Using genomic technology to transform how genetics is used to diagnose and treat disease

Shawn Fayer

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Doug Fowler, Chair

Lea Starita

Daniel Yang

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Using genomic technology to transform how genetics is

used to diagnose and treat disease

Shawn Fayer

Chair of the Supervisory Committee: Doug Fowler

Department of Genome Sciences

Interpreting the clinical significance of rare genetic sequence variants is challenging due limited evidence, and as a result, most newly identified missense variants are interpreted as variants of uncertain significance (VUS). Multiplexed assays of variant effet (MAVEs), where hundreds to thousands of variant effects are measured in a single experiment have significantly accelerated the rate at which functional data are generated. Since functional data can be applied when interpreting variants, MAVEs have the potential to revolutionize clinical genetics by providing functional data at scale to resolve VUS. We systematically evaluated the clinical utility of MAVEs by integrating published MAVE data with clinical interpretations and resolved 49% of VUS for *BRCA1*, 69% for *TP53*, and 15% for *PTEN*. Although we demonstrated the potential for MAVEs to resolve uncertainty in genetic testing, MAVE technologies were limited to genes with phenotypes in utilitarian cancer derived cell lines. We addressed this limitation by developing iPSC-SGE, where variants are edited into iPSCs, enabling phenotyping in differentiated cells. We introduced 498 SNVs into *POLG* and 496 variants into *MYBPC3*. *POLG* variant effects were measured with a growth assay in iPSCs in the context of different background alleles and *MYBPC3* variant effects were measured by variant abundance in cardiomyocytes. iPSC-SGE data was validated with known pathogenic and benign variants and is poised to generate functional data for genes previously inaccessible with MAVEs. Finally, we explored the use of variant effect predictors for variant interpretation, a major factor contributing to the VUS problem. We found that current calibration methods lead to inappropriate evidence for up to 75% of variants and offer a new solution for calibration via clustering VEP data for protein domains on similarity of score distributions. This method enables more accurate evidence strength thresholding while maintaining robust sets of calibration varants. Taken together, cell context specific functional data and variant specific VEP calibration will result in significant reduction to VUS while providing rich phenotypic insight for advancing precision medicine.

Contents:

Chapter 1: Variants of uncertain significance: are functional assays really to the rescue?

Genome guided precision medicine, where genes are sequenced and medical management decisions are made with the context of genomic variants found in a patient is severely limited by our inability to interpret most variants. In particular, rare single nucleotide variants that lead to missene changes in protein sequence are typically interpreted as variants of uncertain significance (VUS) because there is not enough evidence for definitive interpretation. These VUS cannot be used to guide medical management [1] and can cause confusion and distress for both patients and providers [2]. This VUS problem is rapidly expanding as sequencing has become more readily available in the past 10 years and the ClinVar database now has over 1 million unique missense VUS [3].

Clinical classification of genetic variants is performed by combining evidence from various sources including an individual's phenotype, case-control studies, family variant segregation analyses, population frequencies, *in vitro* functional assays, and *in silico* computational variant effect predictors (VEPs). Evidence from these sources is weighted as either supporting, moderate, strong, or very strong based on its accuracy in predicting pathogenicity or benignity. Guidelines set forth by the American College of Medical Genetics and Association for Molecular Pathology (ACMG/AMP) define how multiple pieces of evidence of different strengths can be combined to achieve pathogenic, likely pathogenic, likely benign, or benign interpretations [1]. These guidelines set forth by Richards et al. significantly improved the concordance of variant classification across test laboratories [4]. However, some critical limitations remained, highlighting the need for new technologies and methods for assessing variants more accurately.

Evidence from *in vitro* functional assays was among the strongest evidence in original guidelines where data from "well established functional studies" could receive strong evidence [1]. The ambiguity in this recommendation was problematic for several reasons. First, functional assays were typically conducted on single variants or very small subsets of variants and there was no recommendation on how many variants an assay needed to measure before an assay was considered "well established". Further, there was no recommendation on how to validate a functional assay. As a result, clinical labs had to make judgement calsl on whether to use functional data or not for the interpretation of a given variant. In 2019, the ACMG Sequence Variant Interpretation (SVI) Working Group released guidelines for calculating Bayesian likelihood ratios based on functional assay performance on sets of control variants [5]. These guidelines also provided a mapping from likelihood ratios to evidence strength codes, finally resolving the ambiguity of functional assay application to variant interpretation.

The adoption of Brnich et al. guidelines by the ACMG SVI Working Group was especially timely because the throughput of functional assays had recently been increased dramatically with the advent of multiplexed assays of variant effects (MAVEs) [6]; [7]. Some of the first MAVEs published in human cells used integration of transgene variant libraries via safe harbor landing pads or lentiviral vectors and measured protein abundance or cell survival in the presence of certain drugs [8–10]. A significant advancement in MAVE technology was saturation genome editing (SGE) where variants are introduced into the endogenous locus of genes which are essential to the haploid Hap1 cell line, enabling straightforward and highly accurate cell growth assays [11]. MAVEs transformed functional assays from cumbersome endeavors where variants

were measured one at a time to the measurement of hundreds to thousands of variants in a single experiment. Since MAVEs measure many variants at once, multiple controls can be included in a single experiment and the output can be robustly calibrated based on how well the assay separates pathogenic and benign control variants following Brnich et al. guidelines. However, when Brnich et al. published their guidelines, there was no systematic analysis of the clinical impact of MAVE data with this updated guidance.

Pioneers of MAVE technologies recognized the rapidly expanding VUS problem in clinical genetics and speculated about the clinical impact MAVEs could have to reduce VUS announcing “functional data to the rescue” [12]. This early vision of the potential for MAVE technologies to revolutionize clinical genetics included recognition of the limitations of MAVE technologies. Perhaps most importantly, MAVEs had so far been performed in utilitarian cancer cell lines, preventing their application to disease genes which cause phenotypes only in differentiated cell types. Saturation genome editing (SGE) for example, is limited to the ~2,000 genes essential to the Hap1 cell line [13]; [11]. Most known human disease genes are not essential to Hap1 cells nor are they expressed in the Hap1 cell line. For these reasons, it is essential to expand MAVEs into additional cellular systems, particularly iPSC differentiation models where variants can be measured in the correct cell or tissue context.

Contextualizing MAVE data in variant interpretation is essential for understanding the potential for MAVEs to revolutionize clinical variant interpretation and precision medicine in general. For example, a MAVE that performs perfectly on a set of control variants cannot be used as stand-alone evidence to classify variants as pathogenic or benign under the current ACMG/AMP guidelines. Often the only additional evidence available for rare missense variants is data from computational variant effect predictors (VEPs). VEPs typically score variant pathogenicity based on models built around protein multiple sequence alignments and may be supervised on clinical variants [14–16] or unsupervised [17–20]. In either case, VEP data exist for virtually every possible missense variant. The impact that VEPs have had on variant interpretation, however, has been limited due to the Richards et al. guidelines limiting VEP data to supporting evidence.

Recognizing the limitation imposed on VEP data, the ACMG SVI working group released new recommendations for the use VEPs in variant interpretation [21]. This recommendation involved calibration of VEPs on an aggregate set of ClinVar variants across all human disease genes to provide sufficient clinically labeled variants for calibration of models. This innovative method used local calibration where likelihood ratios were calculated for sliding windows across VEP score ranges defining score thresholds for supporting, moderate, and even strong evidence for some VEPs. This method also solved the problem of too few ClinVar variants per gene, by aggregating all available variants across genes. While this represented a massive step forward in the use of VEPs in variant interpretation, this calibration method neglects the varied performance of predictors across different genes and could result in incorrect evidence for certain thresholds across certain genes. Thus, a new method for VEP calibration is needed to more accurately assign evidence to variants while still maximizing the potential of VEP data for variant interpretation.

In this dissertation, I describe our initiatives to 1) establish the clinical utility of MAVE data by integration of data from several published MAVEs with variant interpretation from a large genetic testing lab, 2) generate a more generalizable MAVE technology, iPSC-SGE where

variant libraries are edited into stem cells and variant effects can be measured in differentiated cell types, and 3) investigate and develop methods for calibration of variant effect predictors for more accurate variant interpretation.

References:

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

2. Mighton, C., Shickh, S., Uleryk, E., Pechlivanoglou, P. & Bombard, Y. Clinical and psychological outcomes of receiving a variant of uncertain significance from multigene panel testing or genomic sequencing: a systematic review and meta-analysis. *Genet Med* **23**, 22–33 (2021).

3. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* **53**, D1313–D1321 (2025).

4. Amendola, L. M. *et al.* Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies. *Am J Hum Genet* **107**, 932–941 (2020).

5. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).

6. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801–807 (2014).

7. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc* **11**, 1782–1787 (2016).

8. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874–882 (2018).

9. Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet* **50**, 1381–1387 (2018).

10. Boettcher, S. *et al.* A dominant-negative effect drives selection of missense mutations in myeloid malignancies. *Science* **365**, 599–604 (2019).

11. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).

12. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet* **101**, 315–325 (2017).

13. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

14. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877–885 (2016).

15. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* **11**, 5918 (2020).

16. Wu, Y., Li, R., Sun, S., Weile, J. & Roth, F. P. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet* **108**, 1891–1906 (2021).

17. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* **53**, D523–D534 (2025).

18. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

19. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

20. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

21. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* **109**, 2163–2177 (2022).

Chapter 2: Closing the gap: systematic integration of multiplexed functional data resolves variants of uncertain significance in *BRCA1*, *TP53* and *PTEN*

Shawn Fayer[1], Carrie Horton[7], Jennifer N. Dines[1], Alan F. Rubin[5,6], Marcy E. Richardson[7], Kelly McGoldrick[7], Felicia Hernandez[7], Tina Pesaran[7], Rachid Karam[7], Brain H. Shirts[3,4], Douglas M. Fowler[1,2,3*], Lea M. Starita[1,3,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

[2]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

[3]Brotman Baty Institute for Precision Medicine, Seattle, WA 98195, USA

[4]Department of Laboratory Medicine and Pathology, University of Washington 98195, Seattle, WA, USA

[5]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia

[6]Department of Medical Biology, University of Melbourne, Melbourne, VIC 3010, Australia

[7]Ambry Genetics, Aliso Viejo, CA 92656, USA

Abstract

Clinical interpretation of missense variants is challenging because the majority identified by genetic testing are rare and their functional effects are unknown. Consequently, most variants are of uncertain significance, and cannot be used for clinical diagnosis or management. Although not much can be done to ameliorate variant rarity, multiplexed assays of variant effect (MAVEs), where thousands of single nucleotide variant effects are simultaneously measured experimentally, provide functional evidence that can help resolve variants of unknown significance (VUS). However, a rigorous assessment of the clinical value of multiplexed functional data for variant interpretation is lacking. Thus, we systematically combined previously published *BRCA1*, *TP53*, and *PTEN* multiplexed functional data with phenotype and family history data for 324 VUS identified by a single diagnostic testing laboratory. We curated 49,281 variant functional scores from MAVEs for these three genes and integrated four different *TP53* multiplexed functional datasets into a single functional prediction for each variant using machine learning. We then determined the strength of evidence provided by each multiplexed functional dataset and reevaluated 324 VUS. Multiplexed functional data was effective in driving variant reclassification when combined with clinical data, eliminating 49% of VUS for *BRCA1*, 69% for *TP53*, and 15% for *PTEN*. Thus, multiplexed functional data, which are being generated for numerous genes, are poised to have a major impact on clinical variant interpretation.

Introduction

Targeted panel testing for cancer predisposition is now widespread, and, as panels increase in size, the likelihood of identifying a rare missense variant of uncertain significance (VUS) also increases[1,2]. As a result, these inconclusive VUS results are commonly returned to individuals. Multigene panel testing for hereditary cancer predisposition, for example, identifies one or more VUS in 40% of individuals who are tested for suspicion of cancer predisposition[2]. Most VUS are missense variants, which can be challenging to interpret as pathogenic or benign according to the American College for Medical Genetics and Genomics Association for Molecular Pathology (ACMG/AMP) guidelines[3]. This challenge is largely due to the fact that missense variants are typically rare, making clinical evidence such as segregation and case-control data scarce. Of the nearly 330,000 missense variants from clinical testing in the ClinVar database, 70% are VUS[4] (**Figure 1A**). The VUS problem has grown exponentially over time (**Figure 1B**) and exists even among extremely well studied cancer predisposition genes such as *BRCA1*, *TP53* and *PTEN*, where the majority of missense variants reported in ClinVar from clinical genetic testing are VUS (*BRCA1* 80%, n = 2,395; *TP53* 64%, n = 589; and *PTEN* 72%, n = 411).

Definitive variant interpretation in *BRCA1*, *TP53* and *PTEN* is critical because morbidity and mortality can be reduced for individuals known to harbor pathogenic variants through increased cancer surveillance and preventative measures[5–7]. In contrast, medical management is not changed for individuals who carry a VUS, and the uncertainty surrounding these variants can provoke anxiety[8] at best or represent a missed opportunity to provide life-saving care at worst. Thus, improved interpretation of VUS directly impacts the well-being of carriers and their families. Furthermore, the ACMG recommends that pathogenic variants in all three of these genes be returned to individuals regardless of the indication for testing[9,10]. However, the recommendations for return of secondary findings are limited to established pathogenic and likely pathogenic variants and secondary VUS are not typically shared with individuals[9]. Thus, individuals may be left in the dark about increased risk if their VUS are reclassified as pathogenic, further highlighting the need for timely VUS resolution.

While little can be done to change the lack of information arising from a variant's rarity, it is now possible to generate variant functional data at scale. Models suggest that functional evidence could lead to the reclassification of most VUS[11], however that has not been tested on a large scale using real-world data. Multiplexed assays of variant effect (MAVEs), where thousands of single nucleotide variants are assayed simultaneously, have been applied to *BRCA1* (MIM: 113705)[12], *TP53* (MIM: 191170)[13,14], and *PTEN* (MIM: 601728)[15,16], producing functional annotations for thousands of variants that can be used as evidence to resolve VUS[17–20]. Recently, guidelines for both generating and using multiplexed functional data have been developed, and create the opportunity to systematically explore the clinical value of multiplexed functional data in the reinterpretation of VUS[21,22]. However, the extent to which multiplexed functional data can result in medically significant variant reinterpretation has not been systematically evaluated.

Thus, we assessed the clinical value of *BRCA1*, *TP53*, and *PTEN* multiplexed functional data by integrating them with existing lines of evidence from clinical variant interpretations for

these genes. First, we curated 49,281 variant functional scores for clinical integration from multiplexed assays across the three genes[21]. Then, we determined the strength of evidence for the functional evidence component of variant interpretation provided by each multiplexed functional dataset based on its ability to predict established pathogenic and benign variants[22] and reevaluated 324 VUS classifications (*BRCA1* = 110, *TP53* = 166, *PTEN* = 48) (**Figure 2**). Multiplexed functional data, when combined with existing lines of evidence, resulted in reclassification of 49% of VUS for *BRCA1*, 69% for *TP53* and 15% for *PTEN.* Thus, multiplexed functional data can help to resolve a large percentage of VUS, highlighting the utility of generating and curating multiplexed functional data. Our analysis revealed two major factors that limited the utility of multiplexed functional data: the modest predictive value of some MAVEs and the scarcity of established pathogenic or benign variants that serve as validation controls for some genes. Based on these findings, we discuss how MAVEs should be designed and piloted with clinical utility in mind, prioritizing genes with established pathogenic and benign variants. We conclude by discussing the overall implications of multiplexed functional data for variant reclassification.

Materials and Methods

Human subjects

This study was approved by the University of Washington (UW) Institutional Review Board STUDY00003598. Data was deidentified for transfer to UW, additional consent was not required.

Clinical data collection:

Data were provided from 9,234 individuals who were found to have a variant in *BRCA1*, *TP53*, and/or *PTEN* by Ambry Genetics in multigene panel testing for cancer predisposition on or before 9/25/2019. Interpretation of sequence variations was performed according to the American College of Medical Genetics and Genomics guidelines[3]. Variants were classified as pathogenic, likely pathogenic, variant of unknown significance (VUS), likely benign, or benign according to the Ambry 5-tier variant classification protocol[23]. Ambry classifications follow a modified version of ACMG/AMP guidelines.

Clinical data filtering:

Individual phenotype and variant data from clinical testing were excluded from this analysis when they met certain exclusion criteria: 1) an individual with a VUS in *BRCA1*, *TP53* or *PTEN* was excluded if that individual was found to have a pathogenic or likely pathogenic variant in another cancer predisposing gene on the multigene panel, 2) an individual with a VUS in *BRCA1*, *TP53* or *PTEN* was excluded if they were found to have one or more additional VUS in other cancer predisposing genes, and 3) an individual with a pathogenic variant in *BRCA1*, *TP53*, or *PTEN* were excluded if they were found to have a pathogenic variant in any other gene. In total, 2,437 of the 9,234 individuals from the clinical data were excluded in our analysis. We used data from 4,723 individuals with a *BRCA1* variant, 1,334 individuals with a *TP53* variant, and 740 individuals with a *PTEN* variant.

Naïve Bayes classifier:
Naïve Bayes classification of *TP53* variants was conducted with the Gaussian naïve Bayes functionality from the Python SciKit Learn library[24]. Prior probabilities were set at 0.5, the default for binary classification, for functionally normal and functionally abnormal classes. Variants used for training were all of the clinically derived ClinVar pathogenic/likely pathogenic and benign/likely benign variants. After training the classifier on function scores from all four *TP53* MAVEs, performance was evaluated with leave-one-out cross-validation. Additional naive Bayes classifiers were trained on function scores from single *TP53* MAVEs and evaluated with leave-one-out cross-validation. Since the four feature classifier had the highest overall accuracy, we used this classifier to make predictions of TP53 variant functional effects.

Variant reinterpretation:
Variants from multigene cancer panels were reinterpreted with multiplexed functional data following ACMG/AMP rules-based guidelines and a Bayesian adaptation for ACMG/AMP rules[3,25]. For *TP53* and *PTEN*, ClinGen Variant Curation Expert Panel (VCEP) adaptations of the ACMG rules were used in reinterpretation with the exception of the functional data evidence code where we exclusively used MAVE data[17,26]. For *BRCA1*, BayesDel[27] predictions were used as computational predictive model data with thresholds of less than 0.147 for benign evidence (BP4) and greater than 0.425 for pathogenic evidence (PP3)[27,28]. For *BRCA1* splice region variants, SpliceAI[29] predictions were used as computational predictive model evidence with a threshold of greater than 0.8 for pathogenic evidence (PP3)[29]. In addition, absence from Gnomad[30] was used as pathogenic population data for *BRCA1* and restricted to supporting level of evidence (PM2_P). All other evidence codes were applied as recommended in the original ACMG guidelines or VCEP adaptations. The Bayesian implementation of the ACMG guidelines was performed as previously described, with prior probability of pathogenicity was set at 0.1 for all variants[25].

Results

Multiplexed functional data curation

First, we evaluated whether each multiplexed functional data set was compatible with three key recommendations for clinical integration[21]. All assays used in this analysis met the first criterion: the function scores generated for each variant must be a direct measure of variant effect. Next, the dynamic range of the assay must be sufficient to separate variant effects targeted by the assay from synonymous variants. For example, an assay designed to detect loss of function variants must have a readout that is able to separate nonsense variants from synonymous variants. When this criterion is met, variants that score like nonsense variants are called “functionally abnormal” and variants that score like synonymous variants are called “functionally normal”. Finally, the assay should have high sensitivity and specificity for clinically ascertained control variants, where benign variants are scored as functionally normal and pathogenic variants are scored as functionally abnormal.

Finally, the strength of evidence that could be applied to multiplexed functional data for variant interpretation was determined following recommendations from the ClinGen Sequence Variant Interpretation (SVI) Working Group[22]. According to these recommendations, the strength

of evidence generated by a functional assay is determined by how well the assay can distinguish between control benign and pathogenic variants and how many control variants are available for this comparison. The resulting odds of pathogenicity (OddsPath) corresponds to strength of evidence codes from the original ACMG/AMP guidelines for variant interpretation: supporting, moderate or strong[3,25].

BRCA1 multiplexed functional data curation

Multiplexed functional data for 3,893 single nucleotide variants (SNVs) of *BRCA1* were generated with saturation genome editing[12] (**Table 1**). We chose this MAVE to assess because it is the largest and most accurate for *BRCA1*. These functional data were suitable for clinical integration because the assay result was directly linked to variant effect. Here, cells with functionally abnormal *BRCA1* variants were depleted relative to wild type after growth selection. Furthermore, the dynamic range of function score distributions of the assay was sufficient to separate functionally abnormal nonsense variants from functionally normal synonymous variants yielding clear thresholds for functionally normal and functionally abnormal variants (**Figure 3A, B**). Using these thresholds, the functional data cleanly separated ClinVar pathogenic and benign variants (AUCPR = 0.97) (**Figure S1**). Thus, we used the published *BRCA1* functional classifications for this study.

To determine the strength of evidence applied to *BRCA1* multiplexed functional data for variant interpretation, we calculated the OddsPath using all clinically derived pathogenic/likely pathogenic and (n=209) and benign/likely benign (n=163) variants from ClinVar as *BRCA1* control variants. The multiplexed functional data correctly assigned 198 of the 209 pathogenic/likely pathogenic controls to the functionally abnormal class and 159 of the 163 benign/likely benign controls to the functionally normal class, resulting in an OddsPath of 52.4 and 0.02, respectively (**Table S1-S2**). These OddsPath values correspond to strong evidence for pathogenic assessment (PS3) of functionally abnormal variants and strong evidence for benign assessment (BS3) of functionally normal variants[22]. Thus, we applied PS3 and BS3 evidence codes to *BRCA1* variants with functionally abnormal and normal scores, respectively.

TP53 multiplexed functional data curation

We chose to explore four existing MAVEs generated to interrogate *TP53's* multiple functions. Here, multiplexed functional data for 8,258 SNVs of *TP53* was generated using two assays that queried loss of function variants and two that queried dominant negative variants[13,14]. Although these *TP53* functional datasets enabled powerful dissections of the molecular mechanisms of *TP53* variant effect, they do not individually meet recommendations for multiplexed functional data clinical integration. In particular, while each assay was conducted in relevant human cell lines and was designed with readouts directly linked to multiple molecular consequences of *TP53* variant pathogenesis, none cleanly separated nonsense and synonymous variants nor known pathogenic from known benign variants (**Figure 3C-J**). Due to the complex landscape of TP53 functional effects, we could not make accurate predictions of variant functional effect for clinical variant interpretation based on any single assay. Therefore, we trained a classifier to predict variant function. Since the function score generated for each variant in each assay represents an independent data point, we used a Gaussian naïve Bayes classifier to make predictions of functional effect for each variant using their scores from all four

assays without transformation. We trained the classifier on the 161 *TP53* missense variants from ClinVar (129 pathogenic/likely pathogenic and 32 benign/likely benign) (**Table S3**) and assessed accuracy of predictions with leave-one-out cross-validation (accuracy = 96%, AUCPR = 0.92). When compared with classifiers trained using function scores from any single assay, the classifier using scores from all four assays performed with greater accuracy (**Figure S2**). Thus, we used the naïve Bayes classifier trained on all four assays to generate predictions of variant functional effect for the 7,893 *TP53* variants with function scores in each of the four assays (**Table S4**).

We determined the strength of evidence yielded by our *TP53* classifier's functional predictions based on cross-validation performance. The overall accuracy of the predictions was 96% and all seven misclassified variants were likely pathogenic variants that were classified as functionally normal (**Tables S4 and S5**). This corresponds to an OddsPath of 30.3 for variants predicted to be functionally abnormal and 0.054 for variants predicted to be functionally normal (**Table S1**). Thus, we applied strong functional evidence (PS3) to the variants predicted to be functionally abnormal and moderate functional evidence (BS3_M) to the variants predicted to be functionally normal[22]. Although the classifier performs with high overall accuracy, some clinically interpreted pathogenic/likely pathogenic variants cannot be detected by the multiplexed assays by which it was trained. Four of these incorrectly classified variants retain partial function in other functional assays which may lead to the functionally normal output in the human cell line overexpression systems used in the *TP53* MAVEs (NM_000546.5(TP53):c.1000G>T (p.Gly334Trp), NM_000546.6(TP53):c.579T>A (p.His193Gln), NM_000546.6(TP53):c.542G>A (p.Arg181His), NM_000546.6(TP53):c.1010G>A (p.Arg337His)[31–33].

***PTEN* multiplexed functional data curation**

Finally, we explored the existing MAVEs for *PTEN*. Effects of 8,088 SNVs were measured using two assays, one for variant effects on protein abundance and the another for variant effects on PTEN phosphatase activity levels[15,16]. Both protein abundance and phosphatase activity scores for *PTEN* variants were direct measures of variant effect. The dynamic range of both assays are sufficient to separate nonsense variants from synonymous variants. However, the sensitivity of both assays is only 0.69 for abundance and 0.71 for activity, respectively when validated against established pathogenic variants in ClinVar (**Figure 3K-N**). Due to the low sensitivity for pathogenic variants of each assay and the dearth of benign variants to calculate OddsPath to determine strength of evidence, the multiplexed functional data do not provide any evidence toward pathogenicity and are capped at BS3 moderate for the phosphatase activity assay and BS3 supporting for the abundance assay (**Tables S1 and S6**). While combining the two datasets with a probabilistic classifier might improve the sensitivity, specificity and strength of evidence, there are too few benign missense *PTEN* variants (n = 2) available for training.

Variant reinterpretation

To understand the impact of adding evidence acquired from multiplexed functional data to the reinterpretation of VUS, we gathered all available evidence for the 324 VUS identified in *BRCA1*, *TP53*, and *PTEN* by a single diagnostic testing laboratory that overlapped these datasets. We then reinterpreted these VUS using multiple established variant interpretation

guidelines to ensure that our conclusions were independent of the approach taken. First, we reinterpreted variants using guidelines either from the American College of Medical Genetics and Association for Molecular Pathology (ACMG/AMP)[3] or from ClinGen Variant Curation Expert Panels (VCEPs)[17,26]. We also reinterpreted variants using a Bayesian adaptation of the ACMG/AMP guidelines[25]. This strategy differs from original, rules-based ACMG/AMP guidelines because the final interpretation is made based on a posterior probability that a variant is pathogenic after combining all evidence in a quantitative framework instead of using evidence codes.

***BRCA1* variant reinterpretation**

We obtained data for 286 *BRCA1* single nucleotide variants identified in clinical multigene cancer panels from 6,490 individuals that have multiplexed functional data. These variants were classified by a single diagnostic laboratory (Ambry Genetics) as pathogenic (n=56), likely pathogenic (n=44), VUS (n=110), likely benign (n=16), and benign (n=60). Of these, 93% were scored as functionally normal or abnormal in the multiplexed functional assay (functionally normal n=156, functionally abnormal n=109). The remaining 7% scored in the intermediate range (n=21) of the assay between the thresholds defining functionally normally and abnormal variants[12]. The clinical interpretations were highly concordant with the multiplexed functional data, with 54 of the 56 pathogenic variants scoring as functionally abnormal and 57 of the 60 benign variants scoring as functionally normal. All five of the discordant pathogenic and benign variants had intermediate functional scores.

Reinterpretation of the *BRCA1* VUS from the laboratory dataset with the multiplexed functional data following the ACMG/AMP guidelines[3] resulted in the reclassification of 49% (54/110) VUS as likely pathogenic (n = 5) or likely benign (n = 49) (**Figure 4A; Table S7**). Of the 110 VUS, 15 scored as functionally abnormal, 82 scored as functionally normal, and 13 as intermediate. In addition to the multiplexed functional data, other existing lines of evidence used to reclassify VUS to likely pathogenic included: missense variant at a position where another missense variant is classified as pathogenic (PM5, n=3), variant absent in population databases (PM2_P, n=4), and agreement between computational predictive models (PP3, n=3). For VUS reclassified as likely benign, the additional evidence was agreement between computational predictive models (BP4, n=49). In cases where variants had enough evidence to be classified as likely benign and absence from population databases was the only conflicting evidence in favor of pathogenicity, we ignored the conflicting evidence and classified these variants as likely benign (n=26)[26]. We note that our reinterpretation of variants with strict adherence to the ACMG/AMP guidelines limits the allowable evidence for variant interpretation that many clinical laboratories routinely use in their interpretations including data siloed to protect PHI and other conflicting lines of evidence. Thus, we expect the application of this ACMG/AMP framework to yield a different proportion of VUS as compared to reinterpretation of the same variants by a clinical laboratory.

We also reinterpreted variants following the Bayesian implementation of the ACMG/AMP guidelines and resolved 80% of the (91/110) *BRCA1* VUS with this method (likely pathogenic, n=5; likely benign, n=84, Table S2). The additional VUS classified as likely benign using the Bayesian method were variants that scored as functionally normal in the multiplexed functional data, but had no other evidence in favor of benign interpretation. Since the Bayesian method

allows for likely benign interpretation for any variant with a posterior probability of pathogenicity <0.1, variants with only functional evidence from an assay achieving strong functional evidence can be classified as likely benign even in the absence of other evidence types. Thus, using either the original or Bayesian adaptation of the ACMG guidelines it is clear that the multiplexed functional data has a high impact in the clinical significance interpretation for *BRCA1* variants.

***TP53* variant reinterpretation**

We obtained data for 294 *TP53* SNVs identified in clinical multigene cancer panels from 1,828 individuals that have multiplexed functional data. These variants were classified by Ambry Genetics as pathogenic (n=37), likely pathogenic (n=60), VUS (n=166), likely benign (n=16), and benign (n=60). Our classifier predictions were largely concordant with these clinical interpretations. Of the 49 variants in this clinical dataset that are absent from the classifier training set, all of the likely benign variants were functionally normal (n=18), 21 of the 26 likely pathogenic variants were functionally abnormal, and all pathogenic variants were functionally abnormal (n=5). All five of the discordant variants have conflicting interpretations in ClinVar, and one is a ClinGen Expert Panel reviewed VUS (**Table S8**). Two of these discordant variants (NM_000546.5(TP53):c.1000G>C (p.Gly334Arg), NM_000546.6(TP53):c.542G>A (p.Arg181His)) have been described as reduced penetrance variants[34–36]. The remaining 166 *TP53* variants from the diagnostic laboratory dataset were VUS, and our classifier predicted 120 to be functionally normal and 46 to be functionally abnormal. To determine the value of *TP53* multiplexed functional data in variant interpretation, we reevaluated the VUS with functional evidence from the classifier.

Reevaluation of *TP53* VUS following an adapted version of the ClinGen VCEP recommendations[26] resulted in the reinterpretation of 69% (115/166) of VUS to likely pathogenic (n = 30) or likely benign (n = 85) (**Figure 4B, Table S9**). Of the 166 *TP53* VUS, 120 were predicted to be functionally normal and 46 were predicted to be functionally abnormal by our classifier. The adapted reinterpretation strategy followed the VCEP guidelines for *TP53* variant interpretation except for the functional evidence component, where we used the functional evidence corresponding only to our classifier predictions. This distinction was made to assess the impact of data from only multiplexed functional assays. In addition to the multiplexed functional data, existing lines of evidence used to reclassify VUS to likely pathogenic included: missense variant at a position where another missense variant is classified as pathogenic (PM5_P, n=3), variant absent in population databases (PM2_P, n=28), missense variant in a mutational hotspot (PM1, n=9), and agreement between computational predictive models (PP3/PP3_M, n=28). For VUS reclassified as likely benign, additional evidence used was agreement between computational predictive models (BP4, n=85) and variant observed in adults unaffected with cancer in population datasets (BS2_P, n=4). We also reinterpreted variants with strict adherence to the *TP53* VCEP functional evidence recommendation, which includes two of the multiplexed datasets used in our analysis, and reclassified 60% of VUS to likely pathogenic (n=19) or likely benign (n=81) (**Table S10**).

Finally, reinterpretation of *TP53* variants following the Bayesian adaptation increased the proportion of resolved VUS to 85% (likely pathogenic (n = 30), likely benign (n = 111)). Similar to *BRCA1*, the majority (25 of 26) of additional VUS reclassified as likely benign using the Bayesian approach were variants where no additional evidence could be applied. Since the

posterior probability of pathogenicity was below 0.1, these variants were classified as likely benign. The lone variant with additional benign evidence (NM_000546.5(TP53):c.328C>A (p.Arg110Ser)) was predicted to be benign by computational predictive models, but was also absent from population databases and occurs at the amino acid position of another pathogenic missense variant. These conflicting pieces of evidence result in a VUS interpretation following rules-based methods, but with the Bayesian method, this variant has a posterior probability of 0.05 and is interpreted as likely benign.

***PTEN* variant reinterpretation**

We obtained data for 74 *PTEN* missense variants identified in multigene cancer panels from 1,061 individuals that had multiplexed functional data from at least one of the *PTEN* assays. These variants were classified by Ambry Genetics as pathogenic (n=7), likely pathogenic (n=17), VUS (n=48), and likely benign (n=2). Both likely benign variants had functionally normal scores from the activity and abundance assays. The seven pathogenic variants were assessed in the phosphatase activity assay. Four scored as low activity and three as intermediate activity. Six of the pathogenic variants were assessed in the abundance assay. Five had low abundance and one had normal abundance. 16 of the likely pathogenic variants were scored in the phosphatase activity assay and 11 were scored as low activity, two as intermediate, and three as normal activity. Of the 13 likely pathogenic variants that were scored in the abundance assay seven had low abundance and 6 had normal abundance. This low sensitivity of each assay for pathogenic/likely pathogenic variants from the diagnostic laboratory dataset is consistent with the assessment using variant annotations from ClinVar.

We attempted to reinterpret *PTEN* VUS with MAVE data following the ClinGen SVI recommendations and were unable to reclassify any of the VUS due to limited strength of evidence (**Tables S1 and S11**). For this reason, we employed guidelines developed by the *PTEN* variant curation expert panel (VCEP) for clinical integration of these datasets[17]. Here, variants deemed functionally abnormal in the phosphatase activity assay receive PS3 strong evidence. Variants deemed functionally abnormal in the multiplexed protein abundance assay receive PS3 supporting evidence. Variants deemed functionally normal in both assays receive BS3 strong evidence. Following the VCEP recommendations[17] with functional evidence restricted to multiplexed functional data resulted in reclassification of 15% of VUS as likely pathogenic (n=7) (**Figure 4C, Table S12**). From the multiplexed functional data, we assigned strong functional evidence for pathogenicity (PS3) to eight of the VUS, supporting functional evidence for pathogenicity (PS3_P) to five VUS, and strong functional evidence for benign effect (BS3) to 22 VUS. In addition to the multiplexed functional data, existing lines of evidence applied to VUS that were reclassified to likely pathogenic include missense variant in a gene with low rate of benign missense variation (PP2, n=7), absence in population databases (PM2, n=7), and missense variant at a position where another missense variant is classified as pathogenic (PM5, n=3). None of the VUS with BS3 evidence could be reclassified to likely benign following *PTEN* VCEP guidelines for two main reasons: 1) there is no consensus for use of *in silico* predictive models for *PTEN* missense variants due to the lack of benign control missense variants and 2) these VUS are present at too low population frequency for BS1 strong evidence. Finally, we note that the *PTEN* VCEP guidelines predate the updated ClinGen SVI

recommendations for the use of functional data, and thus, may not meet current standards for strength of evidence attributable to functional data.

Discussion

Multiplexed functional data is appearing rapidly, but has, so far, not been rigorously evaluated for clinical use. Here, we demonstrated that multiplexed functional data has high utility in clinical variant interpretation by using functional evidence for 49,281 variants derived from MAVEs to reevaluate 324 variants of uncertain significance in three important cancer predisposition genes, *BRCA1, TP53* and *PTEN,* from 774 individuals. Overall, the multiplexed functional data enabled reclassification of 176 (54%) of the 324 VUS, with considerable differences in effectiveness across the genes.

Reinterpretation of *BRCA1* VUS with multiplexed functional data was straightforward because a single functional dataset closely adhered to guidelines for multiplexed functional assay design, the data had high sensitivity and specificity for control variants, and a large number of control variants were available for computing the strength of evidence. These factors allowed PS3 strong and BS3 strong levels of evidence to be used for reinterpretation, resulting in the reclassification of 49% of VUS.

Reinterpretation of *TP53* VUS was complicated by the existence of four distinct multiplexed functional datasets, with no single dataset having a broad enough dynamic range or sufficient sensitivity and specificity to be used alone. We combined these datasets with a naïve Bayes classifier, which was possible because *TP53* has a large number of established benign and pathogenic variants to use for classifier training. Thus, multiple MAVE datasets can be integrated to produce a single, accurate functional prediction that can be used for variant reinterpretation. This approach is particularly important when each functional assay lacks the requisite dynamic range and sensitivity/specificity alone, or when variants in a gene, like *TP53,* have multiple mechanisms of pathogenicity that cannot be probed in a single assay. Ultimately, our machine learning approach, in combination with the large number of available control variants for *TP53*, allowed strong pathogenic (PS3) and moderate benign (BS3_M) levels of functional evidence to be used to reclassify 69% of VUS.

Reinterpretation of *PTEN* VUS highlighted other limitations of multiplexed functional data. Here, lack of benign control variants constrained assay validation, preventing the use of machine learning to combine functional data from multiple assays and also meaningful assessment of the strength of evidence. In addition, this lack of benign control variants inhibited the utility of *in silico* predictors, an important contributor to variant interpretations. Nonetheless, despite these challenges, functional evidence enabled reclassification of 15% of VUS with application of VCEP recommendations for *PTEN* variant interpretation. This analysis highlights the important role of ClinGen VCEPs in developing guidelines for functional data integration for genes where control variants are limited, as we could not reclassify any *PTEN* VUS following the generalized ClinGen SVI recommendations for use of functional data[22].

By analyzing multiple genes and functional data sets, we reveal the key characteristics of each assay and gene that dictate the utility of multiplexed functional data for variant interpretation. Key assay characteristics are the dynamic range of function scores separating functionally normal from abnormal variants, and the predictive value of the assay for correctly identifying known pathogenic and benign variants used as controls. The key gene characteristic

is the availability of pathogenic and benign control variants for assay validation. As for *PTEN*, lack of control variants can severely constrain the strength of evidence arising from functional assays regardless of assay performance.

The MAVE community is making progress in developing assays with high dynamic range and high predictive value[37–40]. However, addressing the lack of benign and pathogenic control variants for validation is more challenging. A perfect assay must have at least 19 control benign *and* pathogenic variants for validation to achieve strong benign and pathogenic functional evidence, whereas at least 11 benign and pathogenic variants are required to achieve moderate evidence[22]. Of the 73 clinically actionable genes on the ACMG Secondary Findings v3.0 list[10], only 23 genes have a sufficient number of control missense variants for a hypothetical MAVE that perfectly distinguishes the pathogenic and benign variants to achieve strong functional evidence (PS3) for interpretation of functionally abnormal variants (**Figure 5B, Table S13**). The 50 genes that do not reach this threshold are limited by the number of control benign variants available for assay validation or lack established pathogenic missense variants. In addition, just 40 of these genes have sufficient control missense variants for a hypothetical MAVE that perfectly distinguishes pathogenic and benign variants to achieve strong evidence for benign interpretation (BS3) of functionally normal variants (**Figure 5A**). Thus, most genes lack the control variants required to deploy variant functional data as strong evidence, and closing this control variant gap will likely require active efforts to generate control variants and changes to variant interpretation practices. For genes like *PTEN* where multiplexed functional data is available but benign control variants are lacking, expert panels could coordinate review of VUS and variants discovered in population sequencing studies. Data generators and clinicians should work together to design, pilot and validate MAVEs, as well as to integrate the resulting functional data into interpretation workflows in order to maximize the potential of these powerful data[21,40]. Ultimately, a confluence of high-quality multiplexed functional datasets, large numbers of control variants and reassessment of interpretation guidelines could transform our ability to definitively interpret VUS.

Supplemental data

Supplemental data include 1 figure and 13 tables.

Declaration of interests

JND is an employee of Adaptive, MER, KM, FH, TP and RK are employees of Ambry Genetics. The remaining authors declare no conflicting interests.

Acknowledgements

We thank Moez Dawood for critical reading of this manuscript and Martha Horike-Pyne for assistance with human subjects. This work was supported by 5T32HG000035 fellowship from the NIH NHGRI administered by UW Genome Sciences to SF, a Catalytic Collaboration award from the Brotman Baty Institute to DMF, a 5U01CA242954 from the NIH NCI to LMS and 5RM1HG010461 from the NHGRI to DMF and LMS. This research benefitted by support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support.

Data and code availability

Code to reproduce the analysis and regenerate all figures is available on GitHub at https://github.com/bbi-lab/VUS-reinterpretation-with-MAVE.

Figure Legends

Figure 1: Missense variants of uncertain significance are a large and growing problem. A. Single nucleotide missense variants colored by ClinVar classifications (Benign = 25,707; likely benign = 16,377; VUS = 227,365, likely pathogenic = 14,716, pathogenic = 22,489, conflicting interpretations = 20,026). ClinVar data downloaded on 10/27/2020. **B.** Missense variants in ClinVar from 2015 to 2020 shown by clinical significance.

Figure 2: Schematic for integration of multiplexed fuctional data into clinical variant interpretation. Top panel: We first collected variant function scores and determined assay dynamic range and sensitivity and specificity for established pathogenic and benign variants. If a single assay had high sensitivity and specificity we used the function scores directly to determine which variants were functionally normal and functionally abnormal. Where possible, we combined multiple MAVE datasets to increase predictive value of function scores and determine the functional class of variants. Finally, we computed the odds of pathogenicity for the assigned functional classes to determine the strength of evidence assigned to each dataset. Bottom panel: Existing evidence for 324 VUS were combined with the MAVE functional evidence to reinterpret variants as either likely pathogenic (orange), likely benign (blue), or VUS (gray).

Figure 3: Function scores for *BRCA1*, *TP53,* and *PTEN* variants of known effect. Histograms of function scores for variants colored by their ClinVar interpretations for each multiplexed functional assay in the left column and nonsense and synonymous variant distributions in the right column. **A, B.** Function scores for *BRCA1* derived from saturation genome editing in a *BRCA1* deficient HAP1 cell line. **C-J.** Function scores for *TP53* derived from four different assays. From top to bottom: *TP53*-null A549 cell line with positive selection for loss of function variants with etoposide. *TP53*-null A549 cell line with negative selection for loss of function variants with nutlin-3. *TP53*-wild-type A549 cell line with positive selection for dominant negative variants with nutlin-3. *TP53*-wild-type AML reporter cell line with positive selection for dominant negative variants with nutlin-3. **K-N.** Function scores for *PTEN* derived from two different assays. From top to bottom: PTEN variant abundance assayed in a HEK293 cell line and PTEN variant phosphatase activity in a humanized yeast system. Histogram color indicates

known clinical effect as reported in the ClinVar database (dark blue = benign, light blue = likely benign, light red = likely pathogenic, dark red = pathogenic).

Figure 4 legend: Reinterpretation of *BRCA1, TP53, and PTEN* VUS using multiplexed functional data. A-C: original variant classifications from Ambry Genetics. **D-F:** Variant classifications after reinterpretation using existing evidence and multiplexed functional data. Dashed sections represent the proportion of VUS reclassified to either likely pathogenic or likely benign.

Figure 5: Strength of evidence that could be assigned to variants of ACMG Secondary Findings v3.0 genes with hypothetical MAVEs that perfectly distinguish between pathogenic and benign controls.

Figures

Figure 1:

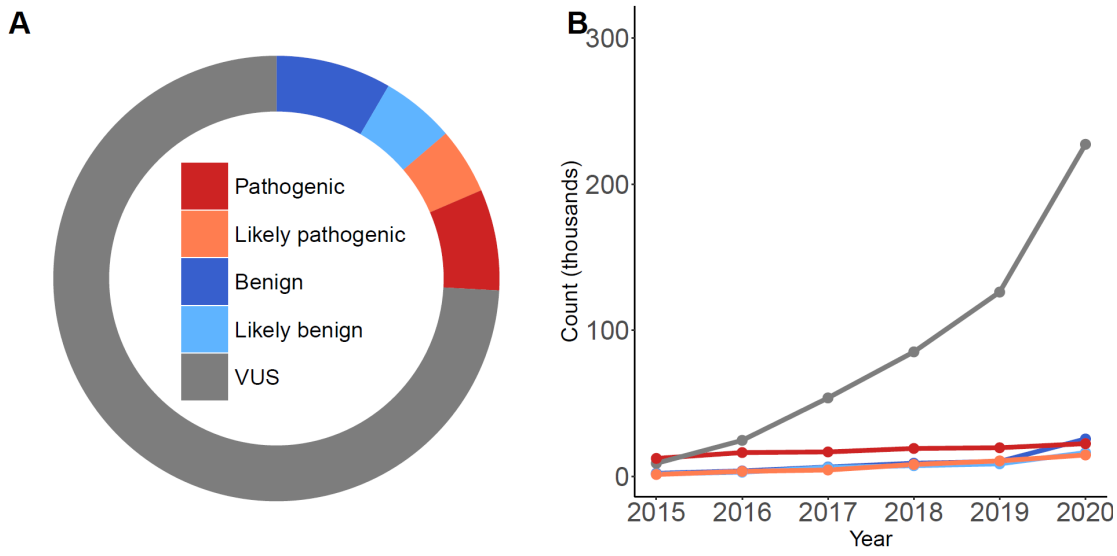


Figure 2:

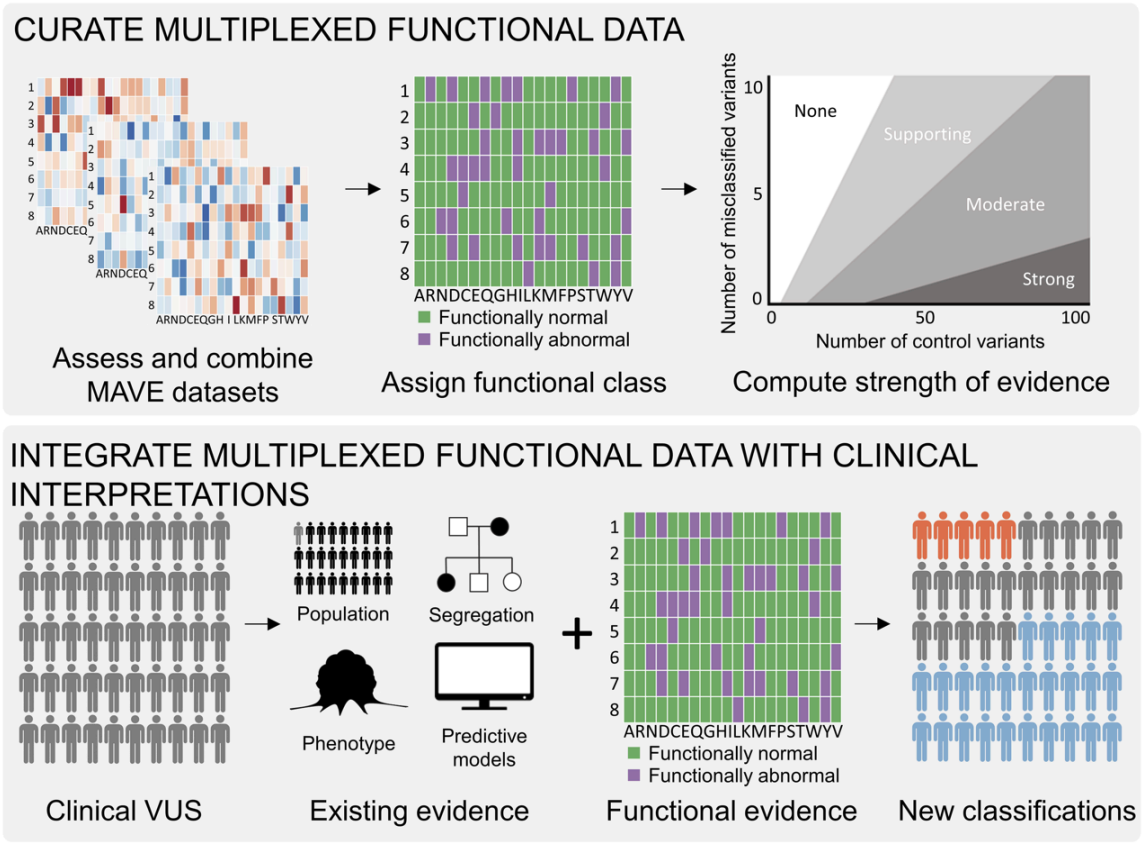

CURATE MULTIPLEXED FUNCTIONAL DATA
ARNDCEQ
ARNDCEQ
ARNDCEQGH I LKMFP STWYV
Assess and combine MAVE datasets
ARNDCEQGHILKMFPSTWYV
Functionally normal
Functionally abnormal
Assign functional class
Number of misclassified variants
None
Supporting
Moderate
Strong
Number of control variants
Compute strength of evidence
INTEGRATE MULTIPLEXED FUNCTIONAL DATA WITH CLINICAL INTERPRETATIONS
Population
Segregation
Phenotype
Predictive models
Clinical VUS
Existing evidence
ARNDCEQGHILKMFPSTWYV
Functionally normal
Functionally abnormal
Functional evidence
New classifications

Figure 3:

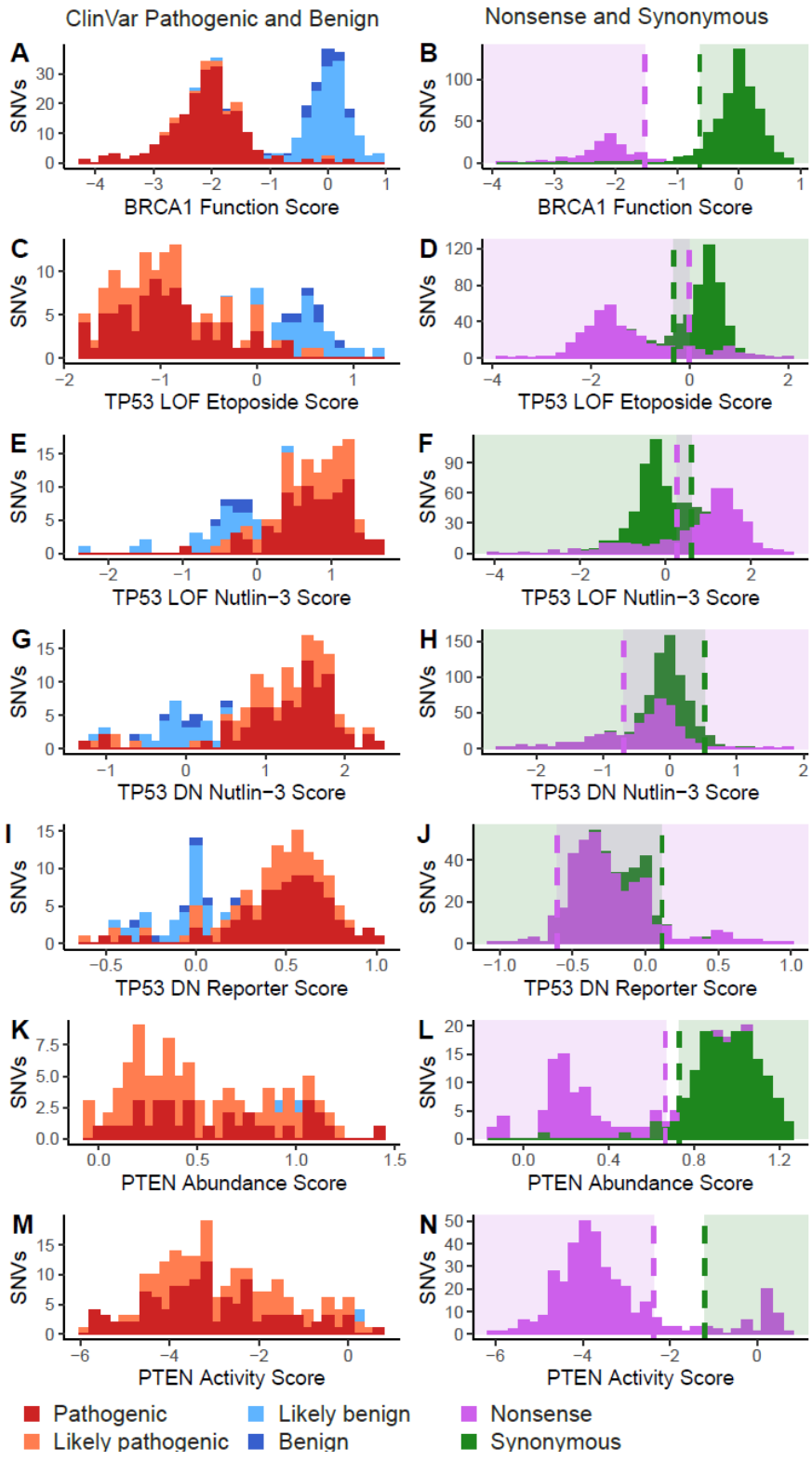

ClinVar Pathogenic and Benign
Nonsense and Synonymous
A
B
C
D
E
F
G
H
I
J
K
L
M
N
SNVs
BRCA1 Function Score
TP53 LOF Etoposide Score
TP53 LOF Nutlin−3 Score
TP53 DN Nutlin−3 Score
TP53 DN Reporter Score
PTEN Abundance Score
PTEN Activity Score
Pathogenic
Likely pathogenic
Likely benign
Benign
Nonsense
Synonymous

Figure 4:

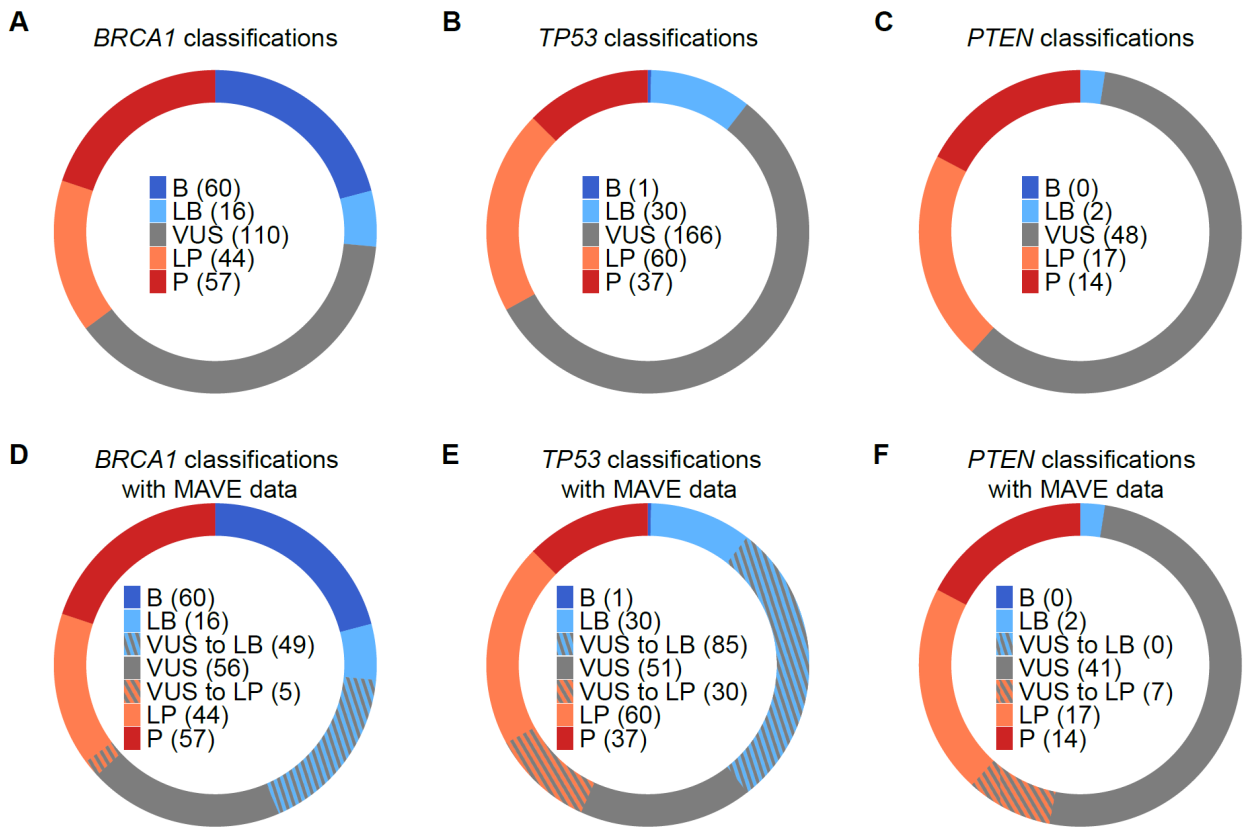


Figure 5:

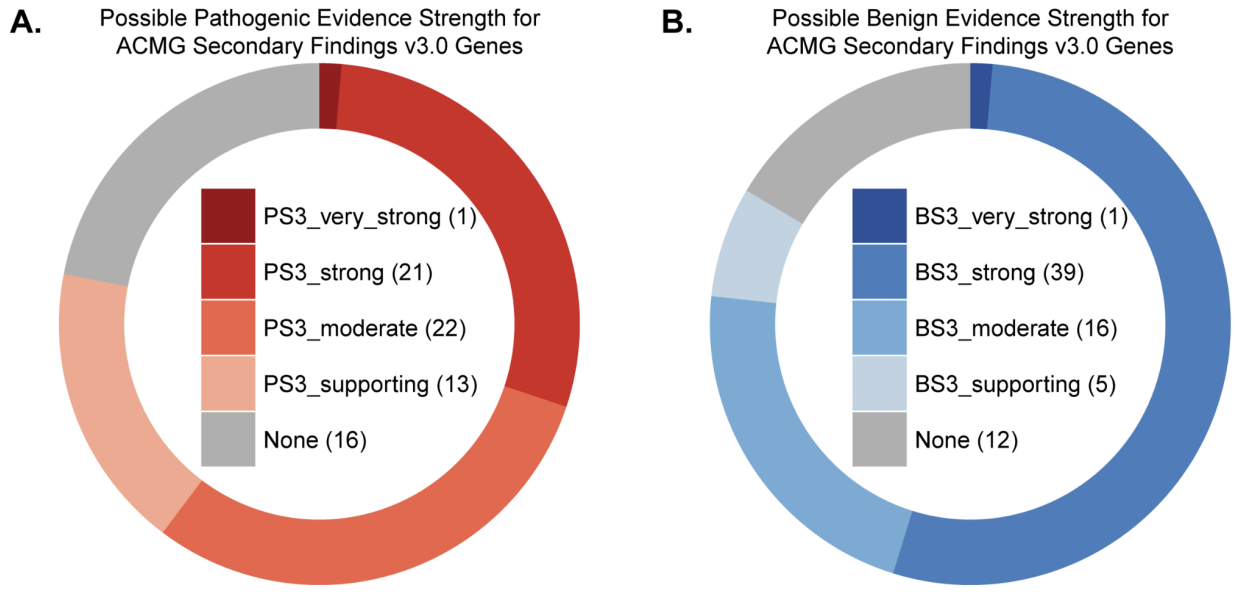


Table 1: Functionally abnormal variant selection: Positive = cells harboring functionally abnormal variants have prolonged survival upon functional selection compared to wild type.

Negative = cells harboring functionally abnormal variants are more rapidly depleted than wild type upon functional selection.

Gene	Number of variants tested	Cell type	Functionally abnormal variant selection	Selection	Functional relevance	Readout	Reference
BRCA1	3,893	HAP1	Negative	Cell growth	*BRCA1* is essential for survival	Sequencing of variants, enrichment after growth	Findlay et al. 2018[12], PMID: 30209399
TP53	8,258	A549	Negative	Etoposide toxicity	Etoposide is more toxic to cells with *TP53* loss of function mutations	Sequencing of variants, enrichment after growth	Giacomelli et al. 2019[13], PMID: 30224644
TP53	8,258	A549	Positive	Nutlin-3 toxicity	Nutlin-3 is less toxic to cells with *TP53* loss of function mutations	Sequencing of variants, enrichment after growth	Giacomelli et al. 2019[13], PMID: 30224644
TP53	8,258	A549	Positive	Nutlin-3 toxicity	Nutlin-3 is less toxic to cells with dominant negative *TP53* mutations	Sequencing of variants, enrichment after growth	Giacomelli et al. 2019[13], PMID: 30224644
TP53	8,258	MOLM13	Negative	Nutlin-3 toxicity	Nutlin-3 is less toxic to cells with dominant negative *TP53* mutations	Sequencing of variants, binned by P21-GFP intensity	Boettcher et al. 2019[14], PMID: 31395785
PTEN	4,112	HEK293	Negative	Cell growth	*PTEN*-GFP fusion protein abundance	Sequencing of variant barcodes, binned by PTEN-GFP intensity	Matreyek et al. 2018[15], PMID: 29785012
PTEN	7,244	YPH-499	Negative	PI3K toxicity	Human *PI3K* is toxic to yeast in absence of functional human *PTEN*	Sequencing of variants, enrichment after growth	Mighell et al. 2018[16], PMID: 29706350

References:

1. Maxwell, K.N., Hart, S.N., Vijai, J., Schrader, K.A., Slavin, T.P., Thomas, T., Wubbenhorst, B., Ravichandran, V., Moore, R.M., Hu, C., et al. (2016). Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. Am. J. Hum. Genet. *98*, 801–817.

2. LaDuca, H., Polley, E.C., Yussuf, A., Hoang, L., Gutierrez, S., Hart, S.N., Yadav, S., Hu, C., Na, J., Goldgar, D.E., et al. (2019). A clinical guide to hereditary cancer panel testing: evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients. Genet. Med. 1–9.

3. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–423.

4. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

5. Kurian, A.W., Sigal, B.M., and Plevritis, S.K. (2010). Survival analysis of cancer risk reduction strategies for BRCA1/2 mutation carriers. J. Clin. Oncol. *28*, 222–231.

6. Villani, A., Tabori, U., Schiffman, J., Shlien, A., Beyene, J., Druker, H., Novokmet, A., Finlay, J., and Malkin, D. (2011). Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: a prospective observational study. Lancet Oncol. *12*, 559–567.

7. Tan, M.-H., Mester, J.L., Ngeow, J., Rybicki, L.A., Orloff, M.S., and Eng, C. (2012). Lifetime cancer risks in individuals with germline PTEN mutations. Clin. Cancer Res. *18*, 400–407.

8. Makhnoon, S., Shirts, B.H., and Bowen, D.J. (2019). Patients' perspectives of variants of uncertain significance and strategies for uncertainty management. J. Genet. Couns. *28*, 313–325.

9. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.

10. Miller, D.T., ACMG Secondary Findings Working Group, Lee, K., Chung, W.K., Gordon, A.S., Herman, G.E., Klein, T.E., Stewart, D.R., Amendola, L.M., Adelman, K., et al. (2021). ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine.

11. Brnich, S.E., Rivera-Muñoz, E.A., and Berg, J.S. (2018). Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. Hum. Mutat. *39*, 1531–1541.

12. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

13. Giacomelli, A.O., Yang, X., Lintner, R.E., McFarland, J.M., Duby, M., Kim, J., Howard, T.P., Takeda, D.Y., Ly, S.H., Kim, E., et al. (2018). Mutational processes shape the landscape of TP53 mutations in human cancer. Nat. Genet. *50*, 1381–1387.

14. Boettcher, S., Miller, P.G., Sharma, R., McConkey, M., Leventhal, M., Krivtsov, A.V., Giacomelli, A.O., Wong, W., Kim, J., Chao, S., et al. (2019). A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. Science *365*, 599–604.

15. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. *50*, 874–882.

16. Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. Am. J. Hum. Genet. *102*, 943–955.

17. Mester, J.L., Ghosh, R., Pesaran, T., Huether, R., Karam, R., Hruska, K.S., Costa, H.A., Lachlan, K., Ngeow, J., Barnholtz-Sloan, J., et al. (2018). Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. Hum. Mutat. *39*, 1581–1592.

18. Mighell, T.L., Thacker, S., Fombonne, E., Eng, C., and O'Roak, B.J. (2020). An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. Am. J. Hum. Genet. *106*, 818–829.

19. Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nat. Protoc. *11*, 1782–1787.

20. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am. J. Hum. Genet. *101*, 315–325.

21. Gelman, H., Dines, J.N., Berg, J., Berger, A.H., Brnich, S., Hisama, F.M., James, R.G., Rubin, A.F., Shendure, J., Shirts, B., et al. (2019). Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. Genome Med. *11*, 85.

22. Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M., et al. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. *12*, 3.

23. Pesaran, T., Karam, R., Huether, R., Li, S., Farber-Katz, S., Chamberlin, A., Chong, H., LaDuca, H., and Elliott, A. (2016). Beyond DNA: An Integrated and Functional Approach for Classifying Germline Variants in Breast Cancer Genes. Int. J. Breast Cancer *2016*, 2469523.

24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research *12*, 2825–2830.

25. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M., Biesecker, L.G., and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet. Med. *20*, 1054–1060.

26. Fortuno, C., Lee, K., Olivier, M., Pesaran, T., Mai, P.L., de Andrade, K.C., Attardi, L.D., Crowley, S., Evans, D.G., Feng, B.J., et al. (2021). Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. Human mutation, *42*, 223-236.

27. Feng, B.-J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization. Hum. Mutat. *38*, 243–251.

28. Tian, Y., Pesaran, T., Chamberlin, A., Fenwick, R.B., Li, S., Gau, C.-L., Chao, E.C., Lu, H.-M., Black, M.H., and Qian, D. (2019). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci. Rep. *9*, 12752.

29. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535–548.e24.

30. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

31. DiGiammarino, E.L., Lee, A.S., Cadwell, C., Zhang, W., Bothner, B., Ribeiro, R.C., Zambetti, G., and Kriwacki, R.W. (2002). A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant p53 tetramer. Nat. Struct. Biol. *9*, 12–16.

32. Kato, S., Han, S.-Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R., and Ishioka, C. (2003). Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc. Natl. Acad. Sci. U. S. A. *100*, 8424–8429.

33. Malcikova, J., Tichy, B., Damborsky, J., Kabathova, J., Trbusek, M., Mayer, J., and Pospisilova, S. (2010). Analysis of the DNA-binding activity of p53 mutants using functional protein microarrays and its relationship to transcriptional activation. Biol. Chem. *391*, 197–205.

34. Powers, J., Pinto, E.M., Barnoud, T., Leung, J.C., Martynyuk, T., Kossenkov, A.V., Philips, A.H., Desai, H., Hausler, R., Kelly, G., et al. (2020). A Rare TP53 Mutation Predominant in Ashkenazi Jews Confers Risk of Multiple Cancers. Cancer Res. *80*, 3732–3744.

35. Zick, A., Kadouri, L., Cohen, S., Frohlinger, M., Hamburger, T., Zvi, N., Plaser, M., Avital, E., Breuier, S., Elian, F., et al. (2017). Recurrent TP53 missense mutation in cancer patients of Arab descent. Fam. Cancer *16*, 295–301.

36. Lolas Hamameh, S., Renbaum, P., Kamal, L., Dweik, D., Salahat, M., Jaraysa, T., Abu Rayyan, A., Casadei, S., Mandell, J.B., Gulsuner, S., et al. (2017). Genomic analysis of inherited breast cancer among Palestinian women: Genetic heterogeneity and a founder mutation in TP53. Int. J. Cancer *141*, 750–756.

37. Da Kuang, Weile, J., Kishore, N., Rubin, A.F., Fields, S., Fowler, D.M., and Roth, F.P. (2021). MaveRegistry: a collaboration platform for multiplexed assays of variant effect. Bioinformatics.

38. Kuang, D., Weile, J., Li, R., Ouellette, T.W., Barber, J.A., and Roth, F.P. (2020). MaveQuest: a web resource for planning experimental tests of human variant effects. Bioinformatics *36*, 3938–3940.

39. Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., and Rubin, A.F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. *20*, 223.

40. AVE Alliance Founding Members (2021). The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution.DOI: 10.5281/zenodo.4989960

Chapter 3: iPSC-SGE: Assessing Variant Effects in Differentiated Cells at Scale

Shawn Fayer[1], Clayton Friedman[2,3,4], Sriram Pendyala[1], Riddhiman Garge[1,5], Ivan Woo[5], Abby McGee[1], Evan McDermot[1], Rachel Powell[1], Pankhuri Gupta[6], Andrew Stergachis[1,6], Sudarshan Pinglay[1], Daniel Yang[2,3,4,7], Douglas M. Fowler[1,5,8*], Lea M. Starita[1,5*]

[1] Department of Genome Sciences, University of Washington, Seattle, WA
[2] Institute for Stem Cell and Regenerative Medicine, University of Washington, School of Medicine, Seattle, WA 98109, USA
[3] Center for Cardiovascular Biology, University of Washington, Seattle, WA 98109, USA
[4] Department of Medicine/Cardiology, University of Washington, Seattle, WA 98109, USA
[5] Brotman Baty Institute for Precision Medicine, Seattle, WA
[6] Division of Medical Genetics, University of Washington, School of Medicine, Seattle, WA 98109, USA
[7] Cardiology/Hospital Specialty Medicine, VA Puget Sound HCS, Seattle, WA 98108, USA
[8] Department of Bioengineering, University of Washington, Seattle, WA

Abstract:

MAVEs have so far been performed in utilitarian cancer cell lines, preventing their application to disease genes which cause phenotypes only in differentiated cells. To overcome this limitation, we developed a method for introducing SNVs into a single allele of a diploid induced pluripotent stem cell line using saturation genome editing (iPSC-SGE). We applied iPSC-SGE to *POLG* and *MYBPC3*. *POLG* encodes a DNA polymerase gamma which is required for replicating the mitochondrial genome. *POLG* variants are a leading cause of inherited mitochondrial depletion syndromes, resulting in a broad range of neurological phenotypes. We introduced variants to exon 16 on both a null *POLG* background and a pathogenic (W748S) *POLG* background. In both cases we cultured iPSCs for 14 days and scored variants based on their depletion. 23 of 438 variants were depleted in W748S cells and 123 of the total 438 exon 16 variants were depleted in null iPSCs. All 16 depleted ClinVar variants were pathogenic and 28 of the 29 functionally normal ClinVar variants were benign. We calibrated the assay for variant interpretation and applied strong functional evidence to delepled variants and moderate benign evidence to functionally normal variants and reclassified 28% of ClinVar variants of uncertain significance (VUS). *MYBPC3* encodes cardiac myosin binding protein C3, where pathogenic variants are the most common cause of familial hypertrophic cardiomyopathy. For this gene, we focused on the C10 domain, which anchors MYBPC3 to myosin heavy chain within the sarcomere and is a hotspot for pathogenic missense variants. We introduced 496 variants into the exon 32 and flanking intronic regions in iPSCs and differentiated these cells to cardiomyocytes for phenotyping. We used FACS to score variants based on their abundance relative to the wild type allele. In total, 87 variants had reduced abundance and 315 variants had normal abundance. All ten reduced abundance ClinVar variants were pathogenic and 22 of 23 normal abundance ClinVar variants were benign. Both POLG and MYBPC3 iPSC-SGE assays receive PS3 strong evidence for functionally abnormal variants and BS3 moderate evidence or functionally normal variants enabling clinical integration of the data. Thus, iPSC-SGE unlocks greater potential of MAVEs by expanding their application to the majority of human disease genes where variants must be phenotyped in differentiated cells.

Introduction:

Genome guided precision medicine is limited due to our inability to classify the majority of variants discovered by genetic testing. Clinical variant classification where the likelihood of variant pathogenicity is assessed has been hindered in many cases because variants are rare and the information required for definitive classification is unavailable. For this reason many newly identified variants are classified as variants of uncertain significance (VUS) and cannot be used to guide medical decisions. This trend is especially problematic for single nucleotide variants (SNVs) that cause missense amino acid substitutions where over 1 million are classified as VUS in ClinVar [1]. This challenge in the interpretation of rare variants represents a major crisis for genomic medicine. In particular, the utility of sequencing genes associated with actionable conditions where interventional recommendations exist to prevent manifestations or reduce morbidity/mortality is severely limited.

Functional assays can provide evidence for the classification of missense variants. The throughput of functional assays has increased significantly through multiplexed assays for variant effect (MAVEs) where advances in DNA sequencing and synthesis technologies have allowed the measurement of hundreds to thousands of variants in a pooled fashion in a single experiment. Because they generate functional data at scale, MAVEs have had an impact on the ability to resolve VUS in clinically relevant genes. For example, adding MAVE-generated functional data to other evidence for BRCA1, TP53, PTEN and MSH2 variants resulted in reclassification of 50%, 69%, 15% and 75% of VUS [2,3]. A class of MAVE called saturation genome editing (SGE) has been particularly powerful for accurately identifying potentially pathogenic variants [4,5]. In SGE all possible variants are directly edited into a region of the genome and then assessed for functional effects. Endogenous genome editing rather than expression of cDNA variant libraries allows assessment of both coding and noncoding variants and splice effects to be measured along with protein effects. Additionally, endogenous editing negates artifacts due to transgene overexpression which may mask subtle protein or splice defects. SGE followed by cell survival as a phenotypic readout has been employed to great effect to assess variants in genes that are essential to the HAP1 cell line [5–11]. HAP1 cells are haploid meaning that only a single allele needs to be edited to assess loss of function variants (LOF)[12]. Loss of function (LOF) variants edited into essential genes become depleted from the cell population and can be tracked by sequencing. SGE has been performed in diploid cell lines but requires upfront cell line engineering [10,13] or more accurately identifies dominant negative (DN) or gain-of-function (GOF) variants rather than LOF variants [14]. Nonetheless, all SGE to date has been performed in cancer or nonhuman cell lines limiting the genes that can be assessed for variant effects.

Assessing variants in the relevant cell type can be key to measuring the functional defect that underlies disease. Cell context is important for assessing the functional effects of coding variation, particularly for genes encoding proteins that function in cell type-specific structures or pathways. Cell context is also critical for assessing noncoding regulatory variation to ensure that the relevant trans-acting factors and chromatin architecture are present. Expression of cell type specific genes is also dependent on local genetic architecture, making endogenous genome editing a suitable method for introducing variants into iPSCs to measure variant effects in differentiated cells. Whereas transgene integration via lentiviral vectors or safe harbor landing

pads silence upon differentiation, editing variants into their endogenous locus ensures they will be expressed in the cell type of interest at more physiologically relevant levels. Many regulatory elements and coding genes also drive cell fate transitions, requiring variants to be assessed in differentiable cell systems. Furthermore, humans are diploid and genetic architecture can affect how alleles interact to express a phenotype. Haploid cells do not allow identification of DN alleles and variants with partial function can be difficult to identify. We therefore developed iPSC-SGE to evaluate variants in the correct cell context and genetic architecture.

Methods

iPSC culture and maintenance

Human induced pluripotent stem cells (iPSCs) were maintained at 37 °C in 5% CO2. All WTC-11, WTC-11-Ngn2 and derivative lines generated for this project were cultured on Matrigel (Corning; cat. no. 354277) and fed every other day with mTeSR Plus and 0.5% penicillin-streptomycin (ThermoFisher; cat. no. 15140122). iPSCs were passaged before reaching confluency by dissociating cells with TrypLE Express (ThermoFisher; cat. no. 12604013) and resuspended in mTeSR Plus supplemented with 10 μM Y-27632 dihydrochloride (Rho kinase [ROCK] inhibitor; Tocris; cat. no. 1254). Clonal lines were isolated via dilution plating in 96 well plates at 0.5 cells per well in mTeSR Plus supplemented with 10 μM Y-27632 and CloneR2 (StemCell Technologies; cat. no. 100-0691). iPSCs were transferred from Matrigel to plates coated with Vitronectin XF (StemCell Technologies; cat. no.07180) at 10ug/ml and conditioned for at least two passages before cardiac differentiation on Vitronectin.

Construction of iPSC-SGE repair templates

Plasmid repair templates were assembled using Twist gene fragments containing the genomic region of interest, homology arms, fluorescent protein fusion, and antibiotic selection markers driven by Ef1α. Homology arms flanked the genomic Cas9 cut site and were at between 200-500 bases in length. Gene fragments were designed with 18-22 base overlaps and assembled together with assembly PCR. Assembled repair templates were then cloned into the pJET vector with directional cloning. SNVs were inserted into mApple fused repair templates with either Gibson assembly or Golden Gate assembly. MYBPC3 exons 32 and 33 SNV libraries were amplified from Twist oligo pools and inserted into the linearized backbone with Gibson assembly with NEBuilder HiFi DNA assembly (NEB; cat. no. E2621S). The *POLG* backbone vector was over 13kb and could not be linearized for Gibson assembly and instead the Twist fragment containing exon 16 was designed with type IIS restriction sites and SNV oligos were amplified with compatible cut sites. SNVs were then inserted using Golden Gate assembly. All repair templates were sequence verified with whole plasmid sequencing via long read sequencing technologies before integration into cells. Assemblies were electroporated into NEB 10-beta electrocompetent cells (NEB; cat. no. C3020K) and plasmids were prepped with the ZymoPURE II Plasmid Maxipre Kit (Zymo; cat. no. D4203).

gRNA design

Synthetic guide RNAs were designed using ccTOP software (https://cctop.cos.uni-heidelberg.de/). Guides were designed for *POLG* intron 10 and *MYBPC3* intron 27. Guides were optimized for predicted cutting efficiency and minimal predicted off target

effects with ccTOP. The guides for inserting the background GFP allele were designed 5' to the guide for SNV integration so that the guide sequence for the SNV allele could be removed from the GFP repair template. Synthetic guide RNAs and spCas9 protein were ordered from Synthego (synthego.com).

Generation of heterozygous GFP lines for SGE

Wild type WTC-11 iPSCs were grown to 70-80% confluency then split 1:5 the day before gene editing. Synthetic gRNAs were pre-complexed into ribonucleoproteins (RNPs) with spCas9 protein for 10 minutes at room temperature per the manufacturer recommendations (Synthego). RNPs and 1ug of plasmid repair template were introduced into 200,000 cells via electroporation with the Lonza Amaxa P3 Primary Cell Kit S (Lonza; cat. No. V4XP-3032) using the Lonza 4D Nucleofector system. Immediately after electroporation, cells were plated into a single well of a 24 well plate coated with Matrigel and supplemented with 1X Clone R2. One week post electroporation, cells were selected with 1ug/ml puromycin for 2 days. Cells were then expanded and sorted for GFP positivity. Since *MYBPC3* is not expressed in stem cells, correct integration of repair templates was determined with CRaTER; CRISPR activation of *MYBPC3* at the endogenous promoter [15] 3 days prior to the sort. GFP positive cells were then plated in 96 well plates at 0.5 cells per well and single cell colonies were selected and expanded for PCR verification of heterozygous integration of GFP repair templates.

Saturation genome editing of iPSCs

Clonal GFP iPSCs were grown to 70-80% confluency then split 1:5 the day before gene editing. Synthetic gRNAs were pre-complexed into ribonucleoproteins (RNPs) with spCas9 protein for 10 minutes at room temperature per the manufacturer recommendations (Synthego). RNPs and SNV containing plasmid repair template (4ug for *POLG* and 8ug for *MYBPC3*) were introduced into 2x10^6 cells via electroporation with the Lonza Amaxa P3 Primary Cell Kit L (Lonza; cat. No. V4XP-3024). Immediately after electroporation, cells were plated into a single well of a 6 well plate coated with Matrigel and supplemented with 1X Clone R2 and 10 μM Y-27632. Cells were expanded and plated sparsely in Matrigel coated 10cm dishes for selection with 10ug/ml blasticidin S HCl (ThermoFisher; cat. no. A1113902). After selection cells were expanded and frozen in 500ul CryoStor CS-10 (Sigma; cat. no. C2874) at 3x10^6 cells per vial. *POLG* cells were sorted for double positive GFP/mApple signal and expanded before cryopreservation.

Cardiac directed differentiation

iPSCs were cultured on 10ug/ml Vitronectin XF prior to cardiomyocyte differentiation. Small molecule directed differentiation was performed as previously described with some modifications [15]. iPSCs were plated in 15ug/ml Vitronectin XF coated 24 well plates at 7.5x10^4 cells per well in mTeSR Plus supplemented with 10 μM Y-27632. Media was changed to fresh mTeSR Plus without 10 μM Y-27632 the next day. Directed differentiation was initiated (D0) when cells reached 60-70% confluency by aspirating mTeSR and replacing media with RBA media: RPMI (Invitrogen; cat. no. 11875135), 0.5 mg/mL bovine serum albumin (Sigma; cat. no. A9418), 0.213 mg/mL ascorbic acid (Sigma; cat. no. A8960) supplemented with 4.5uM CHIR-99021 (TOCRIS; cat. no. 4423). After two days (D2), CHIR-99021 containing media was removed and replaced with RBA supplemented with 2 μM Wnt-C59 (Selleck; cat. no. S7037).

On D4, Wnt-C59-containing media was replaced with RBA. On D6 (and every other day afterwards), media was replaced with cardiomyocyte media: RPMI, B27 plus insulin (Invitrogen; cat. no. 17504044). Cardiomyocyte cultures typically begin beating by D6. After D12, cardiomyocyte media was removed and cardiomyocytes were dissociated to single cells with TrypLE Select 10X (ThermoFisher; cat. no. A1217701) and replated on Matrigel coated 6 well plates at 3x10^6 cells per well in cardiomyocyte media supplemented with 10% FBS and 10 μM Y-27632.

Cardiomyocyte cycloheximide chase and sorting

After D20 and at least 7 days after replating, cardiomyocyte media was replaced with cardiomyocyte media containing 300ug/ml cycloheximide (Sigma; cat. no.) for 3 hours. Cardiomyocytes were then dissociated to single cells with TrypLE Select 1X (ThermoFisher; cat. no. 12563011) and recovered in RPMI. Cells were washed with 1X PBS and fixed witn 4% paraformaldehyde () for 10 minutes at room temperature and washed with 1X PBS. Cardiomyocytes were then resuspended in PBS sort buffer containing 5mM EDTA (cat. no.) and 25mM HEPES (ThermoFisher; cat. no.). Resuspended cells were passed through a 100uM cell strainer into FACS tubes. GFP positive cells were then sorted into 4 quartile bins on mApple to GFP ratio. After the sort, cells were pelleted in Eppendorf tubes a 800 RCF for 10 minutes and supernatant was discarded.

Cardiomyocyte DNA extraction

Cardiomyocyte DNA was extracted from sorted cardiomyocytes using the MagMAX Multi Sample Ultra 2.0 Kits (ThermoFisher; cat. no.A36570) with the MagMAX Cell and Tissue Extraction DNA Extraction Buffer (ThermoFIsher; cat. No. A45469). DNA was extracted to manufacturers specifications with a few modifications. First, cell pellets of no more that 1x10^6 cells were resuspended in extraction buffer and heated to 65C with shaking for 60 minutes with vortexing every 10 minutes. Further, the lysis step on the instrument was performed for 15 minutes. Finally, DNA was eluted from beads in 100ul of water.

Library preparation and sequencing

Genomic DNA from iPSCs was extracted using the Qiagen DNeasy Blood and Tissue Kit (; cat. no. 69504). Sequencing libraries were prepared with a three step nested PCR approach. First, SNV allele specific PCR was performed using primer sets that amplified from mApple to a genomic region outside of the 5' homology arm of the SNV repair template. MYBPC3 allele specific PCR was conducted with Q5 Hot Start HIgh-Fidelity 2x Master Mix (NEB; cat. no. M0494S). POLG allele specific PCR was performed with LongAmp Taq DNA Polymerase (NEB; cat. no.M0323S). gDNA PCRs were split into 4 reactions with 250ng gDNA each and pooled for AMPure XP Bead (Beckman Coulter; cat. no. A63882) cleanup. Exon specific amplicons with Illumina TruSeq and Nextera adaptors were prepared in a second PCR that was AMPure cleaned. The third PCR attached ten base pair barcodes to amplicons for demultiplexing. 2nM libraries were pooled and sequenced on the Illumina Nextseq 2000 instrument.

Calculating iPSC-SGE scores for *POLG* growth assay

SNV log2ratios were calculated by variant frequency at iPSC day 14 vs. pDNA frequency. First, reads were demulitpmexed using Illumina Bcl2Fastq software and paired end reads were assembled using Pear software. Any reads with more than one SNV or an indel was filtered out of the analysis. Reads with a single SNV were mapped to the genomic region of POLG exon 16 and each SNV was assigned a count. Log2ratios were calculated for each experiment relative to pDNA frequences.

Calculating iPSC-SGE scores for *MYBPC3* abundance assay

Abundance scores were calculated as previously described [16]. Briefly, sequence reads were filtered to include only SNVs mapping to the genomic regions of *MYBPC3* exon 32. Weighted average frequencies were calculated by the sum of weighted frequencies for a given variant across all bins divided by the sum of frequencies for that variant across all bins. Bins were weighted as follows: bin 1 (lowest abundance) = 0.1, bin 2 = 0.15, bin 3 = 0.2 and bin 4 (higest abundance) = 1.

CRISPR screen guide cloning

gRNA sequences for the 4,502 human disease genes in our CRISPR screen were from the minimal Cas9 library including the 200 non-targeting controls [17]. Guides were ordered as an oligo pool from Twist and cloned into LentiGuide Blast (Abcam:199622) with Golden Gate assembly. Three replicates of NEB 10-Beta electrocompetent cells were transformed with the assembled gRNA library and plasmids were prepped with ZymoPURE II Plasmid Maxiprep Kit and plasmids were pooled and sequenced with full plasmid sequencing to verify correct assembly and amplicon sequencing of the gRNA region to verify library coverage for lentivirus production.

iPSC neuron CRISPR screen

WTC-11 Ngn2 iPSCs [18] were engineered to integrate a spCas9 expression cassette into the genome via PiggyBac transposition. Lentivirus was produced by co-transfecting HEK-293T cells with gRNA library containing LentiGuide-Blast plasmid along with other lentivirus packaging component plasmids. After three days, lentivirus containing supernatant was collected and lentivirus was concentrated with PEG-it (System Biosciences; cat. no. LV810A-1). Ngn2-Cas9 iPSCs were seeded into Matrigel coated 6 well plates at 2x10^5 cells per well in mTeSR Plus supplemented with 10 μM Y-27632. The following day cells were tranduced with concentrated lentivirus at low MOI (about 30% transduction efficiency). To achieve 200x coverage of gRNA library, 6x10^6 cells were transduced per replicate. Post transduction, cells were selected with blasticidin and expanded to maintain library complexity. Cell pellets were harvested for library screening after blasticidin selection. gRNA containing iPSCs were cultured for 3 weeks, cell pellets were collected, then cells were plated for neuron differentiation following published protocols [18]. CRISPR scores were calculated using MAGeCK-iNC software [19].

Results

Editing nearly 1,500 variants of *POLG* and *MYBPC3* into induced pluripotent stem cells (iPSC) with with iPSC-SGE

Most unique missense variants are classified as VUS in ClinVar[1]. A number that has been growing nearly exponentially since the initiation of the database (**Figure 1A**). SGE is a powerful tool for generating accurate functional data for clinical variant classification[5–8; 9–11]. Since HAP1 SGE relies on a cellular growth phenotype, a major limitation of the approach is the requirement that a gene is essential to the cell line. Of the nearly 800,000 missense VUS in ClinVar within genes that have a well established disease association, only 16% occur within genes that are essential to HAP1. Many of the non-essential disease-associated genes are not expressed in HAP1 and cause phenotypes that manifest in specific cellular and tissue contexts (**Figure 1B**). In addition, essentiality in a haploid cell line neglects the fact that most variants that cause genetic disorders occur in a diploid genetic context where the second allele may modulate the impact of a pathogenic variant. For these reasons, we developed an SGE approach for diploid human induced pluripotent stem cells (iPSC-SGE) where variants are precisely edited into their endogenous locus in iPSCs such that variant effects can be measured after differentiation into cell type of interest.

To assess variants in diploid iPSCs we devised multiple editing strategies depending on the genetic architecture of disease (**Figure 1C**). First, we chose to work with WTC-11, a widely utilized and well characterized episomally derived iPSC line from a healthy adult male donor [1,20]. In addition to wild type WTC-11, we utilized a WTC-11 derived line with a doxycycline inducible mouse *Ngn2* cassette for rapid excitatory neuron differentiation [18]. To engineer WTC-11 lines for SGE, we first modified one allele before targeting the other for SGE. For genes where variants exert their effect in an autosomal recessive fashion we edited the first allele to express a known pathogenic SNV or null allele harboring a large deletion of coding sequence (**Figure 1C**). For genes where variants act in an autosomal dominant architecture we knocked in a reference (wild type) copy to the first allele (**Figure 1C**). For each, we included an antibiotic resistance cassette to select for integrants, fused the engineered gene to GFP at its terminal protein coding exon to identify in-frame edits and created a small intronic deletion to protect the edited copy from re-cutting when a guide for SGE of the second allele is introduced. For genes that are not expressed in the stem cell state we employed CRaTER [15], a strategy to induce expression of fluorescent fusion protein using CRISPR activation to identify cells with in-frame edits. Correctly edited cells containing in-frame reference or variant genes were isolated by screening antibiotic resistant, GFP-positive clones by PCR. We isolated multiple clonal lines for each experiment, sequence verified the GFP allele with nanopore sequencing, and validated differentiation potential where applicable.

We applied iPSC-SGE to introduce libraries of variants into the *POLG* and *MYBPC3* genes in iPSCs, enabling measurement of variant effects in the context of a second allele and in differentiated cell types. *POLG* encodes a DNA polymerase required for replicating the mitochondrial genome. *POLG* variants are a leading cause of inherited mitochondrial depletion syndromes that cause a broad range of neurological disorders that span a spectrum of symptoms and severity[21]. We chose *POLG* as a model for iPSC-SGE because variants can cause disease in both an autosomal dominant and recessive manner. Therefore, we chose to

edit saturation libraries of *POLG* exon 16 into cells with both a null allele to identify LOF variants and a hypomorphic pathogenic allele (W748S) to identify dominant negative variants (**Figure 1C and 1D**). *POLG* exon 16 is 118 bases in length and encodes part of the polymerase's thumb and palm subdomains. We programmed a variant library to make all possible 354 coding SNVs and 84 intronic variants to assess effects on both protein function and splicing. The variant library is delivered to cells in a plasmid-based repair template spanning 13 exons and introns encoding the C-terminus fused to mApple via an internal ribosome entry site (IRES). The repair template has homology arms to target the cassette to the 12th intron and genome integration is selected for with blasticidin and targeting to the correct locus with mApple signal. We assess editing rates by amplifying correctly edited variants from the genome followed by DNA sequencing. In each editing experiment in the W748S line, we recovered 377 of the 384 programmed variants after antibiotic selection. The 98% recovery rate for variants suggests that SGE in iPSCs will be feasible at a scale (**Figure 1D**).

MYBPC3 encodes the cardiac myosin binding protein, which modulates the contraction of heart cells within the sarcomere. Pathogenic variants in *MYBPC3* are the leading cause of familial hypertrophic cardiomyopathy (HCM) ([18,22]. For *MYBPC3* we integrated variants into exons 32 and 33 that encode the C10 domain, which is a hotspot for pathogenic missense variants [23]. Pathogenic missense variants in the C10 domain have a dominant inheritance pattern and cause haploinsufficiency through reduced protein abundance [23,24]. To perform SGE on the MYBPC3 C10 domain we designed repair templates containing the genomic region spanning exons 27-34 fused to mApple. Variant libraries for exon 32 programmed to contain all possible 411 coding SNVs and 87 flanking intronic SNVs and exon 33 which contained 561 coding SNVs and 78 intronic SNVs were separately cloned into the repair template plasmid. We recovered 496 of the 498 programmed variants for exon 32 and 621 of 639 variants for exon 33 after antibiotic selection. The 97-99% recovery rate suggests that editing genes that are not expressed in the stem cell state and are heterochromatinized is as efficient as editing genes that are expressed in stem cells (**Figure 1D**). Taken together, iPSC-SGE represents a greater than 10-fold improvement in scale of variant libraries edited into human stem cells [15,25–27]enabling the scalable assessment of variants across differentiable cell types and in the diploid state.

POLG iPSC-SGE identifies loss of function and dominant negative variants

POLG variants were measured in iPSCs in the context of two different background mutations. First, we introduced exon 16 variants into cells with the pathogenic W748S variant engineered into the GFP allele. After genome editing, cells were selected with blasticidin and sorted for positive GFP and mApple signal to enrich for cells with the pathogenic W748S variant and correctly integrated exon 16 SNVs (**Figure 2A**). Exon 16 variants were sequenced 14 days after sorting and all but 23 missense variants were stably integrated (**Figure 2B**). These missense variants occurred at 12 residues in the polymerase domain spanning positions 847 to 866 at the junction of the thumb and palm subdomains. These residues are involved in binding the polymerase domain to DNA [28] and disruption of POLG polymerase domain binding has been demonstrated to cause a toxic dominant negative mitochondrial DNA depletion phenotype in cultured cells [29]. Thus, we deemed the 23 depleted missense variants in the W748S iPSCs potential dominant negative missense variants.

POLG variants were also measured in *POLG* null iPSCs, following a similar procedure to the W748S cells (**Figure 2C**). In *POLG* null cells, all potential dominant negative variants were depleted by day 14. In addition, all nonsense and canonical splice site variants were depleted. An additional 72 missense variants were depleted in *POLG* null cells compared to W748S background cells (**Figure 2D**). Exon 16 function scores were bimodally distributed and 123 of the total 438 exon 16 variants were depleted in null iPSCs (**Figure 2E**). Functionally abnormal iPSC-SGE scores were broadly in agreement with variant effect predictors (VEPs). We chose to compare function scores with CADD, AlphaMissense and REVEL to validate assay results against different classes of VEPs. With the exception of two splice region variants, all iPSC-SGE depleted variants had deleterious VEP scores across the three predictors. A proportion of functionally normal missense variants were predicted to be deleterious in each of the predictors (CADD = 48%, AlphaMissense = 22%, and REVEL = 65%). This result is consistent with other accurate functional datasets [5,7], which outperform VEPs since predictors tend to have poor sensitivity and/or specificity for benign variants, often scoring them as deleterious [30].

POLG iPSC-SGE scores enables reclassification of 28% of VUS

We calibrated *POLG* exon 16 iPSC-SGE scores for clinical variant classification. First, we fit a two component Gaussian mixture model (GMM) to the full set of function scores. We then set thresholds at the function scores where variants had a greater than 95% chance of being sampled from the left or right distribution based on the component density ratios. Variants below -2.96 were considered functionally abnormal and variants that scored above -2.17 were considered functionally normal. Variants in between these thresholds were considered to have indeterminate function (**Figure 3A**). We then calibrated the functionally abnormal and functionally normal score ranges by calculating pathogenicity likelihood ratios (OddsPath) for each category using ClinVar pathogenic/likely pathogenic and benign/likely benign variants with at least one star [31]. All 16 functionally abnormal ClinVar variants we pathogenic or likely pathogenic resulting in an OddsPath of 24.5 corresponding to PS3 strong pathogenic evidence for variant classification of variants that score functionally abnormal. 28 of the 29 functionally normal variants were ClinVar benign of likely benign variants resulting in an OddsPath of 0.063, corresponding to BS3 moderate benign evidence for functionally normal variants. We chose to calibrate CADD scores for exon 16 since this predictor had the greatest number of overlapping ClinVar variants with the iPSC-SGE data. We applied the same GMM method for CADD scores and found that above 19.8 were predicted abnormal and had an OddsPath of 22.9 corresponding to PP3 strong pathogenic evidence. Variants with CADD scores below 12.5 had an OddsPath of 0.068 and were given BP4 moderate benign evidence.

We applied functional evidence from iPSC-SGE and VEP evidence from CADD to *POLG* VUS in ClinVar. Agreement between the two pieces of evidence in favor of pathogenicity resulted in two pieces of strong evidence applied to 8 of the ClinVar VUS. Agreement in favor of benignity resulted in two pieces of moderate benign evidence applied to 3 of the VUS (**Figure 3E**). Thus, we were able to reclassify 8 VUS as likely pathogenic and 3 VUS as likely benign, resolving 28% of the VUS in this region (**Figure 3F-G**). Beyond VUS, there were a total of 109 exon 16 variants that were both functionally abnormal and had predicted deleterious CADD scores. A total of 118 variants were functionally normal and had predicted tolerated CADD

scores (**Figure 3H-I**). Beyond clinical variant classification, *POLG* iPSC-SGE scores provide insight into patient phenotypes. For example, the potential dominant negative R853Q variant has been described in a patient with severe infantile onset myocerebrohepatopathy spectrum (MCHS) [30,32]. Their second allele was found to harbor the T251I - P587L variants, which are usually associated with less severe POLG related disorders. Additionally, the R853W variant has an atypical presentation and is associated with early onset Parkinsonism when inherited with the G737R variant [33–35]. One additional potential dominant variant has been described in a patient with infantile onset Alpers syndrome where a second variant was not identified [32]. Although there is no clear phenotypic pattern shared between the patients with potential dominant negative variants, these patients tended to have more severe and atypical phenotypes compared to most other pathogenic variants.

MYBPC3 iPSC-SGE identifies reduced abundance variants in iPSC cardiomyocytes

iPSC-SGE solves a major limitation of HAP1 SGE by enabling phenotyping of variants in differentiated cell types. *MYBPC3* is a gene that encodes a sarcomere structural component that is only expressed in cardiomyocytes. Introducing variants into *MYBPC3* in iPSCs enables phenotyping of *MYBPC3* variants after cells with SNV libraries are differentiated into cardiomyocytes. Since *MYBPC3* associated familial hypertrophic cardiomyopathy is inherited in an autosomal dominant fashion, we introduced SNVs via a repair template with an mApple fusion into iPSCs with a wild type GFP fusion on the second allele. After selection of SNV integration with blasticidin, library containing cells were differentiated into cardiomyocytes for phenotyping (**Figure 4A**). Exon 32 is the first exon of the C-terminal MYBPC3 C10 domain, an Ig-like domain that anchors MYBPC3 to myosin heavy chain. An established pathogenic mechanism for C10 missense variants is reduced incorporation into the sarcomere due to loss of protein stability and abundance [23]. Thus, at differentiation day 20, monolayer cardiomyocytes were dissociated into single cells and sorted on mApple to GFP ratio to capture the degree of abundance change for each SNV compared to the wild type GFP within the same cell (**Figure 4 B-C**). Cardiomyocytes were sorted into four bins and variants were scored by calculating weighted average frequency across the four bins (**Figure 4D**). We collected sufficient cells to score 434 of the exon 32 variants and all nonsense and canonical splice site variants had low abundance scores. 57 missense variants also had low abundance scores. All of the missense variants outside of the C10 domain, before residue 1181, had normal abundance (**Figure 4E**), reinforcing the specificity of the loss of abundance phenotype to the C10 domain.

Exon 32 iPSC-SGE scores were bimodally distributed where nonsense and canonical splice site mutations were separated from synonymous and intronic variants (**Figure 4D**). We compared iPSC-SGE scores to VEPs and found broad agreement between predictor scores and function scores. Nearly all functionally abnormal, reduced abundance variants had high CADD scores which correspond to deleterious predictions (**Figure 4G**). AlphaMissense and REVEL predictions also correlated with iPSC-SGE scores, but a small proportion of functionally abnormal variants were predicted to be tolerated by both predictors (**Figure 4H-I**). Nine of the 18 variants with functionally abnormal scores and normal AlphaMissense or REVEL scores occurred at residues A1203, R1205, or K1209. This region is part of the binding interface with MYH7 and the R1205 residue is required for binding the negatively charged MYH7 binding

interface [36,37]. Therefore, it is plausible that variants at these residues allow MYBPC3 to fold and have normal predictor scores, but do not integrate into the sarcomere and have reduced abundance as a result.

MYBPC3 iPSC-SGE leads to reclassification of 35% of exon 32 VUS

We calibrated *MYBPC3* exon 32 iPSC-SGE scores for clinical variant classification. First we fit a GMM to all exon 32 scores to determine functionally abnormal and normal thresholds at 0.231 and 0.326, respectively (**Figure 5A**). OddsPath for functionally abnormal variants was 19.4 and functionally normal variants was 0.08 corresponding to PS3 strong and BS3 moderate evidence. We calibrated CADD since it had the most overlap with iPSC-SGE scores for exon 32. CADD thresholds were calculated at 8.28 for predicted tolerated and 16.119 for predicted deleterious(**Figure 5B**). Tolerated variants received OddsPath of 0.11 and deleterious variants received OddPath of 10 corresponding BP4 moderate and PP3 moderate evidence.

We applied calibrated functional and predictor evidence to the *MYBPC3* exon 32 VUS in ClinVar. Eleven variants received PS3 strong evidence from iPSC-SGE and PP3 moderate evidence from CADD and were reclassified to likely pathogenic. Three variants received BS3 moderate and BP4 moderate evidence from iPSC SGE and CADD and were reclassified to likely benign. The remainder of VUS could not be moved on the basis of functional and computational evidence alone and remained VUS. We expanded this analysis beyond VUS and applied calibration to all SNVs in exon 32 and flanking intronic regions and 82 variants received pathogenic evidence from both sources and 78 variants received benign evidence from both sources. Since these additional variants have not yet been evaluated in patients, this prospective analysis will provide useful evidence in the evaluation of these variants as they are identified in patients in the future.

Identification of 727 Mendelian disease genes essential to iPSC derived neurons

iPSC-SGE in *POLG* and *MYBPC3* demonstrated the power of this technology to measure variant effects in genetic and cell type specific contexts. However, determining a multiplexable phenotype for a gene of interest is often a limiting factor in developing new multiplexed functional assays. Thus, we sought to identify high priority genes that are specifically assayable with iPSC-SGE using differentiated cell type viability as the phenotype. We focused this effort on neurological disease genes since they represent the class of non-Hap1 essential genes with the largest number of VUS in ClinVar (**Figure 1B**). To do this we employed a CRISPR knock-out screen targeting 4,502 genes that have been associated with disease in the GenCC database [38] as well as an additional set of known neuron essential or suspected neuronal disease genes [19,39]. We first integrated Cas9 into WTC11 iPSCs, then transduced the cells with a lentiviral gRNA library containing two guides for each gene [17]. Guides were sequenced after 3 weeks in iPSC culture then cells were differentiated into neurons to assess neuron essential genes (**Figure 6A**). Guides were sequenced in day 21 post transduction iPSCs and in day 21 post induction neurons and fold change was calculated to identify neuron essential genes (**Figure 6B**). Of the 4,502 genes in our screen, gRNAs targeting 714 genes were depleted in neurons compared to iPSC and 13 gRNAs targeting 13 genes were enriched in neurons compared to iPSCs (**Figure 6C**). Our data are consistent with other iPSC

neuron screens. An inducible CRISPRi screen targeting kinases in iPSC derived neurons [19] had 133 depleted genes that overlapped with our data. Of those, 82 were depleted in neurons in our screen, 41 were iPSC essential in our screen, and 10 were not depleted in either neurons or iPSCs in our screen. In total, there were 128,577 ClinVar VUS in the 727 positive hits from our screen, representing 16% of the VUS in ClinVar which are assayable with iPSC-SGE and neuron differentiation.

Discussion

iPSC-SGE solves two major limitations of Hap1 SGE and other multiplexed functional assays performed in utilitarian cell lines where the majority of human disease genes do not have phenotypes and are largely not expressed. First, iPSC-SGE is performed in diploid human iPSCs enabling the phenotyping of variants in the genetic context of a second allele. Since variants are edited into the endogenous locus, variants are expressed with endogenous regulatory machinery more closely representing expression levels in humans and any impact from a second allele. Next, iPSC-SGE enables the differentiation of cells harboring variant libraries into specialized cell types to measure variant effects in correct cell and tissue context. Since genome editing is performed in iPSCs, genes which are not expressed in stem cells can be targeted and variants are expressed upon differentiation into the cell type of interest. We have demonstrated that SNV libraries can be efficiently integrated into regions which are silenced in iPSCs via iPSC-SGE of *MYBPC3*, a gene that is exclusively expressed in cardiomyocytes. For these reasons, iPSC-SGE should be applicable to any gene that has phenotypes in differentiated cells.

iPSC-SGE is a scalable approach for the generation of functional data in differentiated cells even though it requires upstream cell line engineering that is not required in Hap1 SGE. Since iPSC-SGE repair templates contain part or all of a genomic locus for a gene of interest, the backbone SNV repair template is the same for all exons. Therefore, the editing experiment is the same across all exons targeted for a gene of interest utilizing the same guide RNA. Once an experiment is designed and reagents are synthesized, scaling to an entire gene or functional domain is straightforward relative to Hap1 SGE where a new repair template and guide RNA are used for every region of interest.

Beyond variant interpretation, iPSC-SGE allows the measurement of phenotypes in both genetic and cellular contexts that more closely model the phenotypes that occur in humans. This allows for relevant mechanistic insight that cannot be gleaned from Hap1 SGE or transgene based MAVEs performed in other utilitarian cell lines. For example, iPSC-SGE of *POLG* in the presence of two different *POLG* background alleles revealed a subset of loss of function missense variants to be likely dominant negative variants which gave context to the apparent increased severity of phenotypes experienced by patients with these variants. In addition to molecular mechanistic insight, iPSC-SGE allows for the screening of potential therapeutics in the affected cell types in patients with genetic disorders. Since iPSCs with POLG loss of function variants are not deleted in the context of the W748S pathogenic variant, these cells can be differentiated into neurons and relevant phenotypes such as mitochondrial DNA copy number can be measured in the context of different drugs. Thus, iPSC-SGE is poised to be a very powerful method for the screening of therapeutics in the correct cell context with variant specific outputs that may be useful in stratifying individuals for clinical trials.

Since iPSC-SGE allows for differentiation into specialized cell types and repair templates contain intact introns, variant effects can be assessed in intronic regulatory elements and phenotyped in correct cell context. A critique of massively parallel reporter assays (MPRAs) that link expression of a reporter to minimal promoters in the presence of candidate cis regulatory elements, is that these assays either express constructs off of a plasmid or are integrated randomly into the genome of cell line that may not be relevant for their activity. While iPSC-SGE is a more locally constrained assay, since introns remain intact all of the potential regulatory elements in a region of interest can be targeted and measured in the relevant cell context. The combination of coding variant effect measurement in the context of a library of regulatory element variants within an iPSC-SGE experiment has the potential to offer insight into phenotypic differences observed in humans with the same pathogenic coding variant.

A challenge for any multiplexed functional assay is selection of a phenotype for scoring variants. Ideally, deleterious variants cause a cellular growth/fitness disadvantage relative to wildtype, enabling sequencing of variants at early and late time points and scoring based on variant depletion over time. These growth assays are typically the most simple to perform and yield accurate functional data for variant interpretation [2,3]. We demonstrated that iPSC-SGE is applicable to growth phenotypes specific to differentiated cells and identified 727 disease genes with a growth phenotype specific to iPSC derived neurons. Application of iPSC-SGE to these genes has the potential to significantly reduce the VUS burden in clinical genetics as 16% of current ClinVar VUS occur within these genes.

Figure Legends:

Figure 1: iPSC SGE enables phenotyping of variants in differentiated cell types. A) Unique ClinVar missense variants plotted annually by clinical significance. Variants are included from genes in the GenCC database with at least moderate evidence of gene-disease association. Pathogenic/Likely pathogenic and Benign/Likely benign variants were included in the “Likely” category only. **B**) VUS from A are plotted annually and colored by gene panel categories on which they are identified. **C**) iPSC-SGE schematic where the background allele is first edited to fuse a GFP to the target gene and insert a blocking mutation (orange box) so that the guide RNA for allele 2 cannot cut the background allele. SNVs are then edited into allele two with an mApple fusion and antibiotic selection cassette. **D**) SNVs introduced into exon 16 of *POLG* and exons 32 and 33 of *MYBPC3* with iPSC-SGE compared to the maximum number of variants from other iPSC mutagenesis studies.

Figure 2: iPSC-SGE of POLG. A) Schematic of saturation genome editing in W748S iPSCs. **B**) Variants are scored at day 14 post sort and displayed by their position at exon 16 and colored by molecular consequence. 23 potential dominant negative missense variants were depleted and are colored in dark green in all plots. **C**) Schematic of saturation genome editing in POLG null iPSCS. **D**) Variants were scored at day 14 post sort and loss of function and dominant negative variants were depleted. **E**) Histogram of all variants scored in POLG null iPSCs. **F-G**) POLG iPSC-SGE scores on null background compared to variant effect predictors: CADD, AlphaMissense, and REVEL.

Figure 3: Clinical integration of *POLG* iPSC-SGE data. A) Gaussian mixture model fit to full distribution of *POLG* null iPSC-SGE scores. Thresholds are drawn at the density ratio where variants have a 95% or greater chance of being sampled from the left or right distribution. Variants scoring below -2.96 were functionally abnormal and variants scoring above -2.17 were functionally normal. **B**) GMM thresholds fit to CADD scores for exon 16. Variants with scores above 22.9 were considered to be predicted deleterious and variants below 12.5 were predicted normal. **C-D**) ClinVar pathogenic/likely pathogenic and benign/likely benign variants plotted with GMM thresholds. **E**) iPSC-SGE and CADD scores for ClinVar VUS measured in the assay. Dashed lines represent normal and abnormal thresholds for each score range. **F-G**) Donut plots depicting the breakdown of ClinVar variants before (F) and after (G) the application of functional data from iPSC-SGE. H) Histogram of all exon 16 variants colored by variant class. Blue represents functionally normal, gray represents indeterminate, and red represents functionally abnormal. **I**) Functional and predictor scores for all exon 16 variants.

Figure 4: iPSC-SGE of *MYBPC3*. A) Schematic of saturation genome editing of *MYBPC3* with a wild type GFP background allele. For *MYBPC3*, edited iPSCs are differentiated into cardiomyocytes for phenotyping. **B**) Representative cardiomyocytes from exon 32 library experiment. iPSCs are depicted in the top row, and have no MYBPC3 expression. Middle row shows a wild type-like cardiomyocyte where mApple and GFP signals localize to sarcomeres. Bottom row represents a reduce abundance variant where GFP signal localizes to sarcomeres, but mApple is dim and diffuse. **C**) Flow plot from an exon 32 cardiomyocyte sort. GFP is on the y-axis and mAppleo on the x-axis showing a broad distribution of mApple to GFP ratio. **D**) Schematic of sequencing and scoring variants based on frequencies across four bins. **E**) Exon 32 SNV scores by genomic position colored by molecular consequence. Note that *MYBPC3* is on the minus strand. Missense variants with low abundance show the boundary of the C10 domain. **F**) Histogram of all *MYBPC3* exon 32 iPSC-SGE function scores. **G-I**) MYBPC3 iPSC-SGE scores compared with CADD, AlphaMissense, and REVEL.

Figure 5: Clinical integration of MYBPC3 iPSC-SGE data. A) Gaussian mixture model fit to full distribution of *MYBPC3* iPSC-SGE scores. Thresholds are drawn at the density ratio where variants have a 95% or greater chance of being sampled from the left or right distribution. Variants scoring below 0.231 were functionally abnormal and variants scoring above 0.326 were functionally normal. **B**) GMM thresholds fit to CADD scores for exon 16. Variants with scores above 16.12 were considered to be predicted deleterious and variants below 8.28 were predicted normal. **C-D**) ClinVar pathogenic/likely pathogenic and benign/likely benign variants plotted with GMM thresholds. **E**) iPSC-SGE and CADD scores for ClinVar VUS measured in the assay. Dashed lines represent normal and abnormal thresholds for each score range. **F-G**) Donut plots depicting the breakdown of ClinVar variants before (F) and after (G) the application of functional data from iPSC-SGE. **H**) Histogram of all exon 32 variants colored by variant class. Blue represents functionally normal, gray represents indeterminate, and red represents functionally abnormal. **I**) Functional and predictor scores for all exon 32 variants.

Figure 6: CRISPR screen of Mendelian disease genes in iPSC derived neurons. A) Schematic of iPSC neuron CRISPR screen. Cas9 iPSCs are transduced with library of guide RNAs targeting 4,502 genes. After selection for guide integration with blasticidin, iPSCs are

cultured for 3 weeks before differentiation into neurons. Neuron differentiation is carried out until day 21 and guides are sequenced and scored. **B**) CRISPR screen rank plot of average day 21 neuron CRISPR scores for each gene in the screen. C) Volcano plot gene fold change in day 21 neurons vs. iPSCs by -log10 P values. Dashed curved line represents 5% false discovery rate defined by pseudo-gene scores from rand NTC guide pairs. Blue dots represent genes that are significantly depleted in neurons and red dots represent genes that are enriched in neurons. Orange dots are non-significant genes and gray dots are NTC pseudo genes.

Figures

Figure 1:

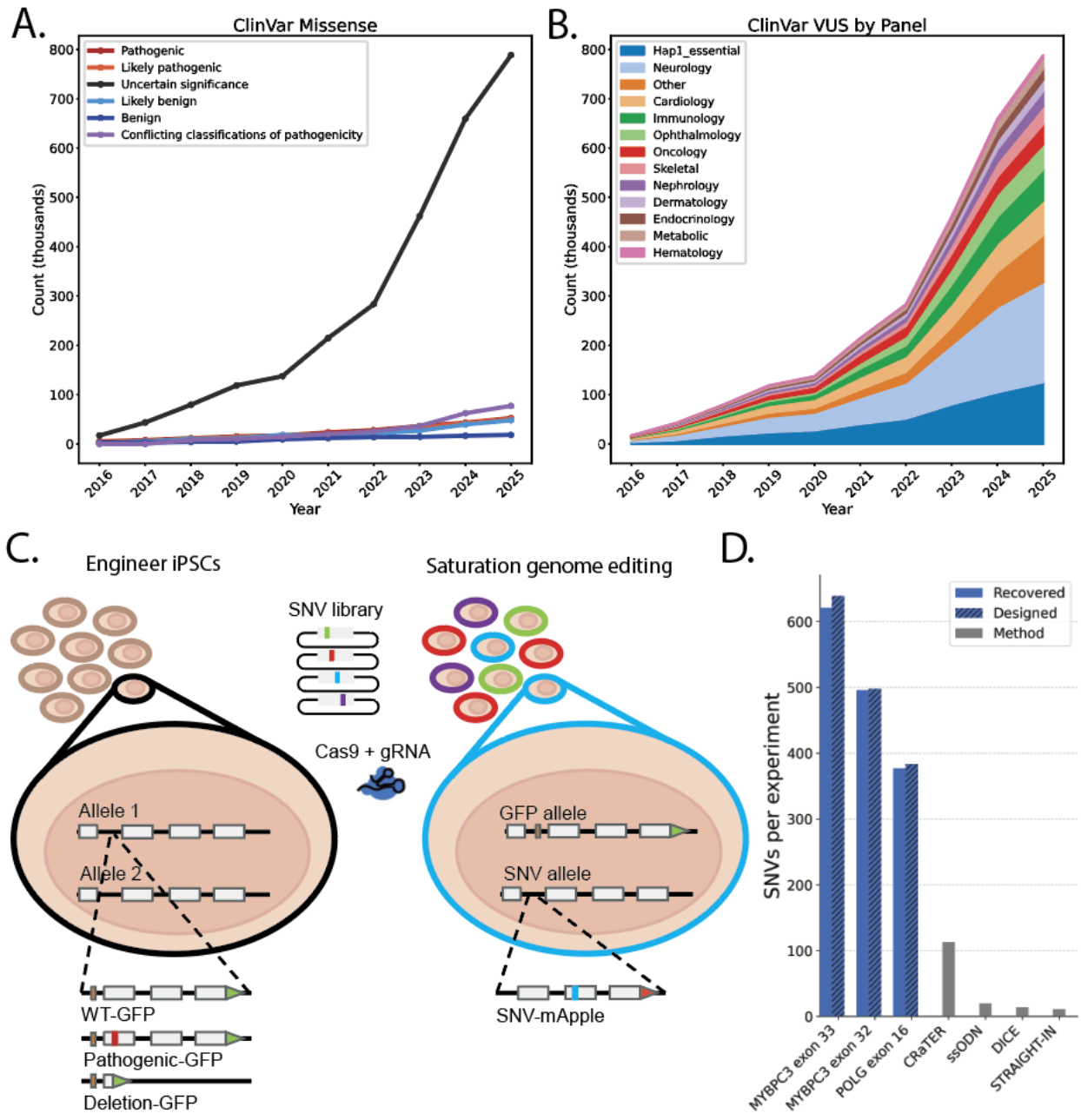


Figure 2:

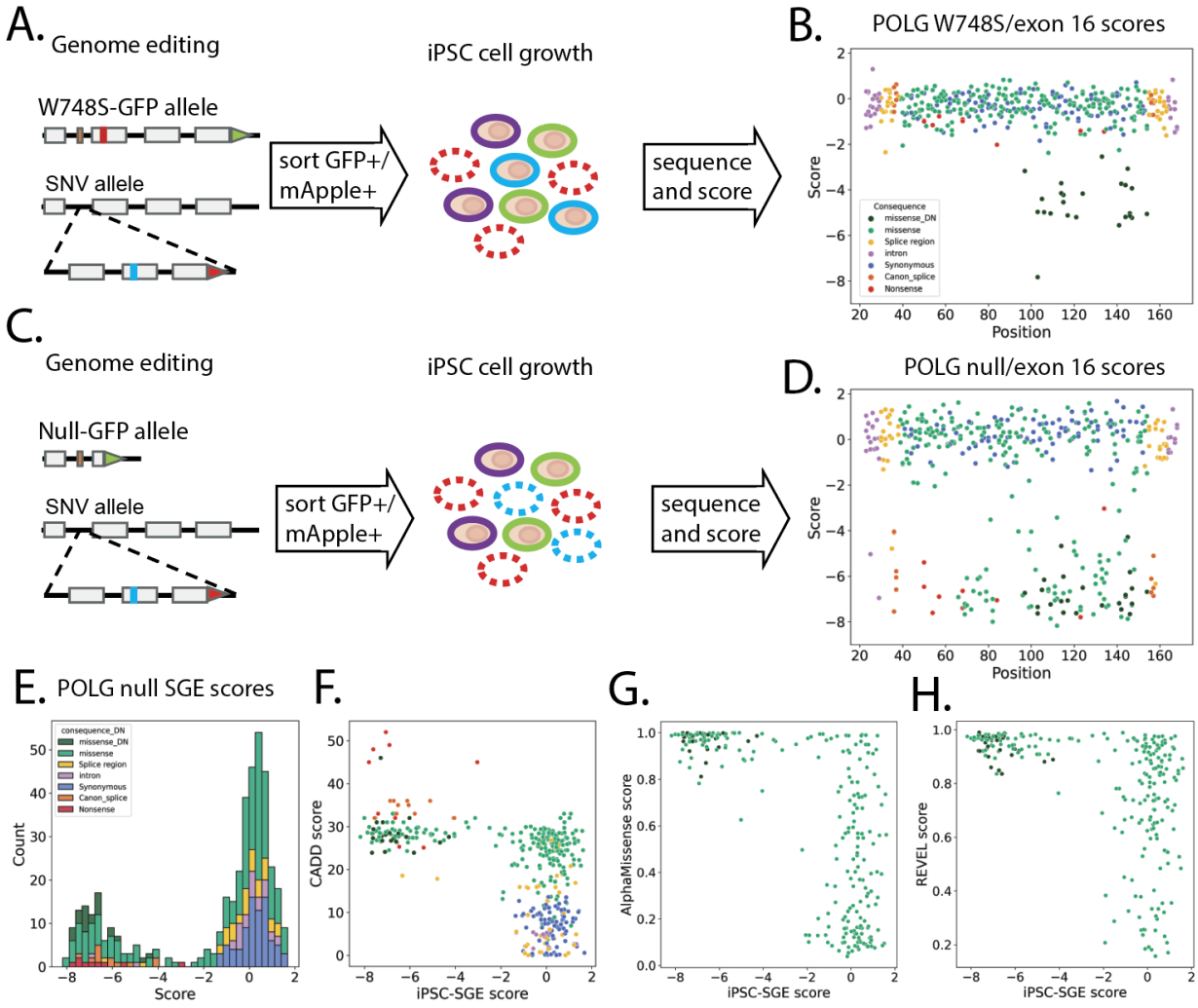

A.
Genome editing
W748S-GFP allele
SNV allele
sort GFP+/
mApple+
iPSC cell growth
sequence
and score
B.
POLG W748S/exon 16 scores
Score
Position
Consequence
missense_DN
missense
Splice region
intron
Synonymous
Canon_splice
Nonsense
C.
Genome editing
Null-GFP allele
SNV allele
sort GFP+/
mApple+
iPSC cell growth
sequence
and score
D.
POLG null/exon 16 scores
Score
Position
E.
POLG null SGE scores
Count
Score
F.
CADD score
iPSC-SGE score
G.
AlphaMissense score
iPSC-SGE score
H.
REVEL score
iPSC-SGE score

Figure 3:

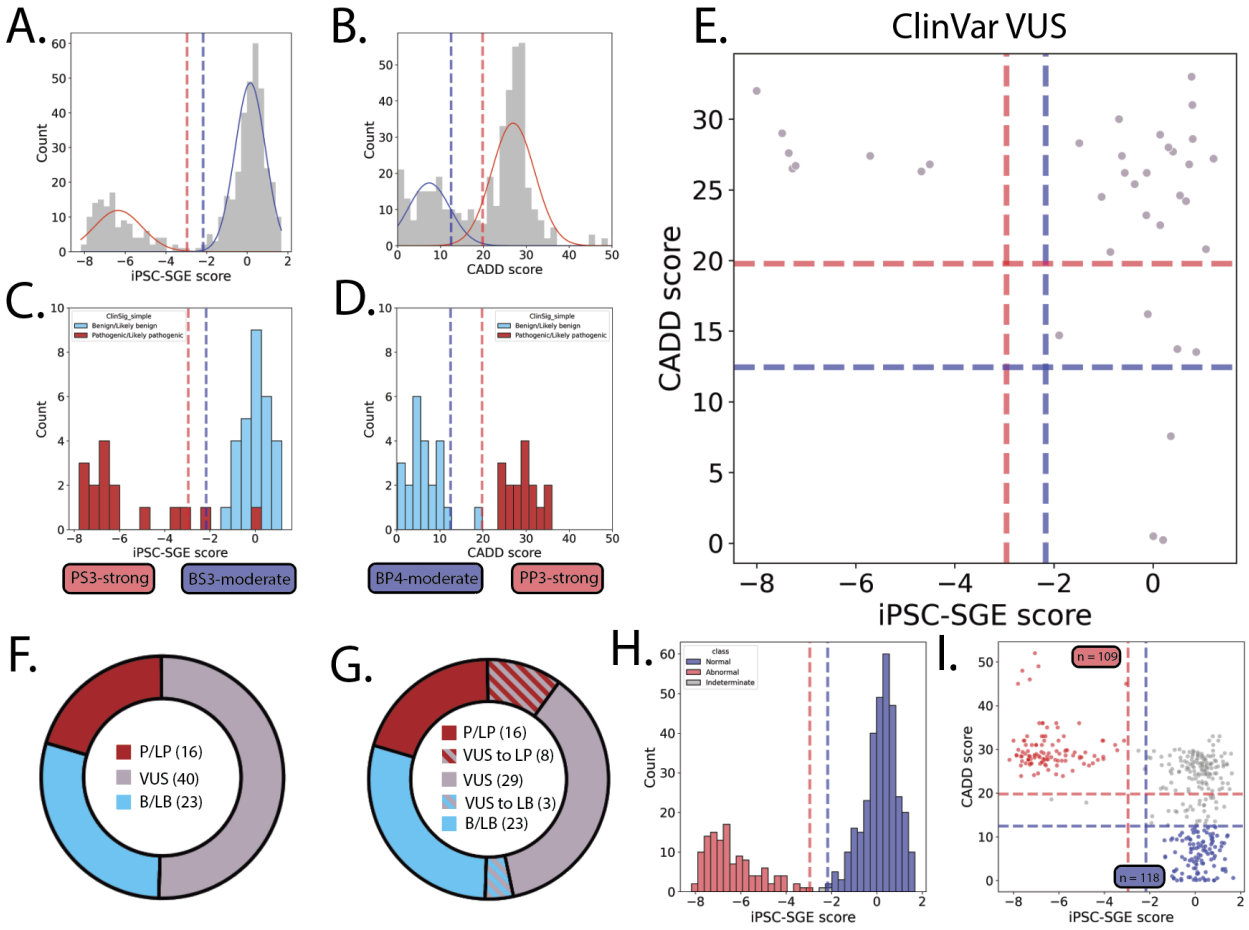

A.
B.
C.
D.
E.
ClinVar VUS
Count
iPSC-SGE score
CADD score
PS3-strong
BS3-moderate
BP4-moderate
PP3-strong
F.
P/LP (16)
VUS (40)
B/LB (23)
G.
P/LP (16)
VUS to LP (8)
VUS (29)
VUS to LB (3)
B/LB (23)
H.
I.
n = 109
n = 118

Figure 4:

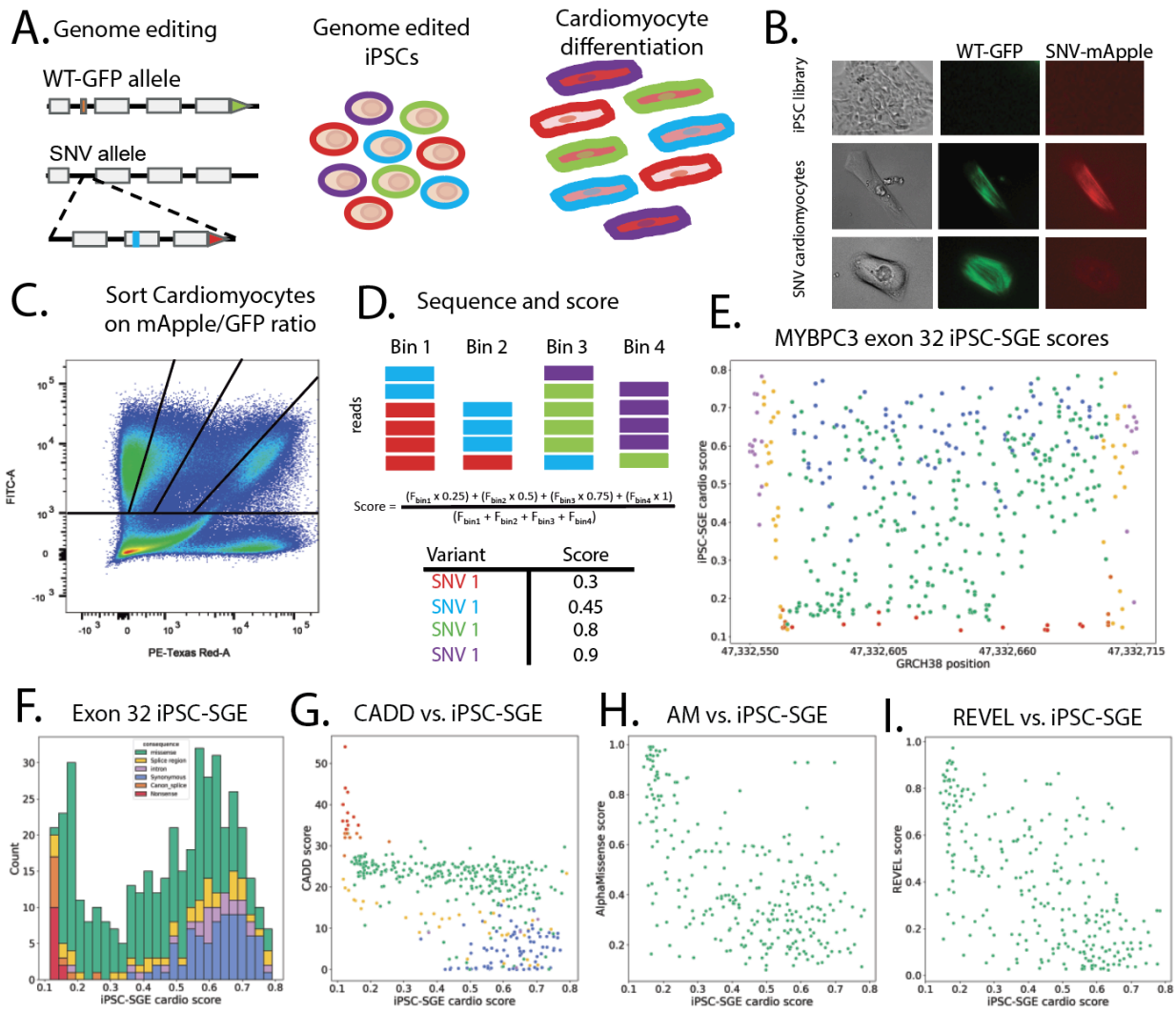

A. Genome editing
WT-GFP allele
SNV allele
Genome edited iPSCs
Cardiomyocyte differentiation
B.
WT-GFP
SNV-mApple
iPSC library
SNV cardiomyocytes
C. Sort Cardiomyocytes on mApple/GFP ratio
FITC-A
PE-Texas Red-A
D. Sequence and score
Bin 1
Bin 2
Bin 3
Bin 4
reads
Score = (F_bin1 x 0.25) + (F_bin2 x 0.5) + (F_bin3 x 0.75) + (F_bin4 x 1) / (F_bin1 + F_bin2 + F_bin3 + F_bin4)
Variant
Score
SNV 1
0.3
SNV 1
0.45
SNV 1
0.8
SNV 1
0.9
E. MYBPC3 exon 32 iPSC-SGE scores
iPSC-SGE cardio score
GRCH38 position
47,332,550
47,332,605
47,332,660
47,332,715
F. Exon 32 iPSC-SGE
Count
iPSC-SGE cardio score
G. CADD vs. iPSC-SGE
CADD score
iPSC-SGE cardio score
H. AM vs. iPSC-SGE
AlphaMissense score
iPSC-SGE cardio score
I. REVEL vs. iPSC-SGE
REVEL score
iPSC-SGE cardio score

Figure 5:

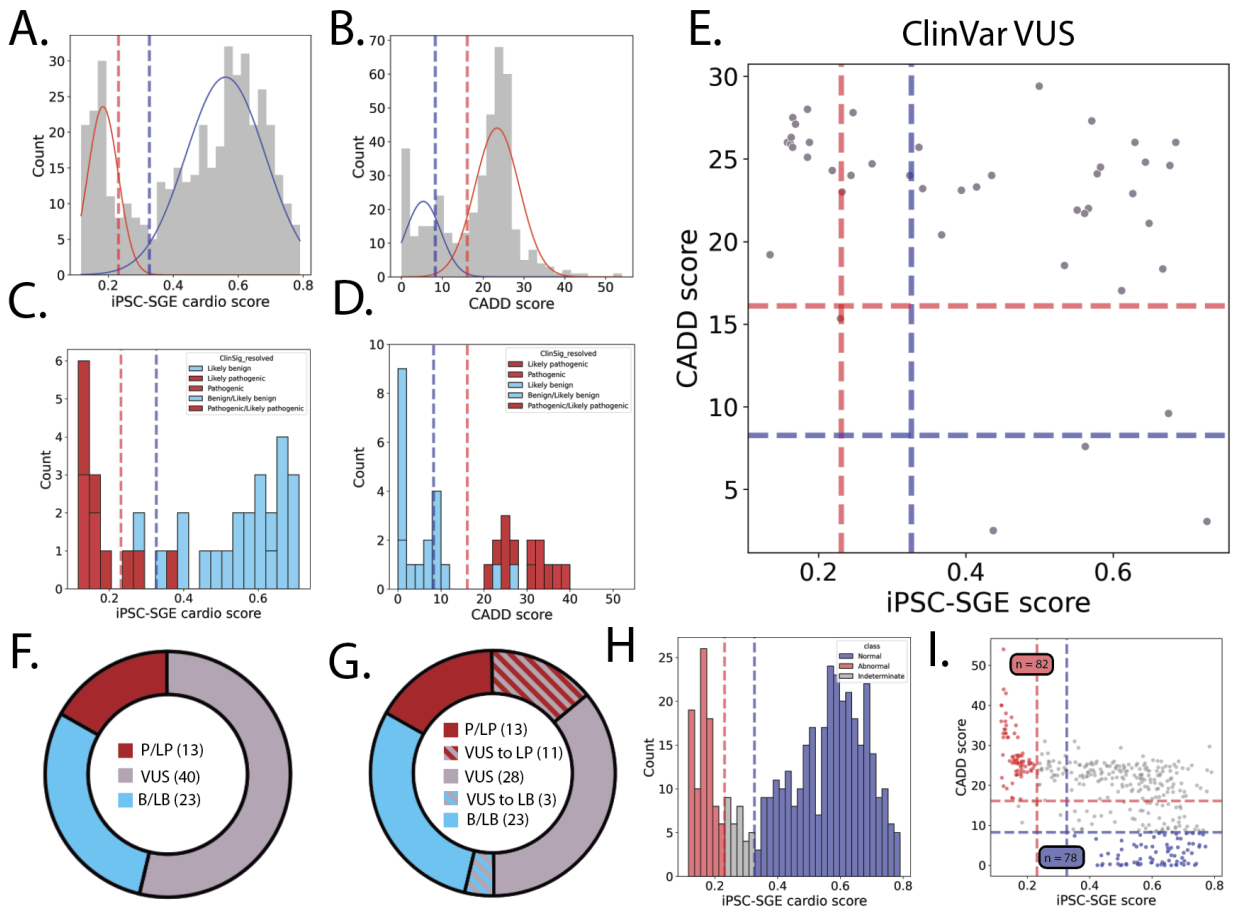
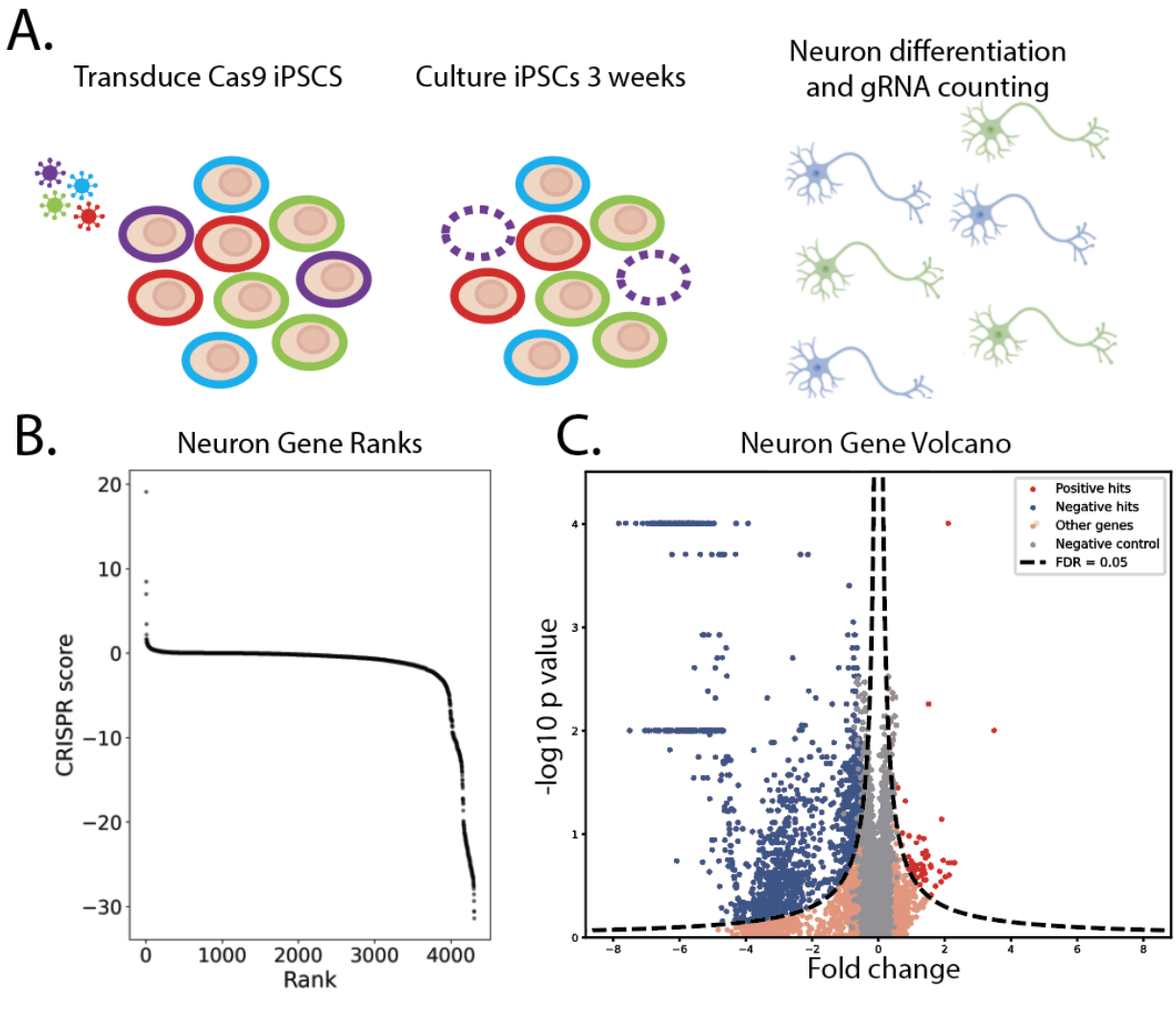

A.
Count
iPSC-SGE cardio score
B.
Count
CADD score
C.
ClinSig_resolved
Likely benign
Likely pathogenic
Pathogenic
Benign/Likely benign
Pathogenic/Likely pathogenic
Count
iPSC-SGE cardio score
D.
ClinSig_resolved
Likely pathogenic
Pathogenic
Likely benign
Benign/Likely benign
Pathogenic/Likely pathogenic
Count
CADD score
E.
ClinVar VUS
CADD score
iPSC-SGE score
F.
P/LP (13)
VUS (40)
B/LB (23)
G.
P/LP (13)
VUS to LP (11)
VUS (28)
VUS to LB (3)
B/LB (23)
H
class
Normal
Abnormal
Indeterminate
Count
iPSC-SGE cardio score
I.
n = 82
n = 78
CADD score
iPSC-SGE score

Figure 6:


A.
Transduce Cas9 iPSCS
Culture iPSCs 3 weeks
Neuron differentiation and gRNA counting
B.
Neuron Gene Ranks
CRISPR score
Rank
C.
Neuron Gene Volcano
-log10 p value
Fold change
Positive hits
Negative hits
Other genes
Negative control
FDR = 0.05

References

1. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* **53**, D1313–D1321 (2025).

2. Fayer, S. *et al.* Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *Am J Hum Genet* **108**, 2248–2258 (2021).

3. Scott, A. *et al.* Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. *Genome Biol* **23**, 266 (2022).

4. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).

5. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).

6. Radford, E. J. *et al.* Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. *Nat Commun* **14**, 7702 (2023).

7. Buckley, M. *et al.* Saturation genome editing maps the functional spectrum of pathogenic VHL alleles. *Nat Genet* **56**, 1446–1455 (2024).

8. Waters, A. J. *et al.* Saturation genome editing of BAP1 functionally classifies somatic and germline variants. *Nat Genet* **56**, 1434–1445 (2024).

9. Olvera-León, R. *et al.* High-resolution functional mapping of RAD51C by saturation genome editing. *Cell* **187**, 5719–5734.e19 (2024).

10. Sahu, S. *et al.* Saturation genome editing-based clinical classification of BRCA2 variants. *Nature* (2025) doi:10.1038/s41586-024-08349-1.

11. Huang, H. *et al.* Functional evaluation and clinical classification of BRCA2 variants. *Nature* (2025) doi:10.1038/s41586-024-08388-8.

12. Carette, J. E. *et al.* Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**, 340–343 (2011).

13. Funk, J. S. *et al.* Deep CRISPR mutagenesis characterizes the functional diversity of TP53 mutations. *Nat Genet* **57**, 140–153 (2025).

14. Meitlis, I. *et al.* Multiplexed Functional Assessment of Genetic Variants in CARD11. *Am J Hum Genet* **107**, 1029–1043 (2020).

15. Friedman, C. E. *et al.* CRaTER enrichment for on-target gene editing enables generation of variant libraries in hiPSCs. *J Mol Cell Cardiol* **179**, 60–71 (2023).

16. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874–882 (2018).

17. Gonçalves, E. *et al.* Minimal genome-wide human CRISPR-Cas9 library. *Genome Biol* **22**, 40 (2021).

18. Wang, C. *et al.* Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening. *Stem Cell Reports* **9**, 1221–1233 (2017).

19. Tian, R. *et al.* CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron* **104**, 239–255.e12 (2019).

20. Kreitzer, F. R. *et al.* A robust method to derive functional neural crest cells from human pluripotent stem cells. *Am J Stem Cells* **2**, 119–131 (2013).

21. Rahman, S. & Copeland, W. C. POLG-related disorders and their neurological manifestations. *Nat Rev Neurol* **15**, 40–52 (2019).

22. Alfares, A. A. *et al.* Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genet Med* **17**, 880–888 (2015).

23. Helms, A. S. *et al.* Spatial and Functional Distribution of Pathogenic Variants and Clinical Outcomes in Patients With Hypertrophic Cardiomyopathy. *Circ Genom Precis Med* **13**, 396–405 (2020).

24. Thompson, A. D. *et al.* Computational prediction of protein subdomain stability in MYBPC3 enables clinical risk stratification in hypertrophic cardiomyopathy and enhances variant interpretation. *Genet Med* **23**, 1281–1287 (2021).

25. Silva, A. C., Moreira, J. N., Lobo, J. M. S. & Almeida, H. *Current Applications of Pharmaceutical Biotechnology*. (Springer Nature, 2020).

26. Lv, W. *et al.* Functional Annotation of TNNT2 Variants of Uncertain Significance With Genome-Edited Cardiomyocytes. *Circulation* **138**, 2852–2854 (2018).

27. Rath, A. *et al.* Functional interrogation of Lynch syndrome-associated MSH2 missense

variants via CRISPR-Cas9 gene editing in human embryonic stem cells. *Hum Mutat* **40**, 2044–2056 (2019).

28. Lee, Y.-S., Kennedy, W. D. & Yin, Y. W. Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations. *Cell* **139**, 312–324 (2009).

29. Jazayeri, M. *et al.* Inducible expression of a dominant negative DNA polymerase-gamma depletes mitochondrial DNA and produces a rho0 phenotype. *J Biol Chem* **278**, 9823–9830 (2003).

30. Niroula, A. & Vihinen, M. How good are pathogenicity predictors in detecting benign variants? *PLoS Comput Biol* **15**, e1006481 (2019).

31. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).

32. Wong, L.-J. C. *et al.* Molecular and clinical genetics of mitochondrial diseases due to POLG mutations. *Hum Mutat* **29**, E150–72 (2008).

33. Rempe, T. *et al.* Early-onset parkinsonism due to compound heterozygous POLG mutations. *Parkinsonism Relat Disord* **29**, 135–137 (2016).

34. González-Vioque, E. *et al.* Association of novel POLG mutations and multiple mitochondrial DNA deletions with variable clinical phenotypes in a Spanish population. *Arch Neurol* **63**, 107–111 (2006).

35. Davidzon, G. *et al.* Early-onset familial parkinsonism due to POLG mutations. *Ann Neurol* **59**, 859–862 (2006).

36. Miyamoto, C. A., Fischman, D. A. & Reinach, F. C. The interface between MyBP-C and myosin: site-directed mutagenesis of the CX myosin-binding domain of MyBP-C. *J Muscle Res Cell Motil* **20**, 703–715 (1999).

37. Kuster, D. W. D. *et al.* A hypertrophic cardiomyopathy-associated MYBPC3 mutation common in populations of South Asian descent causes contractile dysfunction. *J Biol Chem* **290**, 5855–5867 (2015).

38. DiStefano, M. T. *et al.* The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. *Genet Med* **24**, 1732–1742 (2022).

39. Lu, C. *et al.* Essential transcription factors for induced neuron differentiation. *Nat Commun* **14**, 8362 (2023).

Chapter 4: Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes

Malvika Tejura,[1,†] Shawn Fayer,[1,†] Abbye E. McEwen,[1,3,4] Jake Flynn,[5] Lea M. Starita,[1,3,*] Douglas M. Fowler[1,2,3,*]

Abstract

In silico variant effect predictions are available for nearly all missense variants, but played a minimal role in clinical variant classification because they were deemed to provide only supporting evidence. Recently, the ClinGen Sequence Variant Interpretation (SVI) Working Group updated recommendations for variant effect prediction use. By analyzing control pathogenic and benign variants across all genes they were able to compute evidence strength for predictor score intervals, with some intervals generating moderate, strong, or even very strong evidence. However, this genome-wide approach could obscure heterogeneous predictor performance in different genes. We quantified the gene-by-gene performance of two top predictors, REVEL and BayesDel, by analyzing control variants in each predictor score interval in 3,668 disease-relevant genes. Approximately 10% of intervals had sufficient control variants for analysis, and ~70% of these intervals exceeded the maximum number of incorrect predictions implied by the SVI recommendations. These trending discordant intervals arose owing to the divergence of the gene-specific distribution of predictions from the genome-wide distribution, suggesting that gene-specific calibration is needed in many cases. Approximately 22% of ClinVar missense variants of uncertain significance in genes we analyzed (REVEL = 100,629, BayesDel = 71,928) had predictions in trending discordant intervals. Thus, genome-wide calibrations could result in many variants receiving inappropriate evidence strength. To facilitate a review of the SVI's calibrations, we developed a web application enabling visualization of gene-specific predictions and trending concordant and discordant intervals.

Introduction

Clinical genetic testing is of critical importance to precision medicine. Individualized risk assessment and clinical management depend on correctly interpreting sequence variants as either disease-causing (pathogenic) or not disease-causing (benign). When interpreting variants, scientists and clinicians combine evidence from various sources including an individual's phenotype, case-control studies, family variant segregation analyses, population frequencies, *in vitro* functional assays, and *in silico* computational variant effect predictors. Evidence from these sources is weighted as either supporting, moderate, strong, or very strong based on its accuracy in predicting pathogenicity or benignity. Guidelines set forth by the American College of Medical Genetics and Association for Molecular Pathology (ACMG/AMP) define how multiple pieces of evidence of different strengths can be combined to achieve pathogenic, likely pathogenic, likely benign, or benign interpretations[1]. Variants lacking sufficient evidence are interpreted as variants of uncertain significance (VUS) and should not be used to inform medical management[1]. Rare missense variants are particularly challenging to interpret because many key pieces of evidence are missing or uninformative, and consequently often

become VUS[2–4]. Increased genetic testing and the expansion of clinical tests to more genes have resulted in an explosion of VUS and, currently, approximately 86% of missense variants in the ClinVar database are VUS (n=976,946/1,132,728 accessed August 2023)[5]. Timely resolution of this large and rapidly growing VUS problem requires high-quality evidence that can be generated for all variants.

Variant effect predictions from various tools are available for nearly all possible single nucleotide variants in most genes. Thus, variant effect predictors could have a profound impact in reducing VUS. However, predictor evidence was limited to supporting strength in the 2015 ACMG/AMP guidelines because of the relatively low sensitivity and specificity of predictors available at the time[1]. Predictors have substantially improved in accuracy since the publication of these guidelines[6–11], and guidelines for calibrating evidence strength for variant interpretation using a Bayesian framework opened the door to re-evaluating predictor evidence strength[12,13].

Recognizing this opportunity, the ClinGen Sequence Variant Interpretation (SVI) Working Group updated its guidance for the use of predictor evidence for clinical variant interpretation[14]. In this guidance, the predictor score range from an aggregated set of 1,913 genes in the ClinVar database (**Figure 1A, B**) was divided into sliding windows, and a positive likelihood ratio and posterior probability of pathogenicity was calculated from the control variants in the window[5]. From this sliding window analysis, the ClinGen SVI defined predictor score intervals, corresponding to specific evidence strength[12,13]. This genome-wide calibration avoided a key problem that has bedeviled gene-specific evidence strength calibrations, namely that many genes have very few to no known pathogenic and benign variants to use as calibration controls. Genome-wide calibration enabled the calculation of predictor evidence strength from a large number of control variants and thus constituted a massive step forward. Perhaps the most surprising outcome was that REVEL, the best-performing predictor, could generate strong or very strong evidence for approximately 3.5% of possible missense variants in the genome. This outcome is in contrast to the 2015 ACMG guidelines, where REVEL and all other predictors could generate only supporting evidence. Thus, the genome-wide calibration could transform predictor use, leading to the reinterpretation of a substantial fraction of VUS.

Previous studies have shown that predictors perform inconsistently across genes[6,9,15]. Here we investigated how the genome-wide calibration for the two predictors that achieved the highest strength of evidence, REVEL and BayesDel, performed when applied to individual genes. First, we computed the number of correct and incorrect predictions needed in each interval to deem it either trending concordant or trending discordant with the evidence strength assigned by the genome-wide calibration. Then, using previously classified pathogenic and benign variants as controls, we computed the number of incorrect predictions present in each interval for 3,668 disease-relevant genes currently in ClinVar. As expected, 90% of intervals lacked enough control variants to evaluate. For the remaining intervals, approximately 30% were trending concordant, and 70% were trending discordant across both predictors, implying that the evidence strength assigned to that interval for the gene in question was inappropriate. Because strong evidence could drive reinterpretation of VUS, we analyzed more than 350,000 ClinVar missense VUS across both predictors and found that close to a quarter of variants were in trending discordant intervals, while approximately one-fifth of variants were in trending concordant intervals. Analysis of the distribution of variant effect predictions and the disposition of concordant and discordant intervals for each gene revealed that the fundamental assumption

underlying genome-wide calibration, that predictor scores are similarly distributed for all genes, is problematic for some genes. In some genes, the distribution of variant effect predictions matched the all-gene distribution and separated pathogenic and benign variants well. However, for other genes, the distribution of predictions did not match the all-gene distribution, which suggests that some of these genes should be recalibrated. Finally, we developed a web resource that enables clinicians and researchers to visualize the distribution of variant effect predictions for 3,668 genes and view trending concordant and discordant intervals. Thus, we demonstrate the promise of the genome-wide calibration approach taken by the ClinGen SVI for some genes, while revealing this updated guidance leads to a large number of variants receiving inappropriately strong or weak evidence.

Methods

Dataset curation and filtering

To facilitate gene-specific analyses assessing concordance or discordance with the genome-wide calibration, we generated three distinct datasets using supplemental data from Pejaver et al. and downloads from ClinVar[5,14].

ClinGen SVI Calibration Dataset

The ClinGen SVI Calibration Dataset was created to gain insights into the distribution of variants used to develop the genome-wide calibration. The ClinGen SVI Calibration Dataset is a combination of the ClinVar 2019 training dataset and the ClinVar 2020 test dataset from supplemental Data S1 and S3 in Pejaver et al.[14]. The variants in the ClinVar 2019 and 2020 datasets are 1+ stars (variants where at least one submitter has provided assertion criteria for the classification), non-VUS, missense variants, with allele frequencies below 0.01, and are not variants used to train any of the predictors analyzed in the genome-wide calibration including REVEL and BayesDel. The ClinGen SVI Calibration Dataset consists of 20,948 variants from 2,711 genes (**Table S1**).

ClinVar 2023 Dataset

To include the latest ClinVar variants in our analyses, we generated the ClinVar 2023 Dataset by downloading a GRCh37 VCF (variant call file) containing all variants present in ClinVar as of August 23rd, 2023. The VCF was then annotated with REVEL and BayesDel (version 1) predictor scores and gnomAD (version 2.1.1) allele frequencies using the filter-based annotation feature (hg19 assembly, dbsnfp42c for predictor scores and genome_211 and exome_211 for gnomAD allele frequencies) in Annovar (version 2020_06_07)[16]. Like the ClinGen SVI Calibration Dataset, the ClinVar 2023 Dataset was filtered to retain 1+ star, non-VUS missense variants. Genes without any pathogenic variants were excluded, and variants with allele frequencies exceeding 0.01 were also excluded. Allele frequencies were derived from gnomAD exome data unless the variant was not found in exome data, in which case allele frequencies from whole genomes were used[14]. We then mapped Entrez gene IDs to HGNC gene names for use in subsequent analyses. Next, we filtered our dataset on genes classified as having a definitive, strong, or moderate association with disease from the GenCC database (last accessed March 12th, 2024) as performed in Stenton et al.[17]. The final filtered ClinVar 2023 Dataset consisted of 89,947 variants across 3,668 genes (**Table S2**).

ClinVar 2023 Dataset without training variants

To analyze REVEL and BayesDel performance while excluding training variants, we cross-referenced the ClinGen SVI Calibration Dataset with a ClinVar VCF file from December 2020. We considered any variants absent from the ClinGen SVI Calibration Dataset and present in the ClinVar December 2020 download to be training variants because the ClinGen SVI Calibration Dataset was filtered to exclude REVEL and BayesDel training variants. The training variants identified by this analysis were removed from the Clinvar 2023 Dataset (**Table S2**) creating the Clinvar 2023 Dataset without training variants, which consisted of 71,791 variants across 3,623 genes (**Table S3**).

ClinVar 2023 VUS Dataset

To analyze how many variants of uncertain significance were in trending concordant, trending discordant, or insufficient variants intervals, we downloaded a GRCh37 VCF file containing all variants present in ClinVar as of August 23rd, 2023. We then filtered this dataset to retain 1+ star, missense VUS, creating the ClinVar 2023 VUS Dataset (**Table S4**)

REVEL and BayesDel variant effect predictor score distributions for each gene

We downloaded predictions for all possible variants for REVEL (82,100,677 variants, downloaded 02.02.2023) (DataS1, https://zenodo.org/records/11256843) and BayesDel_170824_noAF (82,352,098 variants, downloaded 10.05.2023) (DataS2, https://zenodo.org/records/11256843). We visualized the REVEL and BayesDel predictions for all possible variants as density distributions in Figures 1A and 1B. To generate density distributions for gene-specific REVEL predictions, we mapped each variant to a gene using the provided Ensembl transcripts and the comprehensive gene annotation file from GENCODE release 9, which contained transcript information from Ensembl release 64. The gene-annotated REVEL all variant file contained all 82,100,677 variants **(**DataS3, https://zenodo.org/records/11256843). The BayesDel all variant file had no transcript or strand polarity information, so, to map each variant to a gene, we merged the REVEL all variant annotated file with the BayesDel all variant file on shared chromosome number, genomic coordinates, reference alleles, and alternate alleles. The gene-annotated BayesDel all variant file had 77,814,694 variants (DataS4, https://zenodo.org/records/11256843).

Incorrect prediction tolerance in each evidence strength interval

We defined an “evidence strength interval” as the variant effect predictor score interval denoted in the genome-wide calibration for a given evidence strength. For example, for REVEL, the prediction interval between 0.183 and 0.290 generated supporting benign evidence, and in our analysis corresponds to the “supporting benign evidence strength interval”[14].

We calculated the “incorrect prediction tolerance” for each evidence strength interval in the ClinGen SVI Calibration Dataset (**Table 1**, **Table S1**).

$$incorrect\ prediction\ tolerance\ = \frac{number\ of\ incorrectly\ predicted\ variants\ in\ evidence\ strength\ interval}{total\ number\ of\ variants\ in\ evidence\ strength\ interval} \times 100$$

The incorrect prediction tolerance for each interval defines the maximum percentage of incorrectly predicted variants that can occur in the interval, as implied by the genome-wide calibration. For example, the genome-wide supporting pathogenic interval for REVEL had 392 incorrectly predicted benign or likely benign control variants out of 1,534 total variants, leading to an incorrect prediction tolerance of 25.88%.

Next, to quantify the extent of incorrect predictions within each evidence strength interval in each gene, we calculated the "incorrect prediction percentage". The incorrect prediction percentage is the percentage of variants in each gene in the ClinVar 2023 Dataset (**Table S2**) classified as pathogenic, likely pathogenic, benign, or likely benign that were incorrectly predicted by either REVEL or BayesDel within each evidence strength interval.

The incorrect prediction tolerance is the maximum percentage of allowed incorrect predictions in the ClinGen SVI Calibration Dataset, whereas the incorrect prediction percentage is the percentage of incorrect predictions in a given interval for a given gene. Both the incorrect prediction tolerance and the incorrect prediction percentage are calculated in the same way; however, the incorrect prediction tolerance remains the same for each interval throughout our analysis, whereas the incorrect prediction percentage changes on a gene and interval-wise basis.

$$incorrect\ prediction\ percentage\ =\ \frac{number\ of\ incorrectly\ predicted\ variants\ in\ evidence\ strength\ interval}{total\ number\ of\ variants\ in\ evidence\ strength\ interval}\times 100$$

For example, the supporting pathogenic interval for REVEL in *BRCA1* had 31 incorrectly predicted benign or likely benign control variants out of 82 total variants, leading to an incorrect prediction percentage of 37.8%. This calculation was repeated for every interval in each of the 3,668 genes in the ClinVar 2023 Dataset (REVEL: **Table S5,** BayesDel: **Table S6)**.

Assessing evidence strength interval concordance or discordance for individual genes.

To determine whether an evidence strength interval in a gene was concordant with the genome-wide calibration, we compared the percentage of incorrectly predicted control variants observed in the interval to the incorrect prediction tolerance for that interval. We employed a statistical approach based on the binomial distribution to account for the often small number of control variants in each evidence strength interval. Here, for each evidence strength interval, we used the binomial distribution and the incorrect prediction tolerance (**Table 1**) to calculate the minimum number of control variants in the interval necessary to have an 80% chance of observing at least one incorrectly predicted variant in that interval (**Table 2**).

$$x_{incorrect}\ =\ number\ of\ incorrectly\ predicted\ variants\ in\ the\ evidence\ strength\ interval$$
$$n\ =\ minimum\ number\ of\ control\ variants\ needed\ in\ the\ evidence\ strength\ interval$$
$$p_{incorrect}\ =\ probability\ of\ an\ incorrectly\ predicted\ variant\ (e.g.\ incorrect\ prediction\ tolerance)$$

$$P\left(x_{incorrect}\geq 1\right)=\ 1\ -\ P(x\ =\ 0)$$
$$0.8\ =\ 1\ -\ n\,C\,x_{incorrect}\times p_{incorrect}{}^{x_{incorrect}}\times\left(1\ -\ p_{incorrect}\right)^{n-x_{incorrect}}$$

$$n = \frac{\log log\,(0.2)}{\log log\left(1-p_{incorrect}\right)}$$

For example, for the moderate pathogenic interval in the REVEL predictor, we set $p_{incorrect}$ equal to the incorrect prediction tolerance of 10.75% or 0.1075, yielding a minimum of 15 variants needed to have an 80% chance of observing at least one incorrectly predicted variant in the interval. We repeated this calculation for all the evidence strength intervals for both REVEL and BayesDel. If the number of control variants in the interval equaled or exceeded the minimum number of control variants, and the incorrect prediction percentage of the interval was less than or equal to the incorrect prediction tolerance, the interval was deemed "trending concordant" with the genome-wide calibration. We deemed such intervals trending concordant in recognition of the fact that, as more control variants accrue, some of the concordant intervals could become discordant.

To determine whether an evidence strength interval in a gene was discordant with the genome-wide calibration, we again compared the percentage of incorrectly predicted pathogenic or benign control variants observed in the interval to the incorrect prediction tolerance for that interval. Here, we used the binomial distribution to calculate the minimum number of control variants necessary to have an 80% chance of observing at least one correctly predicted variant in the evidence strength interval (**Table 3**).

$x_{correct} = number\ of\ correctly\ predicted\ variants\ in\ the\ evidence\ strength\ interval$
$n = minimum\ number\ of\ control\ variants\ needed\ in\ the\ evidence\ strength\ interval$
$\left(n - x_{correct}\right) = number\ of\ incorrectly\ predicted\ variants\ in\ evidence\ strength\ interval$
$p_{correct} = probability\ of\ a\ correctly\ predicted\ variant\ or\ (1 - incorrect\ prediction\ tolerance)$

$$P\left(x_{correct} \geq 1\right) = 1 - P(x = 0)$$

$$0.8 = 1 - n\,C\,x_{correct} \times p_{correct}^{\ x_{correct}} \times \left(1 - p_{correct}\right)^{n - x_{correct}}$$

$$n = \frac{\log log\,(0.2)}{\log log\left(1-p_{correct}\right)}$$

For example, for the moderate pathogenic interval in the REVEL predictor, we set $p_{correct}$ equal to one minus the incorrect prediction tolerance of 10.75% (1 - 0.1075, or 0.8925) yielding a minimum of 1 variant needed to have an 80% chance of observing at least one correctly predicted variant in the interval. We repeated this calculation for all evidence strength intervals. If the number of control variants in the interval equaled or exceeded the minimum number of control variants, and the incorrect prediction percentage of the interval was greater than the incorrect prediction tolerance, the interval was deemed "trending discordant" with the genome-wide calibration. We deemed such intervals trending discordant in recognition of the fact that, as more control variants accrue, some of the discordant intervals could become concordant.

While the number of variants needed for concordance varied by evidence strength interval, the number of variants needed for discordance was almost always one, except for the supporting pathogenic intervals, which required two variants. This dichotomy is a result of the binomial theorem. If the probability of success (e.g. correctly predicting a variant's effect) is high, then the number of data points required to observe at least one success is low. Therefore, the number of variants needed to have an 80% chance of observing at least one correctly predicted variant is much lower than the number of variants needed to have an 80% chance of observing at least one incorrectly predicted variant. Intervals that contained too few variants to be deemed either trending concordant or trending discordant were deemed to have insufficient evidence.

Since the incorrect prediction tolerance for the BP4 very strong and BP4 strong evidence strength intervals for REVEL was 0%, we were unable to quantify the number of variants needed in those intervals for concordance or discordance using the binomial-based method described above. As an alternative, we tallied the number of BP4 very strong and BP4 strong evidence strength intervals for REVEL that contained only correctly predicted variants vs. the number that contained any incorrectly predicted variants.

Results

To analyze the genome-wide calibrations for REVEL and BayesDel, we first used the incorrect prediction percentage equation (**Figure 1C**) to quantify the proportion of incorrect predictions in each evidence strength interval in the ClinGen SVI Calibration Dataset[14] (**Table S1**). This proportion defined the incorrect prediction tolerance for each interval, representing the maximum allowable incorrect predictions while remaining consistent with the posterior probability for that interval in the genome-wide calibration. For example, the supporting pathogenic interval for REVEL in the ClinGen SVI Calibration Dataset (**Table S1**) had a total of 1,534 variants, 397 of which were incorrectly predicted, making the incorrect prediction tolerance for this interval 25.88%. The incorrect prediction tolerance was calculated across all intervals for both REVEL and BayesDel, and we found that the incorrect prediction tolerance was similar across both predictors at each evidence strength interval (**Table 1**).

Many genes had an excess of incorrectly predicted variants across the range of evidence strength intervals

To determine if the genome-wide calibrations are appropriate for individual genes, we computed the percentage of incorrectly predicted variants (incorrect prediction percentage) for each evidence strength interval for all 3,668 genes in the ClinVar 2023 Dataset (**Figure 1C**, **Table S2**). In intervals where the number of variants equaled or exceeded the minimum number of control variants required for our analysis (see Methods), we compared each interval's incorrect prediction percentage to the interval's genome-wide incorrect prediction tolerance. Intervals were deemed trending concordant when the incorrect prediction percentage was less than or equal to that interval's genome-wide incorrect prediction tolerance. Intervals were deemed trending discordant when the incorrect prediction percentage exceeded that interval's genome-wide incorrect prediction tolerance (**Figure 1D)**.

For example, for *BRCA1* (MIM: 113705), the REVEL supporting pathogenic interval had 82 variants, enough to determine whether the interval was trending discordant or trending

concordant (**Table S1, S2**). 31 of the 82 variants (37.8%) were incorrectly predicted, exceeding REVEL's supporting pathogenic incorrect prediction tolerance of 25.88%. Thus, the REVEL *BRCA1* supporting pathogenic interval was deemed trending discordant.

We applied this approach to all evidence strength intervals across 3,668 genes for REVEL and BayesDel. The REVEL calibration produced seven evidence strength intervals per gene, meaning that across the 3,668 genes, there were a total of 25,676 intervals. 23,235 (90%) of these intervals had insufficient variants to determine concordance or discordance and, of these, 14,958 (60%) had no variants (see Methods). Of the 2,441 intervals with sufficient variants for analysis, 1,783 (73%) were trending discordant and 658 (27%) were trending concordant. The BayesDel calibration produced five evidence strength intervals per gene, meaning that across the 3,668 genes, there were a total of 18,340 intervals. 15,692 (86%) of these intervals had insufficient variants, and of these, 7,794 (50%) had no variants. Of the 2,648 intervals with sufficient variants, 1,916 (72%) were trending discordant and 732 (28%) were trending concordant. Thus, most intervals in most genes lacked sufficient control variants to assess concordance or discordance, with many intervals containing no variants. Of intervals with sufficient variants, most were trending discordant with respect to the genome-wide calibrations.

Over half of all pathogenic evidence intervals with sufficient variants were trending discordant for both predictors (~65% for REVEL and ~67% for BayesDel). The proportion of discordant pathogenic intervals increased as the evidence strength increased (REVEL: supporting = 60%, moderate = 66%, strong = 83%; BayesDel: supporting = 62%, moderate = 68%, strong = 81%). Moreover, ~82% of intervals with sufficient variants were discordant across all of the benign evidence intervals for both predictors (REVEL: supporting = 80%, moderate = 87%, BayesDel: supporting = 79%, moderate = 84%; **Figure 2 A, B**). We saw similar trends when the analysis was repeated on the ClinVar 2023 dataset without training variants (**Figure S1 A, B**). In general, these findings underscore the variability in gene performance when applying the genome-wide calibrations.

Intervals that were trending discordant in our analysis were largely driven by genes with very few control variants. However, several important genes with many control variants had trending discordant intervals (**Figure 3**). Multiple intervals in these genes tended to trend discordant. For example, multiple pathogenic intervals in *MSH2* (MIM: 609309), *BRCA1*, and *BRCA2* (MIM: 600185) trended discordant for both predictors, while multiple benign intervals in *ENG* (MIM: 131195), *USH2A* (MIM: 608400), and *NF1* (MIM: 613113) trended discordant. Genes with many control variants were also more likely to trend discordant in pathogenic intervals than benign intervals, highlighting the potential to overestimate variant pathogenicity.

The benign very strong and strong evidence strength intervals for the REVEL predictor have an incorrect prediction tolerance of 0% since there were no pathogenic variants in these intervals at the time of the SVI calibration. Therefore, we were unable to evaluate these intervals using the method described above. Instead, we tallied the number of correctly and incorrectly predicted variants in these intervals to assess how the SVI calibrations performed. In the REVEL benign strong interval, there were a total of 1,378 variants across 648 genes. Of these, 1,379 variants in 645 genes were correctly predicted, and nine variants in the benign strong intervals of nine different genes were incorrectly predicted. The genome-wide calibration suggests that these intervals should be free of incorrectly predicted variants, so we deemed

them trending discordant. To ensure that pathogenic splicing variants mis-annotated as missense variants in ClinVar were not responsible for the discrepancy between REVEL predictions and control variant classifications, we analyzed all nine discrepant control variants using SpliceAI[18]. One variant with a REVEL score in the very strong benign interval but classified as pathogenic in ClinVar is *ODAD1* (MIM: 615038), NM_001364171.2(*ODAD1*):c.853G>A (p.Ala285Thr). This variant had a splice donor gain score of 0.54, indicating this variant has a high probability of affecting splicing, which could explain the discrepancy. Indeed, RNA studies from individuals homozygous for this variant revealed transcripts with aberrant splicing that resulted in premature truncation, likely causing *ODAD1* loss of function[19,20]. The other eight incorrectly predicted variants had donor/acceptor loss and donor/acceptor gain scores of less than 0.2 in SpliceAI, suggesting that these variants have a low probability of affecting splicing, and thus that splicing does not account for the discrepancy between their REVEL predictions and ClinVar classifications.

In the REVEL benign very strong interval, there were a total of 81 variants across 66 genes. Of these, 80 variants across 65 genes were correctly predicted, and one variant in one gene was incorrectly predicted. We deemed this very strong interval trending discordant. The incorrectly predicted variant had donor/acceptor loss and donor/acceptor gain scores of less than 0.2 in SpliceAI, suggesting that this variant has a low probability of affecting splicing, and thus that splicing does not account for the discrepancy between the REVEL prediction and ClinVar classification. Overall, the appearance of these pathogenic variants in the most stringent benign strong and very strong intervals underscores the shortcomings of genome-wide calibration and highlights the need for gene-specific approaches that can provide more granularity.

To better understand the clinical impact of using the genome-wide calibration, we investigated concordance across the 78 genes recommended by the ACMG for reporting secondary findings (ACMG SF). These genes are associated with disorders with significant morbidity and mortality with a high lifetime penetrance and have established medical or surgical interventions[21,22]. Of the 78 ACMG SF genes, 24 had at least one trending discordant interval for REVEL, and 31 had at least one trending discordant interval for BayesDel. In total, 73/148 (~50%) of REVEL intervals with sufficient variants and 82/165 (~50%) of BayesDel intervals with sufficient variants were trending discordant with the genome-wide calibrations (**Figure 2C, D**). However, 75/148 (~50%) of REVEL intervals and 83/165 (~50%) of BayesDel intervals were trending concordant. ~70% of intervals with sufficient variants for both REVEL and BayesDel provide pathogenic evidence (**Figure 2C, D**).

We note that our ACMG SF analysis was conducted with predictor training variants included, though we observed similar trends when training variants were excluded (**Figure S1C, D**). Since the training sets for REVEL and BayesDel included variants in the ACMG SF genes, these predictors are expected to perform well on them. Indeed, the ACMG SF genes had more trending concordant intervals than other genes, but the genome-wide calibrations still yielded inappropriate evidence strength in many cases. One example is *MLH1* (MIM:120436), where four of five REVEL evidence strength intervals were trending discordant with the genome-wide calibrations. Thus, the ACMG SF genes highlight a key limitation of the genome-wide calibration strategy, and, because variants in these genes are clinically actionable, this limitation could have serious consequences.

The distributions of predictions for individual genes reveal causes of trending discordant intervals

We next investigated trending discordant intervals to understand the origin of discordance. We identified three distinct patterns, two of which are potential causes of trending discordant intervals. In some genes, the genome-wide calibration was not discordant with any of the evidence strength intervals, such as in *CSF1R* (MIM: 164770) for REVEL and *COL7A1* (MIM: 120120) for BayesDel (**Figure 4A, B**). As previously noted, most intervals that were not trending discordant lacked sufficient variants to be deemed trending concordant. Nonetheless, for these genes, the control variants appeared to match the genome-wide predictor score distribution, suggesting that the genome-wide evidence strength calibration intervals may be correct.

In other genes, the predictions separated benign and pathogenic control variants, but the distribution of predictions was shifted to the right, meaning that the genome-wide calibration of evidence strength intervals did not capture the true distribution of pathogenic and benign variants for those genes (e.g. *MSH2*, **Figure 4C, D**). In extreme cases like *MSH2*, benign control variants were shifted so far to the right that the supporting pathogenic intervals contained far more benign control variants than pathogenic variants, suggesting that these intervals in fact would provide benign evidence. In other genes, the predicted variant effect scores separated benign and pathogenic control variants, but the distribution was shifted to the left (e.g. *ENG*, **Figure 4E, F**). In these genes, predictions in the benign evidence strength interval would actually be predictive of pathogenicity. Thus, analysis of distributions of predictions for individual genes revealed distinct patterns of discordance between control variants and the genome-wide calibration, highlighting the necessity for gene-specific recalibration in many cases. The removal of REVEL and BayesDel training variants does not change the conclusion of this analysis (**Figure S2**).

Many variants of uncertain significance receive incorrect evidence strength

Variants of uncertain significance have an unknown relationship to disease and should not be used for clinical decision-making. Our analysis suggests that the genome-wide calibration could lead to many VUS receiving inappropriate evidence strength. Reclassification of these VUS, based on an inappropriate evidence strength, has the potential to do harm. To determine the extent of this problem, we calculated the number of VUS in the 3,668 genes in the ClinVar 2023 Dataset in trending discordant intervals. For REVEL, 177,078 (42%) of VUS were in intervals with sufficient control variants. Of these, 76,449 (18% of total VUS) variants had predictions in trending concordant intervals and 100,629 (24%) had predictions in trending discordant intervals (**Figure 5A**). For BayesDel, 127,409 (35%) of VUS were in intervals with sufficient control variants. 55,481 (15% of total VUS) had predictions in trending concordant intervals and 71,928 (20%) had predictions in trending discordant intervals (**Figure 5B**). Removal of REVEL and BayesDel training variants yielded a nearly identical distribution of VUS across interval classes (**Figure S3**). Thus, as expected, the majority of VUS had predictions in evidence-strength intervals with insufficient control variants. However, more than half of the VUS with predictions in intervals with sufficient control variants were in trending discordant intervals

and thus would receive inappropriate evidence. While genome-wide calibrations are a major step forward, they could result in inappropriately high or low evidence strength for many VUS.

A web application to view genome-wide calibrations for all genes in ClinVar

We developed a web application to enable facile visualization of REVEL or BayesDel prediction distributions for all possible single nucleotide variants, individual control variants, and genome-wide calibration intervals for the 3,668 genes in our ClinVar 2023 Dataset and 3,623 genes in our ClinVar 2023 Dataset without training variants. Our web app can be accessed at https://calibration.gs.washington.edu. With variant interpretation workflows in mind, our app enables clinicians and researchers to plot a predictor score distribution for a gene of interest annotated with the evidence strength intervals from the genome-wide calibration. This functionality facilitates the interpretation of variants within each gene, presenting graphs of both the whole genome and gene-specific variant effect prediction distributions. Genome-wide calibration intervals are labeled according to their concordance or discordance, aiding users in understanding the predictive performance of these intervals. Additionally, users can download a table containing the variants used to generate the graphs, enhancing accessibility, and enabling further analysis.

Discussion

Calibration of variant effect predictors has the potential to revolutionize the use of predictor evidence in variant interpretation. The ClinGen SVI developed a calibration method enabling predictor evidence to be evaluated rigorously, demonstrating that, in principle, predictors can yield strong and moderate evidence. Their innovative method relies on aggregating control variants across genes, which enables calibration for genes with few control variants. This genome-wide calibration is, however, at odds with the ClinGen SVI group's recommendations for the calibration of functional data which requires gene-specific calibration[12]. The underlying assumption of the functional evidence calibration recommendation is that functional evidence for different genes will have different distributions, thus necessitating gene- and dataset-specific calibration. However, the genome-wide calibration of predictor evidence assumes that gene-specific differences in the variant effect prediction distributions either do not exist or do not appreciably impact evidence strength calculations. Thus, while the ClinGen SVI showed that genome-wide calibration improves evidence strength accuracy, especially for genes with few control variants, they also note that gene-specific calibration is appropriate when sufficient control variants are available[14].

To assess the degree to which the genome-wide calibration is appropriate for individual genes, we evaluated the ClinGen SVI-defined evidence strength intervals by calculating the incorrect prediction tolerance for each evidence strength interval for REVEL and BayesDel. We focused our analysis on these two predictors since they achieve the highest level of evidence in the genome-wide calibration and are the most commonly recommended by variant curation expert panels[9,14,23–26]. We found that predicted variant effect distributions were not uniform across genes and, consequently, many genes did not conform to the assumptions of the genome-wide calibration method. Because predicted variant effect distributions could be shifted either right or left, the genome-wide calibration results in variants in some genes receiving evidence that is either stronger than indicated by control variants, weaker than indicated by

control variants, or even conflicting (e.g. receiving pathogenic evidence when benign is indicated). Moreover, we show that if the genome-wide calibrations were applied to existing variants of uncertain significance, a large fraction would receive inappropriate evidence.

We found that some genes are incorrectly calibrated and have shifted variant effect predictor score distributions compared to the distribution of all possible predictor scores. Although most trending discordant intervals for these genes have very few control variants, the addition of more clinically interpreted variants over time is unlikely to resolve discordance for these genes. This is because the distribution of control variants in these genes is also shifted when compared to the distribution of all control variants used for calibration. This effect is illustrated by genes like *MSH2* where the distribution of all scores is right-shifted and moves benign predictions to the right, and *ENG* where most predictions are left-shifted and moves the pathogenic distribution to the left. For this reason, we suggest that variant curation expert panels and variant review scientists use our web application (https://calibration.gs.washington.edu) to examine the distribution of all REVEL or BayesDel predictions for compatibility with genome-wide calibration before using either predictor for variant classification.

Further, we found that some genes with shifted distributions have many control variants and can be recalibrated individually. For example, *MSH2* is a gene with many control variants but is trending discordant in all pathogenic evidence intervals for both REVEL and BayesDel. However, *MSH2* has sufficient control variants to enable gene-specific calibration. Therefore, for cases like *MSH2*, which trend discordant in a single direction and have sufficient control variants, we suggest individual gene calibration. Other genes that do not align with evidence strength intervals assigned by the genome-wide calibration include *BRCA1* and *BRCA2* (https://calibration.gs.washington.edu). Indeed, the hereditary breast, ovarian, and pancreatic cancer variant curation expert panel also made this observation for their recommended predictor, BayesDel. They noted that scores are shifted to the right, which would lead to most known benign variants being classified as indeterminate. Thus, they recalibrated the BayesDel predictor for these genes, assigning BayesDel predictions moderate evidence strength in their recent guidance[27].

The main limitation of our analysis of gene-level concordance with genome-wide calibration is the paucity of control pathogenic and benign variants for most genes. To account for this, we used the binomial theorem to define the minimum number of control variants in each interval to assess individual gene concordance with genome-wide calibration. Following this framework, we found that the majority of evidence strength intervals that had sufficient control variants were trending discordant with genome-wide calibration. This finding was largely driven by genes with very few control variants since the number of variants required to be trending discordant was far less than the number required to be trending concordant. The addition of more correctly predicted control variants as new pathogenic and benign assessments are deposited into ClinVar could lead to trending discordant intervals becoming trending concordant, but this outcome is unlikely. For example, *GJB3* (MIM: 603324) is trending discordant in the REVEL supporting and moderate pathogenic intervals. Both intervals contain three control variants, with two of the three incorrectly predicted. To become trending concordant for the supporting interval, an additional five correctly predicted pathogenic control variants and no incorrectly predicted variants would be required. To resolve the *GJB3* moderate pathogenic

interval, an additional 16 correctly predicted pathogenic control variants would be required. Although it is possible that the intervals in *GJB3* are correct and all new control variants would be correctly predicted, such an outcome is unlikely given the current number of incorrectly predicted variants. For these reasons, we conclude most discordant intervals will not resolve with the addition of new control variants.

Because we showed that ~35% of genes had at least one trending discordant interval, our findings have considerable clinical implications. Variants in these genes may be given an inappropriate evidence strength. While variant effect predictions comprise just one line of evidence used in clinical variant classification, inappropriate evidence strength could contribute to inappropriate variant classification. For genes like *MSH2*, where the distribution of control benign variants was shifted to the right, the majority of control variants in the pathogenic supporting interval were benign. *MSH2* variants with scores in this interval would receive inappropriate evidence for pathogenicity and possibly inappropriate classifications. Our gene-specific analysis highlighted the scale of this problem and identified genes that are not compatible with genome-wide calibration. Thus, we suggest a cautious approach when using genome-wide calibrations of predictor scores and, at minimum, validating that individual gene distributions mirror the genome-wide distribution used for calibration.

Our analysis revealed that gene-specific differences in predictor performance reduce the efficacy of the genome-wide predictor calibration approach. Addressing this shortcoming will require taking gene-specific differences into consideration. Indeed, predictors have emerged that leverage human variant frequency information in a gene-specific fashion. For example, popEVE transforms scores based on gene-specific constraints reflected in variant frequencies[28], which may make them better-suited for genome-wide evidence strength calibration. However, our results strongly suggest that future genome-wide calibrations should, at minimum, include gene-specific evaluation. Moreover, gene-specific predictor calibration approaches should be developed to reduce gene-specific effects. Genes that already have sufficient control variants should be individually calibrated, as has already been done by some variant curation expert panels[23,27,29]. In genes lacking sufficient control variants, a potential solution is to include control variants from other sources, such as biobanks. For example, the TP53 variant curation expert panel used variants found in unaffected individuals at comparatively high frequency and never found in an individual with cancer as presumptive benign missense controls when calibrating functional assays[23]. A complementary approach would be to use the calibration strategy recently developed for functional data[12] where evidence strength is computed for just two intervals, one benign and one pathogenic. This approach aggregates control variants over larger score ranges, gaining power to compute evidence strength but assigning the same strength to divergent scores. Thus, the ClinGen SVI genome-wide predictor calibration approach is a massive step forward, and the next step is development of methods that enable accurate, gene-specific predictor calibration.

Figure legends:

Figure 1 Gene-specific analysis workflow

Histograms depict REVEL and BayesDel predictions for all possible single nucleotide variants and control pathogenic and benign variants from all genes in the ClinGen SVI Calibration dataset. Dashed vertical lines indicate the genome-wide evidence strength intervals (A, B). The

equation used to calculate the number of incorrectly predicted variants in each evidence strength interval is shown (C). The incorrect prediction tolerance for each evidence strength interval from the ClinGen SVI Calibration Dataset, which is the maximum percentage of incorrectly predicted variants allowed in each evidence strength interval, is shown (D). An example gene illustrating three different results of our analysis: an evidence strength interval trending concordant, an evidence strength interval trending discordant, or an evidence strength interval having an insufficient number of variants for analysis (E).

Figure 2 Analysis of evidence strength interval concordance with the genome-wide calibration
We analyzed predictions for each evidence strength interval in different sets of genes in the ClinVar 2023 dataset and determined whether each interval had too few variants for analysis or, whether those with sufficient variants were trending concordant or discordant. A stacked bar graph depicts our evidence strength interval class for all disease-relevant genes for the REVEL predictor (A); all disease-relevant genes for the BayesDel predictor (B); ACMG secondary findings genes for the REVEL predictor (C) and ACMG secondary findings genes for the BayesDel predictor (D). The number of evidence strength intervals in each class is also shown.

Figure 3 Some trending discordant intervals arise from genes with many control variants
The relationships between the number of variants in either REVEL (A, all pathogenic intervals; B, all benign intervals) or BayesDel (C, all pathogenic intervals; D, all benign intervals) evidence strength intervals and the percent of incorrectly predicted variants in each interval are shown as scatterplots. Each dot represents an evidence strength interval in a particular gene colored by evidence strength. Horizontal dashed lines indicate the incorrect prediction tolerance as defined by the genome-wide calibration for each interval, colored by evidence strength. Dots above the horizontal dashed line of the same color represent trending discordant intervals. All trending concordant intervals are colored gray and intervals with insufficient variants are not shown. Select trending discordant intervals with many control variants are labeled by gene in each plot. Since each plot represents either all pathogenic intervals or all benign intervals for the indicated predictor, each gene is represented multiple times per plot.

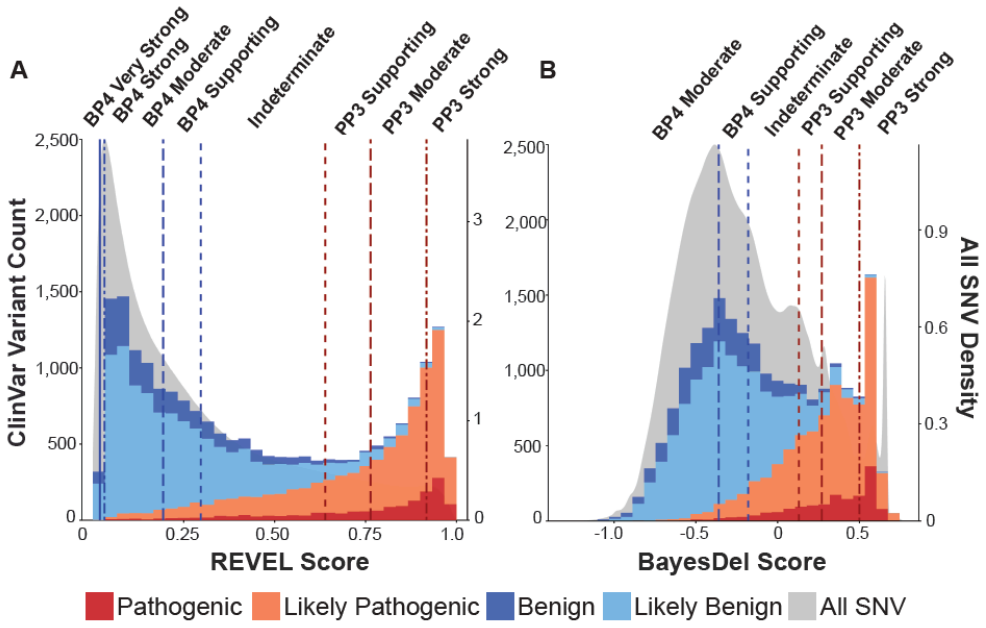
Figure 4 Gene-specific predictor score distributions reveal causes of discordance
Gene-specific histograms depict REVEL and BayesDel predictor scores for all possible single nucleotide variants and control pathogenic and benign variants from the Clinvar 2023 dataset. Genes were chosen to illustrate genome-wide calibration alignment with gene-specific distributions with no discordance (A, B); misalignment of genome-wide calibrations with gene-specific distributions owing to a right-shift (C, D); or misalignment of genome-wide calibrations with gene-specific distributions owing to a left-shift. Dashed vertical lines indicate the genome-wide evidence strength intervals articulated by the ClinGen SVI Working Group. REVEL BP4 strong and very strong evidence strength intervals are not depicted.

Figure 5 Many ClinVar variants of uncertain significance are in trending discordant intervals

Donut plots depict the number of one-star, missense VUS from the ClinVar 2023 VUS Dataset that were in trending concordant, trending discordant, or insufficient variants intervals across all 3,668 disease-relevant genes analyzed for REVEL (A) and BayesDel (B).

Figures

Figure 1:



C

$$\text{incorrect prediction percentage} = \frac{\text{number of incorrectly predicted variants per evidence strength interval}}{\text{total number of variants per evidence strength interval}} \times 100$$

D Gene-specific incorrect prediction percentage for all evidence strength intervals

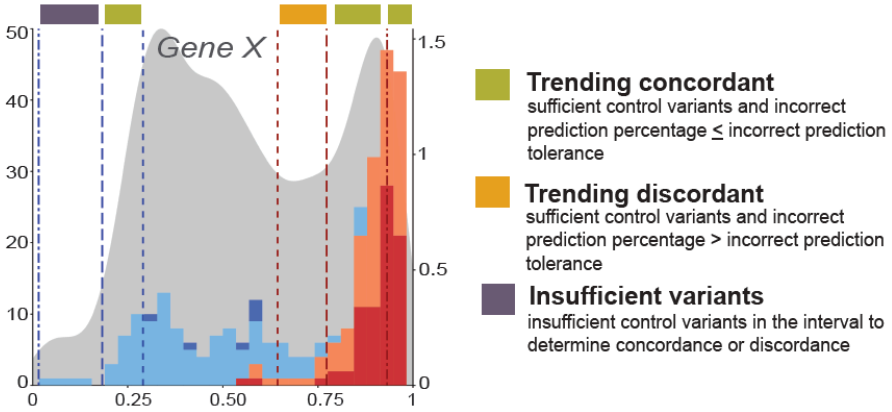


Figure 2:

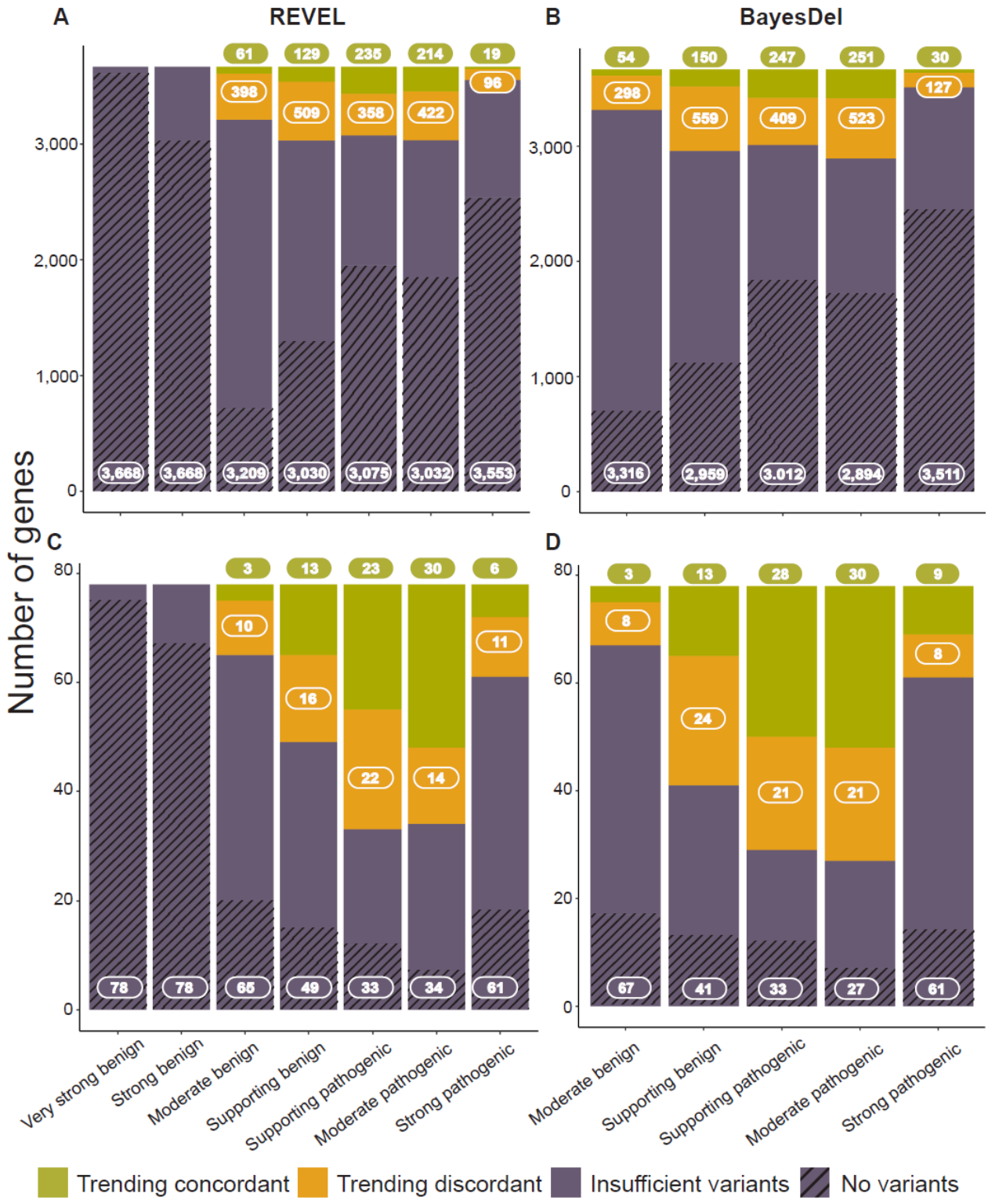

A
REVEL
B
BayesDel
C
D
Number of genes
3,000
2,000
1,000
0
80
60
40
20
61
129
235
214
19
398
509
358
422
96
3,668
3,668
3,209
3,030
3,075
3,032
3,553
54
150
247
251
30
298
559
409
523
127
3,316
2,959
3,012
2,894
3,511
3
13
23
30
6
10
16
22
14
11
78
78
65
49
33
34
61
3
13
28
30
9
8
24
21
21
8
67
41
33
27
61
Very strong benign
Strong benign
Moderate benign
Supporting benign
Supporting pathogenic
Moderate pathogenic
Strong pathogenic
Moderate benign
Supporting benign
Supporting pathogenic
Moderate pathogenic
Strong pathogenic
Trending concordant
Trending discordant
Insufficient variants
No variants

Figure 3:

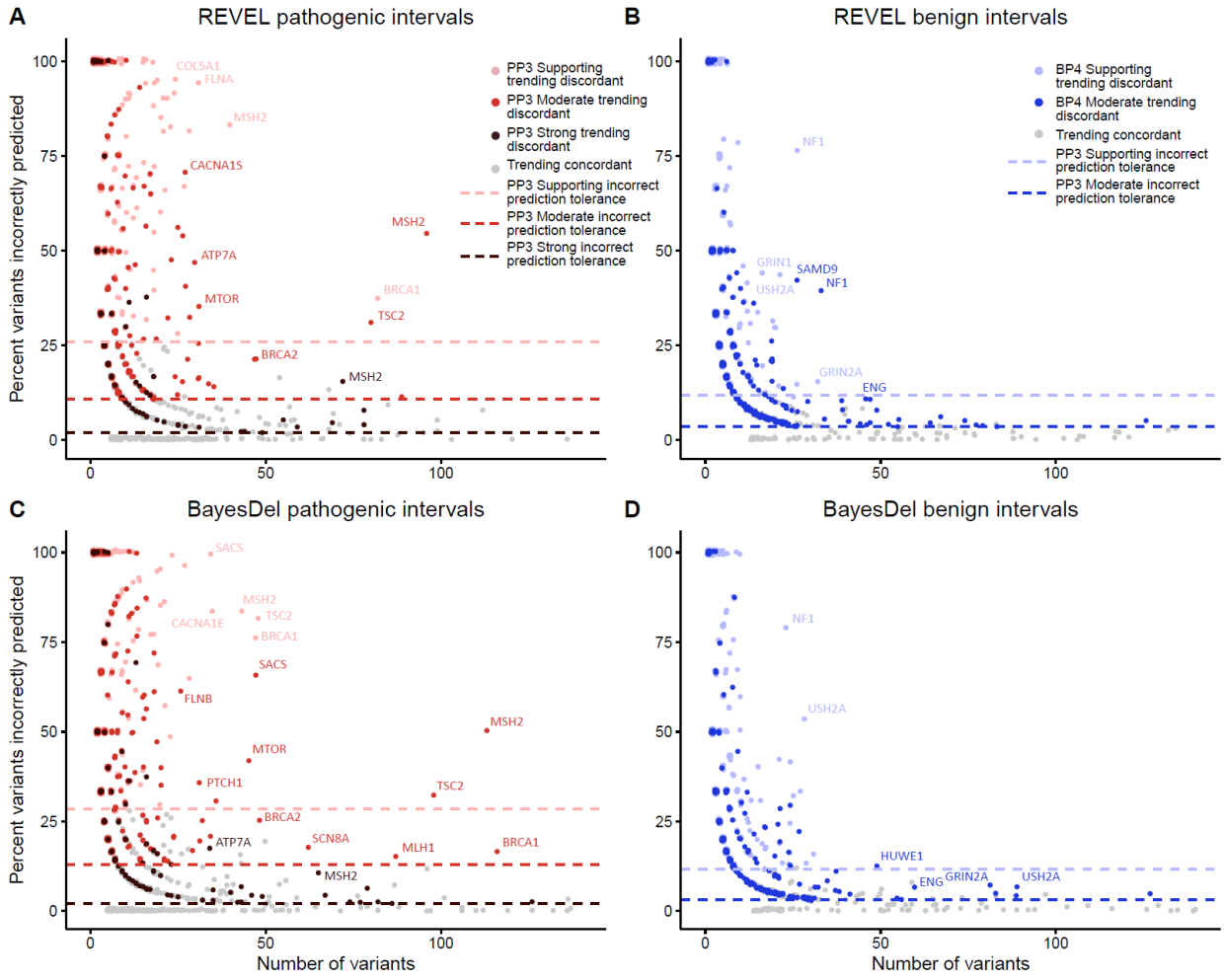

A
REVEL pathogenic intervals
B
REVEL benign intervals
C
BayesDel pathogenic intervals
D
BayesDel benign intervals
Percent variants incorrectly predicted
Number of variants
PP3 Supporting trending discordant
PP3 Moderate trending discordant
PP3 Strong trending discordant
Trending concordant
PP3 Supporting incorrect prediction tolerance
PP3 Moderate incorrect prediction tolerance
PP3 Strong incorrect prediction tolerance
BP4 Supporting trending discordant
BP4 Moderate trending discordant
Trending concordant
PP3 Supporting incorrect prediction tolerance
PP3 Moderate incorrect prediction tolerance
COL5A1
FLNA
MSH2
CACNA1S
ATP7A
MTOR
BRCA2
TSC2
BRCA1
NF1
GRIN1
SAMD9
USH2A
GRIN2A
ENG
SACS
CACNA1E
FLNB
PTCH1
SCN8A
MLH1
HUWE1

Figure 4:

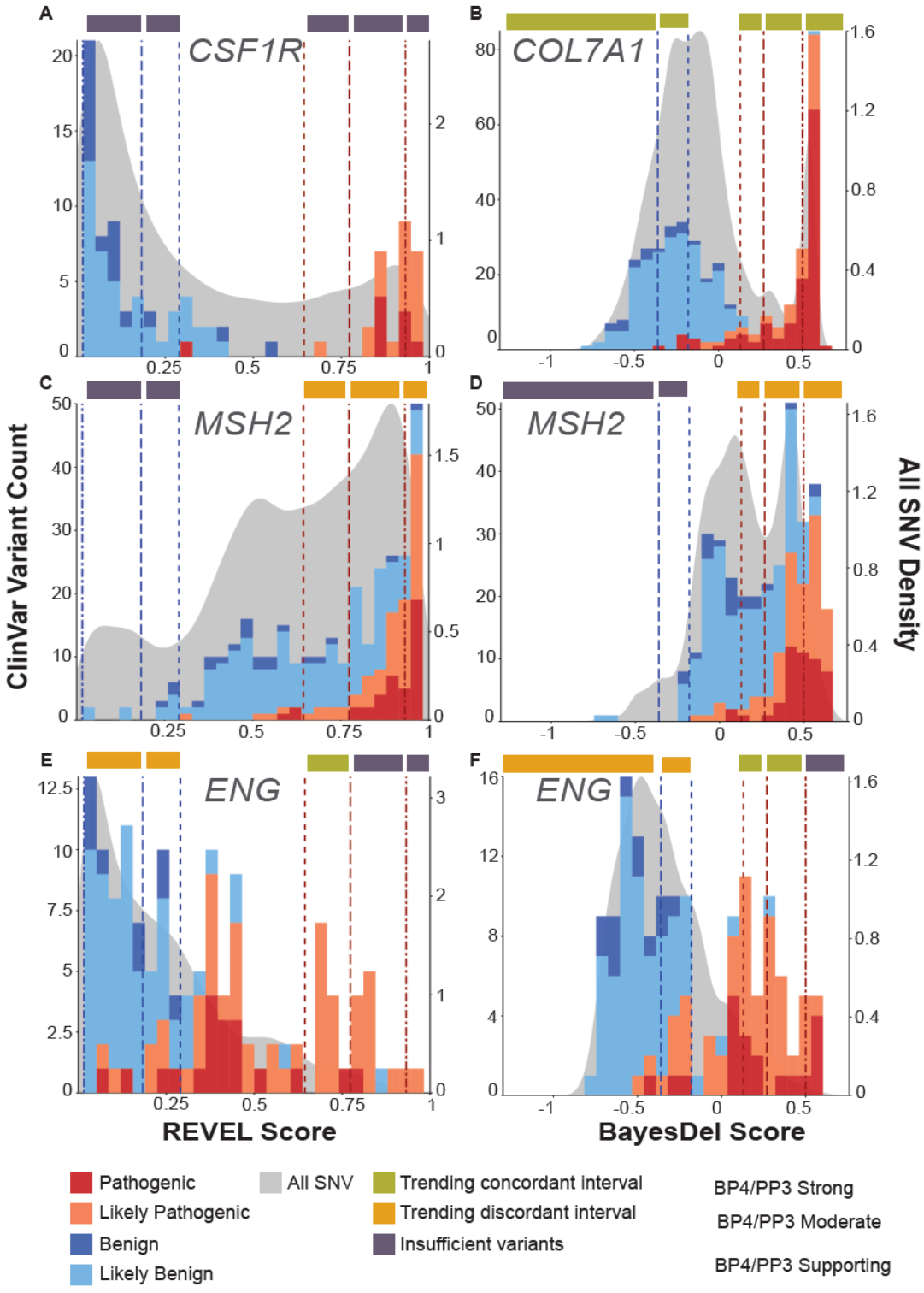

A
CSF1R
B
COL7A1
C
MSH2
D
MSH2
E
ENG
F
ENG
ClinVar Variant Count
All SNV Density
REVEL Score
BayesDel Score
Pathogenic
Likely Pathogenic
Benign
Likely Benign
All SNV
Trending concordant interval
Trending discordant interval
Insufficient variants
BP4/PP3 Strong
BP4/PP3 Moderate
BP4/PP3 Supporting

Figure 5:

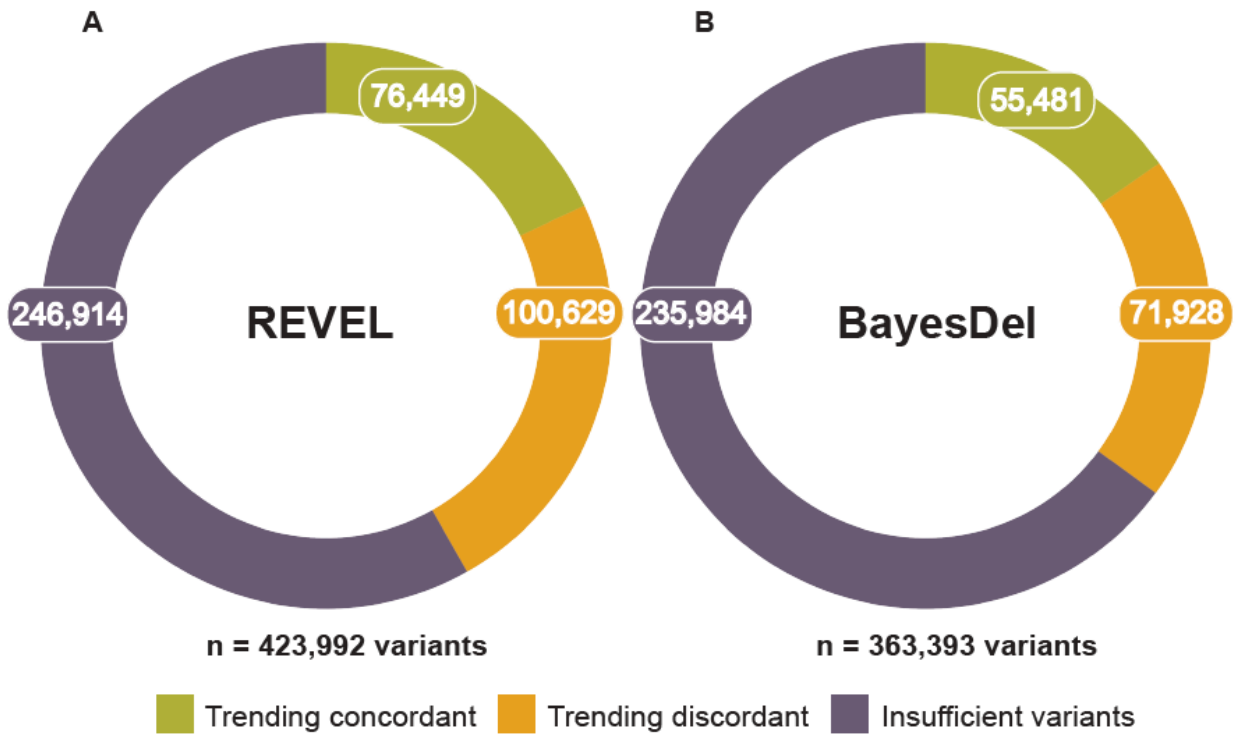


Table 1: Incorrect prediction tolerance per evidence strength interval for REVEL and BayesDel

Evidence strength interval	BP4 Very Strong/Strong	BP4 Moderate	BP4 Supporting	PP3 Supporting	PP3 Moderate	PP3 Strong
REVEL	0%	3.60%	11.76%	25.88%	10.75%	1.93%
BayesDel	N/A	3.17%	11.63%	28.52%	12.86%	2.02%

Table 2: Number of variants needed per evidence strength interval for concordance

Evidence strength interval	BP4 Very Strong/Strong	BP4 Moderate	BP4 Supporting	PP3 Supporting	PP3 Moderate	PP3 Strong
REVEL	N/A	1	1	2	1	1
BayesDel	N/A	1	1	2	1	1

Table 3: Number of variants needed per evidence strength interval for discordance

Evidence strength interval	BP4 Very Strong/Strong	BP4 Moderate	BP4 Supporting	PP3 Supporting	PP3 Moderate	PP3 Strong
REVEL	N/A	44	13	6	15	83
BayesDel	N/A	50	14	5	12	79

Acknowledgments We would like to thank members of the Fowler and Starita labs, Dr. Andrew B. Stergachis, and Dr. Brian H. Shirts for helpful feedback. This work was supported by the NHGRI Center for Multiplexed Assessment of Phenotype RM1HG010461 (DMF, LS, AEM, SF, MT), Center for Actionable Variant Analysis UM1HG011969 (DMF, LS, SF, MT), and R01HG013025 (DMF, LS, AEM), the NIH/NIGMS Medical Genetics Postdoctoral Training Grant 5T32GM007454 (AEM), the NIH/NHGRI Genome Training Grant T32-HG000035-25 (SF), the Early Career Award Alex's Lemonade Stand for Childhood Cancer and RUNX1 foundation 21-25037 (AEM), the Brotman Baty Institute Catalytic Collaborations Grant CC28 (AEM) and Catalytic Collaboration award (SF).

Data and code availability

Supplemental files contain all necessary datasets to reproduce the analysis and figures. Large datasets are hosted on Zenodo at https://zenodo.org/records/11256843. Code and additional files can be found at https://github.com/FowlerLab/VEP-calibrations.

References

1. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

2. Chen, E., Facio, F.M., Aradhya, K.W., Rojahn, S., Hatchell, K.E., Aguilar, S., Ouyang, K., Saitta, S., Hanson-Kwan, A.K., Capurro, N.N., et al. (2023). Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. JAMA Netw Open *6*, e2339571.

3. Fayer, S., Horton, C., Dines, J.N., Rubin, A.F., Richardson, M.E., McGoldrick, K., Hernandez, F., Pesaran, T., Karam, R., Shirts, B.H., et al. (2021). Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. Am. J. Hum. Genet. *108*, 2248–2258.

4. Horton, C., Hoang, L., Zimmermann, H., Young, C., Grzybowski, J., Durda, K., Vuong, H., Burks, D., Cass, A., LaDuca, H., et al. (2024). Diagnostic Outcomes of Concurrent DNA and RNA Sequencing in Individuals Undergoing Hereditary Cancer Testing. JAMA Oncol *10*, 212–219.

5. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J.,

Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

6. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95.

7. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

8. Wu, Y., Li, R., Sun, S., Weile, J., and Roth, F.P. (2021). Improved pathogenicity prediction for rare human missense variants. Am. J. Hum. Genet. *108*, 1891–1906.

9. Tian, Y., Pesaran, T., Chamberlin, A., Fenwick, R.B., Li, S., Gau, C.-L., Chao, E.C., Lu, H.-M., Black, M.H., and Qian, D. (2019). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci. Rep. *9*, 12752.

10. Feng, B.-J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization. Hum. Mutat. *38*, 243–251.

11. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492.

12. Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M., et al. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. *12*, 3.

13. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M., Biesecker, L.G., and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet. Med. *20*, 1054–1060.

14. Pejaver, V., Byrne, A.B., Feng, B.-J., Pagel, K.A., Mooney, S.D., Karchin, R., O'Donnell-Luria, A., Harrison, S.M., Tavtigian, S.V., Greenblatt, M.S., et al. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am. J. Hum. Genet. *109*, 2163–2177.

15. Fortuno, C., James, P.A., Young, E.L., Feng, B., Olivier, M., Pesaran, T., Tavtigian, S.V., and Spurdle, A.B. (2018). Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants. Hum. Mutat. *39*, 1061–1069.

16. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

17. Stenton, S.L., Pejaver, V., Bergquist, T., Biesecker, L.G., Byrne, A.B., Nadeau, E., Greenblatt, M.S., Harrison, S., Tavtigian, S., Radivojac, P., et al. (2024). Assessment of the evidence yield for the calibrated PP3/BP4 computational recommendations. https://doi.org/10.1101/2024.03.05.24303807

18. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535–548.e24.

19. Onoufriadis, A., Paff, T., Antony, D., Shoemark, A., Micha, D., Kuyt, B., Schmidts, M., Petridi, S., Dankert-Roelse, J.E., Haarman, E.G., et al. (2013). Splice-site mutations in the axonemal outer dynein arm docking complex gene CCDC114 cause primary ciliary dyskinesia. Am. J. Hum. Genet. *92*, 88–98.

20. Knowles, M.R., Leigh, M.W., Ostrowski, L.E., Huang, L., Carson, J.L., Hazucha, M.J., Yin, W., Berg, J.S., Davis, S.D., Dell, S.D., et al. (2013). Exome sequencing identifies mutations in CCDC114 as a cause of primary ciliary dyskinesia. Am. J. Hum. Genet. *92*, 99–106.

21. Miller, D.T., Lee, K., Gordon, A.S., Amendola, L.M., Adelman, K., Bale, S.J., Chung, W.K., Gollob, M.H., Harrison, S.M., Herman, G.E., et al. (2021). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *23*, 1391–1398.

22. Miller, D.T., Lee, K., Abul-Husn, N.S., Amendola, L.M., Brothers, K., Chung, W.K., Gollob, M.H., Gordon, A.S., Harrison, S.M., Hershberger, R.E., et al. (2023). ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *25*, 100866.

23. Fortuno, C., Lee, K., Olivier, M., Pesaran, T., Mai, P.L., de Andrade, K.C., Attardi, L.D., Crowley, S., Evans, D.G., Feng, B.-J., et al. (2021). Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. Hum. Mutat. *42*, 223–236.

24. Chora, J.R., Iacocca, M.A., Tichý, L., Wand, H., Kurtz, C.L., Zimmermann, H., Leon, A., Williams, M., Humphries, S.E., Hooper, A.J., et al. (2022). The Clinical Genome Resource (ClinGen) Familial Hypercholesterolemia Variant Curation Expert Panel consensus guidelines for LDLR variant classification. Genet. Med. *24*, 293–306.

25. Wu, D., Luo, X., Feurstein, S., Kesserwan, C., Mohan, S., Pineda-Alvarez, D.E., Godley, L.A., and collaborative group of the American Society of Hematology - Clinical Genome Resource Myeloid Malignancy Variant Curation Expert Panel (2020). How I curate: applying American Society of Hematology-Clinical Genome Resource Myeloid Malignancy Variant Curation Expert Panel rules for RUNX1 variant curation for germline predisposition to myeloid malignancies. Haematologica *105*, 870–887.

26. Johnston, J.J., Dirksen, R.T., Girard, T., Gonsalves, S.G., Hopkins, P.M., Riazi, S., Saddic,

L.A., Sambuughin, N., Saxena, R., Stowell, K., et al. (2021). Variant curation expert panel recommendations for RYR1 pathogenicity classifications in malignant hyperthermia susceptibility. Genet. Med. *23*, 1288–1295.

27. Parsons, M.T., de la Hoya, M., Richardson, M.E., Tudini, E., Anderson, M., Berkofsky-Fessler, W., Caputo, S.M., Chan, R.C., Cline, M.C., Feng, B.-J., et al. (2024). Evidence-based recommendations for gene-specific ACMG/AMP variant classification from the ClinGen ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel. https://doi.org/10.1101/2024.01.22.24301588

28. Orenbuch, R., Kollasch, A.W., Spinner, H.D., Shearer, C.A., Hopf, T.A., Franceschi, D., Dias, M., Frazer, J., and Marks, D.S. (2023). Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. medRxiv. https://doi.org/10.1101/2023.11.27.23299062

29. Plazzer, J.P., Macrae, F., Yin, X., Thompson, B.A., Farrington, S.M., Currie, L., Lagerstedt-Robinson, K., Frederiksen, J.H., van Overeem Hansen, T., Graversen, L., et al. (2024). Mismatch repair gene specifications to the ACMG/AMP classification criteria: Consensus recommendations from the InSiGHT ClinGen Hereditary Colorectal Cancer / Polyposis Variant Curation Expert Panel. https://doi.org/10.1101/2024.05.13.24307108

Chapter 4 addendum: Clustering variant effect predictor score distributions to improve accuracy of calibration

Introduction

We demonstrated that calibration of variant effect predictors (VEPs) using the aggregate of all pathogenic and benign variants in ClinVar is problematic. Since genes encode proteins with vastly different structures that are expressed in many different contexts, it is reasonable to expect that a VEP score could have different consequences for different genes. We evaluated evidence strength intervals from Pejaver et al. [1] and identified that on the individual gene level, evidence strength is inappropriate for 73% of evidence strength intervals. As a potential consequence, variants that are interpreted with these thresholds may be incorrectly interpreted as likely pathogenic or likely benign. One possible solution described by Pejaver et al. is individual gene calibration where gene specific thresholds can be derived. However, most genes have too few ClinVar pathogenic or benign variants to be calibrated individually. Thus, we developed a new method for calibrating VEP data based on clustering protein domains on VEP score distributions. This method ensures that calibration of cluster data is based on functional elements with similar VEP score distributions. Thus, aggregation of ClinVar variants within clusters allows for more accurate calibration, while still increasing the total number of variants available as compared to individual gene calibration. We applied this method to AlphaMissense [2] since the model is untrained on clinical labels, allowing for more robust calibration on all ClinVar variants.

Methods

Filtering AlphaMissese data

Alphamissese data was filtered to contain only predictions from curated human disease genes with moderate or above evidence from the GenCC database [3]. Variants were then filtered out of the AlphaMissese score sets if they had a SpliceAI score above 0.2 for any of the predicted splice alterations [4]. After likely splice altering variants were filtered out we limited our analysis to genes where missense variants are known to cause disease by filtering out any gene with zero pathogenic variants in ClinVar [2,5]. The filtered AlphaMissese scores were then mapped to Pfam domains [6].

Clustering AlphaMissense score distributions

Alphamissese score distributions for all predictions for gene-pfam pairs were clustered on Jensen-Shannon divergence. The optimal clustering distance was determined with K-means clustering where the optimal number of clusters was the point at which the within cluster sum of square errors only minimally decreases as more cluster are added.

Local calibration of clusters

We calibrated Alphamisses score distribution clusters as in Pejaver et al.[1]. Briefly, a posterior is calculated for all possible sliding windows fo ClinVar pathogenic/like pathognenic and

benign/likely benign variants in a given cluster. We use the precomputed prior estimation from Pejaver et al. to map calibration curves to AMCG evidence strength intervals.

Results

Clustering AlphaMissese distributions on protein domains defines optimal calibration clusters

We chose to focus our calibration efforts on Pfam domains because we found score distributions of Pfam domains within genes to be significantly divergent. Thus, Pfam domains were likely to be a driving factor in the diversity of predictor shape distributions. Compared to predictions outside of Pfam domains, those within Pfam domains tended to have AlphaMissense scores closer to 1, and were more likely to be predicted deleterious (**Figure 1A, C**). Additionally, the majority of pathogenic missense variants occur within Pfam domains (**Figure 1B, D**). We constructed an all by all heatmap of gene Pfam pairs clustered by Jesnen-Shannon distance to identify clusters of domains with similar AlphaMissese score distributions. We then optimized the number of clusters using K-means clustering to minimize the amount of within cluster sum of square error, resulting in 27 clusters for calibration (**Figure 1F**). Score distributions for ClinVar variants within each cluster closely mirror the full score distributions used for clustering (**Figure 2**). Additionally, when the full score distribution of predictions is left skewed, the pathogenic variants in the tail tend to also be left shifted as in cluster 20. Conversely, when score distributions are right skewed, the benign variants in the cluster also tend to be right skewed as in cluster 10. Since these skewed clusters tend to be dominated by benign or pathogenic variants that are tightly distributed toward the extremes for the score range, the remainder of the score range becomes evidence for the opposite class which fill the tail of the score distribution. For example, in cluster 10, an AlphaMissese score of 0.7 corresponds to benign evidence for the small proportion of variants that have scores as low as 0.7 in that cluster. In contrast, a score of 0.7 in a more bimodally distributed cluster typically represents pathogenic evidence.

Local calibration of clusters defines diverse cluster specific PP3/BP4 thresholds

We applied sliding window local calibration to the 27 optimally derived clusters based on JSD. 16 of the 27 cluster had sufficient variants for calibration. Even though calibration was not possible for 11 clusters, the 16 that were calibrated represented over 98% of the AlphaMissense predictions from clustering data. We show that bimodally distributed clusters tend to produce balance calibration curves for both the benign and pathogenic evidence regions of the score distribtoin (**Figure 3A-C**). Right-skewed distributions tended to have more pathogenic varints which were also heavily right skewed and benign variants that filled the remainder of the score distribution. For these reasons, local calibration yielded extremely right shifted pathogenic evidence and broadly right shifted benign evidence (**Figure 3D-F**). Finally, left skewed distributions tended to have left shifted benign and pathogenic evidence windows (**Figure 3G-I**).

Discussion

Genome wide calibration of VEPs has revolutionized how predictor evidence is used in clinical interpretation. By aggregating all ClinVar variants, sliding window local calibration can

identify predictor score thresholds corresponding to different strengths of evidence. Additionally, this calibration can be performed on any gene, even those with little to no ClinVar variants. However, we demonstrated that genome wide predictor calibration is inappropriate for up to 73% of genes [7]. We showed that the reason calibration was inappropriate for some genes was not because predictors performed poorly, but rather becuase underlying predictor score distributions varied widely across genes. Thus, we developed a new method for predictor calibration that clusters predictor score distributions for Pfam domains and performs local, sliding window calibration on those clusters.

Cluster calibration solves two important problems in the calibration and use of VEP data in variant interpretation. First, the problem of too few ClinVar variants per gene is solved by aggregating the ClinVar variants within clusters. Since clusters have similar underlying score distributions, the aggregation of ClinVar variants within clusters is less likely to have varied performance across genes within a cluster vs. the whole genome. Second, using Pfam domains as the unit for clustering resulted in more accurate calibration suggesting that the driving factor in variable AlphaMisses scores distributions is protein domain architecture across genes. By clustering on domains we captured this diversity and appropriately applied evidence across dominas.

Looking forward, this calibration method should be rigorously validated against known gold standard ClinGen Variant Curation Expert Panel (VCEP) interpretations as well as unbiased biobank datasets. Improved calibration should lead to likelihood ratios that more closely correlate with known interpretations and more closely correlate with patient phenotypes in large scale biobank studies. Since most variants fall into clusters with many ClinVar variants, a significant proportion of variants may receive greater than supporting evidence for variant interpretation. Application of this clustering approach along with scaling up novel MAVE approaches should result in a substantial reduction in the uncertainty currently plaguing clinical genetic testing.

Figure Legends:

Figure 1: Clustering AlphaMissense distributions on Jensen-Shannon divergence. A) Distribution of all AlphaMissense score for the all regions annotated within Pfam domains. **B**) Distribution of AlphaMissense scores for all 1+ star ClinVar variants within Pfam domains. **C**) Distribution of all AlphaMissense scores for the all regions outside of annotated Pfam domains. **D**) Distribution of AlphaMissense scores for all 1+ star ClinVar variants outside of annotated Pfam domains. **E**) All by all heatmap of gene-Pfam pairs colored by Jensen-Shannon divergence (JSD). Darker colors have JSD closer to zero and thus, have similar shapes. Lighter colors have JSD values closer to one and have more divergent distribution shapes. **F**) K-means clustering elbow plot for determining optimal clustering.

Figure 2: Optimally clustered distributions of the top 16 clusters by number of variants. Gray plots represent score distributions for all predictions within a cluster and, and the red and blue plots below represent the ClinVar variant distributions for the same cluster.

Figure 3: Local calibration of optimal JSD clusters. A-C) calibration of cluster 18, the cluster with the larger number of variants. Left plot represents benign evidence calibration, middle plot

depicts the ClinVar distribution in red and blue and the density plot for all variants in gray, and the right plot represents the pathogenic evidence calibration. **D-F**) Calibration of the significantly right shifted cluster 10. **G-I**) Calibration of the left shifted cluster 21 demonstrating how clustering on score distributions influences evidence thresholding.

Figures

Figure 1:

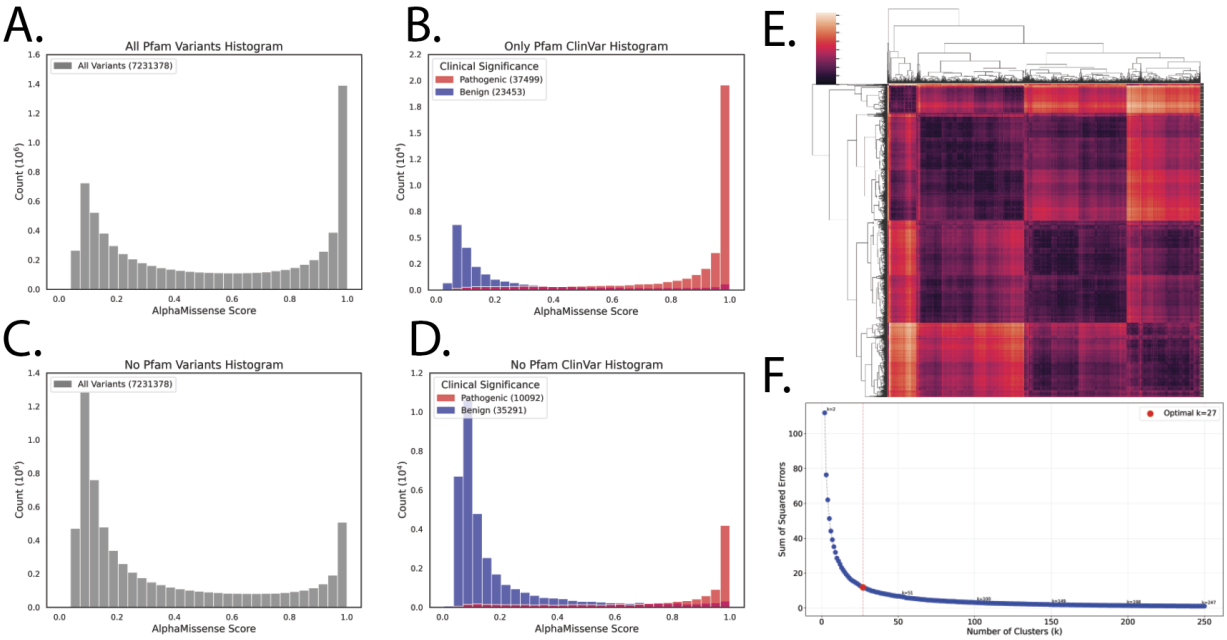


Figure 2:

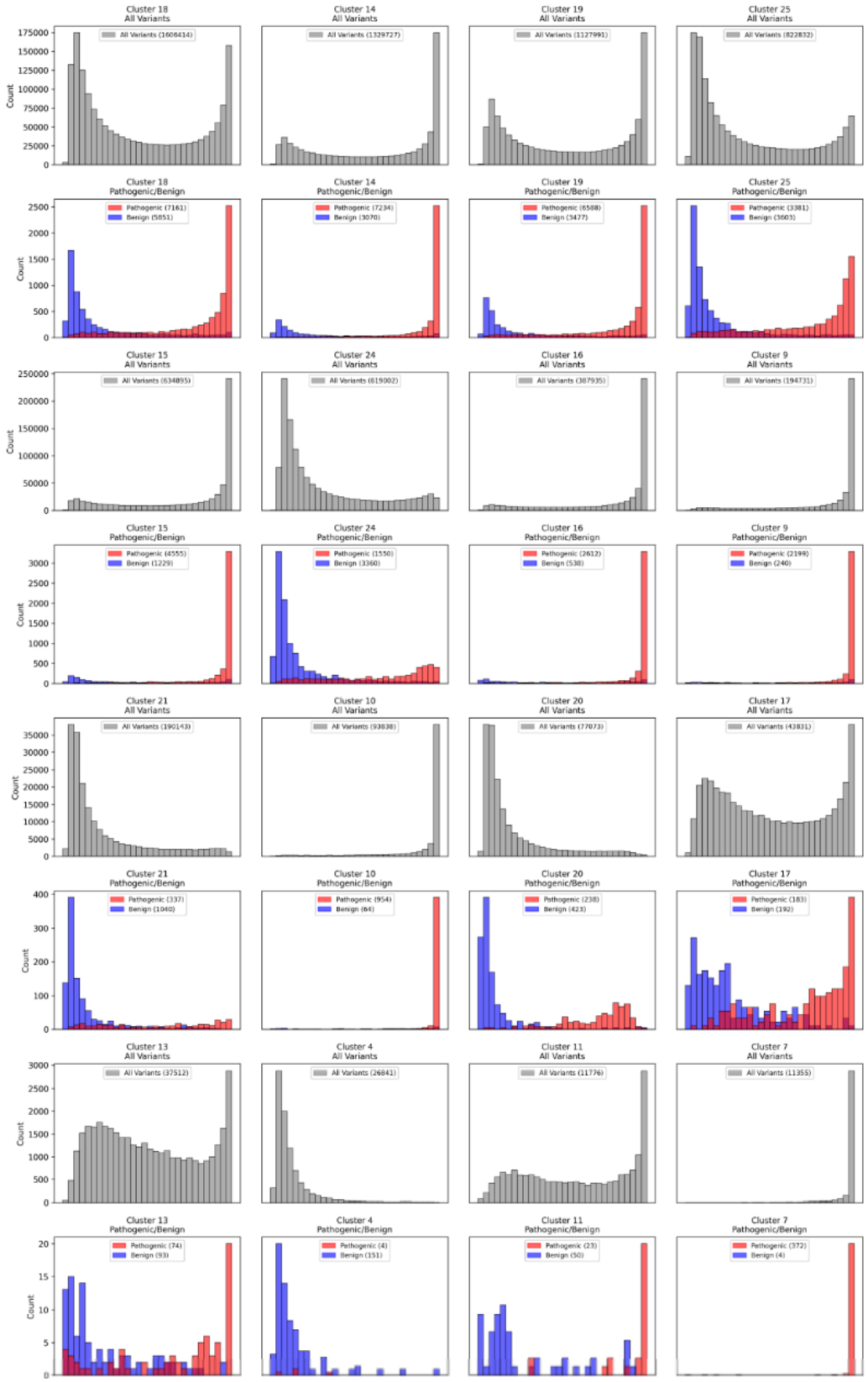
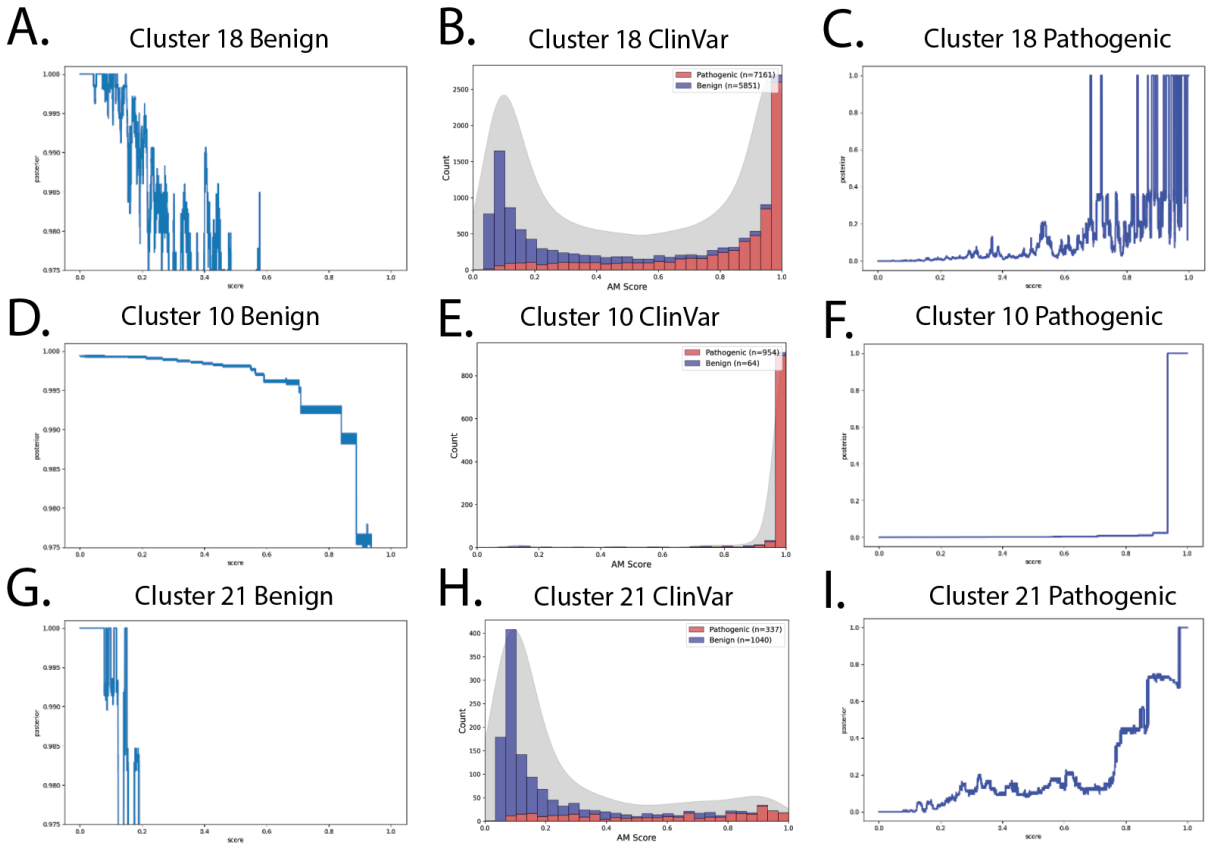

Cluster 18
All Variants
All Variants (1606414)
Cluster 14
All Variants
All Variants (1329727)
Cluster 19
All Variants
All Variants (1127991)
Cluster 25
All Variants
All Variants (822832)
Count
Cluster 18
Pathogenic/Benign
Pathogenic (7161)
Benign (5851)
Cluster 14
Pathogenic/Benign
Pathogenic (7234)
Benign (3070)
Cluster 19
Pathogenic/Benign
Pathogenic (6588)
Benign (3477)
Cluster 25
Pathogenic/Benign
Pathogenic (3381)
Benign (3603)
Cluster 15
All Variants
All Variants (634895)
Cluster 24
All Variants
All Variants (619002)
Cluster 16
All Variants
All Variants (387935)
Cluster 9
All Variants
All Variants (194731)
Cluster 15
Pathogenic/Benign
Pathogenic (4555)
Benign (1229)
Cluster 24
Pathogenic/Benign
Pathogenic (1550)
Benign (3360)
Cluster 16
Pathogenic/Benign
Pathogenic (2612)
Benign (538)
Cluster 9
Pathogenic/Benign
Pathogenic (2199)
Benign (240)
Cluster 21
All Variants
All Variants (190143)
Cluster 10
All Variants
All Variants (93838)
Cluster 20
All Variants
All Variants (77073)
Cluster 17
All Variants
All Variants (43831)
Cluster 21
Pathogenic/Benign
Pathogenic (337)
Benign (1040)
Cluster 10
Pathogenic/Benign
Pathogenic (954)
Benign (64)
Cluster 20
Pathogenic/Benign
Pathogenic (238)
Benign (423)
Cluster 17
Pathogenic/Benign
Pathogenic (183)
Benign (192)
Cluster 13
All Variants
All Variants (37512)
Cluster 4
All Variants
All Variants (26841)
Cluster 11
All Variants
All Variants (11776)
Cluster 7
All Variants
All Variants (11355)
Cluster 13
Pathogenic/Benign
Pathogenic (74)
Benign (93)
Cluster 4
Pathogenic/Benign
Pathogenic (4)
Benign (151)
Cluster 11
Pathogenic/Benign
Pathogenic (23)
Benign (50)
Cluster 7
Pathogenic/Benign
Pathogenic (372)
Benign (4)

Figure 3:



References:

1. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* **109**, 2163–2177 (2022).

2. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

3. DiStefano, M. T. *et al.* The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. *Genet Med* **24**, 1732–1742 (2022).

4. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).

5. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* **53**, D1313–D1321 (2025).

6. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic*

Acids Res **53**, D523–D534 (2025).

7. Tejura, M. *et al.* Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. *Am J Hum Genet* **111**, 2031–2043 (2024).

Chapter 5: Transforming the future of clinical genetics with technology development

Variant interpretation remains a critical barrier to the implementation of precision medicine as the majority of newly identified rare missense variants are still interpreted as VUS in ClinVar. This work aimed to help resolve this problem in three distinct ways. First, we demonstrated the substantial impact MAVEs have in clinical variant interpretation, offering a solution for resolving about 50% of VUS across different tumor suppressor genes. As the MAVE community continues to generate more assays with improved technologies, we will undoubtedly be able to make significant advances in further reduction of the uncertainty in clinical genetics. While MAVEs could offer substantial hope, technologies were still limited by their reliance on utilitarian cancer derived cell lines and inability to assess variants in proper cellular contexts.

Next, to address the cell context limitation, we developed iPSC-SGE where variant libraries are edited into iPSCs at the endogenous locus of the gene of interest. iPSC-SGE is advantageous over traditional MAVEs for two reasons. First, variant effects are measured in the context of a second allele and genetic interaction between engineered variants and the background allele can be measured. Second, iPSCs harboring variant libraries can be differentiated and variant effects measured in differentiated cell types. This offers the benefit of variant assessment in correct cell context and allows access to phenotypes from genes which are not expressed in utilitarian cell lines. Beyond use in variant interpretation, iPSC-SGE enables investigators to make deeper mechanistic discoveries and to use that information to gain a deeper understanding of phenotypes in patients and to select possible therapeutics to screen in the presence of variants.

Finally, even with a near perfect functional assay, as was the case with *BRCA1*, VUS reclassification remains constrained by limitations in other forms of evidence. In particular, data from variant effect predictors (VEPs) has been historically underutilized. The updated ACMG guidance on the calibration and use of VEP data sought to solve this problem by calibrating a set of widely used VEPs with a genome-wide aggregate set of ClinVar variants across genes. Our comprehensive evaluation of these guidelines revealed that many genes have discordant ClinVar distributions compared with genome-wide calibration thresholds. To address this limitation, we clustered predictor score distributions from Pfam domains to achieve more accurate calibration while maintaining robust sets of ClinVar variants across domains. This calibration method, in combination with the latest VEP models, is primed to assign stronger evidence for most variants while improving accuracy of calibration over current methods.

Taken together, these innovations are poised to have a significant impact on genome guided precision medicine. iPSC-SGE fills the gaps of conventional MAVEs conducted in utilitarian cell lines and improved VEP calibration will provide more accurate and stronger evidence for variant interpretation across all genes. Investment in scaling of MAVEs is already happening as part of the Impact of Genomic Variant on Function (IGVF) consortium, where hundreds of thousands of variant effects are being generated with Hap1 SGE and trangene based variant abundance assays. The data from these assays will be used to resolve VUS from clinical testing across 40 genes. Application of the same level of investment into iPSC-SGE will undoubtedly offer similar levels of VUS resolution with the added benefit of contextual insight.

Beyond variant interpretation with standard ACMG guidelines, iPSC-SGE data represents an opportunity to train machine learning models on variant effect data in context.

This will enable the generation of context aware and context specific variant effect predictors. At present most predictors lack context awareness and make blanket predictions about variant pathogenicity. Since many variants only cause phenotypes in specific cell types, the generation of context aware predictors is appealing. Further, measuring varints in specific cell context allow for the computation design of variant specific therapies like protein mini-binders that could stabilize certain interactions. iPSC-SGE allows for the screening of such molecules at scale and in the presence of multiple variants. Thus, iPSC-SGE in combination with better calibrated VEPs will usher us into a new era of precision medicine, where VUS are substantially reduced in traditionally difficult to assess genes and variant specific therapies are tested at scale for prioritization in clinical trials.