
NCBoost	Pathogenic non-coding variants using ensemble scoring	28.55%
ncER	Essential regulatory variants using ensemble scoring	61.94%

(Complete reference: see Kim et al. 2020)

Evaluating different heritability models

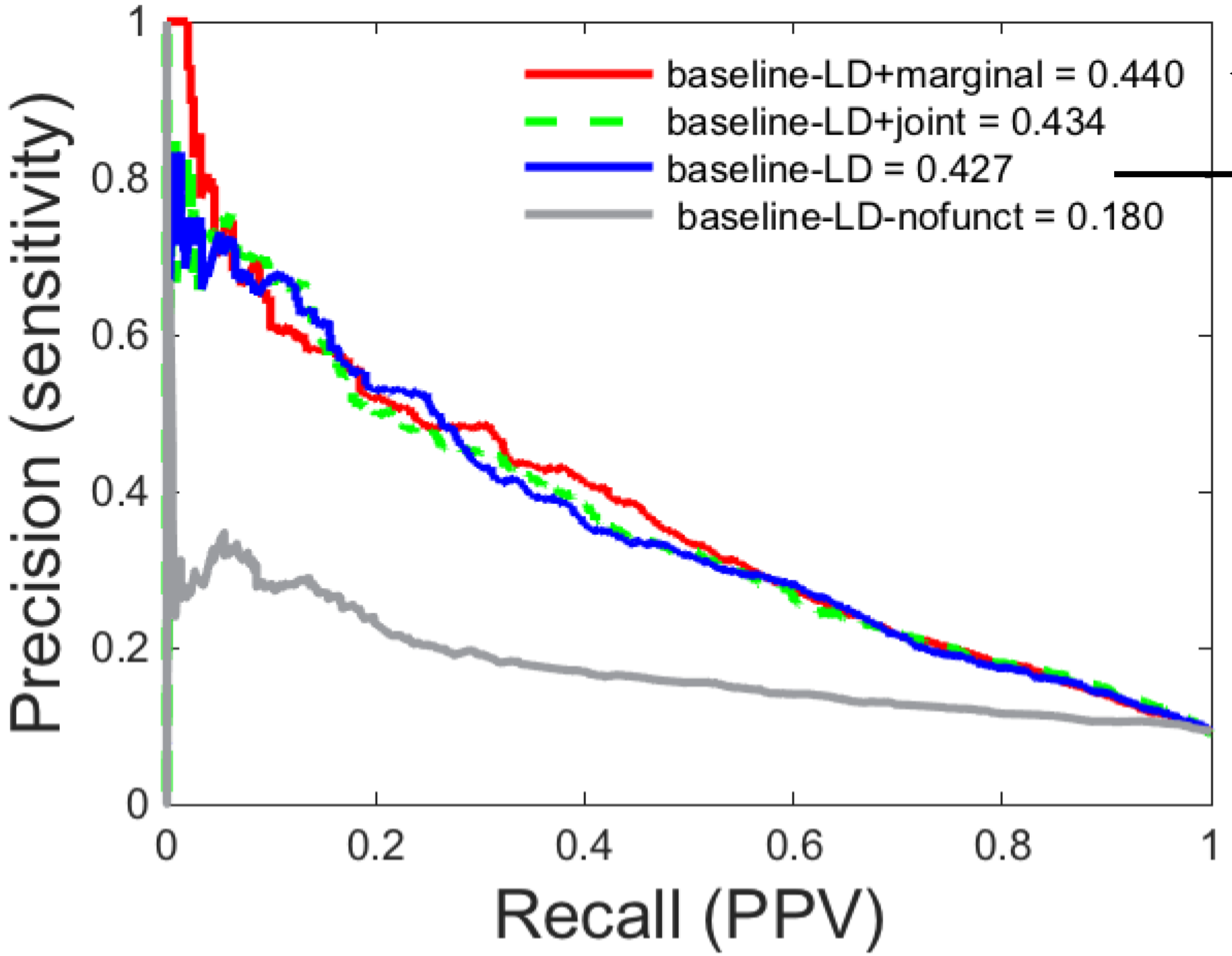
- baseline-LD: 86 existing annotations
- baseline-LD+joint: +11 new jointly significant annotations
- baseline-LD+marginal: +64 new marginally significant annotations
- Improvement: relative to baseline-LD-nofunct (only MAF/LD annotations)

Score	# scores	# marginally significant annotations		# significant annotations in a combined joint model	
		published	boosted	published	boosted
Mendelian missense	11	2*	10	1*	2
Genome-wide Mendelian	6	3	6	2	3
Additional scores	18	6**	13	0**	0
Baseline-LD model annotations	47	n/a	24	n/a	3

82 scores analyzed   64 new annotations   11 new annotations

Improved heritability model better predicts disease-associated fine-mapped SNPs by +4.9% to +21.3%

Weissbrod et al. fine-mapped SNPs across 49 UKBB traits



+4.9%, relative to baseline-LD-nofunct

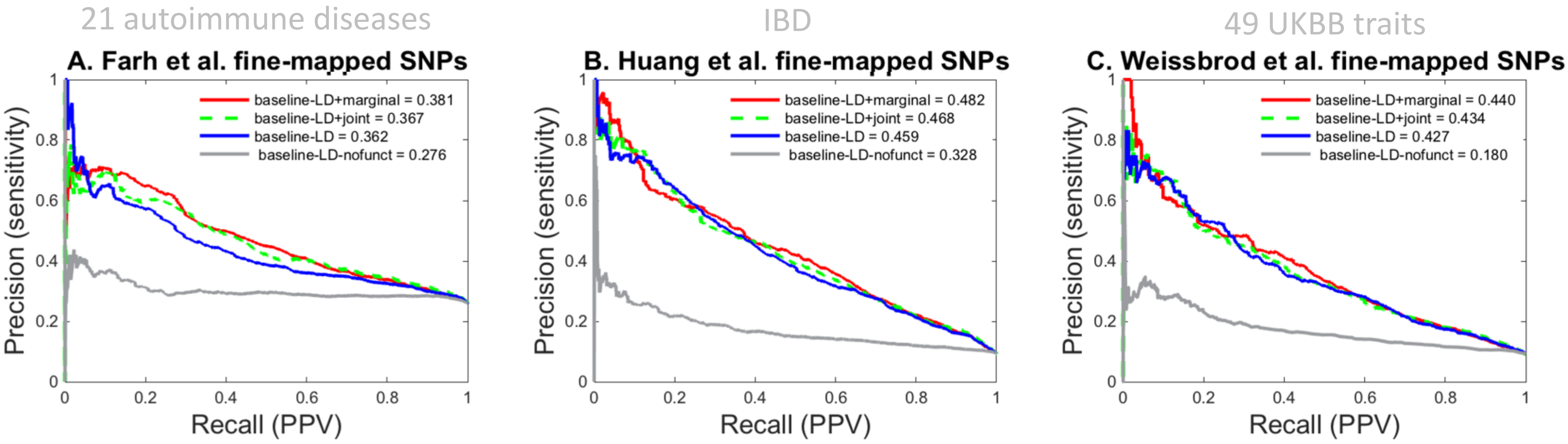
+21.3% using Farh et al. fine-mapped SNPs for ADs

+17.8% using Huang et al. fine-mapped SNPs for IBD

(Weissbrod et al. biorxiv 2020 – accepted in principle Nat Genet)

ASHG
American Society of Human Genetics

Improved heritability model better predicts disease-associated fine-mapped SNPs by +4.9% to +21.3%



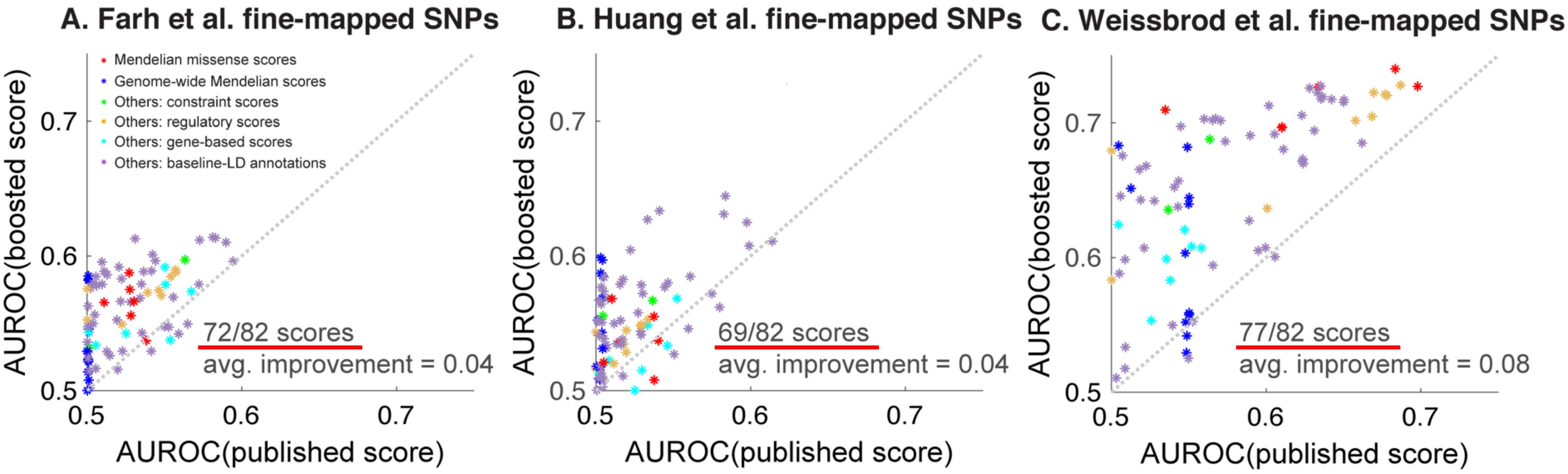
P(baseline-LD, baseline-LD+marginal) < 1e-100

- baseline-LD+marginal significantly improves classification accuracy of fine-mapped SNPs

(Farh et al. Nature 2015, Huang et al. Nature 2017, Weissbrod et al. biorxiv 2020 – accepted in principle Nat Genet)

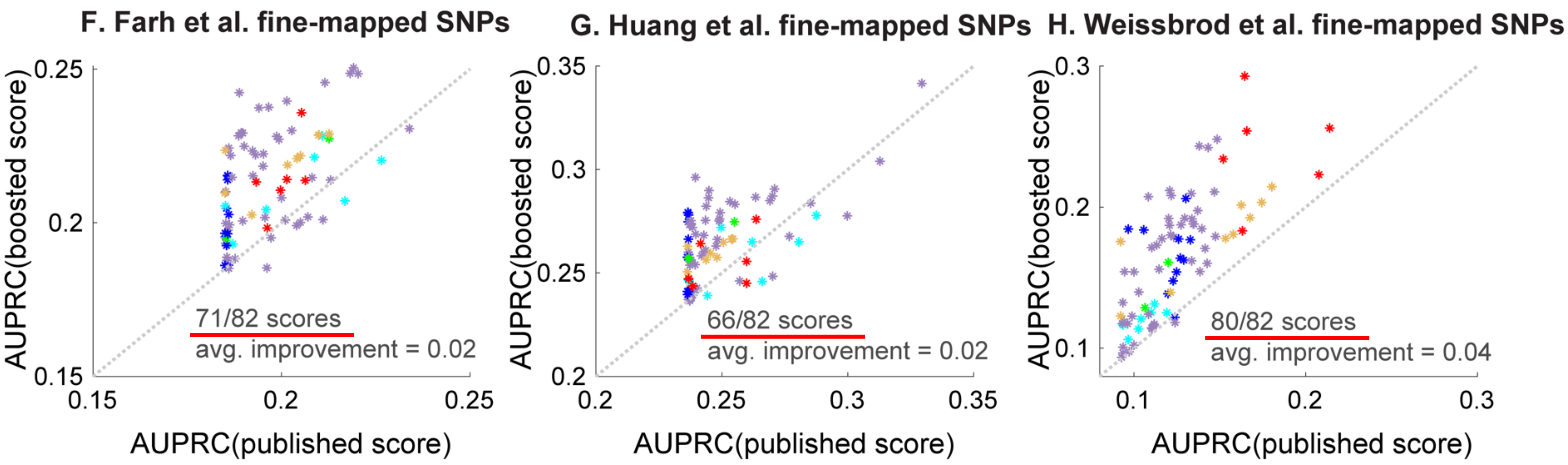
Boosted scores better classifies fine-mapped SNPs

- Compared 82 published vs. 82 boosted scores in classifying fine-mapped SNPs from LD-, MAF-, genomic-element-matched control SNPs.
- r(AUROCs, S-LDSC τ*) = 0.38 – 0.48

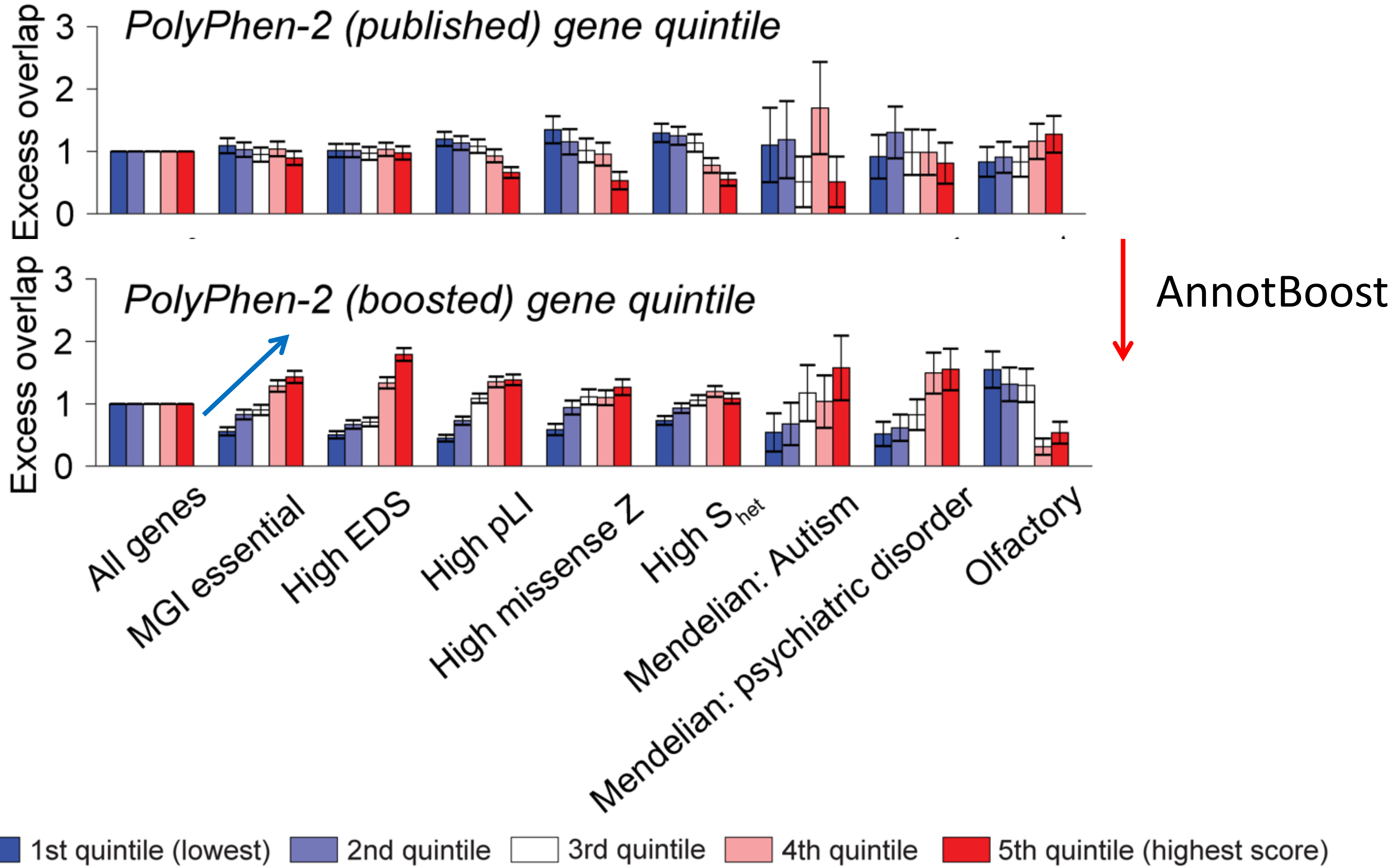


Boosted scores better classifies fine-mapped SNPs

- Similar findings using AUPRCs instead of AUROCs.



AnnotBoost can help identify biologically important genes



(Lek et al. Nature 2016, Samocha et al. Nat Genet 2014, Cassa et al. Nat Genet 2017)