Title: REVELations: Uncovering Hidden Heterogeneity Lurking in the Generalized Calibration of Computational Tools

Authors: Scott T Pew[1], Julie L Boyle[2], Marcy E Richardson[3], Madison B Wiffler[4], Russel Bell[1], Emily Nadeau[5], Colin C Young[3], Predrag Radivojac[6], Steven E Brenner[7], Marc Greenblatt[4], and Sean V Tavtigian[1]

1) Department of Oncological Sciences and Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84132.

2) Department of Family and Preventive Medicine and Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84132.

3) Ambry Genetics, Aliso Viejo, CA. USA.

4) Department of Neurobiology, University of Utah, Salt Lake City, UT 84132, USA.

5) Department of Medicine and University of Vermont Cancer Center, University of Vermont, Larner College of Medicine, Burlington, VT 05405, USA.

6) Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA.

7) Department of Plant and Microbial Biology and Center for Computational Biology, University of California, Berkeley, CA 94720, USA.

Background/Rationale: Accurate classification of genetic variants poses a significant challenge in clinical genetic testing. Computational tools address this issue by providing predictions of pathogenicity for any possible single nucleotide missense substitution (snMS). A generalized calibration of several of these tools into Bayesian points was an important step towards accelerating variant classification, with the calibrations now being incorporated into variant curation expert panel (VCEP) guidelines. However, there is concern that these genome-wide calibrations may hide significant gene-to-gene heterogeneity. Here we examine the performance of several of these tools towards snMS classification in the cancer susceptibility genes *ATM*, *BRCA1*, *CHEK2*, and *MSH2*.

Methods: Prediction scores from the tools Align-GVGD AlphaMissense, BayesDel, MutPred2, REVEL, and VEST4 were obtained for all possible snMS variants in *ATM*, *BRCA1*, *CHEK2*, and *MSH2*. Raw and rankscore values were obtained from nine additional programs (FATHMM, GERP, LRT, Mutation Assessor, MutationTaster, PhyloP, Polyphen, SIFT, and SiPhy) that are components of both BayesDel and REVEL. Odds ratios were calculated from case-control mutation screening data by logistic regression.

Results: On average, BayesDel and REVEL predicted far fewer benign snMS for *BRCA1* and *MSH2* relative to all other predictive tools combined (*BRCA1*: 11.5% vs 78.4% and *MSH2*: 1.9% vs 33.6%). The larger groupings of predicted benign variants from other callers yielded odds ratios close to 1, consistent with evidence of benignity. *ATM* and *CHEK2* do not show this behavior, with percentages of benign predictions from BayesDel and REVEL more closely matching those of other callers (*ATM*: 32.4% vs 46.4 and *CHEK2*: 34.6% vs 41.1%). Decomposition of the shared individual components between BayesDel and REVEL showed that FATHMM, the most heavily weighted component in both tools, consistently over-predicted evidence of pathogenicity for snMS in *BRCA1* and *MSH2* (65% and 94% respectively) but not in *ATM* or *CHEK2* (12% and 22%).

Discussion: Holistic review of computational tool performance across several genes, paired with case-control analysis, can reveal gene-specific limitations or strengths. Such analyses provide an alternative to the "re-call method" for comparing calibrations, and should become routine to guide decisions to accept, modify, or reject pre-calibrated genome-wide computational score thresholds.