

CAGI slide pack guidelines

I agree to my slides being used according to the following terms: _____ *Author Name*

Here _____ *Sunduz Keles* _____

- 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 2013. 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.



CAGI causal eQTL prediction challenge

Sunyoung Shin

Rene Welch

[Sündüz Keleş](#)

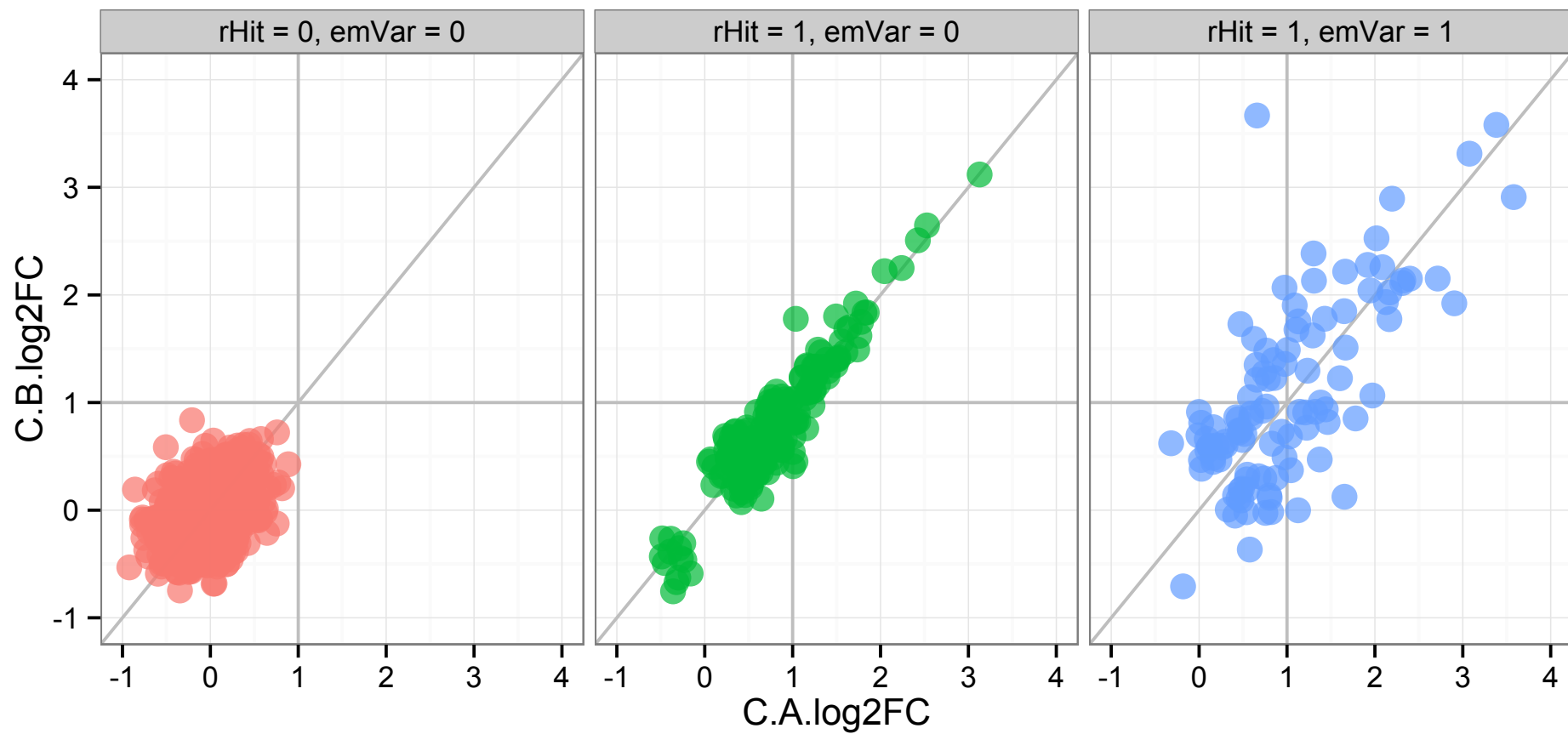
Department of Statistics

Department of Biostatistics and Medical Informatics

University of Wisconsin, Madison

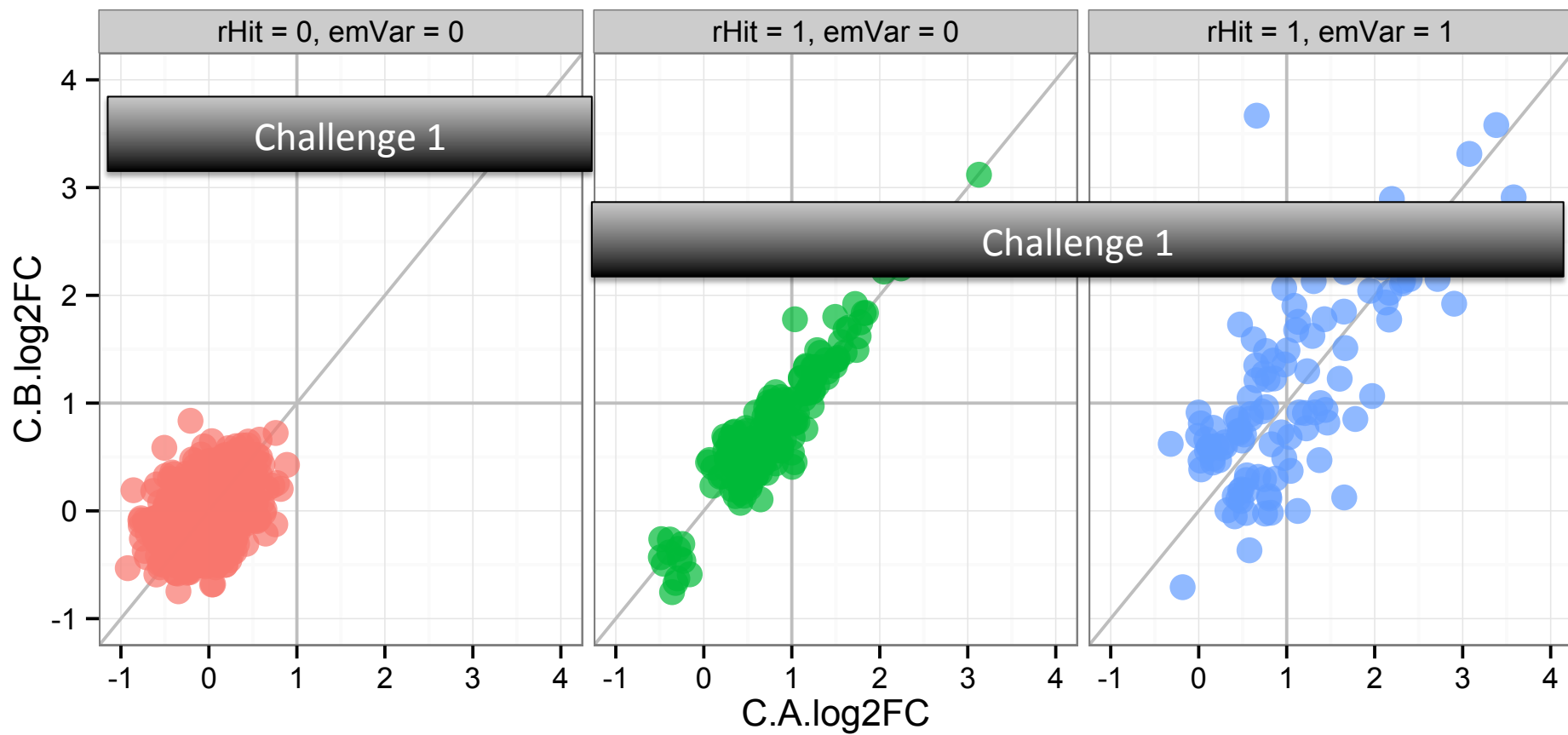
Data

● rHit = 0, emVar = 0 ● rHit = 1, emVar = 0 ● rHit = 1, emVar = 1



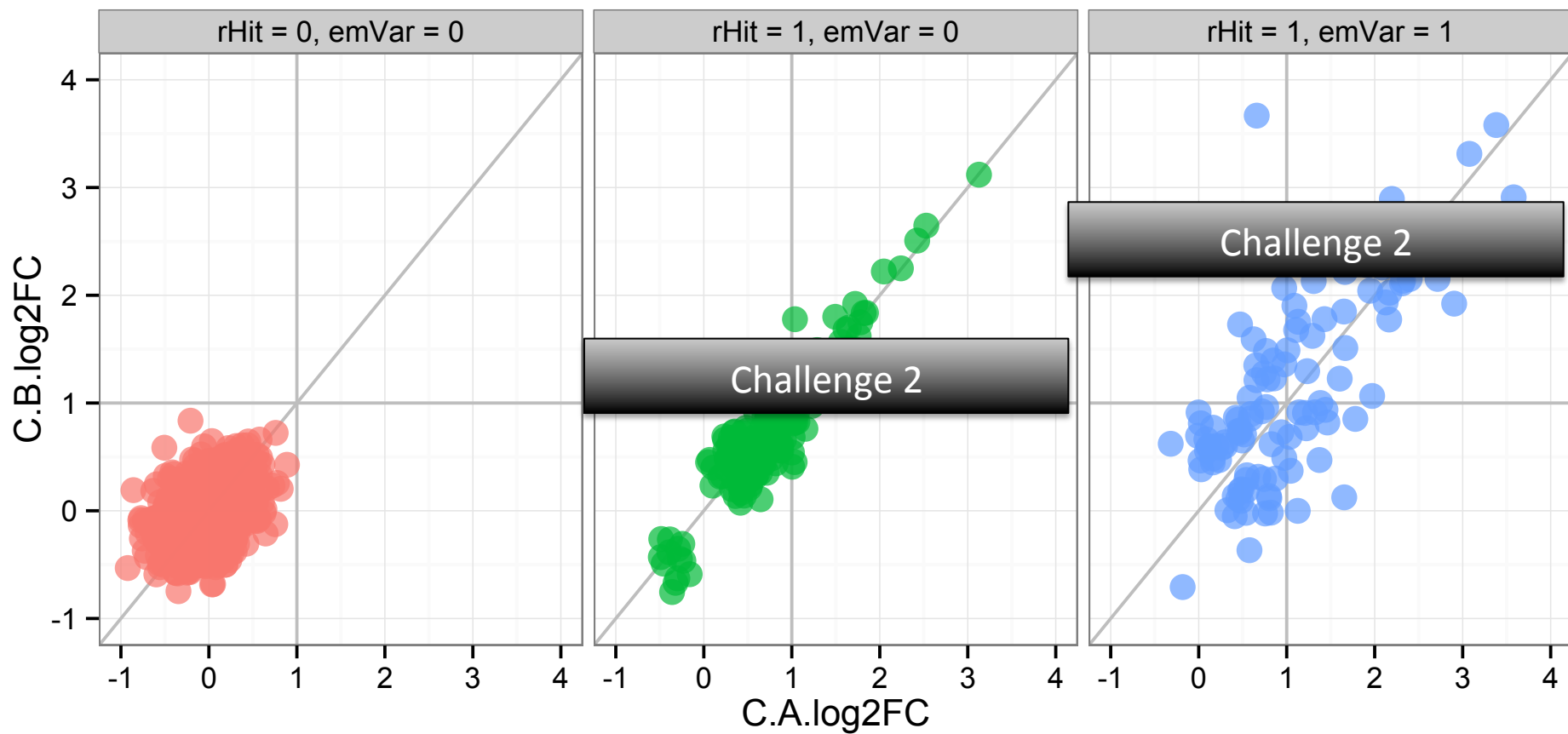
Data

● rHit = 0, emVar = 0 ● rHit = 1, emVar = 0 ● rHit = 1, emVar = 1

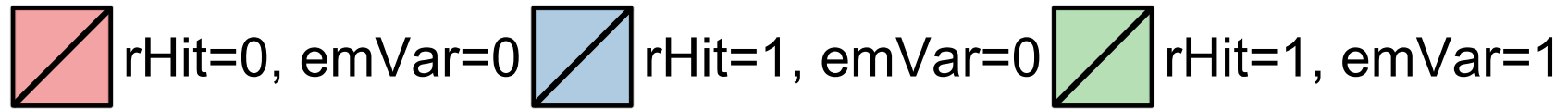


Data

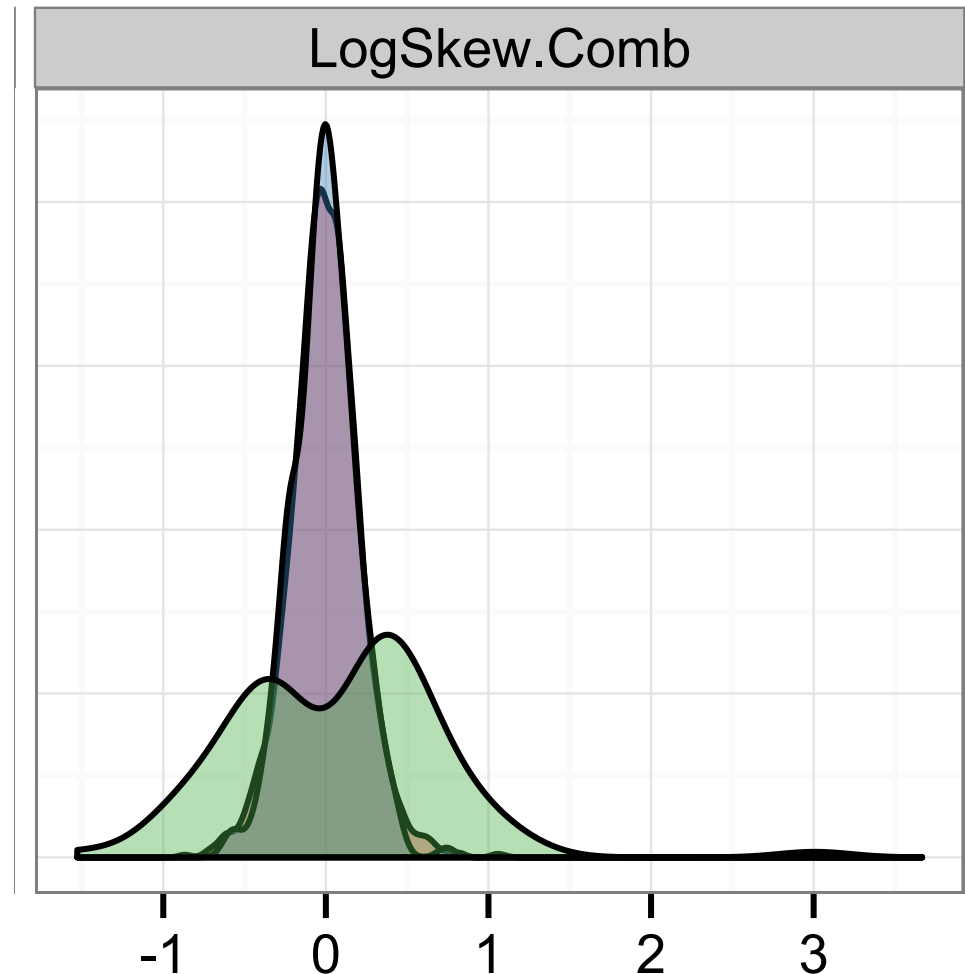
● rHit = 0, emVar = 0 ● rHit = 1, emVar = 0 ● rHit = 1, emVar = 1



Exploratory analysis of the response



Log ratio of the relative
enrichment of the Ref
and the variant



Features & Prediction

- Sequence features: Static
- Chromatin features: Cell type specific
- Random forest

Features: Sequence TF motif match (C1) Impact on TF motif match (C2)

... ACAATTAGAAGGA**G**TTCGA...
... ACAATTAGAAGGA**A**TTCGA...

C1



FIMO (MEME Suite)

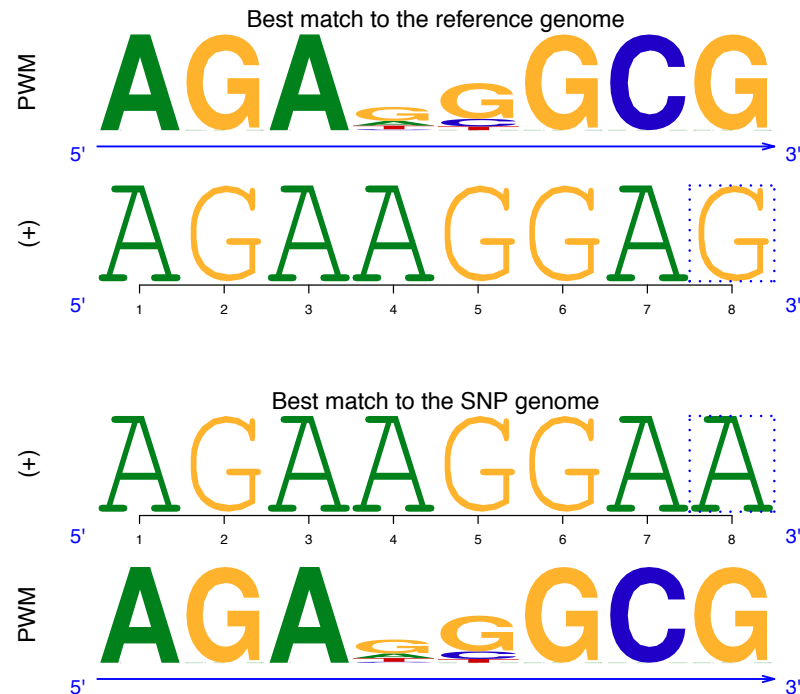
DeepBind (Alipanahi et al., 2015)

Features: Sequence TF motif match (C1) Impact on TF motif match (C2)

... ACAATTAGAAGGA^GTCGA...

... ACAATTAGAAGGA^ATCGA...

C2



Features: Sequence

TF motif match (C1)

Impact on TF motif match (C2)

PWM scans with
FIMO using
motif libraries

Deep learning
methods
(DeepBind)

UniPROBE
8mers

ENCODE
JASPAR
FACTORBOOK

Features:

GM12878 specific chromatin features from ENCODE

GM12878
Methylation

GM12878
DNase

GM12878
Histone
Modifications

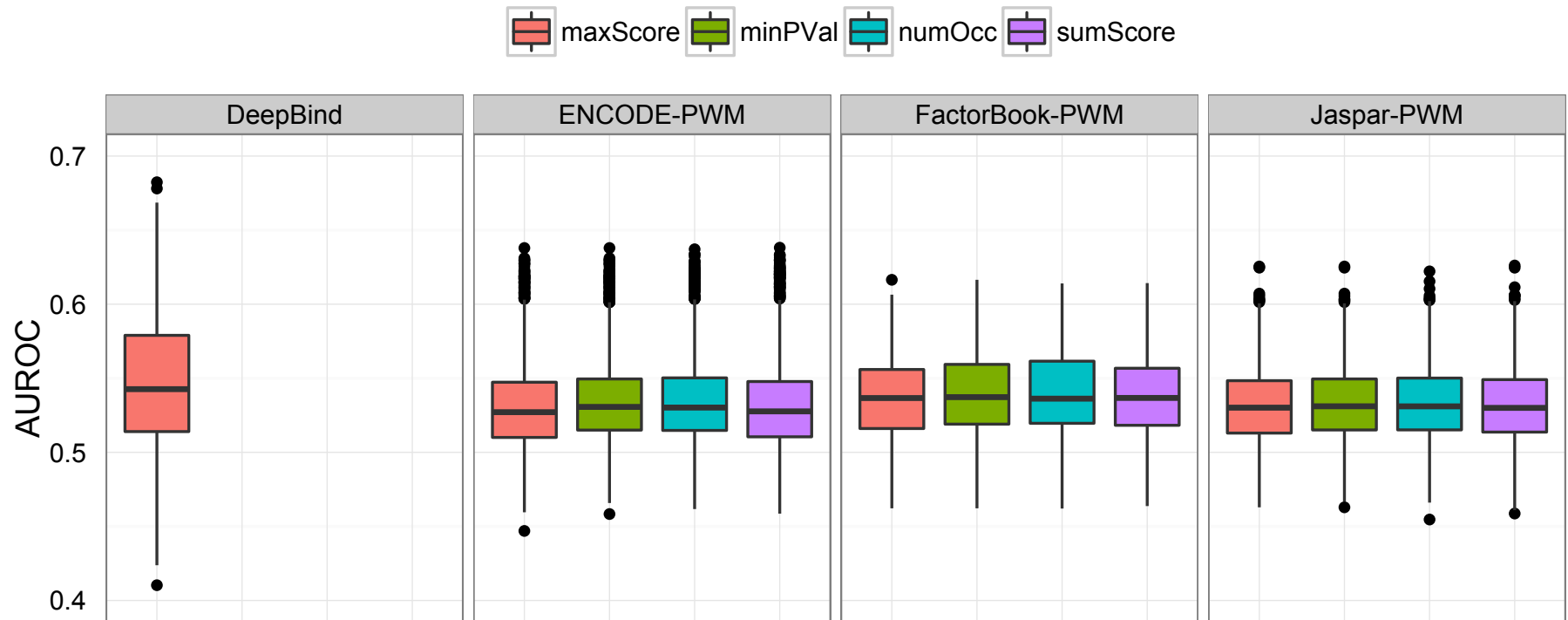
GM12878
TF Occupancy
(aggregated)

Features

Data type	Number of features	Description	Variations		Summary between Ref and Alt
Position weight matrix scores	2065 (ENCODE) 1082 (JASPAR) 79 (FACTORBOOK)	Scan the entire oligo construct with PWMs using FIMO	maxScore minP numOcc sumScore	ENCODE JASPAR FACTORBOOK	C1: Max C2: Diff
UniProbe k-mer scores	320	Use UniPROBE scores of the kmers in the oligo construct	sum, mean, max. sum_bp21, mean_bp_21, max_bp_21		C1: Max C2: Diff
GM12878 Specific ENCODE Features	13 (Methy DNase 10 Histone TF aggregate)	Overlap genomic coordinates of the oligo constructs with ENCODE processed data	Original peak width +/- 150 bps around the summit Number of overlaps Any overlap		
DeepBind	927	Apply DeepBind with their provided models			C1: Max C2: Diff

PWM and variants

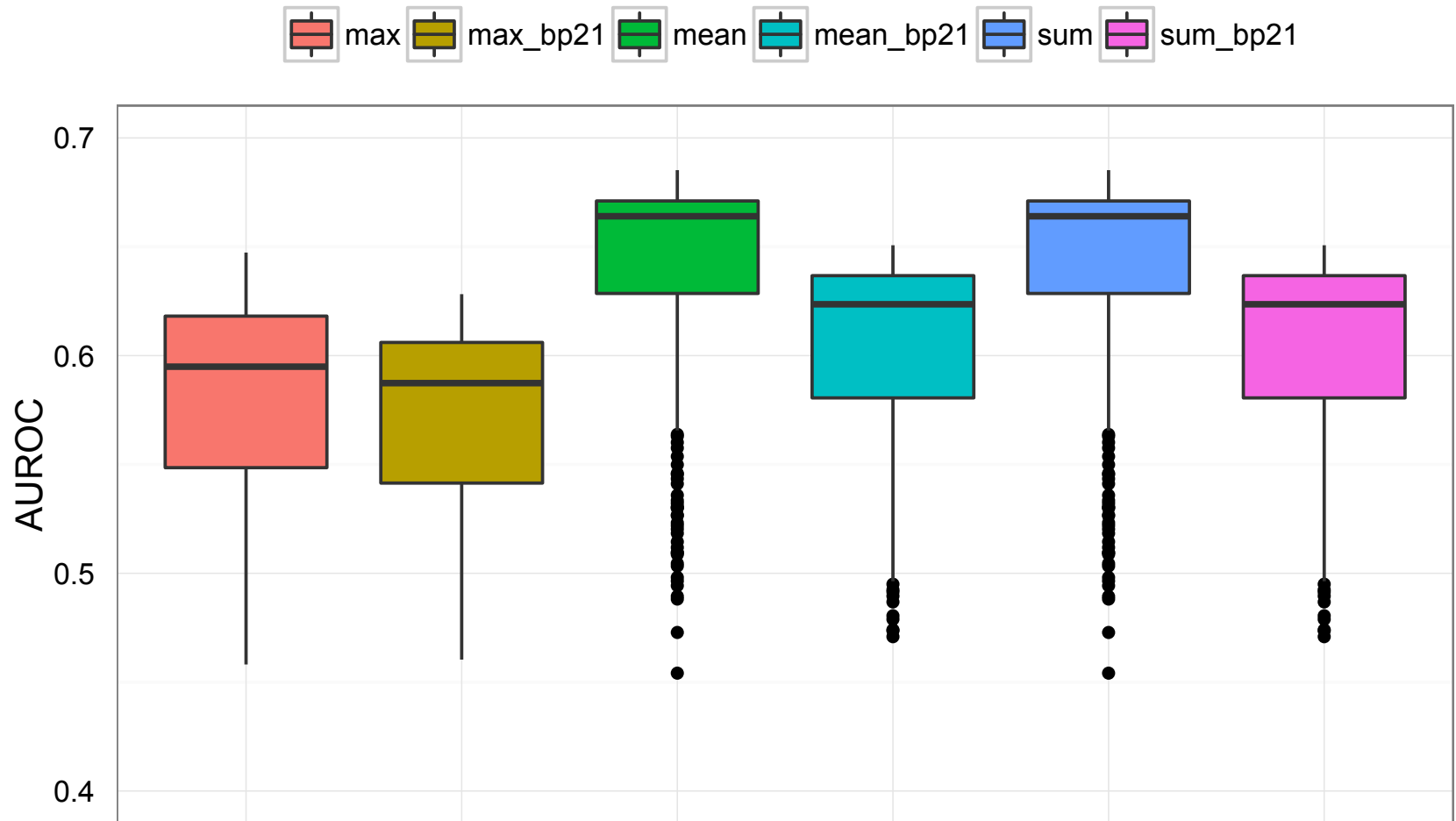
Challenge 1: One feature at a time



~0.52-0.54

UniPROBE k-mers (8mers)

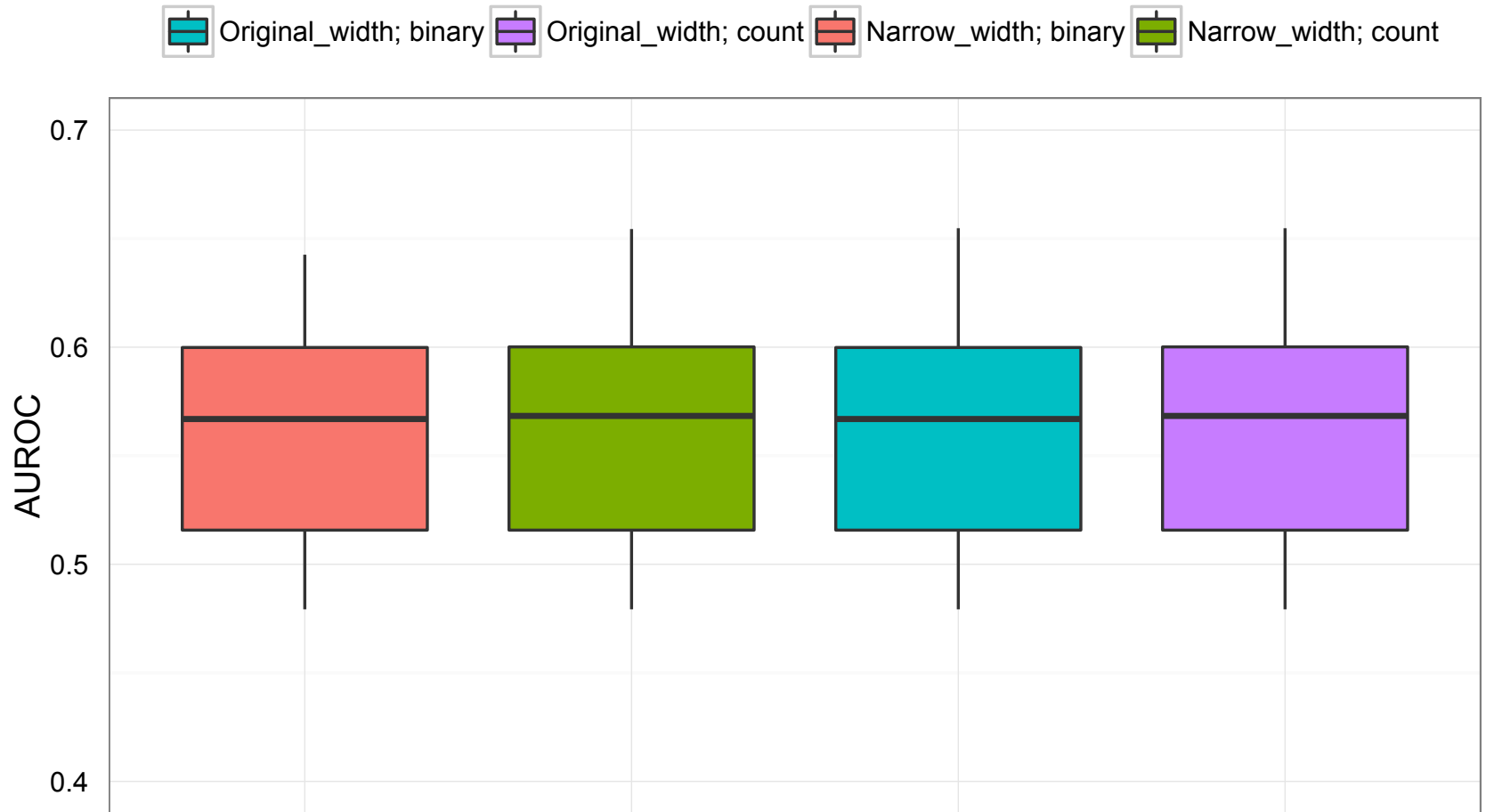
Challenge 1: One feature at a time



~0.59-0.66

ENCODE GM12878 TF/Histone/DNase/Methy

Challenge 1: One feature at a time



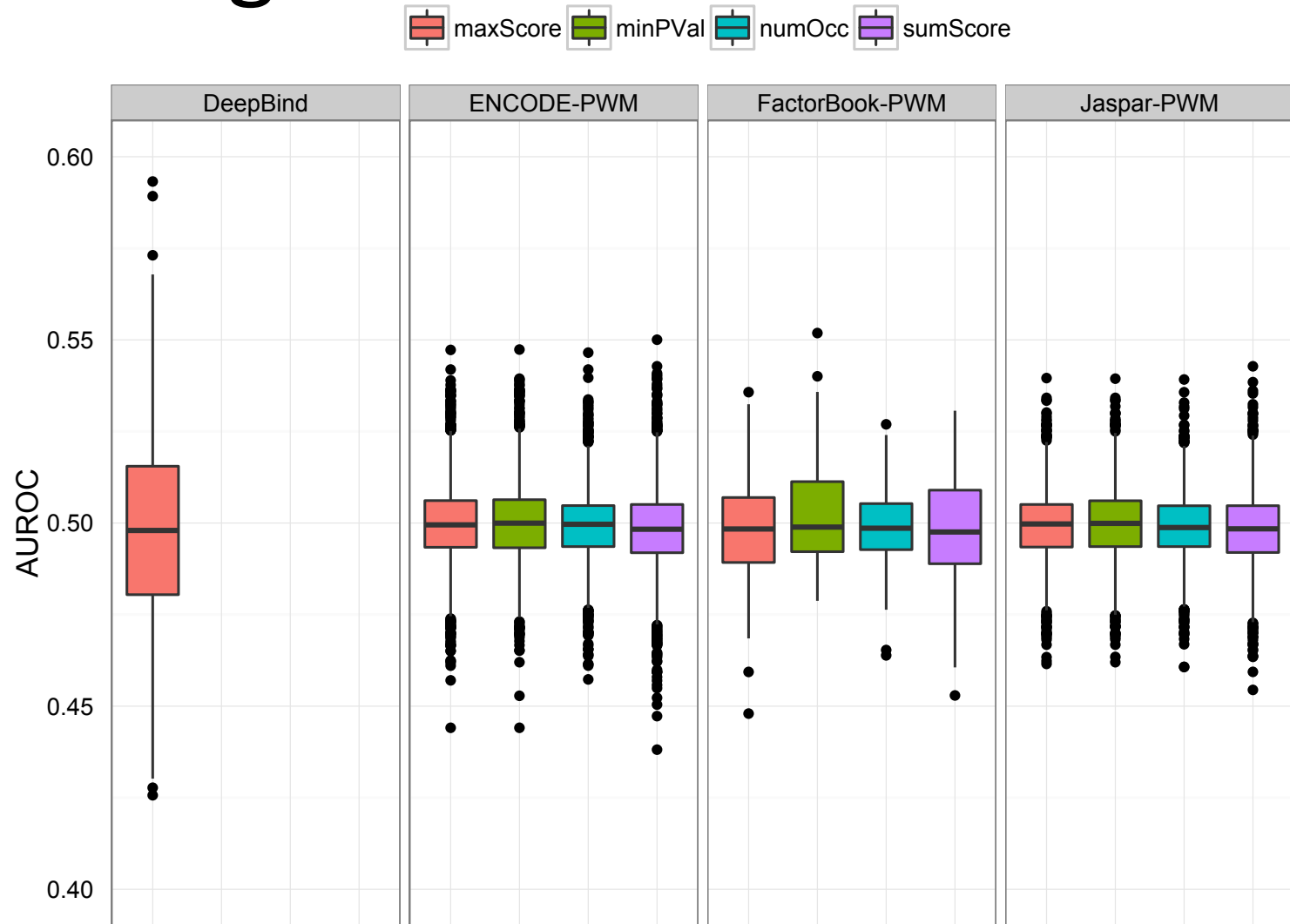
~0.57

C1: Results: predict rHit w Random Forest

	AUROC	AUPR
UniPROBE	0.6888	0.2712
GM12878 (ENCODE) Chromatin	0.6559	0.3086
DeepBind	0.7619	0.4054
UniPROBE+DeepBind	0.7656	0.3989
DeepBind+GM12878 (ENCODE) Chromatin	0.7766	0.4395
UniPROBE+DeepBind +GM12878 (ENCODE) Chromatin	0.7803	0.4483
UniPROBE+ENCODE PWM +GM12878 (ENCODE) Chromatin	0.7867	0.4584

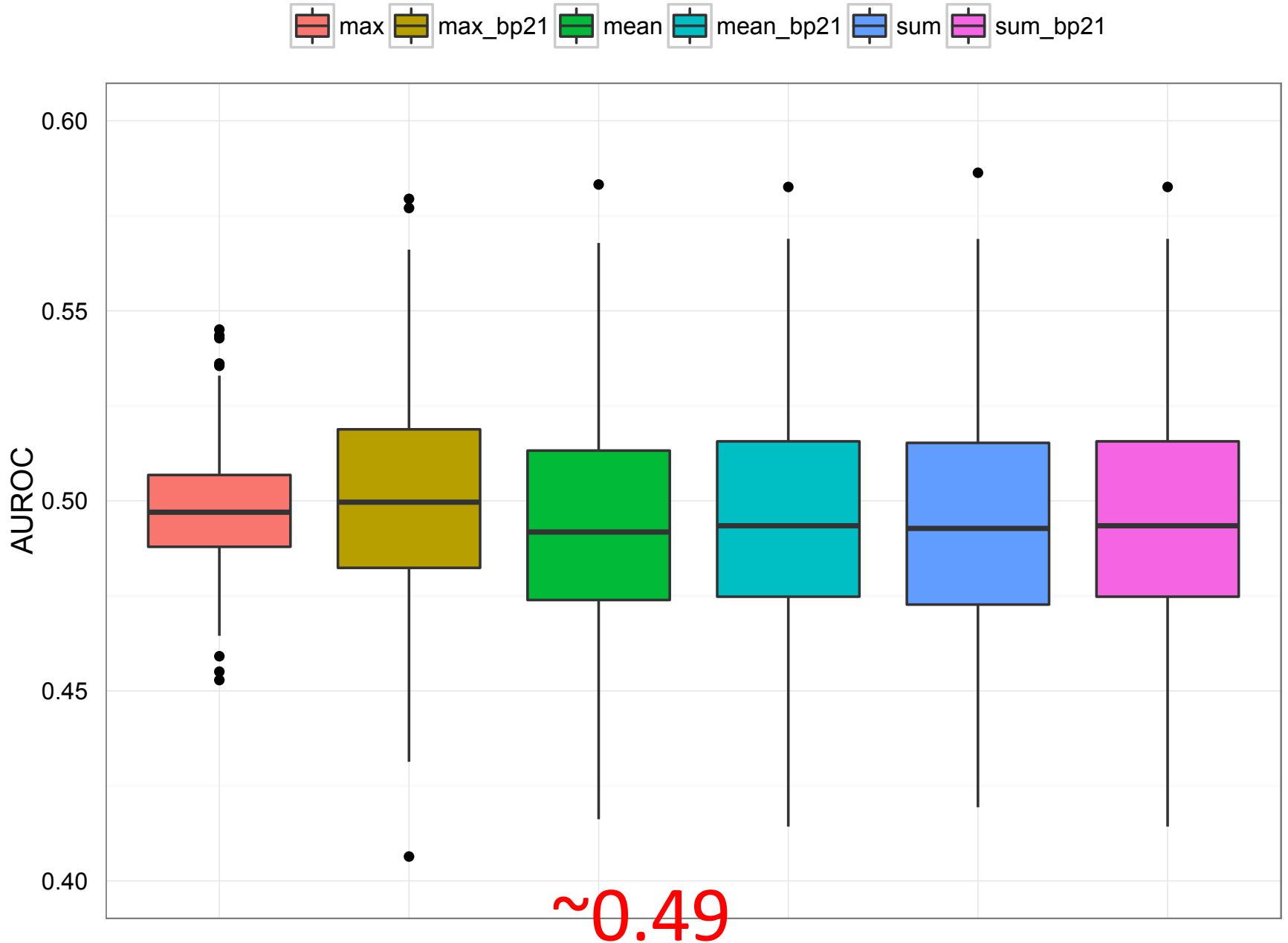
PWM and variants

Challenge 2: One feature at a time

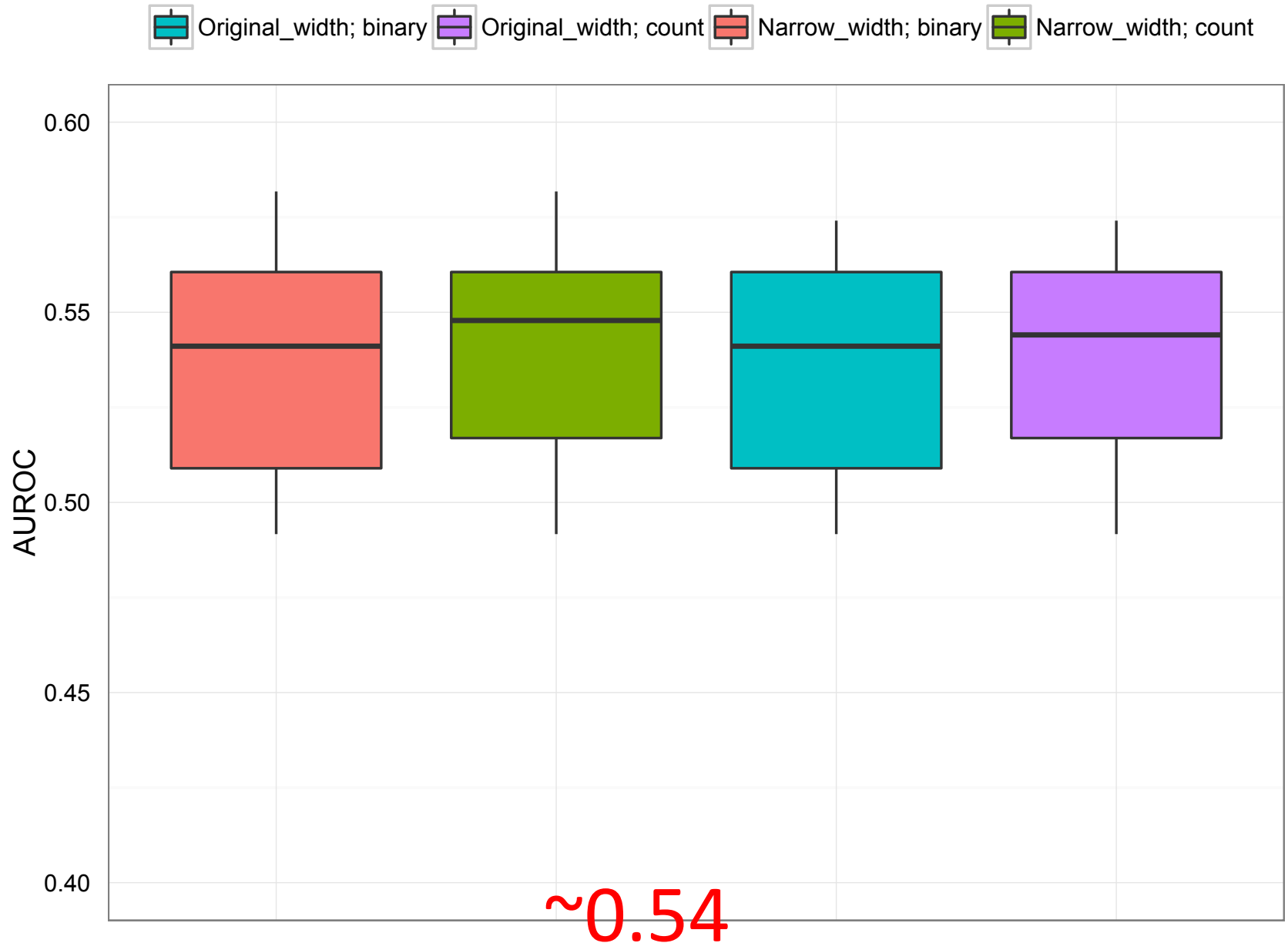


~0.49

UniPROBE k-mers (8mers)



ENCODE GM12878 TF/Histone/DNase/Methy



Results: Challenge 2: predict emVar w Random Forest

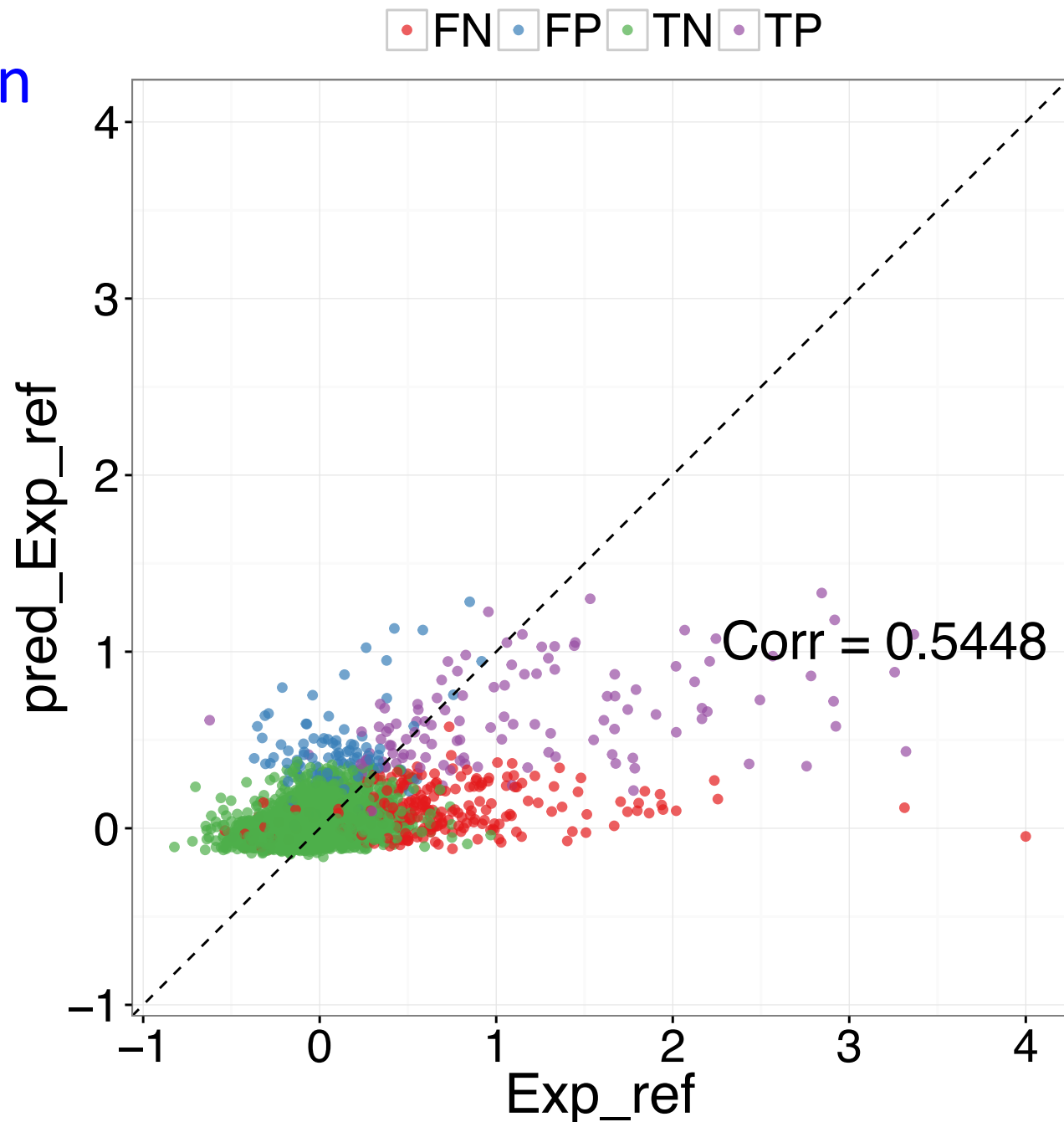
Model	AUROC	AUPR
DeepBind	0.521	0.311
ENCODE_PWM (sum)+ GM12878 (ENCODE) Chromatin	0.562	0.344
UniPROBE (Sum)	0.579	0.330
GM12878 (ENCODE) Chromatin	0.570	0.319
ENCODE_PWM (Sum)+ UniPROBE (Sum)	0.594	0.368
ENCODE_PWM (sum)+ GM12878 (ENCODE) Chromatin	0.562	0.344

Results: Challenge 2: predict emVar w Random Forest

Model	AUROC	AUPR
DeepBind	0.560	0.374
ENCODE_PWM (sum)	0.603	0.399
UniPROBE (Sum)	0.612	0.406
GM12878 (ENCODE) Chromatin	0.610	0.415
UniPROBE (Sum)+ GM12878 (ENCODE) Chromatin	0.616	0.411
DeepBind+ UniPROBE (Sum)+ GM12878 (ENCODE) Chromatin	0.588	0.386

Including predicted expression change improves all models for C2.

Prediction of relative expression from challenge 1



Summary

- **Univariate analysis:**
UniPROBE features are best for rHit,
Cell-type specific chromatin features are best for emVar.
- **Multivariate analysis:**
Significant improvement for rHit,
Little improvement for emVar.
- Including **predicted expression change** improves emVar prediction.

Things we thought about but did not have time to do

- **Expression levels** of TFs in the two GM cell lines so that the model can utilize interactions.
- **DNA shape** - initial exploration did not seem promising.
- **Boosting** and acknowledgement of **correlated predictors**.
- **LD information** – SNPs are clustered.

Acknowledgements

- Sunyoung Shin, Ph.D.
- Rene Welch
- Jose A. Rodriguez-Martinez, Ph.D. (Ansari Group, UW Madison)