@CAGInews

CAGI

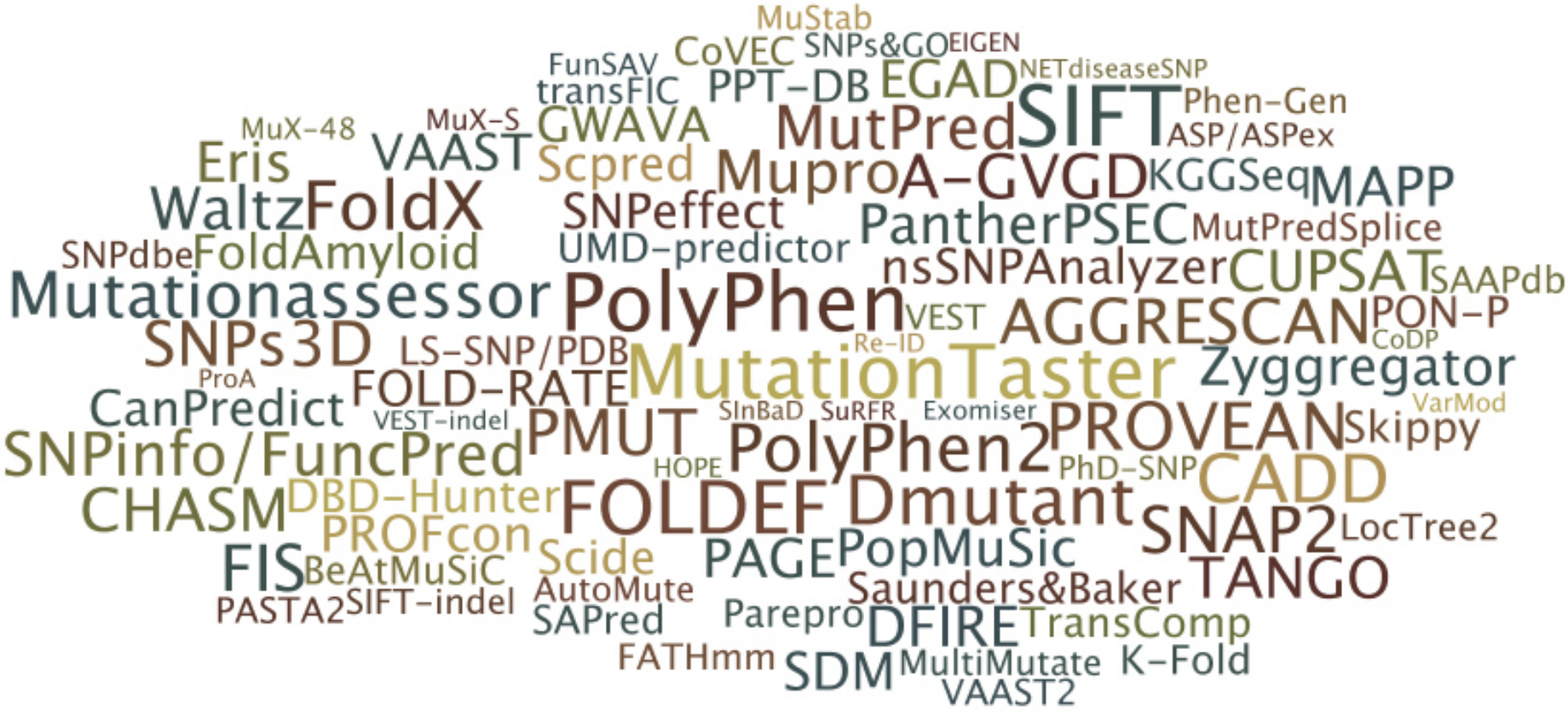
The Critical Assessment of Genome Interpretation

www.genomeinterpretation.org

Gaia Andreoletti

April 4, 2017

Some variant effect prediction methods



https://genomeinterpretation.org/impact

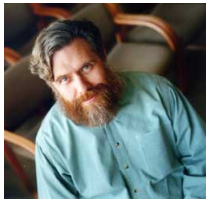
Goals of the CAGI experiment

- Understand the current capabilities for interpreting known and unclassified variants.
- Reveal bottlenecks and identify progress.
- Show where effort may best be focused.
- Highlight innovations.
- Engage and connect diverse disciplines necessary for genome interpretation.

CAGI 4 (2016) Challenges

NAGLU variants. Predict the effect of naturally occurring missense mutations in N-acetyl-glucosaminidase on cellular enzymatic activity.
Data provided by: Jon LeBowtiz, Biomarin Pharmaceuticals

NPM-ALK variants. Predict the effect of mutations in the kinase domain of the Nucleophosmin - Anaplastic Lymphoma Kinase fusion protein on kinase activity and Hsp90 binding affinity.
Data provided by: Paolo Bonvini, Padua Children's Hospital

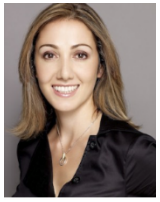


SUMO ligase variants. Predict the effects of missense mutations in human SUMO ligase (UBE2I) on competitive growth in a high-throughput yeast complementation assay.
Data provided by: Frederick "Fritz" Roth, University of Toronto

Pyruvate kinase variants. Predict the effects of single-amino-acid mutations on enzyme activity, and allosteric activation and inhibition in cell extracts.
Data provided by: Aron Fenton, University of Kansas Medical Center

PGP. Match the genomes of 21 consenting Personal Genome Project participants to their respective phenotypic profiles from a set of 239 binary traits and 71 “decoys”.
Data provided by: George Church, Harvard University

eQTL-causal SNPs. Identify regulatory sequences and eQTL-causal variants, and estimate their effects on activation of transcription in a massively parallel reporter assay.
Data provided by: Pardis Sabeti, Broad Institute



Crohn’s exomes. Distinguish between Crohn's disease patients and healthy individuals, and predict age of disease onset in a sample of 111 exomes.
Data provided by: Andre Franke, Christian-Albrechts-Universität zu Kiel

Warfarin exomes. Estimate patients’ therapeutic warfarin doses from their exomes.
Data provided by: Russ Altman, Stanford University

Bipolar exomes. From exomes, identify which individuals have BD and which individuals are unaffected.
Data provided by: Peter Zandi and colleauges, Johns Hopkins University

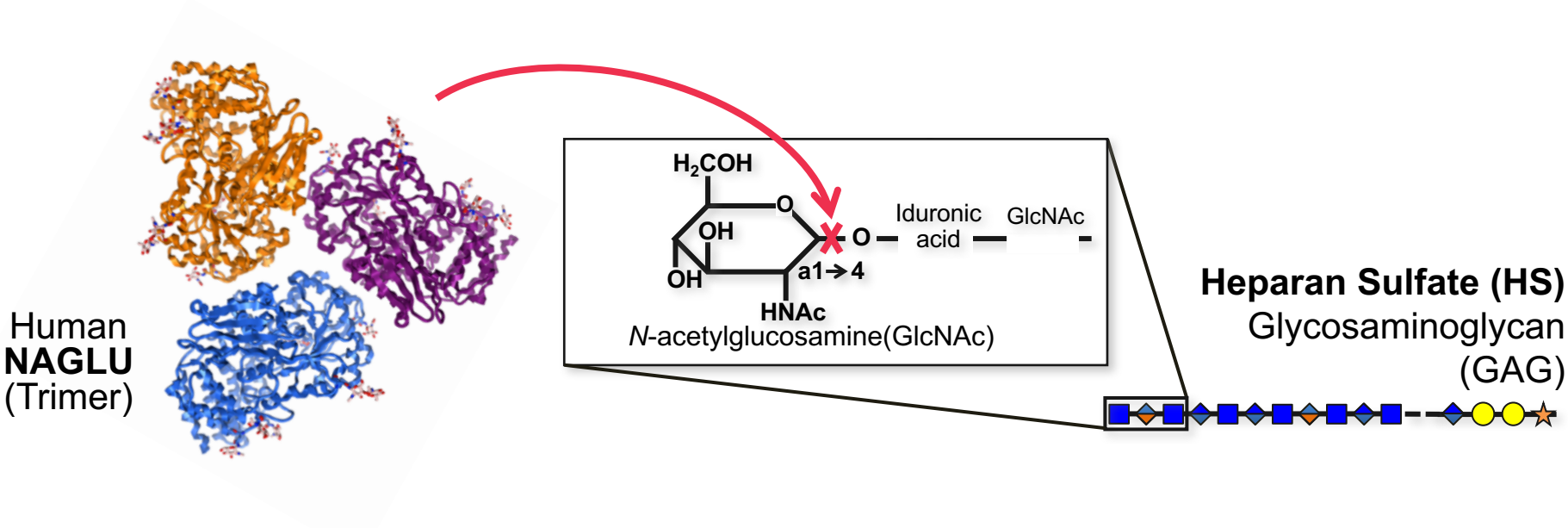
Hopkins clinical gene panel. Match the patients’ gene panel sequences to their clinical descriptions and predict the causal pathogenic variants.
Data provided by: Garry Cutting, Johns Hopkins University

SickKids clinical genomes. Match the patients’ genomes to their clinical descriptions and predict the causal pathogenic variants. *Data provided by: Stephen Meyn & colleagues, SickKids*

An example of CAGI challenge of nonsynonymous variants and targeted assays

Some background about NAGLU:

- NAGLU deficiencies autosomal recessive lysosomal storage disease: San Filippo disease.
- Severe neurological disease.
- Birth incidence: ~0.5 per 100,000
- Deleterious variants are rare.



The NAGLU challenge

Aim:

Predict the effect of naturally occurring missense mutations on cellular enzymatic activity

Dataset:

153 missense NAGLU mutations found in ExAC.

Predictions:

17 submissions from 10 groups.

Assessment:

Predictions were compared with enzyme activity assayed in transfected cell lysates

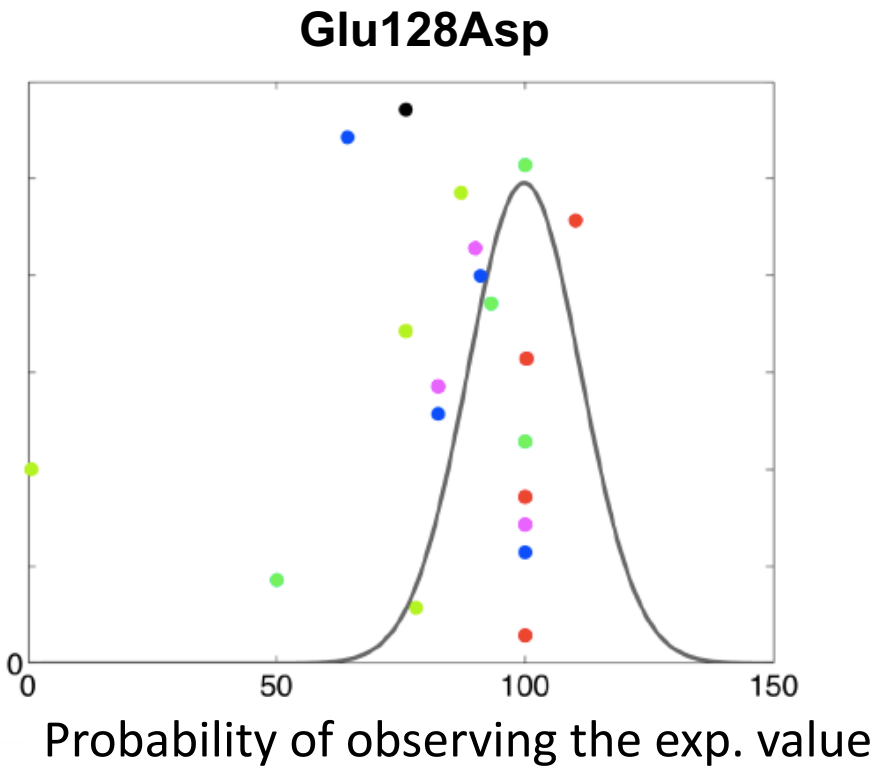
https://genomeinterpretation.org/content/4-NAGLU

Data provided by Jonathan H. LeBowitz, Wyatt T. Clark, BioMarin Pharmaceutical

Assessing NAGLU mutation predictions

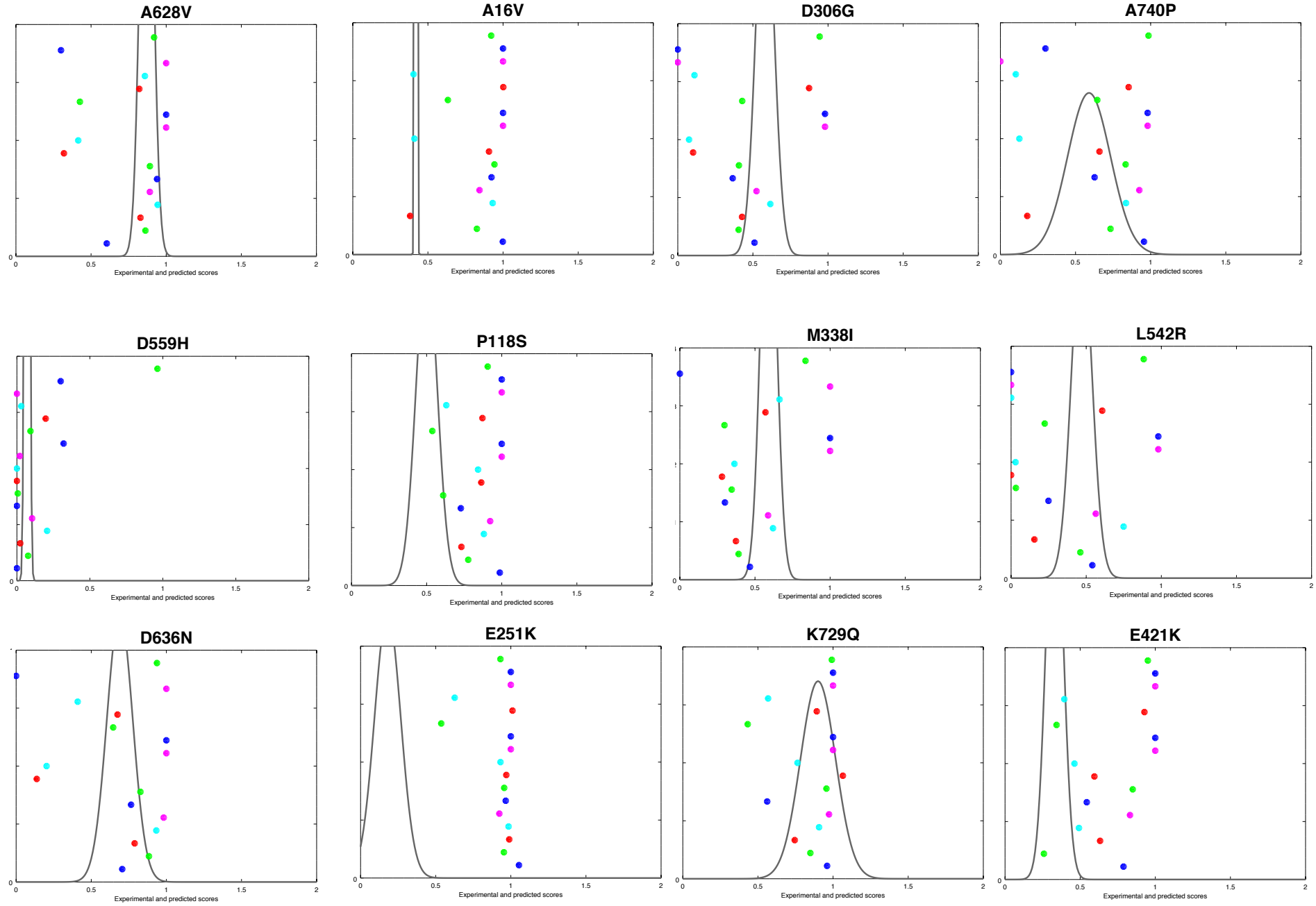
Probability of observing the experimental value

Predictions



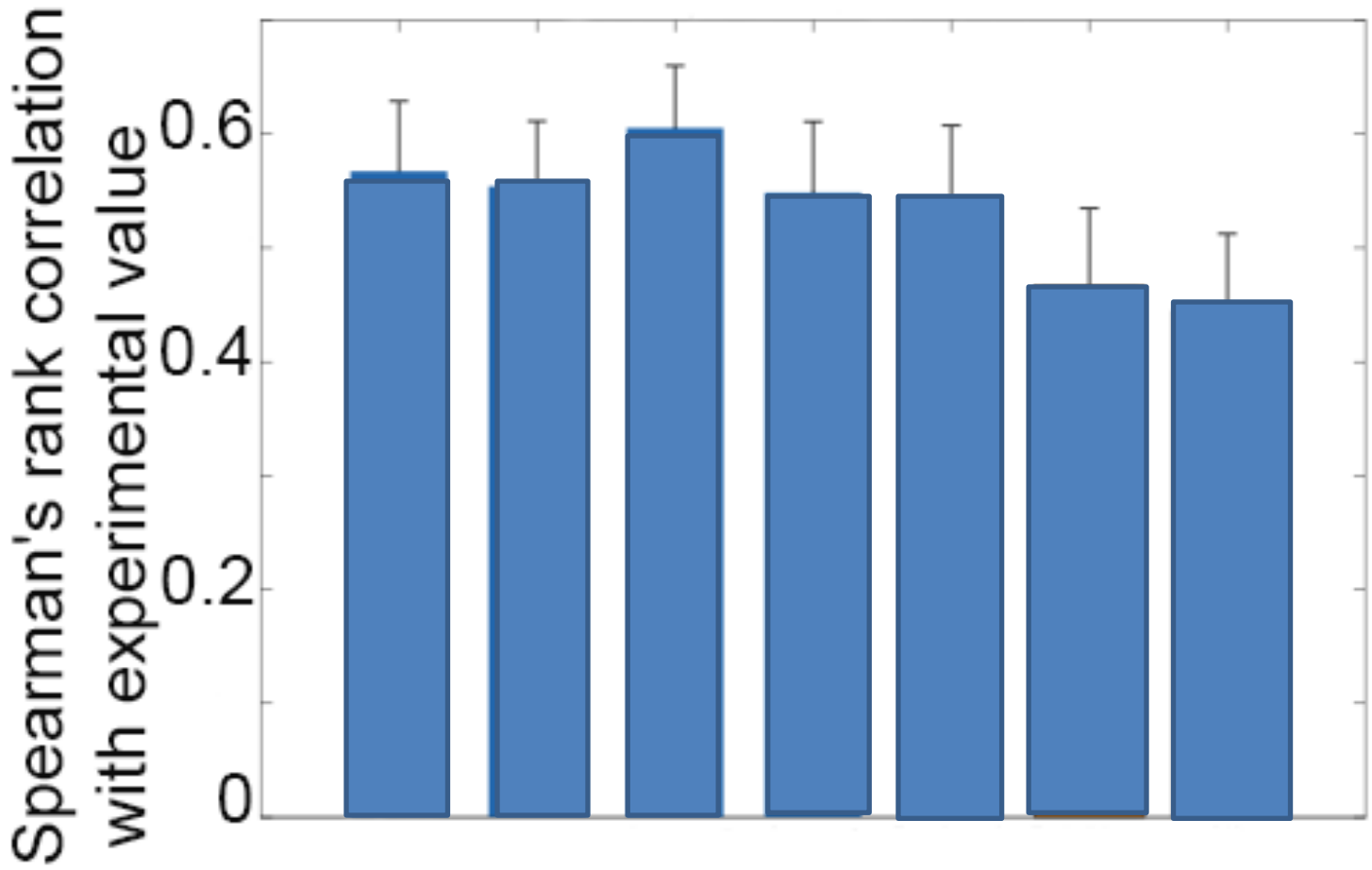
Experimental relative growth rate

Predicted relative growth rate



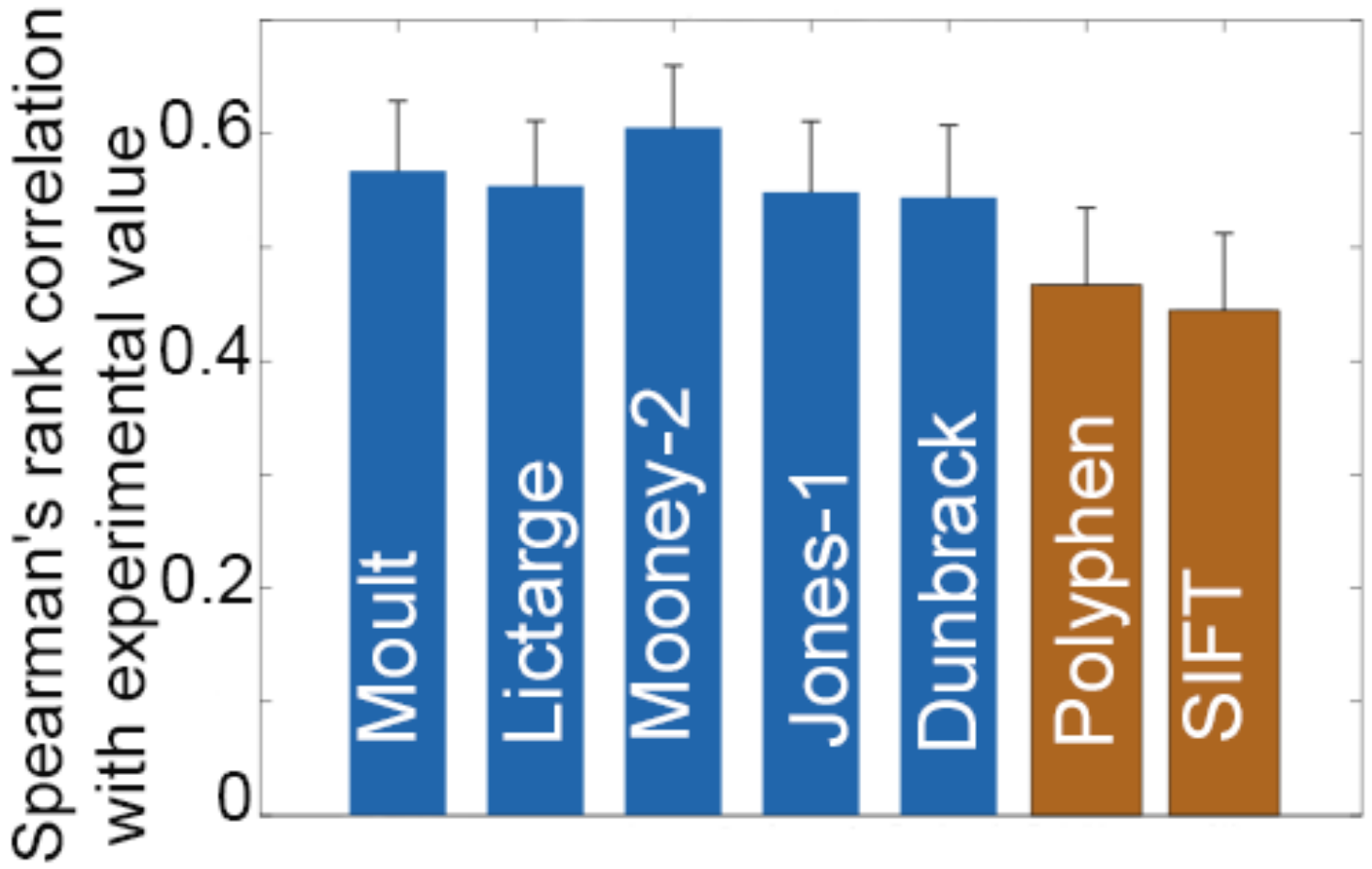
Probability of observing the exp. value

Top missense prediction methods are highly statistically significant



Evidence from this theme also seen in: CBS, SUMO, L-PYK, MRE11 and NBS1, BRCA1 and BRCA2, P16, RAD50, CHEK2,and SCN5A.

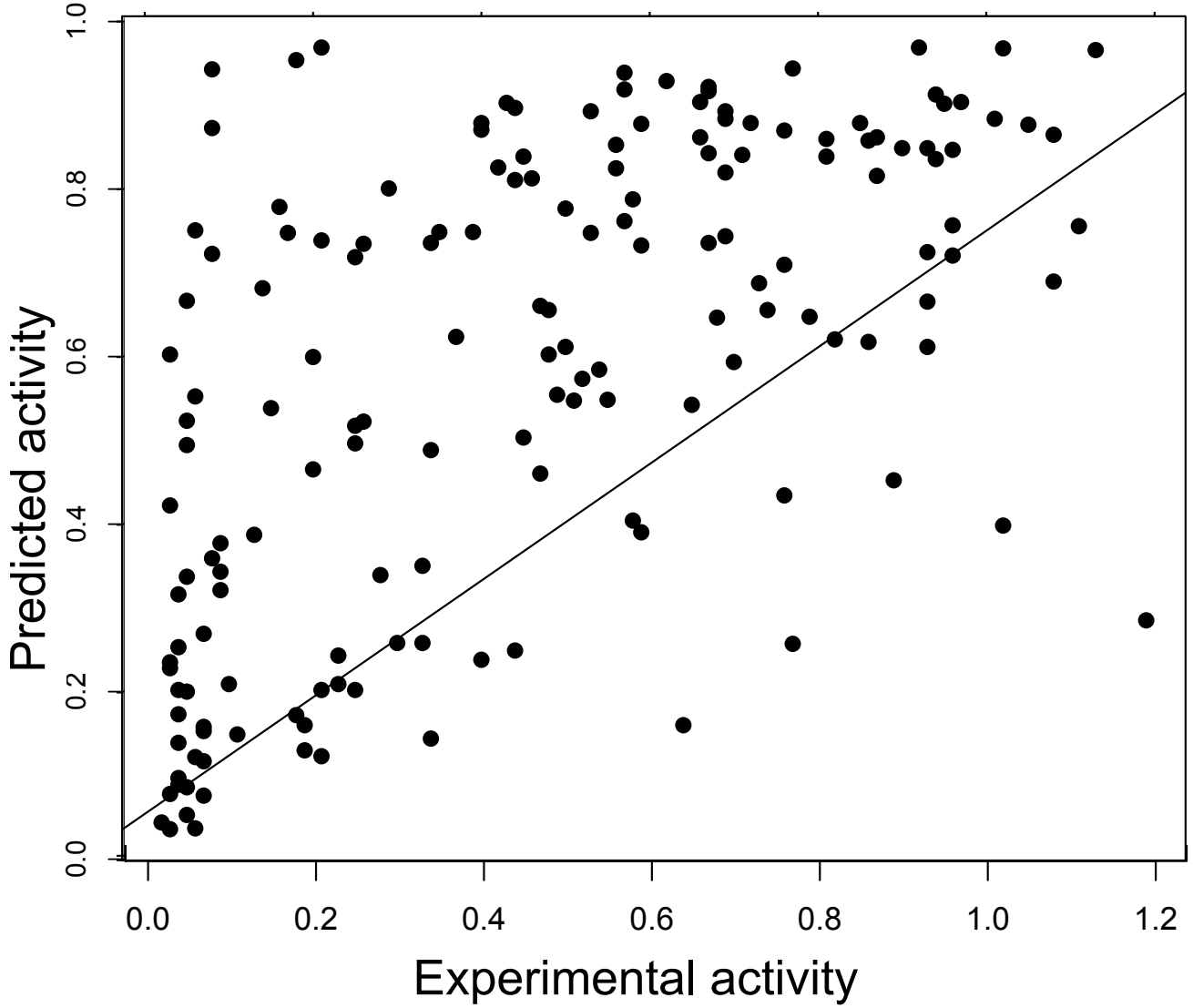
Predictors performed better than baseline methods



Evidence from this theme also seen in: CBS, SUMO, L-PYK, MRE11 and NBS1, BRCA1 and BRCA2, P16, RAD50, CHEK2,and SCN5A.

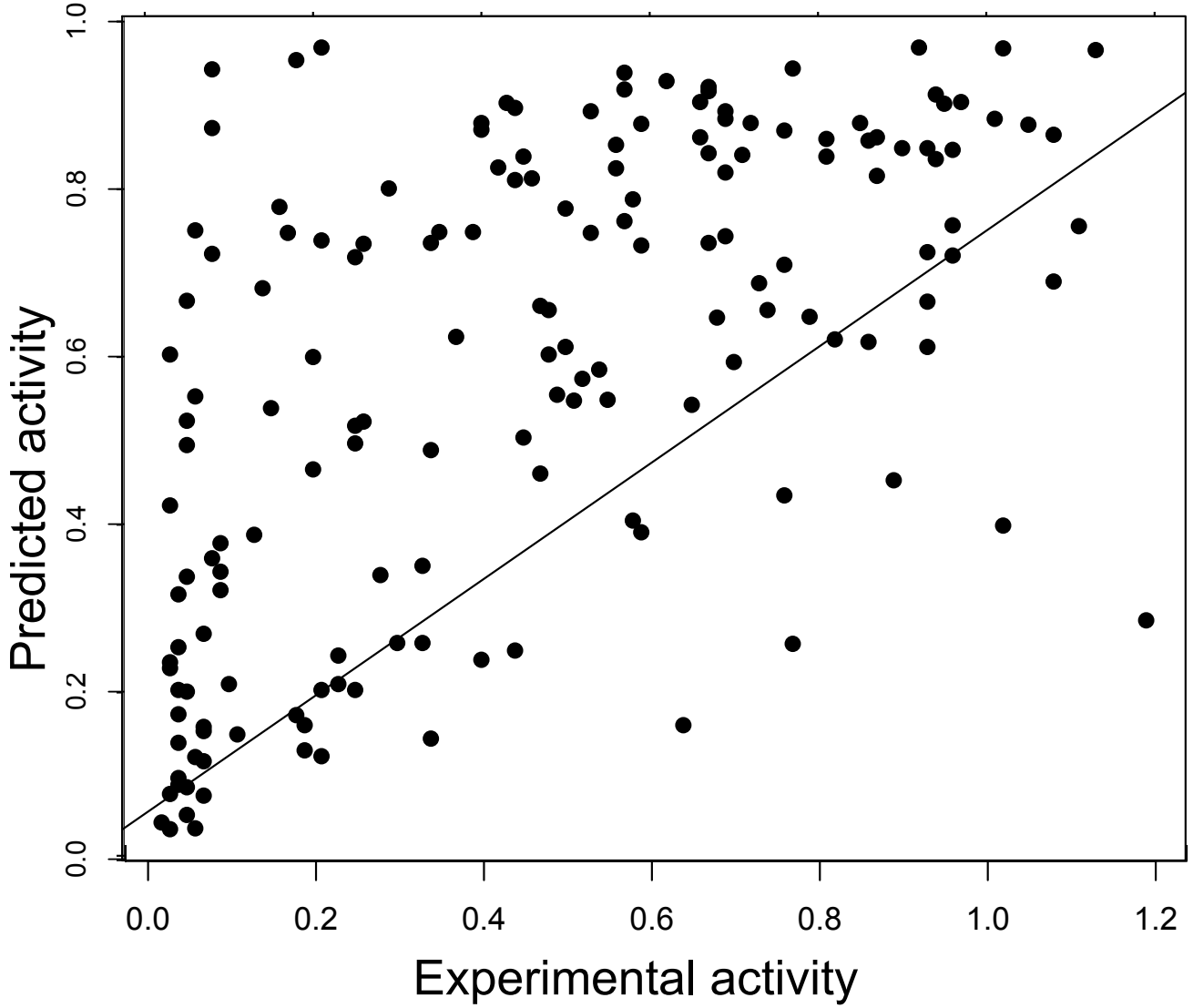
Individual variant accuracy is limited

NAGLU challenge top performing method. R2 = 0.36; Corr.coff = 0.60



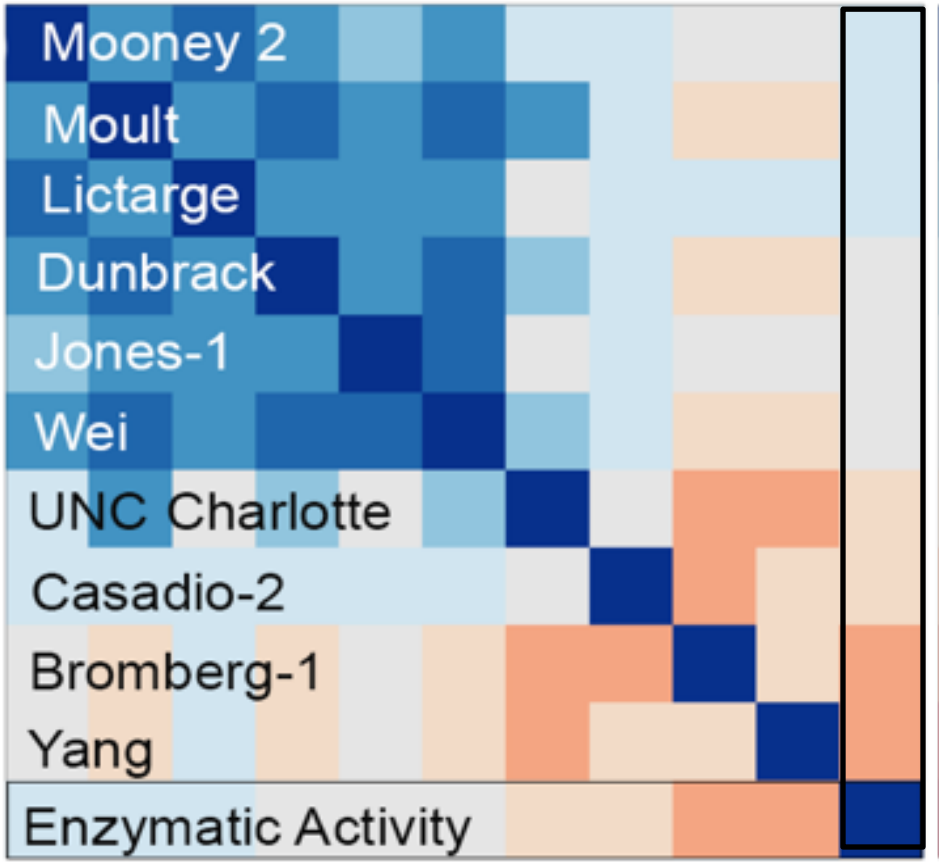
Evidence from this theme also seen in: CBS, SUMO, L-PYK, MRE11 and NBS1, BRCA1 and BRCA2, P16, RAD50, CHEK2,and SCN5A.

There might be potential for missense interpretation at the extreme of the distribution



Evidence from this theme also seen in: CBS, SUMO, L-PYK, MRE11 and NBS1, BRCA1 and BRCA2, P16, RAD50, CHEK2,and SCN5A.

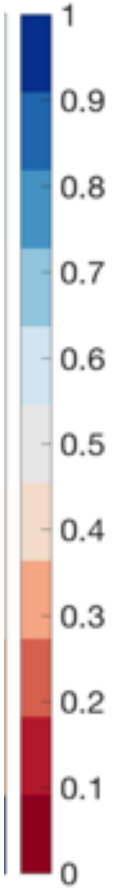
Missense impact prediction methods tend to correlate better with each other than with experiment



High correlated

Intensity represents absolute value of Pearson's r for predicted activity of NAGLU variants.

Not correlated



Evidence from this theme also seen in: CBS,SUMO, BRCA1 and BRCA2, P16, and RAD50.

Predictors were able to identify causal variants that were overlooked by a clinical laboratory

Aim:
Predict patients' clinical descriptions and pathogenic variants from their genome sequences.

Dataset:
Genomes and phenotypic descriptions of 25 undiagnosed pediatric genetics patients.

Result:
- Submissions from 4 teams - 187 variants proposed.
- Two-thirds related to phenotype, one-third not obviously related.
- In two cases, these have been validated as causative.
- Two further compelling variants are being validated.

https://genomeinterpretation.org/content/4-SickKids_clinical_genomes
Data provided by Stephen Meyn, the Hospital for Sick Children, Toronto

Some of the lessons learned from CAGI

- In general predictions are statistically significant. However, predictive accuracy for specific variants is low
- Bespoke approaches often enhance performance.
- Missense methods tend to correlate better with each other than with experiments.
- Use of multiple uncalibrated missense impact prediction methods in the clinic is not advised.
- Predictors were able to identify causal variants that were overlooked by a clinical laboratory.
- Predicting complex traits from exomes is fraught.
- Interpretation of non-coding variants shows promise but is not at the level of missense.
- The CAGI community has potential for addressing unanswered questions in genome interpretation.

Acknowledgements

Website: genomeinterpretation.org

National Human Genome Research Institute

NATIONAL CANCER INSTITUTE

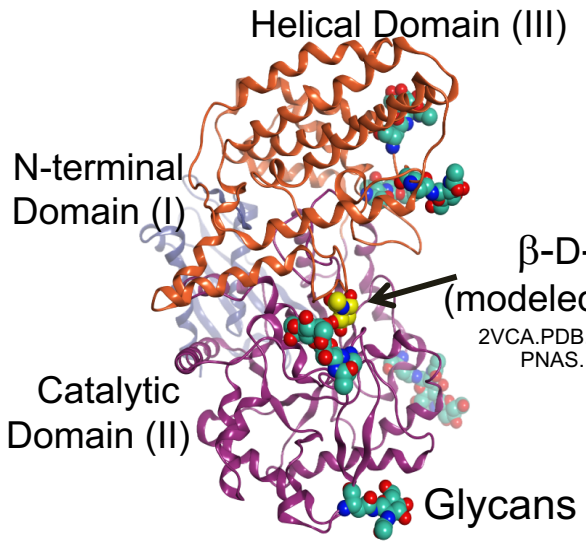
TATA CONSULTANCY SERVICES

Predictors: Allison Abad, Ogun Adebali, Ivan Adzhubey, Talal Amin, Johnathan R. Azaria, Giulia Babbi, Eraan Bachar, Benjamin Bachman, Minkyung Baek, Greet De Baets, Michael Beer, Violeta Beleva-Guthrie, Bonnie Berger, Brady Bernard, Rajendra Bhat, Rohit Bhattacharya, Samuele Bovo, Marcus Breese, Aharon S. Brodie, Yana Bromberg, Binghuang Cai, Colin Campbell, Chen Cao, Emidio Capriotti, Marco Carraro, Rita Casadio, Billy H. W. Chang, Shann-Ching Chen, Yun-Ching Chen, Chien-Yuan Chen, Melissa Cline, Andrea Corredor, Carla Davis, Mark Diekhans, Rezarta I. Dogan, Christopher Douville, Ian Driver, Roland Dunbrack, Joost van Durme, Andrea Eakin, Matthew Edwards, Gokcen Eraslan, Hai Fang, Carlo Ferrari, Anna Flynn, Lukas Folkman, Colby T. Ford, Adam Frankish, Zaneta Franklin, Yao Fu, Alessandra Gasparini, Tom Gaunt, David Gifford, Manuel Giollo, Nina Gonzaludo, Valer Gotea, Julian Gough, Yuchun Guo, Jennifer Harrow, Marcia Hasenahuer, Lim Heo, Ramin Homayouni, Raghavendra Hosur, Cheng L. V. Huang, Peter Huwe, Sohyun Hwang, Tadashi Imanishi, Jules Jacobsen, Chan-Seok Jeong, Yuxiang Jiang, David T. Jones, Daniel Jordan, Beomchang Kang, Rachel Karchin, Panagiotis Katsonis, Sunduz Keles, Manolis Kellis, Nikki Kiga, Dongsup Kim, Eiru Kim, Jack F. Kirsch, Michael Kleyman, Andreas Kraemer, Anshul Kundaje, Kunal Kundu, Pui-Yan Kwok, Ernest Lam, Dae Lee, Gyu R. Lee, Insuk Lee, Pietro Di Lena, Emanuela Leonardi, Andy Li, Jun Li, Yue Li, Biao Li, Olivier Lichtarge, Chiao-Feng Lin, Rhonald C. Lua, Angel Mak, Pier L. Martelli, David Masica, Zev Medoff, Aziz M. Mezlini, Rahul Mohan, Alexander M. Monzon, Sean D. Mooney, Matthew Mort, John Moult, Steve Mount, Eliseos Mucaki, Jonathan Mudge, Nikola Mueller, Chris Mungall, Katsuhiko Murakami, Yoko Nagai, Noushin Niknafs, Abhishek Niroula, Conor M. L. Nodzak, Yanay Ofran, Ayodeji Olatubosun, Kymberleigh Pagel, Lipika R. Pal, Taeyong Park, Nathaniel Pearson, Vikas Pejaver, Jian Peng, Alexandra Piryatinska, Catherine Plotts, Predrag Radivojac, Aditya R. Rao, Aliz Rao, Graham Ritchie, Peter Rogan, Frederic Rousseau, Jana M. Schwarz, Joost Schymkowitz, Chaok Seok, George Shackelford, Sohela Shah, Maxim Shatsky, Ron Shigeta, Hashem A. Shihab, Jung E. Shim, Junha Shin, Sunyoung Shin, Ilya Shmulevich, Bradford R. Silver, Nasa Sinnott-Armstrong, Ben Smithers, Yesim A. Son, Mario Stanke, Nathan Stitziel, Andrew Su, Laksshman Sundaram, Paul Tang, Nuttinee Teerakulkittipong, Natalie Thurlby, Janita Thusberg, Kevin Tian, Collin Tokheim, Silvio C. E. Tosatto, Yemliha Tuncel, Tychele Turner, Ron S. Unger, Aneeta Uppal, Gurkan Ustunkar, Jouni Valiaho, Mauno Vihinen, Mary Wahl, Michael Wainberg, Meng Wang, Maggie Wang, Yanran Wang, Xinyuan Wang, Li-San Wang, Liping Wei, Qiong Wei, Rene Welch, Stephen Wilson, Chunlei Wu, Lijing Xu, Qifang Xu, Yuedong Yang, Christopher Yates, Yizhou Yin, Chen-Hsin Yu, Dejian Yuan, Jan Zaucha, Haoyang Zeng, Maya Zuhl **Data Providers**: Russ Altman, Adam P. Arkin, Madeleine P. Ball, Jason Bobe, Paolo Bonvini, Bethany Buckley, George Church, Garry R. Cutting, Emma D'Andrea, Lisa Elefanti, Aron W. Fenton, Andre Franke, Nina Gonzaludo, Joe W. Gray, Linnea Jannson, John P. Kane, Pui-Yan Kwok, Rick Lathrop, Jonathan H. LeBowitz, Federica Lovisa, Angel C. Y. Mak, Mary J. Malloy, Richard McCombie, Chiara Menin, M. Stephen Meyn, John Moult, Robert Nussbaum, Lipika R. Pal, Britt-Sabina Petersen, Mehdi Pirooznia, James B. Potash, Clive R. Pullinger, Jasper Rine, Frederick Roth, Pardis Sabeti, Jeremy Sanford, Maria C. Scaini, Nicole Schmitt, Jay Shendure, Molly Sheridan, Michael Snyder, Tim Sterne-Weiler, Paul L. F. Tang, Sean Tavtigian, Ryan Tewhey, Silvio C. E. Tosatto, Jochen Weile, G. Karen Yu, Peter Zandi **Assessors**: Aashish Adhikari, Marco Carraro, John-Marc Chandonia, Rui Chen, Wyatt T. Clark, Roxana Daneshjou, Roland Dunbrack, Iddo Friedberg, Gad Getz, Nick Grishin, Rachel Karchin, Anat Kreimer, Stephen. Meyn, Sean D. Mooney, Alexander A. Morgan, John Moult, Robert Nussbaum, Jeremy Sanford, David B. Searls, Artem Sokolov, Josh Stuart, Shamil Sunyaev, Sean Tavtigian, Silvio C. E. Tosatto, Qifang Xu, Nir Yosef **Organization and Management**: Daniel Barsky, Steven E. Brenner, John-Marc Chandonia, Ajithavalli Chellappan, Flavia Chen, Navya Dabbiru, Roger A. Hoskins, Melissa K. Ly, John Moult, Andrew J. Neumann, Gaurav Pandey, Sadhna Rana, Susanna Repo, Rajgopal Srinivasan, Stephen Yee, Sri Jyothsna Yeleswarapu, Maya Zuhl **Advisory Board**: Russ Altman, George Church, Tim Hubbard, Scott Kahn, Sean D. Mooney, Pauline Ng, Susanna Repo, John Shon **Scientific Council**: Patricia Babbitt, Atul Butte, Garry R. Cutting, Laura Elnitski, Reece Hart, Ryan Hernandez, Rachel Karchin, Robert Nussbaum, Michael Snyder, Shamil Sunyaev, Joris Veltman, Liping Wei.

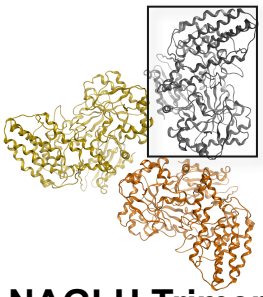
Join CAGI 5!

@CAGInews

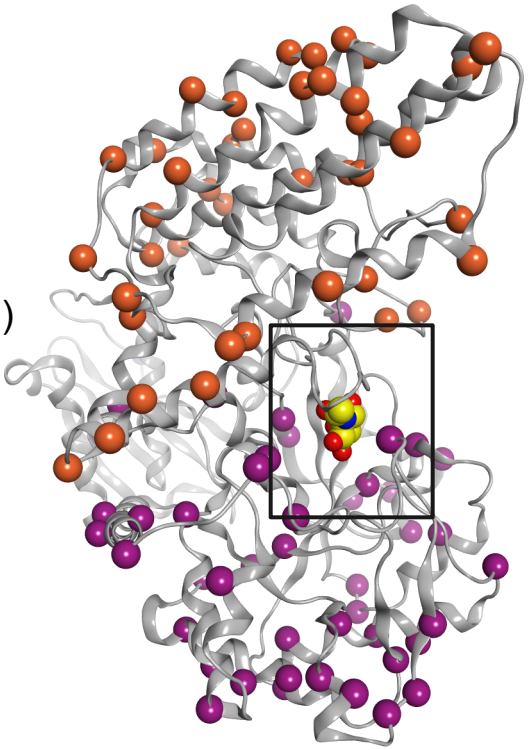
Mutations resulting in <10% wild-type NAGLU activity are scattered throughout the catalytic domain, and relatively *near* (NOT within) the active site, compared to less deleterious mutations



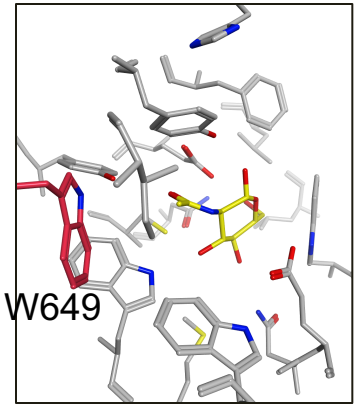
NAGLU Monomeric Unit



NAGLU Trimer



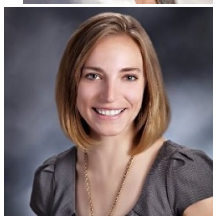
Mutation sites that lead to **>30%** wild-type NAGLU enzyme activity



Mutations are found **near, but not within** the active-site (except for W649)

Meiyappan et al. US 8,775,146 B2, 2014

Hopkins clinical panel challenge



Predict patients' clinical descriptions and pathogenic variants from a gene panel sequences in the absence of additional clinical information

Dataset:

- VCF file for exomes of 83 genes in 106 patients with a range of clinical presentations (n=14) for which physicians ordered genetic testing.

Result:

- 5 groups submitted 11 distinct submissions.
- Predictors correctly identified the disease class in 36 of 43 patients (84%) where the Hopkins laboratory found a variant.
- Predictors correctly identified the disease class in in 39 of 63 patients (62%) where the Hopkins laboratory did not find a variant.
- Each prediction group correctly diagnosed at least one patient that was not successfully diagnosed by any other groups
- CAGI participants found deleterious mutations in genes that were not in the panel that was ordered. These were associated with a disease different from the one indicated by referral physician.

https://genomeinterpretation.org/content/4-Hopkins_clinical_panel

Data provided by Bethany Buckley, and Garry R. Cutting, The Johns Hopkins University

File location

- Box:/CAGI/OtherConferences/Presentations/NGS 2017 Barcelona/NGS 17 v07 3 Apr 2017 GA.pptx

- **15 min + 3min questions**
- **Meeting: Barcelona NGS'17: Structural Variation and Population Genomics. April 3-5 2017**

- Session 4: Genomics

(chair: Cedric Notredame) @ 14:48-15:06 (GMT+1)