

CAGI slide pack guidelines

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

_____ *Ken Chen* _____

- 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 *Prioritizing pathogenic variants using functional genomics data and DNA, protein sequence-derived features*, 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.

Prioritizing pathogenic variants using functional genomics data and DNA, protein sequence-derived features

Ken Chen¹, Maolin Ding¹, Huiying Zhao², Yuedong Yang¹

1. School of Computer Science and Engineering, Sun Yat-sen University
2. Sun Yat-sen Memorial Hospital, Sun Yat-sen University

May 16, 2022

Analysis of the CAGI6-Sherloc data

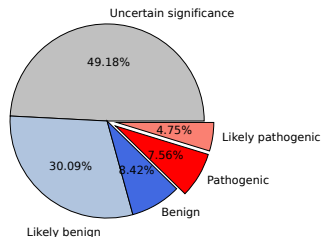


Fig: The proportion of different clinical classifications

Table: Distribution of variants in different regions

Percentage	Train (n=414 970)	Test (n=121 711)
CDS	87.96 %	71.28 %
intronic	7.90 %	24.90 %
splicing	3.77 %	2.95 %
UTR	0.37 %	0.84 %
upstream	0.00 % (n=21)	0.03 % (n=31)
downstream	0.00 % (n=2)	0.00 % (n=0)
intergenic	0.00 % (n=2)	0.00 % (n=5)

annotated by ANNOVAR¹; **CDS**: coding sequences;

splicing: within 2bp of a splicing junction;

UTR: untranslated regions;

up-/downstream: the 2000-bp up-/downstream of transcripts

- unbalanced dataset: ~ 10 % of variants are pathogenic
- Almost all (> 99 %) the variants reside in transcripts
- > 70 % of variants are in coding regions

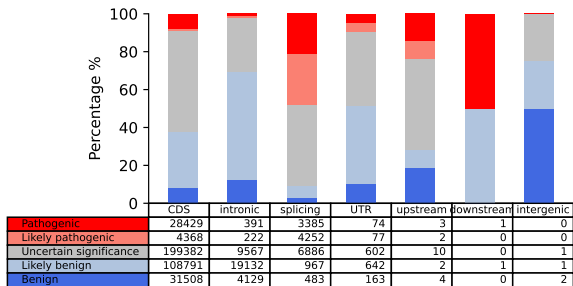


Fig: The proportion of clinical classifications in different regions (training data)

- **splice site variants:** RNA-level features (splicing, RBP binding sites, ...)
- **coding variants:** protein-level features (evolutionary conservation, structure, ...)
- **intronic variants:** epigenetic®ulatory features (TFBS, eQTL, chromatin accessibility, histone marks, ...)

Methods: input features

basic annotation

- length
- frameshift
- start gain/loss
- stop gain/loss
- synonymous

mutation database

- Allele Frequency (AF)
from gnomAD(v3)
- eQTL
merging 49 tissues from GTEx

conservation scores

- phyloP and phastCons
in mammal, primate &
vertebrate from UCSC

RNA features

- predicted $\Delta\Psi$
by MMSplice
- RBP binding sites
in K562 & HepG2
from ENCODE

protein features

- PSSM(protein-level conservation)
psiblast against Uniref90
- disorder
predicted by SPOT-disorder

other genomics features

- TFBS: CTCF/EP300/POLR2A
from ENCODE
- chromatin accessibility
DNase-seq from ENCODE
- core histone marks
from Roadmap

FASVAR: Integrating **F**unctional genomics data **A**nd DNA/protein **S**equence-derived features to prioritize pathogenic **V**ARiants

Position-specific scoring matrix (PSSM)

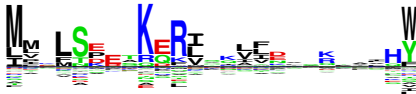
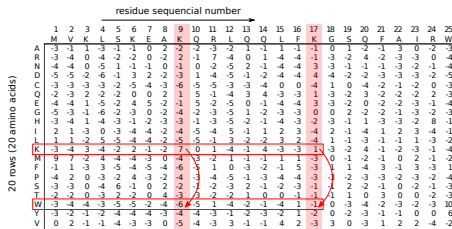
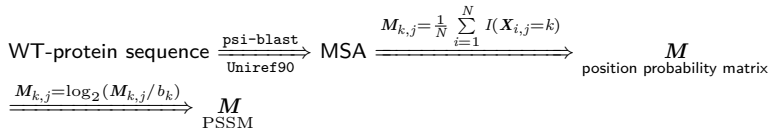


Fig: PSSM & sequence logo²

PSSM-based mutation score:

- single substitution (SNV),
the j -th AA k_1 is substituted by k_2 :
- multiple substitutions (MNV),
from the j -th AA to the j' -th AA:

$$S_{k_1 \rightarrow k_2, j} = |M_{k_2, j} - M_{k_1, j}|$$

$$S_{k_*, j-j'} = \sum_{t=j}^{j'} |M_{k_2, t} - M_{k_1, t}|$$

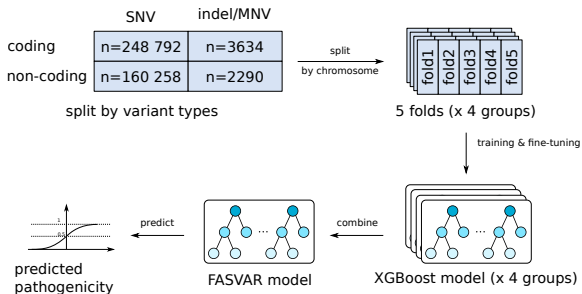
- deletion, $j \sim j'$ AAs deleted:

$$S_{\text{del}, j-j'} = \sum_{t=j}^{j'} |M_{k_1, t}|$$

- insertion, a d -AA fragment
inserted at j :

$$S_{\text{ins}, j-j'} = \sum_{t=j}^{j+l-1} |M_{k_1, t}|$$

Methods: model training



- training separated models for coding/non-coding SNVs/indels(MNVs)
- avoid overfitting: **chromosome-split** 5-fold cross-validation
- FASVAR-meta: a meta-model combining FASVAR with CADD, JARVIS, ReMM, MVP, and M-CAP

Results: feature analysis based on training data

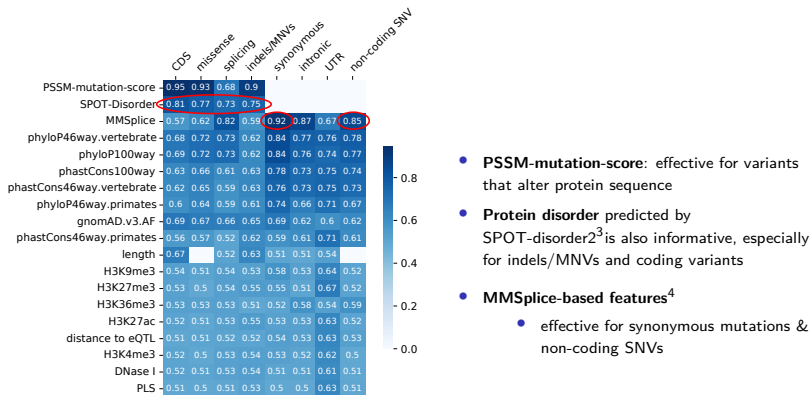


Fig: The AUC of top-20 features on training data in different groups (*a mutation could be categorized to multiple groups*)

Results: evaluation

- Test data labels: taken from ClinVar ⁵ (date: 2022-04-03)
 - 68 432 labeled variants (8856 pathogenic/likely pathogenic variants)
- Methods for comparison
 - whole-genome methods: CADD⁶, JARVIS⁷, ReMM⁸
 - missense-only methods: MVP⁹, m-CAP¹⁰
- A more challenging scenario:
 - *data leakage*¹¹: genes with rich labeled variants in training data may result in an inflated evaluation of performance
 - A hard test dataset: excluding genes with more than 10 variants in the training data from the test dataset
⇒ 7026 variants on 904 genes (329 pathogenic and 6697 non-pathogenic (benign/uncertain significance) variants)

⁵<https://ftp.ncbi.nlm.nih.gov/pub/clinvar/>

⁶Kircher et al., 2014; Rentzsch et al., 2019.

⁷Vitsios et al., 2021.

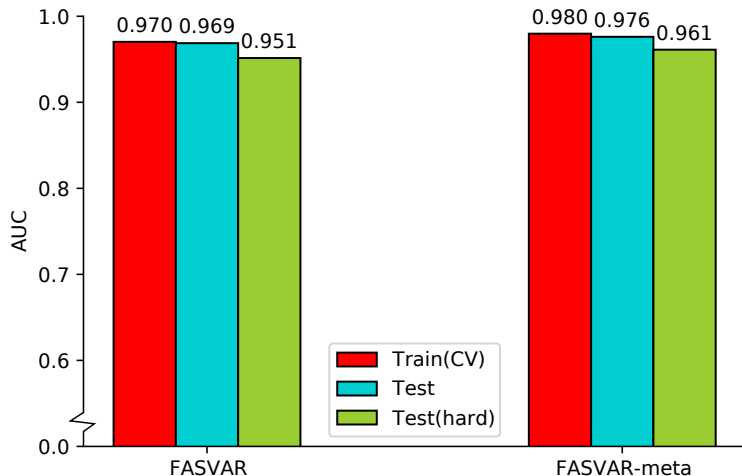
⁸Smedley et al., 2016.

⁹Qi et al., 2021.

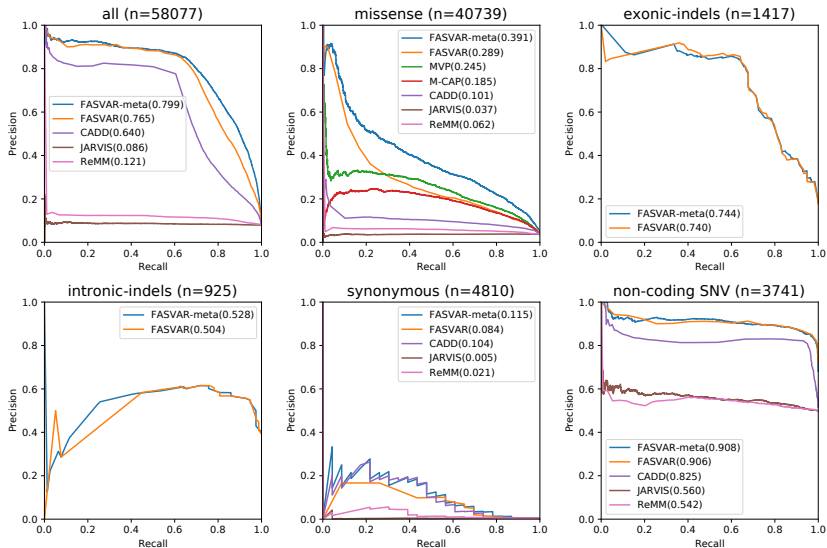
¹⁰Jagadeesh et al., 2016.

¹¹Grimm et al., 2015.
Created for CAGI 6

Results: performance on different datasets

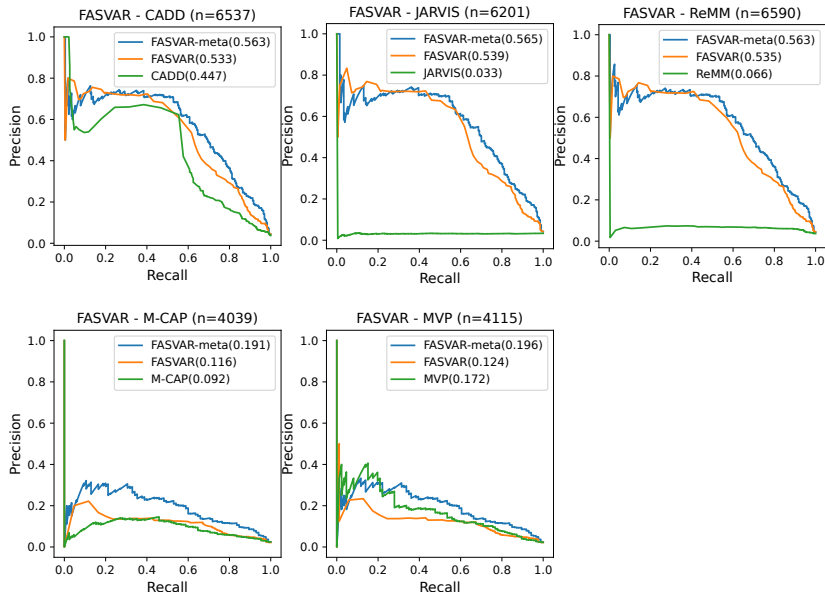


Results: comparing FASVAR with other methods



NOTE: Only the variants that can be predicted by all the models in each group are retained

Results: evaluation on the hard test dataset



Summary

- **Evolutionary information** contributes the most to the prediction
 - Protein evolutionary information: PSSM substitution score
 - DNA evolutionary information: conservation scores
- **Protein disordered regions** are *likely*¹² to be more tolerant to mutations
- Disrupting **splicing** may be one mechanism of the pathogenicity of synonymous and non-coding mutations
- The **chromosome-split cross-validation strategy** effectively avoids data leakage and makes the model robust for genes with rare labeled variants

Thanks!

References I



Ahmed, S. S., Rifat, Z. T., Lohia, R., Campbell, A. J., Dunker, A. K., Rahman, M. S., & Iqbal, S. (2022). Characterization of intrinsically disordered regions in proteins informed by human genetic diversity [Publisher: Public Library of Science]. *PLOS Computational Biology*, 18(3), e1009911. <https://doi.org/10.1371/journal.pcbi.1009911>



Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, 20(1), 48. <https://doi.org/10.1186/s13059-019-1653-z>



Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., & Borgwardt, K. M. (2015). The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22768>]. *Human Mutation*, 36(5), 513–523. <https://doi.org/10.1002/humu.22768>



Hanson, J., Yang, Y., Paliwal, K., & Zhou, Y. (2016). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks [eprint: <https://academic.oup.com/bioinformatics/article-pdf/33/5/685/25148412/btw678.pdf>]. *Bioinformatics*, 33(5), 685–692. <https://doi.org/10.1093/bioinformatics/btw678>



Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, 48(12), 1581–1586. <https://doi.org/10.1038/ng.3703>



Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3), 310–315. <https://doi.org/10.1038/ng.2892>

References II



Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J. J., Chung, W. K., Guan, Y., & Shen, Y. (2021). MVP predicts the pathogenicity of missense variants by deep learning [Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Mutation;Sequence annotation Subject_term_id: machine-learning;mutation;sequence-annotation]. *Nat Commun*, 12(1), 510. <https://doi.org/10.1038/s41467-020-20847-0>



Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, 47(D1), D886–D894. <https://doi.org/10.1093/nar/gky1016>



Smedley, D., Schubach, M., Jacobsen, J. O., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet*, 99(3), 595–606. <https://doi.org/10.1016/j.ajhg.2016.07.005>



Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B., & Petrovski, S. (2021). Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning [Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genomics;Machine learning;Population genetics;Statistical methods Subject_term_id: genomics;machine-learning;population-genetics;statistical-methods]. *Nat Commun*, 12(1), 1504. <https://doi.org/10.1038/s41467-021-21790-4>

References III



Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16), e164–e164.
<https://doi.org/10.1093/nar/gkq603>