

CAGI slide pack guidelines

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

_____ *Mabel Furutsuki & Wyatt Clark* _____

- 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.



Evaluation of predictions of the impact of missense mutations on enzymatic activity in the GAA Challenge (CAGI 5)

Mabel Furutsuki

Mentor: Gaia Andreoletti

Brenner Lab, University of California, Berkeley

CAGI5 Conference, Chicago, IL

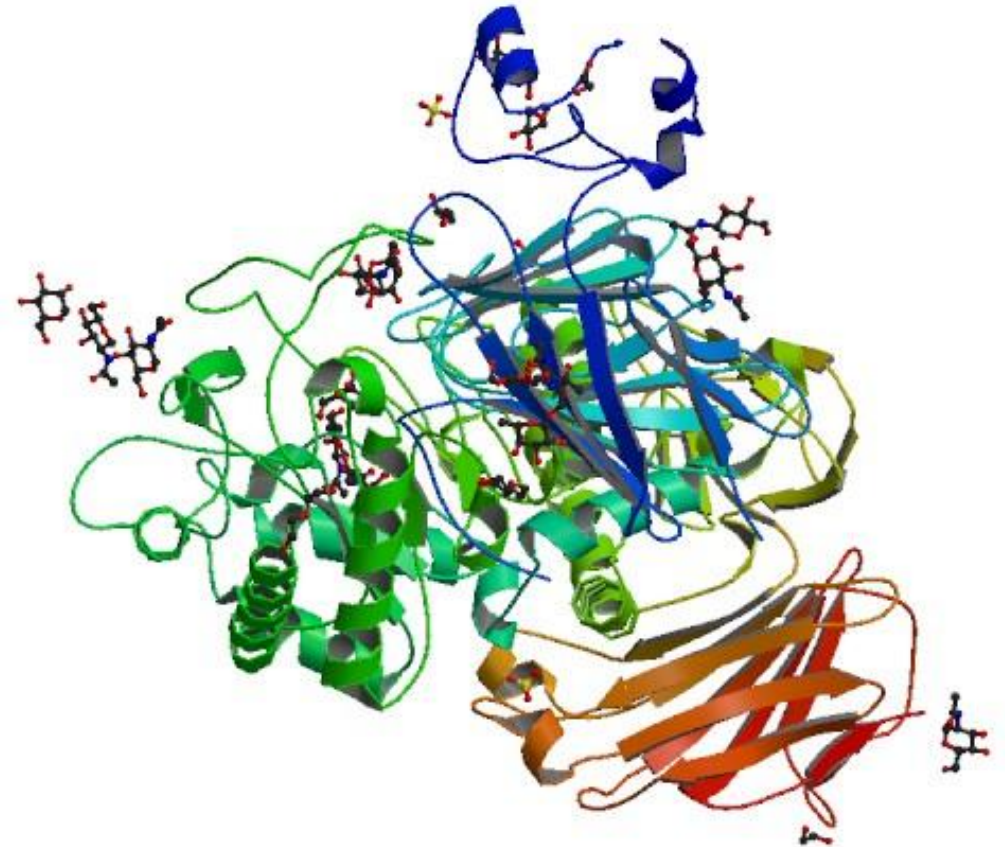
5th July, 2018

GAA Challenge: Overview

- BioMarin has functionally assessed the enzymatic activity (in transfected cell lysates) of the 357 novel missense mutations in the ExAC dataset
- ExAC dataset contains a large number of missense mutations whose contribution to disease incidence is unknown
- Participants predicted effects of missense mutations in GAA on enzymatic activity
 - Participants were asked to predict a percentage of wild-type activity along with a standard deviation.
- 7 Groups and 26 Submissions
 - Lichtarge Lab (Baylor College of Medicine, United States)
 - BolognaBiocomputing (University of Bologna, Italy)
 - ANA (University of California Berkeley, United States)
 - Ostrokach (University of Toronto, Canada)
 - Mooney Radivojac 2018 (Indiana University Bloomington, United States)
 - BioFold (University of Alabama Birmingham, United States)
 - Bromberg Lab (Rutgers University, United States)

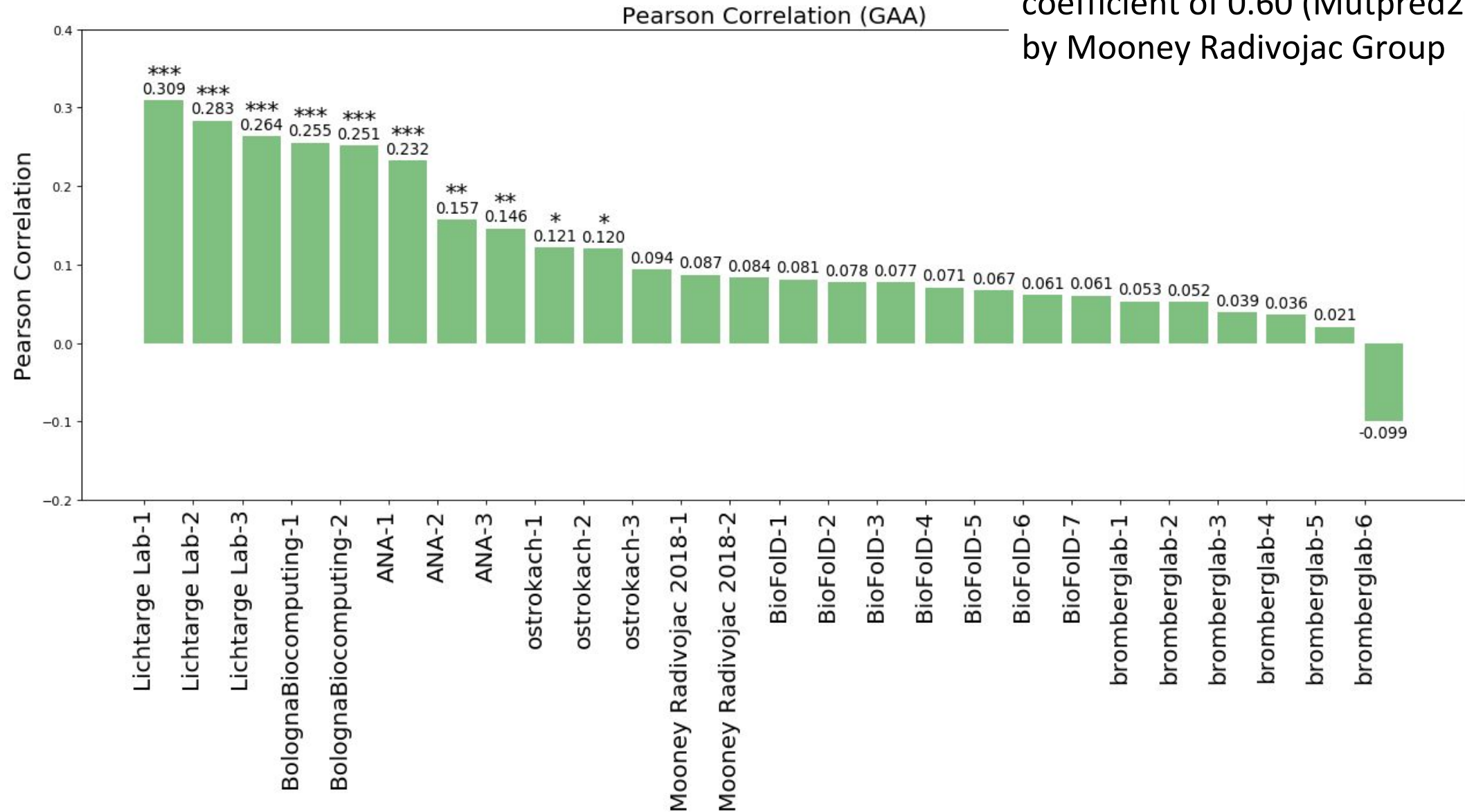
Deficiencies in GAA cause Glycogen Storage Disease II (Pompe's Disease)

- Acid alpha-glucosidase (GAA, NP_000143.2) is a lysosomal enzyme that hydrolyzes the terminal, non-reducing end of glycogen releasing a single alpha-D-glucose residue.
- Pompe's Disease can have a spectrum of severity depending on the residual activity of GAA.



GAA Challenge: Correlations

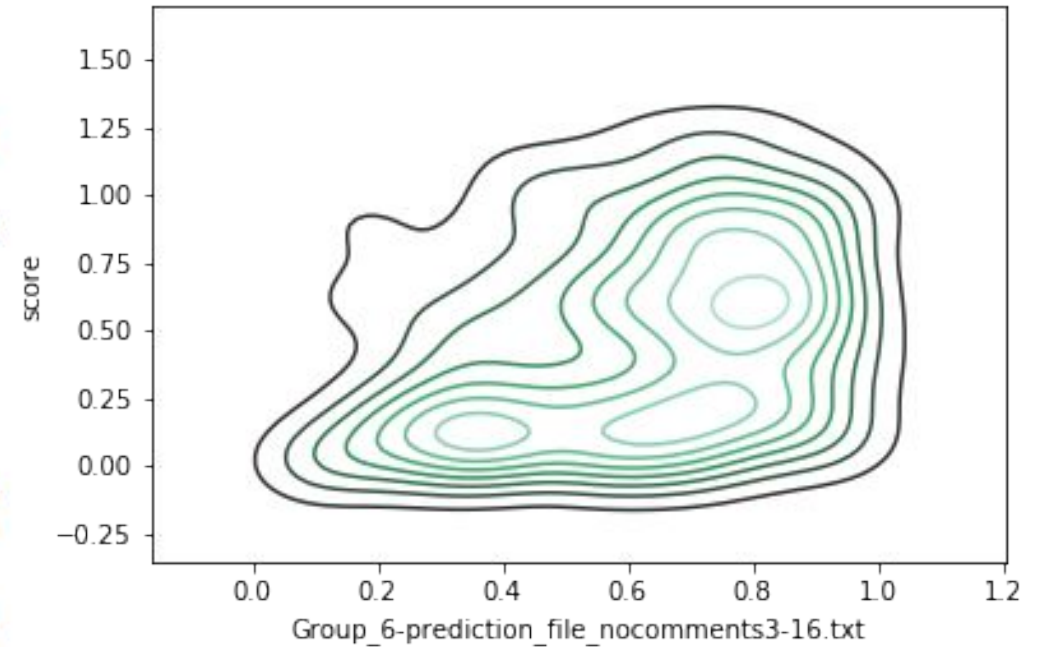
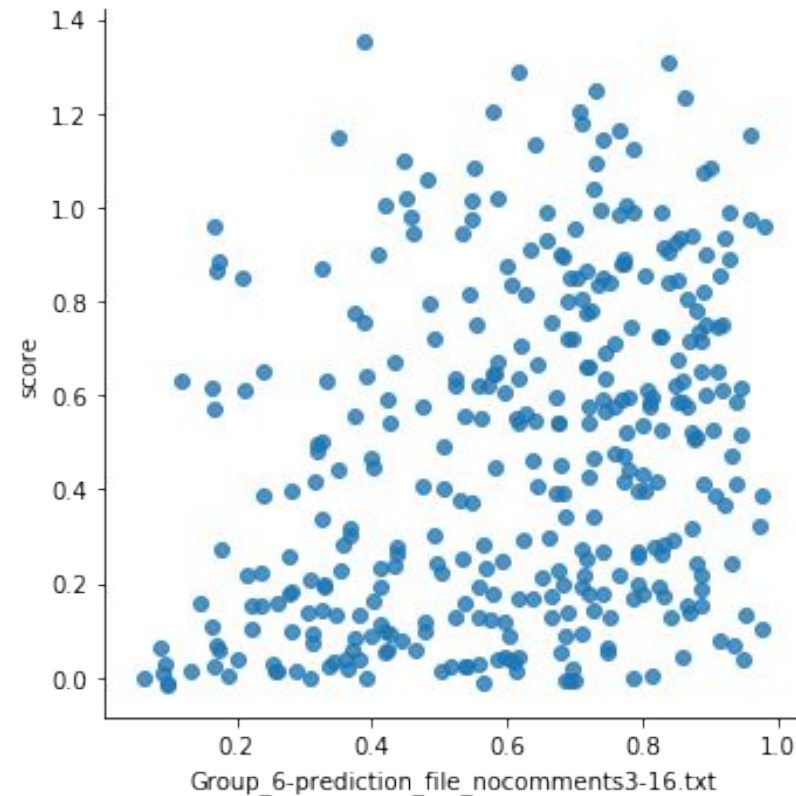
NAGLU had highest correlation coefficient of 0.60 (Mutpred2 Beta) by Mooney Radivojac Group



GAA

Group 6 File 3 Scatterplot

→ Best-performing group on all 357 variants based on correlations



Smoothing with KDE

Conclusions

- GAA correlation coefficients were observed to be lower than those of past challenge (NAGLU) - indicating much lower level of performance.
- Strong correlation of how well teams predict on variants with high vs. low experimental variation in their activity levels
 - Therefore we can identify easy/hard to predict variants
- For GAA, Group 6 submissions performed best and are highly correlated with each other, but none are highly correlated with the true value

Acknowledgements

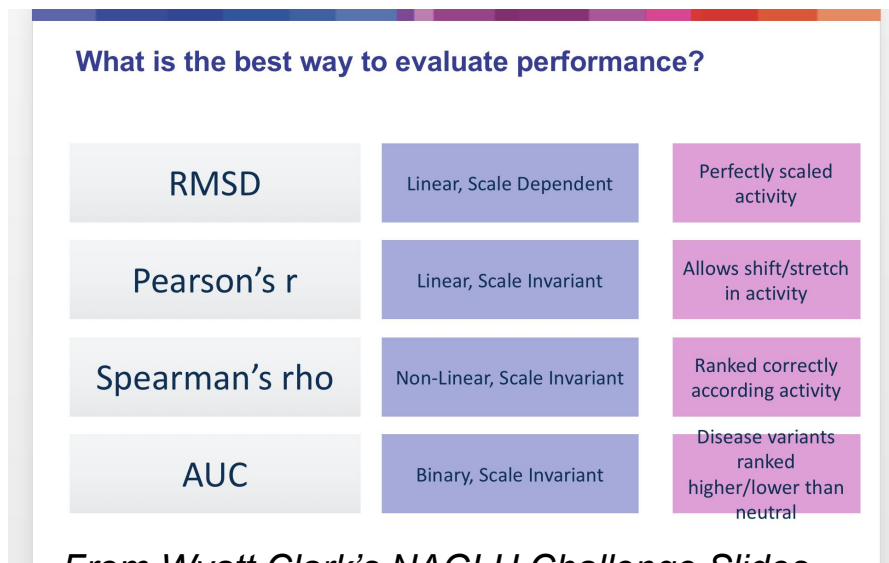
- Dr. Wyatt Clark
- Brenner Lab
- CAGI Participants and Data Providers
- Dr. Gaia Andreoletti
- Dr. Zhiqiang Hu

Extra Slides

In Progress, To Do, and For Later

R Package CAGIEval to Evaluate Performance

- Flexible package to perform submission validation and exploratory data analysis for missense mutation datasets
- Evaluates and plots RMSD, AUC/ROC (threshold between positive and negative variants is user-inputted with default 0.5), Density Distance (of 2 normal distributions), Correlations between methods, Spearman/Pearson/Kendall Correlation, PCA, Scatterplots.



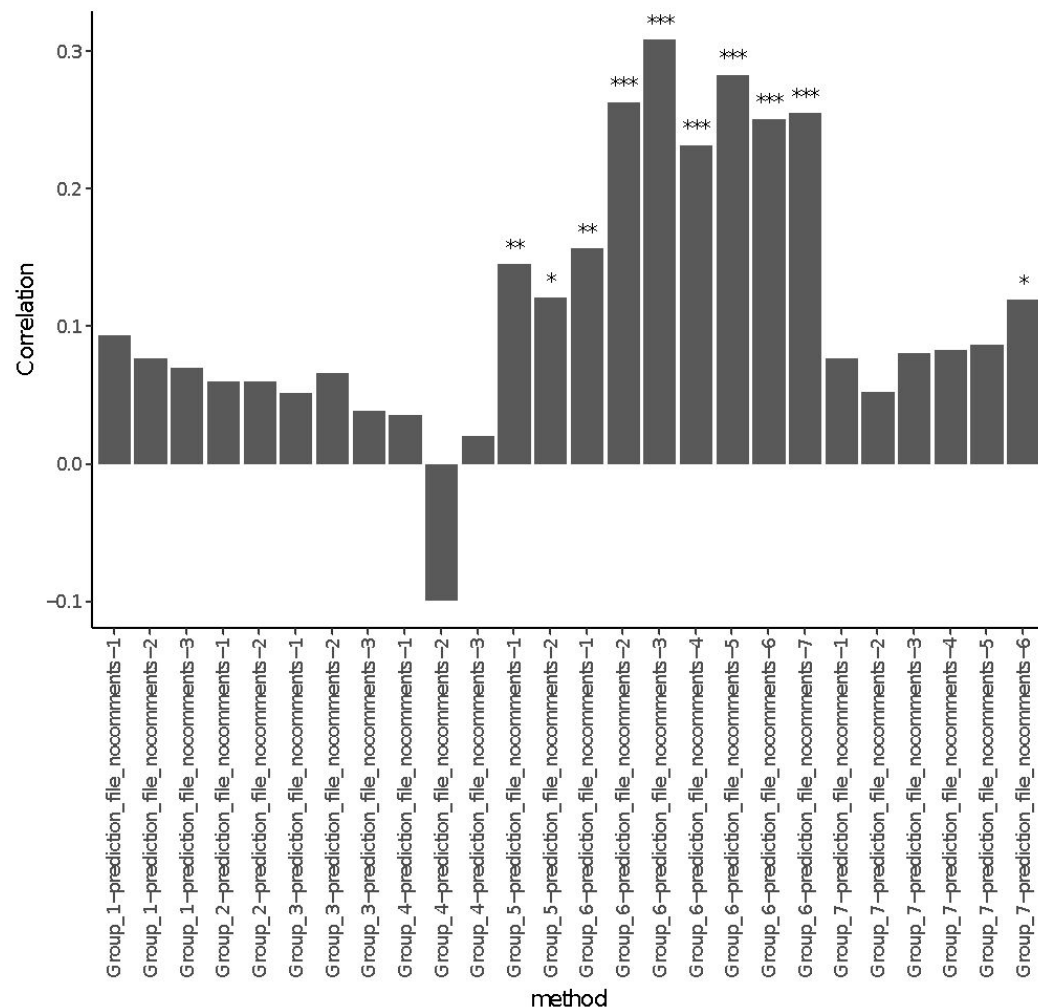
What is the best way to evaluate performance?		
RMSD	Linear, Scale Dependent	Perfectly scaled activity
Pearson's r	Linear, Scale Invariant	Allows shift/stretch in activity
Spearman's rho	Non-Linear, Scale Invariant	Ranked correctly according activity
AUC	Binary, Scale Invariant	Disease variants ranked higher/lower than neutral

From Wyatt Clark's NAGLU Challenge Slides

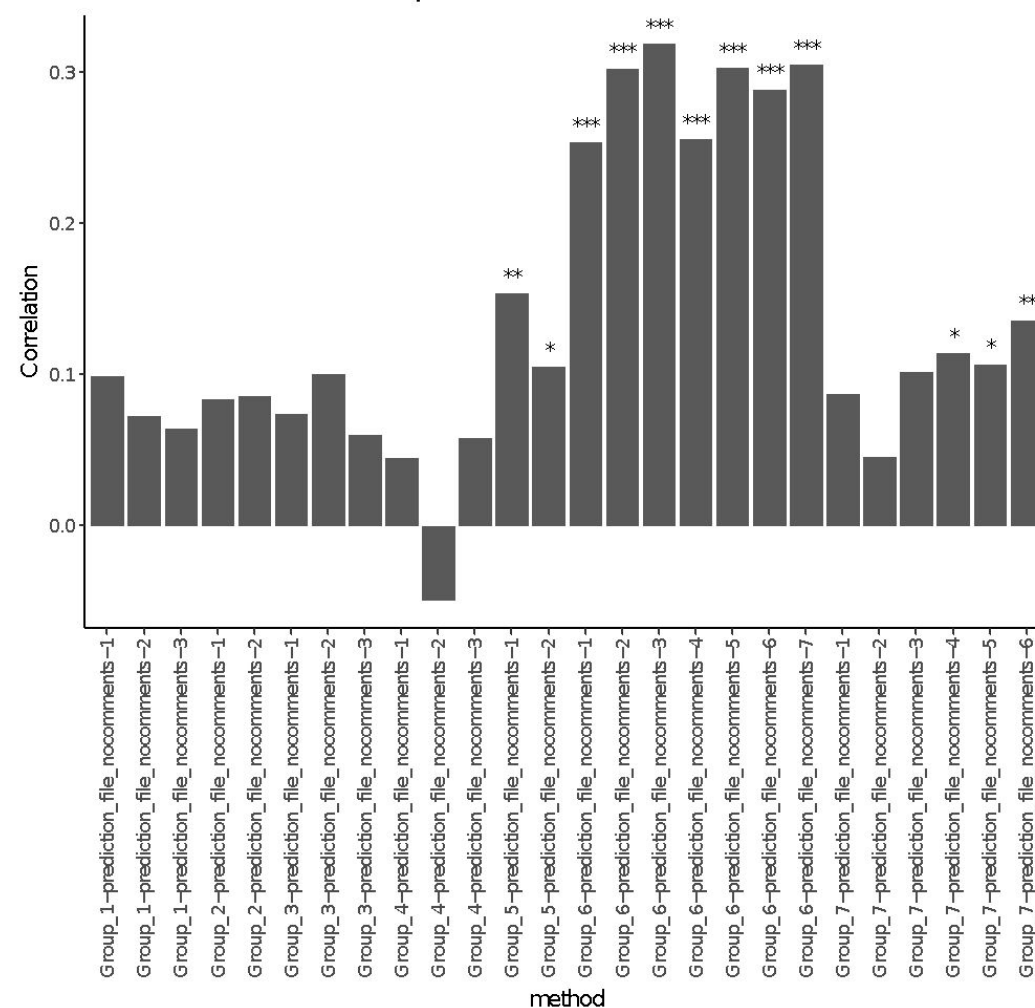
- Python scripts also used in evaluation of predictions
- The assessment of this challenge will include evaluations that recognize novelty of approach

Pearson and Spearman Correlation

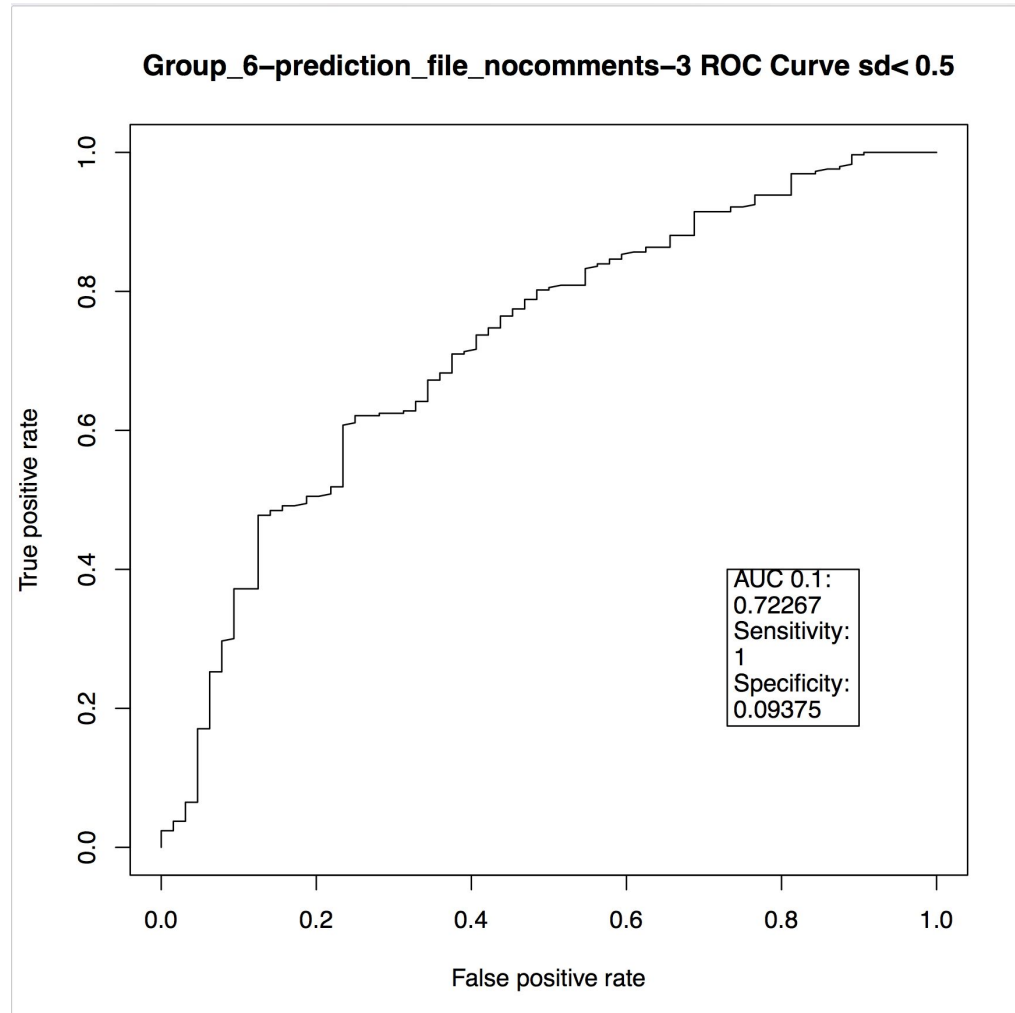
Plot of Pearson Correlation vs. Predictor



Plot of Spearman Correlation vs. Predictor



AUC/ROC Curves for Group 6-3

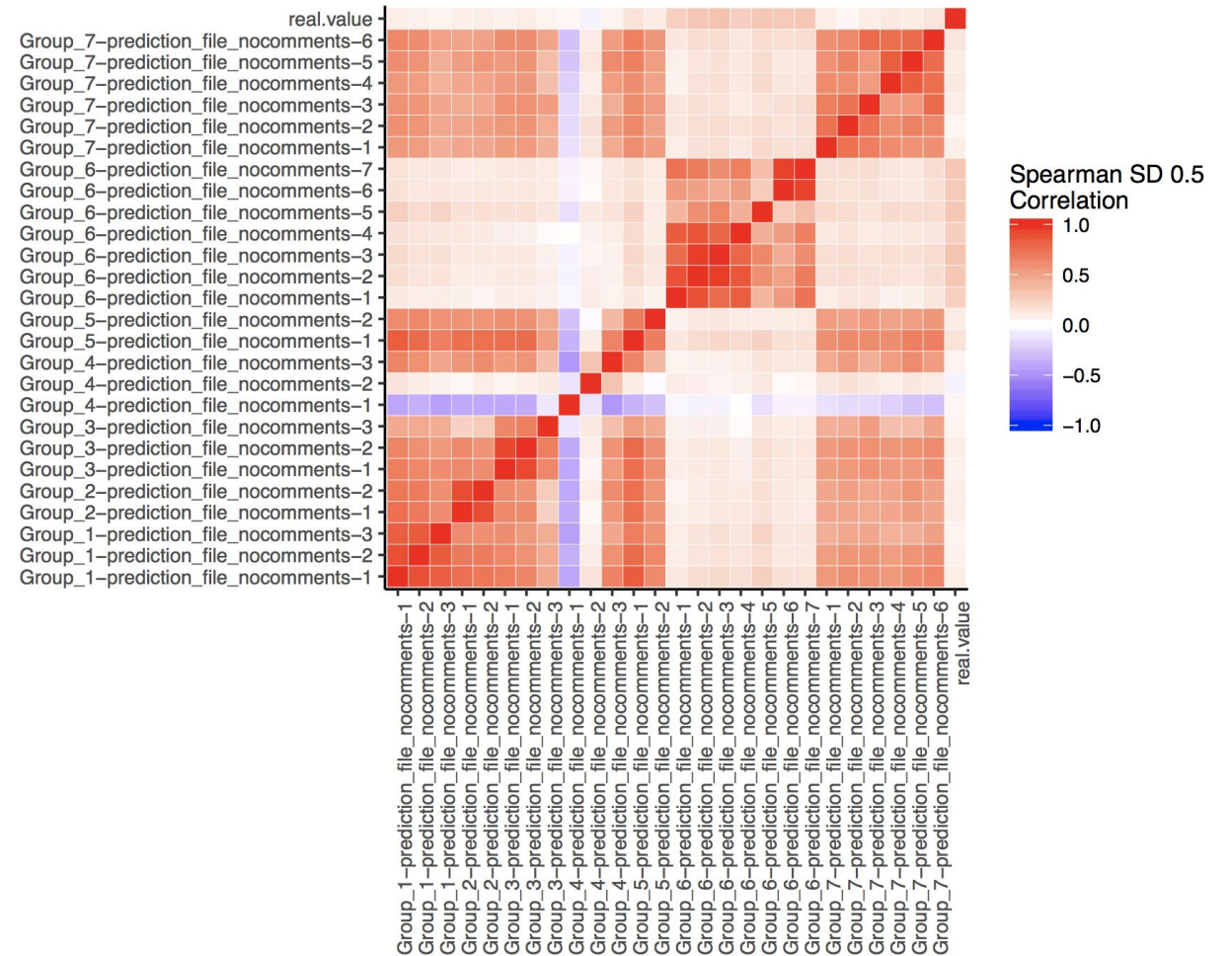


Binary Cutoff at 10% Activity

“severe” : <3%
“less_severe” :
3-10%
“mild” : 10-15%
“benign” : >15%

GAA

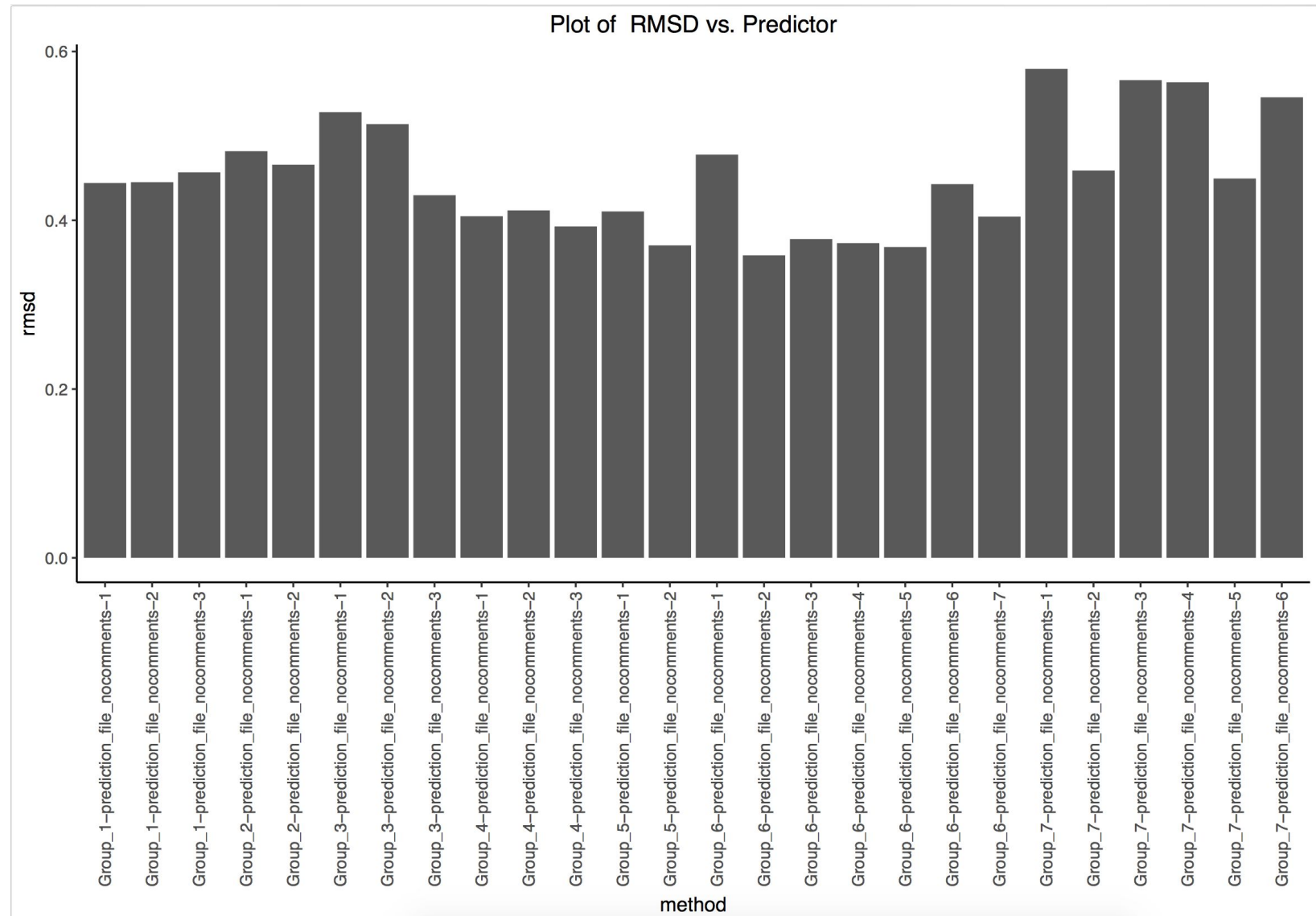
Between-Method Correlation



GAA

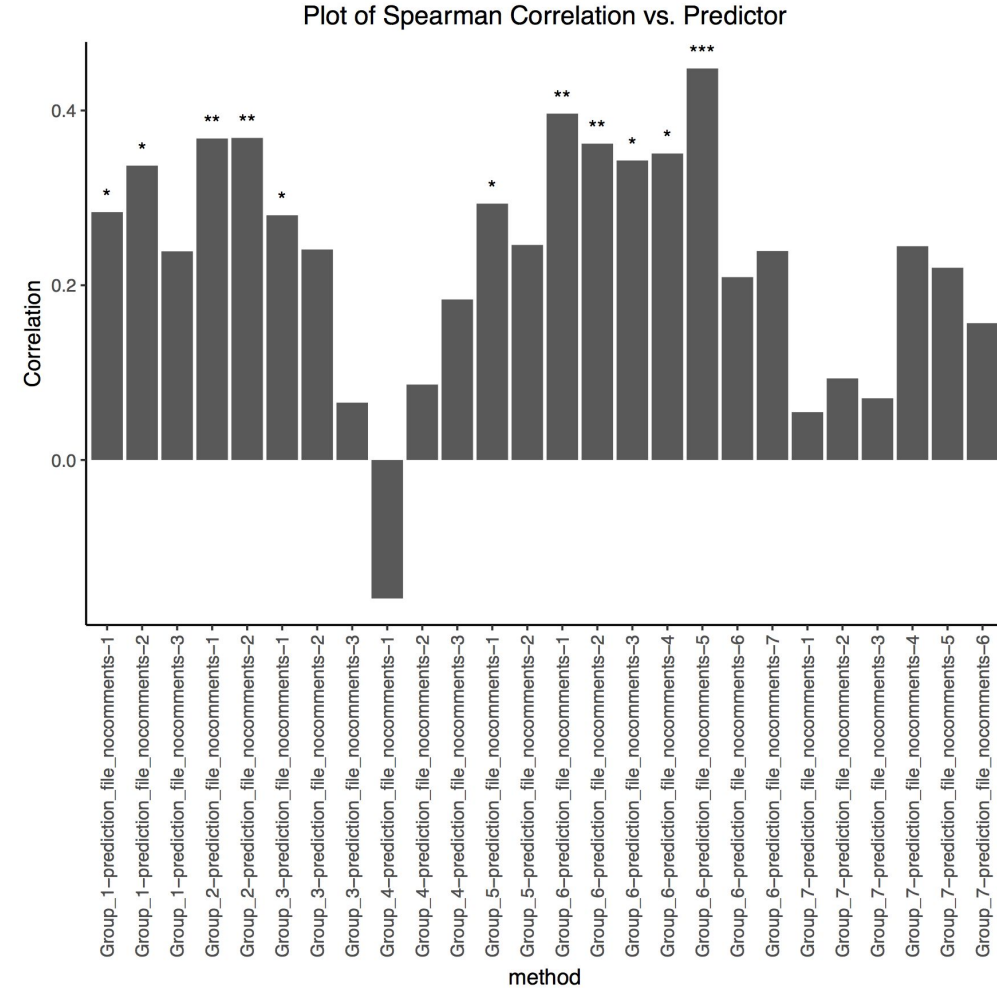
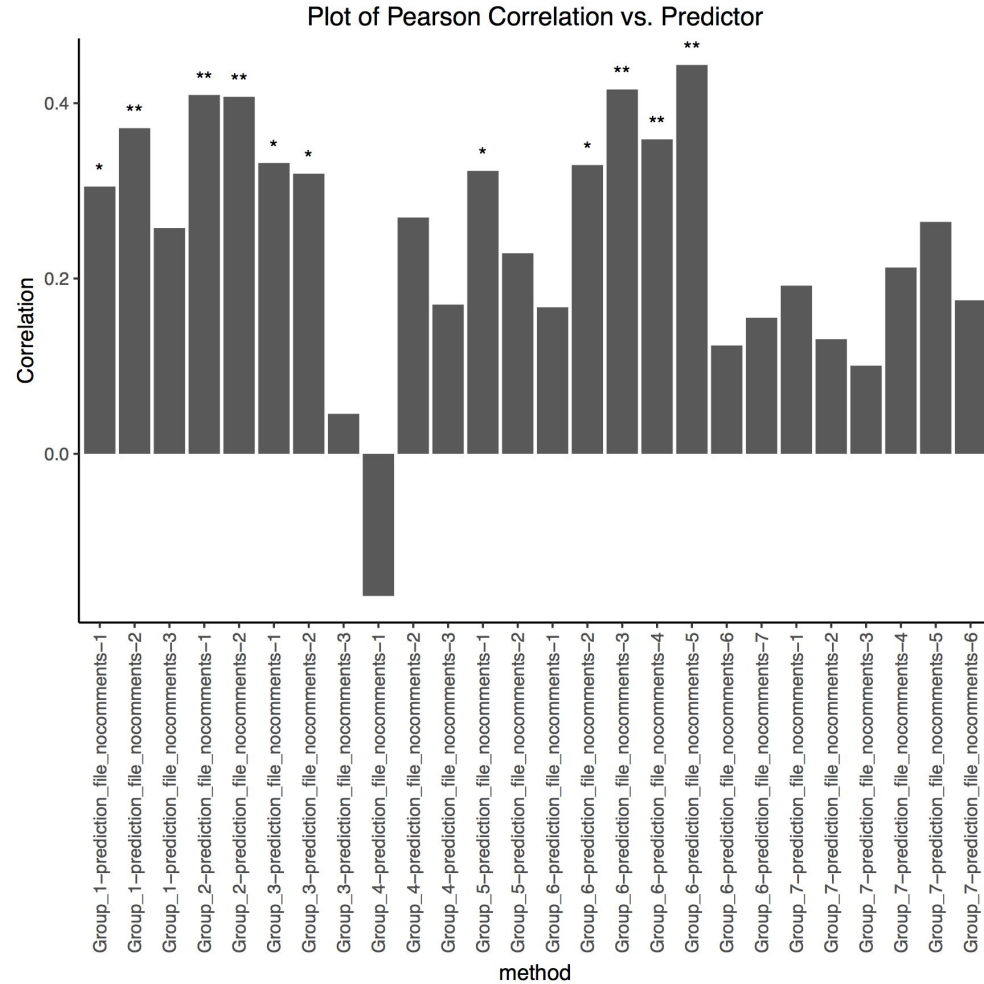
RMSD

- RMSD with extreme SD filtering showed similar pattern



GAA C-Terminal Variants (n=51)

Pearson and Spearman Correlation



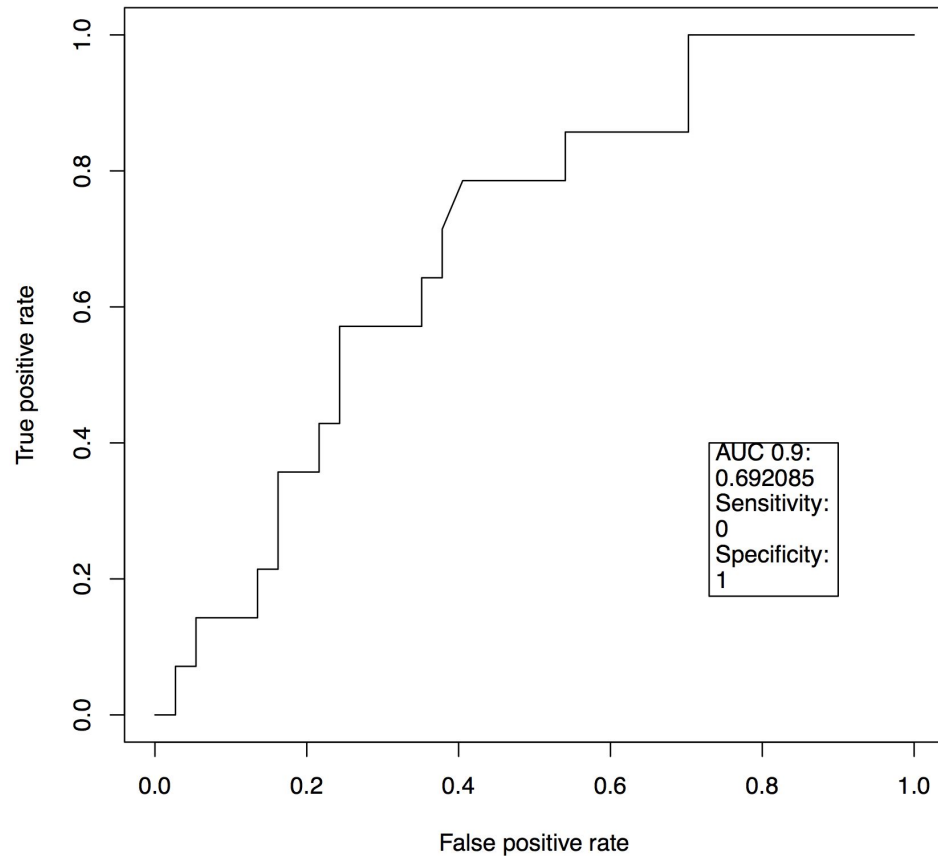
GAA C-Terminal Variants (n=51)

AUC/ROC Curves for Group 6-5

“severe” : <3%
“less_severe” :
3-10%
“mild” : 10-15%
“benign” : >15%

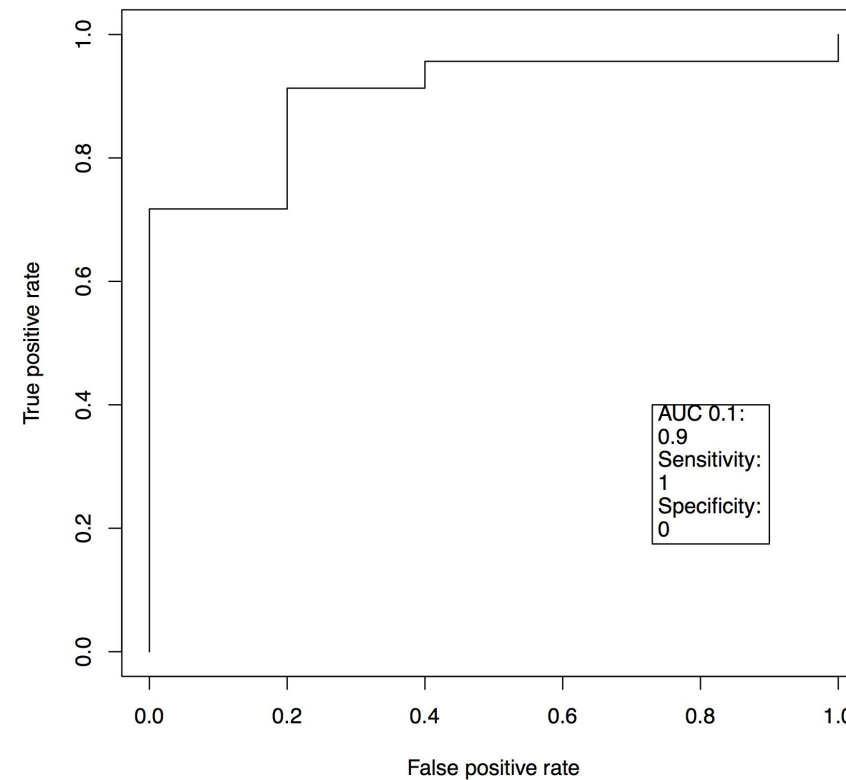
Package Default Threshold 0.9

Group_6-prediction_file_nocomments-5 ROC Curve

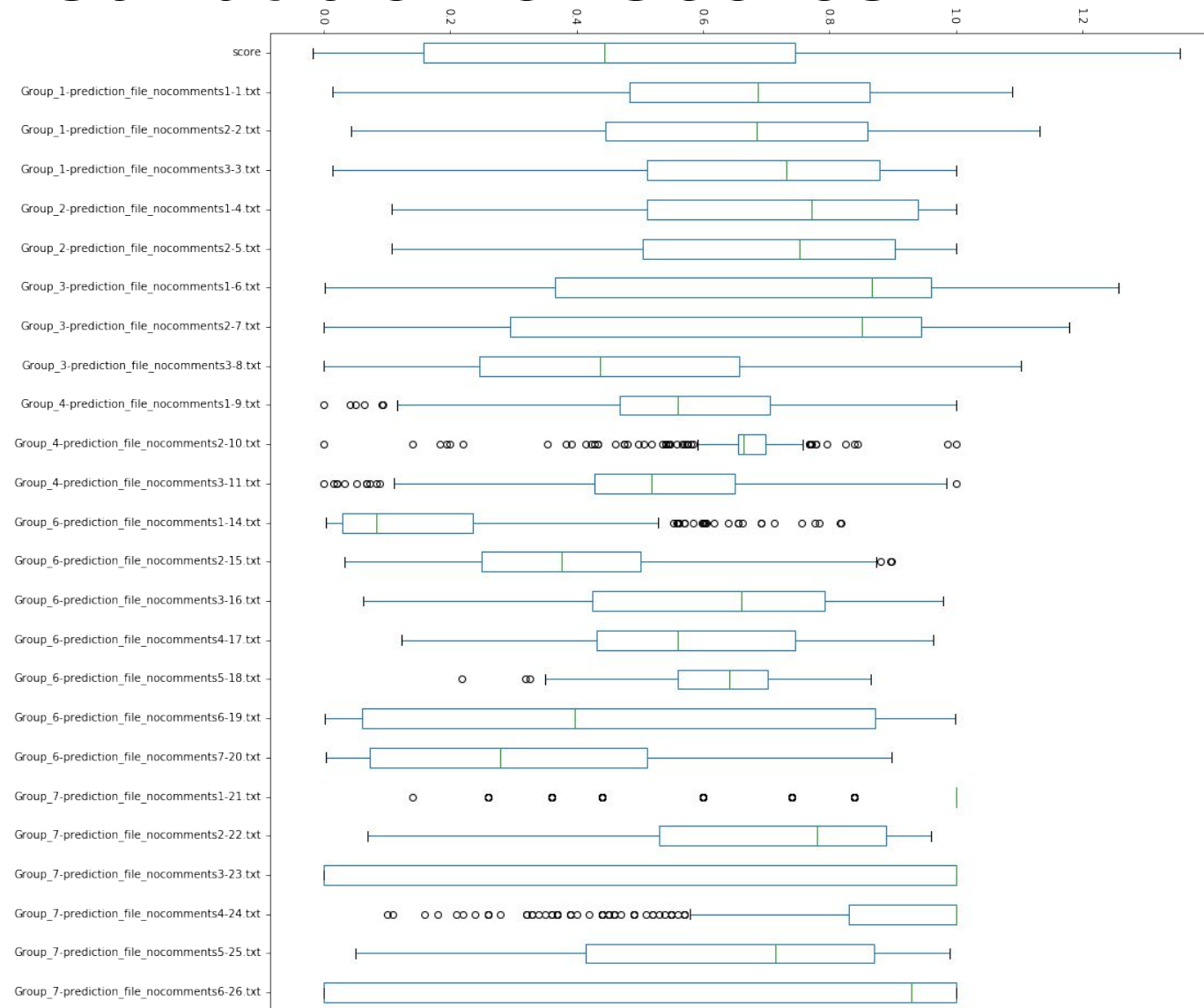


Binary Cutoff at 10% Activity

Group_6-prediction_file_nocomments-5 ROC Curve sd< 0.5



GAA - Distribution of Scores



GAA - Ease of Prediction (RMSD) vs. SD

Variant-Wise RMSD
Hardest to Predict

Easiest to Predict

GAA - Ease of Prediction (Z) vs. SD

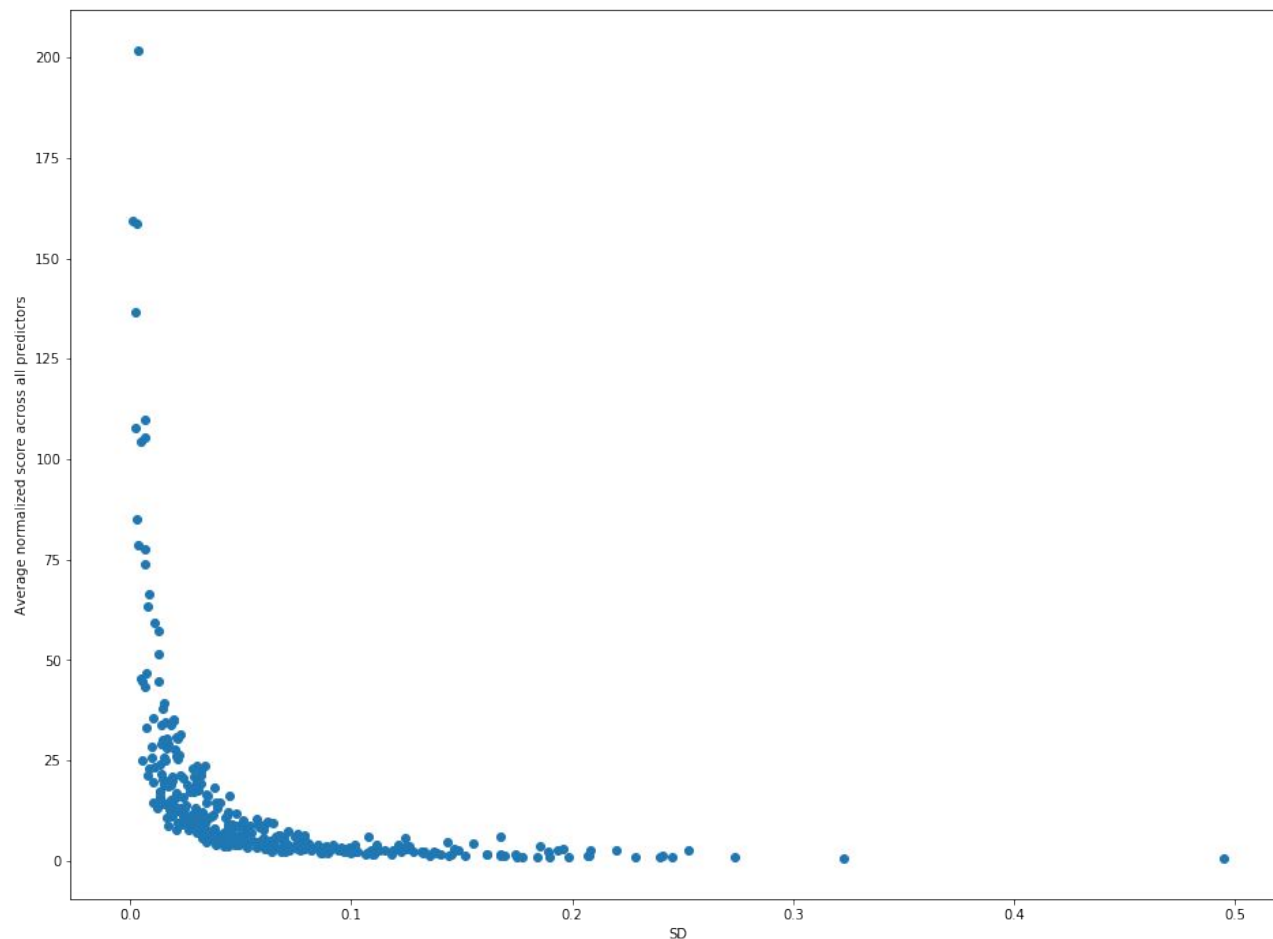
Z-score normalization

Hardest to Predict

Variant	SD	Average normalized score across all predictors
p.Pro690Leu	0.003656	201.723600
p.Tyr609His	0.001464	159.271762
p.Arg154His	0.003279	158.737793
p.Ser20Phe	0.002458	136.677486
p.Glu212Lys	0.006668	109.970696
p.Arg190Cys	0.002779	107.849901
p.Ala93Thr	0.006646	105.266844
p.His56Tyr	0.005048	104.267690

Easiest to Predict

Variant	SD	Average normalized score across all predictors
p.Gln69Glu	0.322895	0.513198
p.Gly920Ser	0.494842	0.534823
p.Leu948Phe	0.189893	0.759861
p.Leu12Phe	0.273565	0.850250
p.Val15Ile	0.174928	0.891478
p.Val816Ile	0.239582	0.913875
p.Ile257Val	0.184342	0.928933
p.Ala120Gly	0.198415	0.936777
p.Ala761Gly	0.228712	0.977124



HGMD Dump

With ROC Curve Fixed

Further Research

- Provide explanations for counterintuitive observations:
 - Predictions tend to be worse for variants with lower experimental variance, for all predictors
 - Predictions tend to be better for C-terminal variants for GAA
- Do teams perform poorly on variants with high activity that are at the same position as known pathogenic variants?
- Add more confirmatory statistical analysis tests and plotting capabilities to the R package