Systematic Evaluation of Functional and Computational Evidence for Large-Scale Resolution of Variants of Uncertain Significance

Malvika Tejura

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Douglas M. Fowler, Chair

Lea M. Starita, Chair

Danny E. Miller

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Systematic Evaluation of Functional and Computational Evidence for Large-Scale Resolution of Variants of Uncertain Significance

Malvika Tejura

Chairs of the Supervisory Committee:

Douglas M. Fowler & Lea M. Starita

Department of Genome Sciences

Interpreting missense variants remains a major challenge in clinical genomics, with approximately 90% of observed missense variants classified as variants of uncertain significance (VUS). These uncertain classifications limit the clinical utility of genetic testing and hinder medical management for individuals. In this thesis I show that computational evidence is gene dependent: variant-effect predictors (VEPs) vary substantially across genes in their ability to distinguish benign and pathogenic variation. This gene-level heterogeneity demonstrates that gene-agnostic evidence thresholds can overestimate or underestimate evidence strength for clinical variant classification, motivating the development of gene-specific calibration frameworks within the IGVF Consortium. A central contribution of this thesis addresses the growing challenge of VUS, where I integrate calibrated large-scale functional measurements from multiplexed assays of variant effect (MAVEs) with gene-specific calibrated variant-effect predictions to build a scalable framework for resolving VUS. Applied across 40 clinically relevant genes, this framework resolves ~75% of 16,115 ClinVar VUS while maintaining an error rate of <1% among benign and pathogenic controls. In addition, applying this framework to >90,000 unobserved variants enable preemptive classification of 62% of variants before they are observed in clinical databases, further reducing the future burden of VUS. Together, these results show that production-scale functional data, combined with computational evidence, can substantially reduce VUS and help realize the promise of precision medicine.

Table of Contents

Table of Contents 1
Acknowledgements 5
Dedication 11
Chapter 1: Introduction 12
1.1 Setting the stage: The Growing Gap between Sequencing and Variant Classification 12
1.2 Standardized Guidelines for Clinical Variant Classification in Mendelian Disease Genes 13
1.3 The Need for Scalable Evidence: MAVEs and Variant Effect Predictors to the Rescue ... 14
1.4 Frameworks for Calibrating MAVE Measurements and VEP Predictions 17
1.5 Combining Functional and Predictive Data to Resolve VUS 19
Figure 1.1 Adapted from McEwen et al. Nat Rev Genet (2025): Variants of uncertain significance are an exponentially growing clinical problem 21
1.6 Figures 21
Figure 1.2 : Combination of evidence types and their respective strengths leads to VUS reclassification as outlined in Richards et al. Genet. Med. (2015) 22
Figure 1.3 Adapted from McEwen et al. Nat Rev Genet (2025) Overview of multiplexed assays of variant effect workflows 23
Figure 1.4 Figure 1A Adapted from Tejura et al. AJHG (2024) 24
Figure 1.5 Data curation and harmonization workflow 25
1.7 Tables 26
Table 1: Adapted from Tavtigian et al. *Human Mutat* (2020) Point values for ACMG/AMP strength of evidence categories 26
Table 2: Adapted from Tavtigian et al. *Human Mutat* (2020) Point based variant classification categories 26
1.8 References 27
Chapter 2: Multiplexed assays of variant effect for clinical variant interpretation 33
2.1 Abstract 33
2.2 Introduction 34
2.3 MAVE technologies and design 36
2.4 Three eras of germline variant classification 38
2.4.1 Early era: low-throughput assays and little guidance 38
2.4.2 Middle era: advent of MAVEs and limited guidance 39
2.4.3 Current era: wide use of MAVEs and clear guidance 43
2.5 MAVEs for non-coding germline variants 46
2.6 MAVEs for tumour variants and therapy resistance 48

2.7 MAVEs for drug development and personalization 49
2.7.1 Drug target identification and personalization 50
2.7.2 Developing and refining drugs, vaccines and cell therapies 51
2.8 MAVEs and AI 52
2.8.1 Improving MAVE data sets with AI 53
2.8.2 MAVE data sets for benchmarking and training AI models 54
2.9 Conclusions and future perspectives 55
2.10 Figures 58
Figure 2.1: Variants of uncertain significance are an exponentially growing clinical problem 58
Figure 2.2 Overview of multiplexed assays of variant effect workflows. 59
Figure 2.3 Timeline of key events in the use of multiplexed assay of variant effect data in clinical contexts. 61
Figure 2.4 An example of multiplexed assays of variant effect data in action. 62
Figure 2.5 Reclassification of variants of uncertain significance with multiplexed assays of variant effect data. 63
Figure 2.6 Multiplexed assays of variant effect data and artificial intelligence. 64
2.11 Tables 65
2.11.1 Table 1: Clinical implications of assay design choices 65
2.11.2 Table 2: Multiplexed assays of variant effect data sets that have been clinically calibrated or used to reclassify variants of uncertain significance 69
2.12 References 71
Chapter 3: Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes 92
3.1 Abstract 92
3.2 Introduction 93
3.3 Methods 96
3.4 Results 104
3.5 Discussion 111
3.6 Data and code availability 117
3.7 Figures 118
Figure 3.1 Gene-specific analysis workflow 119
Figure 3.2 Analysis of evidence strength interval concordance with the genome-wide calibration 121
Figure 3.3 Some trending discordant intervals arise from genes with many control variants 122

Figure 3.4 Gene-specific predictor score distributions reveal causes of discordance 124
Figure 3.5 Many ClinVar variants of uncertain significance are in trending discordant intervals 124
3.8 Tables 125
3.8.1 Table 1 Incorrect prediction tolerance per evidence strength interval for REVEL and BayesDel 125
3.8.2 Table 2 Number of variants needed per evidence strength interval for concordance 125
3.8.3 Table 3 Number of variants needed per evidence strength interval for discordance . 125
3.9 References 126
Chapter 4: A scalable approach to resolving variants of uncertain significance 131
4.1 Abstract 132
4.2 Main 132
4.2.1 Production-scale MAVEs provide variant effect measurements for 62,215 variants across 10 genes 135
4.2.2 Curating experimental data from the community to generate an integrated variant effect dataset 137
4.2.3 Transforming experimental and predictive data into clinical evidence with improved calibration methods 138
4.2.4 Variant classification workflow using experimental and predictive evidence is highly accurate 141
4.2.5 Experimental and predictive evidence alone can reclassify 75% of existing VUS and preclassify 62% of unobserved variants 144
4.2.6 Identifying protein properties disrupted by variants guides future MAVEs and provides mechanistic insights into pathogenicity 147
4.2.7 An ecosystem of resources to disseminate and visualize experimental and predictive data and evidence 148
4.3 Discussion 150
4.4 Methods 153
4.5 Data Availability 173
4.6 Code Availability 173
4.7 Supplementary Notes 174
4.8 Figures 176
Figure 4.1 Integrated framework for experimental and computational characterization of coding variant effects 177
Figure 4.2 IGVF-produced MAVE data for 62,000 variants 179
Figure 4.3: Curating an expanded set of experimental data from the MAVE community.. 180

Figure 4.4: Improved gene- and variant-specific calibration methods improve calibration of experimental and predictive data 181
Figure 4.5 99.3% of PLP and BLB classifications from experimental and predictive evidence match ClinVar 184
Figure 4.6 Experimental and predictive evidence can resolve ~75% of VUS and preclassify 62% of unobserved variants 186
Figure 4.7 Profiling protein properties to guide future MAVEs and provide mechanistic insights into pathogenicity 188
4.9 References 189
Chapter 5: Conclusions and Future Directions 204
5.1 References 207

Acknowledgements

When I first joined Genome Sciences, I knew I was drawn to variant functionalization, and it felt natural to gravitate toward labs generating large-scale data on genetic variation. What I didn't expect, coming in as a bench scientist who spent long hours in tissue culture and entire weeks cloning and sequencing constructs, was that within a year I would barely touch a pipette. I had never coded a single line, yet somehow, I ended up taking on a project that relied almost entirely on computational work. If someone had told me that ahead of time, I would have laughed. This thesis exists only because of the people who helped me navigate that shift, and I begin by thanking my advisors, Lea M. Starita and Douglas M. Fowler, whose guidance shaped both this work and my growth as a scientist.

Doug, I knew I wanted to work with you from the moment I heard how highly my colleagues at Maze spoke about your science. I didn't yet understand how much your approach to thinking about the big picture, the details in the experiments, and data analysis would influence my own. Your ability to follow ideas wherever they lead and find meaningful connections in the noise changed the way I try to understand problems. Clinical variant interpretation was completely unfamiliar to me at the start, and you opened that field with a level of curiosity and intellectual energy that pushed me to see what was possible. Thank you for giving me room to grow into this work and for the encouragement that helped carry me through the moments when I questioned whether I could.

Lea, I was excited to work with you from the moment you introduced me to optimizing COVID sequencing workflows during a time when catching emerging variants felt urgent and meaningful. I've always admired your willingness to be hands on, whether it is loading sequencers or running gels, while also steering complex projects with clarity. Your leadership

stands out in the way you balance hands-on work with high-level coordination, and in how you create direction in environments that can otherwise feel uncertain. You helped me understand how science operates not only through the details of experiments and data, but also through the networks of collaboration, communication, and decision-making that sustain it. I've learned from the way you handle challenging situations, and how you make space for your trainees to grow while still offering support when needed. Thank you for supporting my work and for guiding me through spaces I didn't always know how to navigate. I am grateful for the confidence you placed in me, even when I was still learning how to find it in myself.

Beyond my primary advisors, I want to acknowledge all the Fowler lab, Starita lab and BBI, specifically two mentors who helped me find my footing and build a space of my own in this field: Shawn and Abbye.

Shawn, I am grateful for the time I spent working with you during my rotation and the paper we co-authored. You were patient in a way that made it easy to ask questions, and you always answered them with clarity and depth. You challenged me to take my ideas further than I initially thought possible and helped me understand how to think more critically and creatively about scientific problems.

Abbye, you have been such an important mentor throughout my time here. From having the best "sniffer" for whether my results passed the sniff test, to our Friday lunches after lab meeting, you made the process of growing as a scientist feel grounded, supportive, and genuinely fun. You pushed me when I needed it, talked me through uncertainty and celebrated the wins with me. Your mentorship has meant a great deal, and I'm grateful for the confidence and perspective you helped me develop.

In addition, I would like to thank my committee members: William S. Noble and Danny E. Miller, for their guidance, insight, and continued support throughout my graduate training. I am also grateful to Judit Villén for her support during my early years in the program.

I would like to acknowledge the IGVF Coding Variant Focus Group, who have played a huge role in my growth and success as a graduate student. I've learned so much from being part of this consortium, and I truly couldn't have asked for a better group of people to work with. The collaboration, support, and everything you've taught me have meant the world, thank you for shaping both my science and my experience.

I would also like to shout out my 2021 cohort members; you all made those first couple of years pass by so much easier. But I especially want to acknowledge Shruti Jain, Lizzie Plender, Syd Sattler, and our honorary cohort member Leah Anderson. You four have been the heart of my graduate school experience: the people who turned an overwhelming new chapter into something joyful and unforgettable. From late girls' nights spent dreaming up our own women-founded companies and imagining big futures, to turning our genome hackers meetings into full-blown venting sessions, to all our shenanigans in the middle of Estes Park, you've been my safe space. You are the ones who understood, truly understood, the highs of graduate school that felt electric and the lows that felt unbearable. You held space for all of it with grace and love. This journey would not have been the same without you. Thank you for standing with me, for celebrating my milestones, and for letting me be my truest self. I'm so grateful we got to grow through this together.

I would also like to acknowledge a few other significant individuals who have deeply influenced my academic journey. Dr. Stephen Floor, my very first academic mentor and the first lab I ever joined; you set the standard for what exceptional mentorship looks like. I have never

again encountered a mentor who matched your clarity and dedication to training young scientists. So much of how I think, learn, and show up in science comes from the foundation you gave me. I would also like to thank everyone I worked with at Ultima Genomics, as well as Maze Therapeutics. You all supported my growth at such a pivotal time and played a huge role in helping me transition confidently into graduate school. I'm endlessly grateful for your guidance.

To all my friends who have supported me throughout this entire journey, I am so grateful. Kaanchana, you have been a constant source of stability for me since high school. Thank you for believing in me even when I didn't believe in myself. When I first moved to Seattle, one of the biggest comforts was knowing that my best friend was there waiting to welcome me. From exploring all the new places (for me) in Seattle, to literally living a block away, to driving down to Oregon to pick up my baby girl, I appreciate all our memories together. Even when I was going through the ups and downs of graduate school, you showed up with empathy and tough love when I needed it, and a kind of unwavering friendship that only we understand. Thank you for everything.

Thank you, Rishma (future Dr. Rishma!), for being there at the very beginning of my academic journey. You were the person who taught me how to focus and how to study with intention. I will never forget our times at De Anza, when my curiosity for science peaked. You were there at the exact moment my interest in this field took shape, and so much of my path, especially applying to graduate school, only happened because you were right there with me every step of the way. Thank you for believing in me long before I believed in myself.

Thank you to Shriti and Aida for all the love and support over the years. From our endless FaceTimes to you both flying out to visit me, I've truly been blessed by your presence since our Berkeley days. You two have been constants through every phase of my life,

celebrating my wins, and reminding me who I am outside of grad school. Three lovely bunnies, always and forever.

I also want to thank Suzanne Braun for keeping life fun since our middle school days. I'm so grateful that we got to share a little piece of Seattle together; having you close for a few months meant a lot to me.

Thank you so much, Shruti Prasanna, for making such a lasting impact on my life in Seattle. From the very first day I met you and Jaadu, I felt nothing but gratitude. You quite literally taught me how to live, how to be generous, and how to show up for people. Your ready-to-help nature is a blessing, and I'm thankful I got to experience it firsthand. I could not have survived this chapter of my life without you. Of course both you and Pranav kept it so real and so grounded during a time when I needed it most. And of course, my baby boy Jaadu, thank you for every hug, every cuddle, every wiggly butt welcome that instantly lifted my mood. You brought joy into so many of the Seattle gloomy days.

Thank you so much to Nivedha Ravi for being a rock during my time in graduate school. From our long "chai talks" where we sat together contemplating the meaning of life, to watching the entire Marvel Universe in timeline order, you've given me some of my absolute favorite memories of this chapter. Thank you for always listening with such intention and for offering the kind of advice that comes from real wisdom and real care. I appreciate being able to grow with you, you made the gloomy days a bit more bearable, and I'm so grateful we got to reconnect again after middle school.

And now, saving the most precious people for last: my family.

Thank you, Mom and Dad, for being a constant source of inspiration, strength, and unconditional love. I haven't known a life without you, and I hope I never will. My resilience

and fire come directly from the two of you. Thank you for teaching me how to stand up for myself, but also how to be vulnerable when I need it most. Thank you for coming to Seattle at the drop of a hat whenever I needed you, for taking my stress-fueled meltdowns with grace, and for supporting me through every phase of this journey. I truly couldn't have done this, nor anything else in my life, without you. No matter what I said I wanted to pursue, you backed me fully, and that kind of support is a gift that not everyone gets. I am rich because I have parents like you.

To my brother, Shashank, although we constantly fight, there's a love there that everyone can see. Thank you for being a steady source of support for me. And yes, from now on, that's Dr. Tejura to you. Jokes aside, I am genuinely grateful that you moved to Seattle. I wasn't holding it together nearly as well as I pretended to and having you close meant everything. From helping me take care of Masha to helping me manage life in general, having an older brother like you is a blessing I don't take for granted.

And to my baby girl, my chocolate girl, Masha, my heart. I truly could not have survived graduate school without you. Waking up every morning to your smiling face, your cuddles, your insistence that we go outside and breathe fresh air even when I was drowning in stress, you kept me grounded and sane. You are genuinely the reason I am graduating. Your wiggly-butt greetings and your unconditional love, there is no version of this journey where I make it through without you. You are my love, my joy, and everything in between.

Dedication

I dedicate this thesis to my family: Manisha (Mom), Manish (Pop), Shashank, and Masha. Love you all to the moon and back.

Chapter 1: Introduction

1.1 Setting the stage: The Growing Gap between Sequencing and Variant Classification

Large-scale sequencing technologies have reshaped genomics research, with one of their most profound impacts being in clinical genetics. Understanding how genetic variants impact us is important to advancing precision medicine, and enables clinicians to diagnose genetic disorders, estimate disease risk, and guide medical management [1–3]. The emergence of next-generation sequencing has made it possible to sequence the exomes or genomes of individuals with genetic conditions at an unprecedented rate, revealing an enormous landscape of human variation (0.4% of our 3.2 billion base pair genome)[4] and has helped connect genotype to phenotype.

While our capacity to identify variants has grown exponentially, our ability to interpret their clinical significance is lagging. This gap between sequencing our genome and interpreting it is one of the central bottlenecks in realizing the full promise of precision medicine. For example, almost 90% of missense variants in the clinical variant database (ClinVar) remain as variants of uncertain significance (VUS), which are variants for which the given evidence is insufficient to confidently determine their risk of pathogenicity (Figure 1)[5,6]. The presence of a VUS is unhelpful for both individuals and clinicians, offering no clear answers about disease risk or medical management [7,8]. Sequencing alone is not enough, underscoring the need for scalable strategies that determine the functional and clinical impact of genetic variants. This urgent need for improved clinical variant classification motivates the work described in this thesis.

1.2 Standardized Guidelines for Clinical Variant Classification in Mendelian Disease Genes

Prior to 2015, variant classification in clinical laboratories was often conducted using variable approaches and criteria. However, in 2015, the American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) introduced guidelines for variant interpretation in Mendelian genes to establish a standardized framework across the field. These guidelines aimed to ensure consistency and improve the accuracy of variant classification[9].

The ACMG/AMP variant interpretation guidelines introduced different types of evidence that can be utilized. Evidence included using family history, patient phenotype, population frequency, and data from experimental assays and variant effect predictors to determine the pathogenicity of variants[9]. Proband and segregation data have traditionally been the most definitive forms of evidence in evaluating variants, however, obtaining such data is costly and labor-intensive, limiting its scalability in large-scale clinical variant analysis[10].

Each evidence type was assigned a specific evidence code, reflecting its strength and relevance to variant classification (Figure 2). To classify variants, the guidelines outlined a system of evidence code combinations. By considering multiple evidence types and their respective evidence codes, variants could be categorized into one of five categories: pathogenic, likely pathogenic, variants of uncertain significance (VUS), likely benign, or benign. These categories were based on the likelihood of the variant causing disease given the evidence provided, ranging from greater than a 99% chance of causing disease in the pathogenic category to less than a 0.1% chance of causing disease in the benign category[8,9]. By introducing these guidelines, the ACMG and AMP aimed to promote consistency and standardization in variant interpretation, enabling better comparison and sharing of variant data among laboratories.

In 2018, the ClinGen Sequence Variant Interpretation Working Group demonstrated that the ACMG/AMP evidence codes introduced in 2015 follow a Bayesian framework, and the qualitative categories and combinations introduced earlier have an underlying mathematical foundation [11]. This work helped quantify the guidelines and introduced a path where more systematic refinements to these guidelines could be made.

In 2020, the Bayesian framework for variant classification was further formalized by scaling evidence strengths to a points-based system (Table 1), allowing an easier combination of evidence codes and providing a clearer way to resolve conflicting evidence from different sources. Under this system, the total points accumulated from all evidence translated directly into an overall classification, helping improve consistency and reproducibility in variant classification and interpretation[12].

1.3 The Need for Scalable Evidence: MAVEs and Variant Effect Predictors to the Rescue

Although the variant interpretation guidelines and their updates to a more quantitative framework provided valuable guidance, traditional sources of strong evidence such as segregation studies and detailed clinical phenotyping are still needed to classify variants. However, these sources of evidence are slow, expensive, and cannot keep up with the rate of variants discovered. Therefore, there is a need for evidence types that can help interpret variants at scale.

One such type of evidence are multiplexed assays of variant effect (MAVEs) which are assays that enable the functional assessment of thousands to tens of thousands of genetic variants

in parallel [13,14]. Unlike low-throughput assays that test one variant at a time, MAVEs use a pooled experimental design combined with sequencing to quantify variant effects across entire coding (and often noncoding) regions. Depending on the given assay, MAVEs can measure protein abundance, enzymatic activity, cell fitness or drug sensitivity[15–18]. MAVEs generate dense variant function maps and provide the type of experimental evidence capable of scaling with modern sequencing technologies (Figure 3). These data offer direct biological insight into variant consequences, making MAVEs a powerful tool for reclassifying VUS and understanding mechanisms of disease[19].

MAVEs have been previously used to systematically resolve VUS in well-known cancer predisposition genes such as *TP53, BRCA1,* and *PTEN,* where approximately 60%, 50% and 15% of VUS, respectively, were resolved in these genes[20]. Given their utility in clinical variant interpretation a comprehensive review of MAVE technologies, applications, and their clinical relevance, including their limitations is presented in chapter two of this thesis, based on a published review article for which I am an author on[5].

In addition to MAVEs, variant effect predictors (VEPs), machine-learning models that estimate the deleterious impact of genetic variants, represent another scalable source of evidence for variant classification. VEPs predict the impact of genetic mutations on the function of a protein for nearly all possible missense variants in protein coding genes, illustrating the scale at which these models can be used for clinical variant classification. VEPs generally use features derived from multiple sequence alignment, protein structure information databases, and population databases to predict the pathogenicity of variants. Historically, most variant effect predictors relied on evolutionary conservation, building on substitution scoring concepts such as the BLOSUM62 matrix [21] and using features derived from multiple sequence alignments (MSAs)

to quantify how intolerant a protein position is to variation [22,23]. Tools such as SIFT [23] and PolyPhen-2 [24] demonstrated early success in leveraging conservation and protein structure features to predict variant deleteriousness, establishing conservation-based modeling as a foundational approach in computational variant interpretation. Later generations of predictors incorporated more structural features such as solvent accessibility, hydrogen-bonding networks, and estimates of stability changes (ΔΔG) [25–28]. Many missense variants destabilize proteins, showing that models involving protein structure information could capture an added layer of protein function that conservation-only models could not [29,30]. However, models involving protein structure information, perform poorly in intrinsically disordered regions which lack stable tertiary structure [31,32].

More recently, the field has shifted toward models that integrate both sequence and structure-derived features, as well as meta-predictors that combine the outputs of multiple individual VEPs to improve prediction accuracy [33,34]. The modern VEP landscape now includes supervised predictors trained on labeled variants [33,35], unsupervised or generative models that infer variant pathogenicity directly from evolutionary sequence families, such as EVE [36], semi-supervised models [37], or approaches based on protein language models [38]. Despite their power, VEPs face several limitations that constrain their clinical utility. Predictor performance varies substantially across genes, and inconsistent benchmarking practices can make comparisons between tools challenging [39,40]. In addition, circularity, where training and evaluation datasets overlap, can lead to an overly optimistic performance of variant predictions [41]. Together, these limitations underscore the need for systematic evaluation of VEPs and for integrating them into gene-specific, evidence-based frameworks for clinical variant classification.

1.4 Frameworks for Calibrating MAVE Measurements and VEP Predictions

While MAVE and VEP data offer highly scalable sources of evidence for variant classification, the original ACMG/AMP guidelines left uncertainty about how these data types should be used in clinical variant classification. MAVEs and VEPs both generate quantitative scores that reflect either functional impact or predicted deleteriousness; however, these values cannot be directly used in clinical settings. Instead, they need to be calibrated, which means translating how much evidence we can give either a MAVE or a VEP for clinical variant classification. Without such calibration, clinicians and variant review scientist can potentially interpret the outputs from both sources inconsistently. In the 2015 guidelines, "well-established" functional assays (i.e. MAVEs) could contribute strong evidence toward pathogenicity, but the criteria for determining whether an assay was well established was not well defined[9]. As a result, functional data for variant classification could be applied subjectively. Computational predictors (i.e. VEPs) were also constrained: regardless of their underlying method, they were limited to supporting evidence only for variant classification. This created inconsistencies and underutilized two evidence types that, in principle, could scale with the rapid growth of sequencing. To reduce the ambiguity surrounding the use of functional evidence, Brnich et al. introduced the Odds of Pathogenicity (OddsPath), a Bayesian statistic that quantifies how well a functional assay separates pathogenic from benign variants[42]. OddsPath provided a formal, quantitative method for assigning evidence strength to functional assays, and helped objectively separate "well-established" assays that could be used for variant classification and created a path towards standardizing functional evidence. Building on this quantitative direction, the ClinGen SVI later developed calibrations for computational predictors by aggregating ClinVar missense

variants across the genome and formalizing a quantitative framework to define score thresholds corresponding to supporting, moderate, strong, or very strong evidence[43](Figure 4). This approach allowed predictors such as REVEL to contribute more than supporting evidence for variant classification, paving the way for using VEPs at their calibrated potential. However, because these thresholds were derived genome-wide, they do not fully capture the underlying score distribution variations per gene, and as I show in Chapter 3, genome-wide calibration can lead to both underestimation and overestimation of predictive strength depending on the gene[40]. Despite their utility, both calibration frameworks have limitations. OddsPath requires large sets of pathogenic and benign controls, which are not always available for many genes. In addition, OddsPath assigns a single evidence strength (in either the pathogenic or benign direction) for an entire assay, even when variants differ substantially in their measured effects. For example, a variant that scores as lowly non-functional is given the same evidence strength as a variant that scores as highly non-functional. Similarly, genome-wide VEP calibrations treat all genes the same, obscuring important gene-specific differences. These limitations motivated the development of new calibration strategies. In Chapter 4, I briefly describe a functional calibration framework developed within the IGVF consortium, specifically the coding variant focus group, that uses a skew normal mixture model drawing on distributions from pathogenic, benign, synonymous, and gnomAD variants to generate variant-specific evidence strengths[44]. While I did not create the model, I played an integral role in curating and harmonizing the datasets that this model was tested on. In Chapter 4, I also briefly introduce a parallel effort within the consortium that calibrates computational predictors on a gene-specific basis or within clusters of similar domains across the genome to better capture biological heterogeneity across genes. Together, these efforts demonstrate how functional and predictive evidence can be

translated into appropriate evidence strengths using robust quantitative frameworks, ultimately improving the consistency with which we used these two evidence types in clinical variant interpretation.

1.5 Combining Functional and Predictive Data to Resolve VUS

While properly calibrated functional and predictive evidence can help resolve large numbers of VUS, being able to apply these two pieces of evidence systematically required an enormous amount of upstream work. Fortunately, working within a consortium [45] not only enabled the development of new calibration methods, but also made teamwork essential for identifying and curating published functional data available for clinically relevant genes. Curation efforts involved extracting and standardizing metadata according to the minimum information standards for MAVEs[46], assessing assay design and appropriateness for clinical utility, and determining whether each dataset could be meaningfully applied to variants observed in ClinVar. In addition to metadata curation, much of this functional evidence had never been harmonized, and individual studies reported variants and their functional scores in a variety of ways[17,18,47]. Substantial effort went into mapping these datasets to their relevant transcripts and genomic coordinates so that they could be integrated into one seamless framework for further analysis (Figure 5). Over the past year, a major focus of my work has been building this harmonized resource, where I linked functional measurements, computational predictions[33,35,37], ClinVar assertions[6], population frequencies[48], and other metadata into a framework that enabled systematic calibration of the functional and predictive data and their application to current VUS in ClinVar.

In chapter 4 of my thesis, I demonstrate how we used this resource with calibrated functional and predictive data, to resolve nearly 75% of existing VUS across 40 clinically relevant genes. Importantly, the same framework allowed us to pre-classify approximately 62% of unobserved variants, providing provisional interpretations for variants that have never been seen in the population. These results illustrate the power of harmonizing and calibrating scalable evidence types and represent the core analytical work of this thesis. The impact of these efforts extends beyond this dissertation. The harmonized datasets, calibration procedures, and variant-level interpretations will be disseminated through MaveMD, a clinician-facing platform described briefly in the final chapter of this thesis.

In summary, the challenges outlined in this chapter highlight the need for scalable, quantitative, and standardized approaches to variant classification, approaches that leverage functional and predictive data while addressing their limitations. The following chapters build on this foundation: first by introducing the clinical utility of MAVEs, then by establishing the need for gene-specific calibrations of variant effect predictors, and finally by demonstrating how calibrated functional and predictive evidence can be combined to resolve VUS and disseminated for clinical use. Much of this work was carried out in close collaboration with members of the IGVF consortium, reflecting the collective effort required to generate, harmonize, and apply new calibration methods to these large-scale datasets.

1.6 Figures

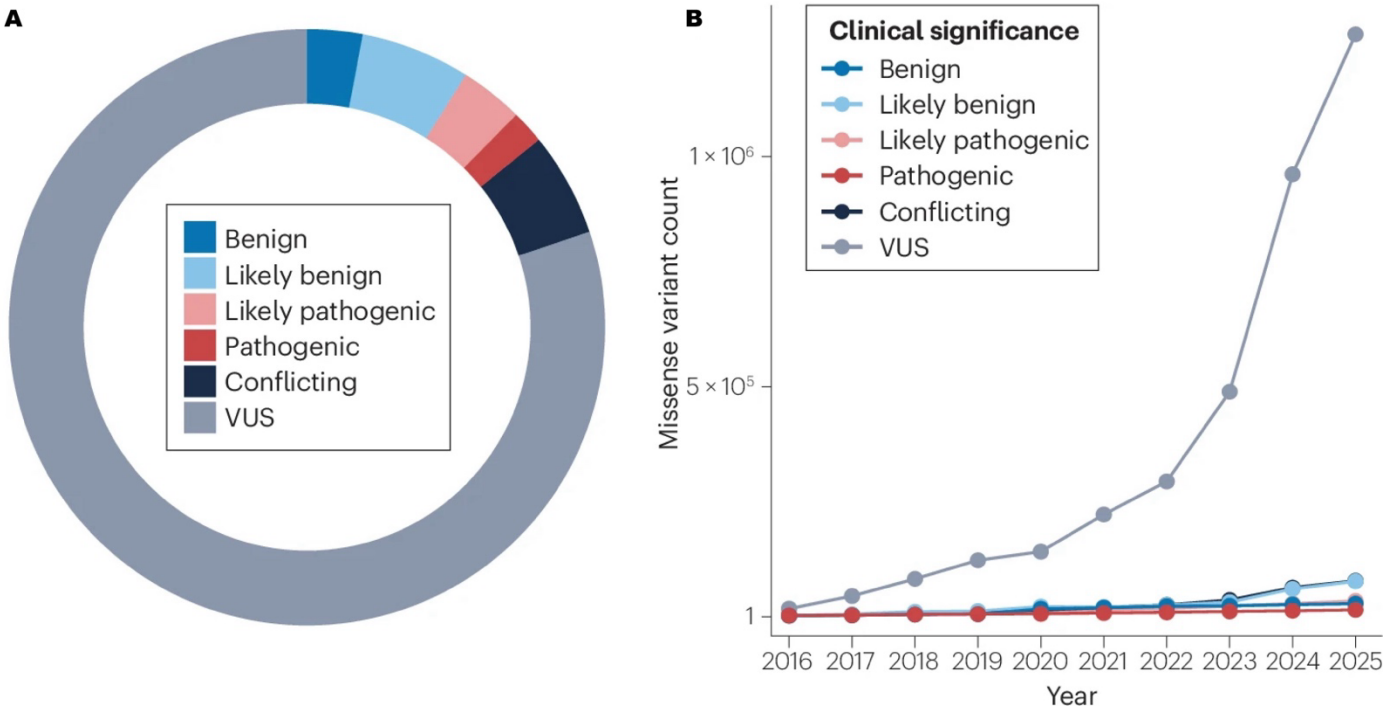


Figure 1.1 Adapted from McEwen et al. Nat Rev Genet (2025): Variants of uncertain significance are an exponentially growing clinical problem

A Single-nucleotide missense variants coloured by ClinVar classifications: benign (dark blue) = 28,616; likely benign (light blue) = 77,190; variants of uncertain significance (VUS) (grey) = 1,265,600; likely pathogenic (light red) = 34,378; pathogenic (dark red) = 14,985; conflicting interpretations (black) = 782,856. ClinVar data downloaded on 12 December 2024. **B** Missense variants in ClinVar from 2015 to 2024, shown by clinical significance. See Supplementary information for more details on how ClinVar data were processed.

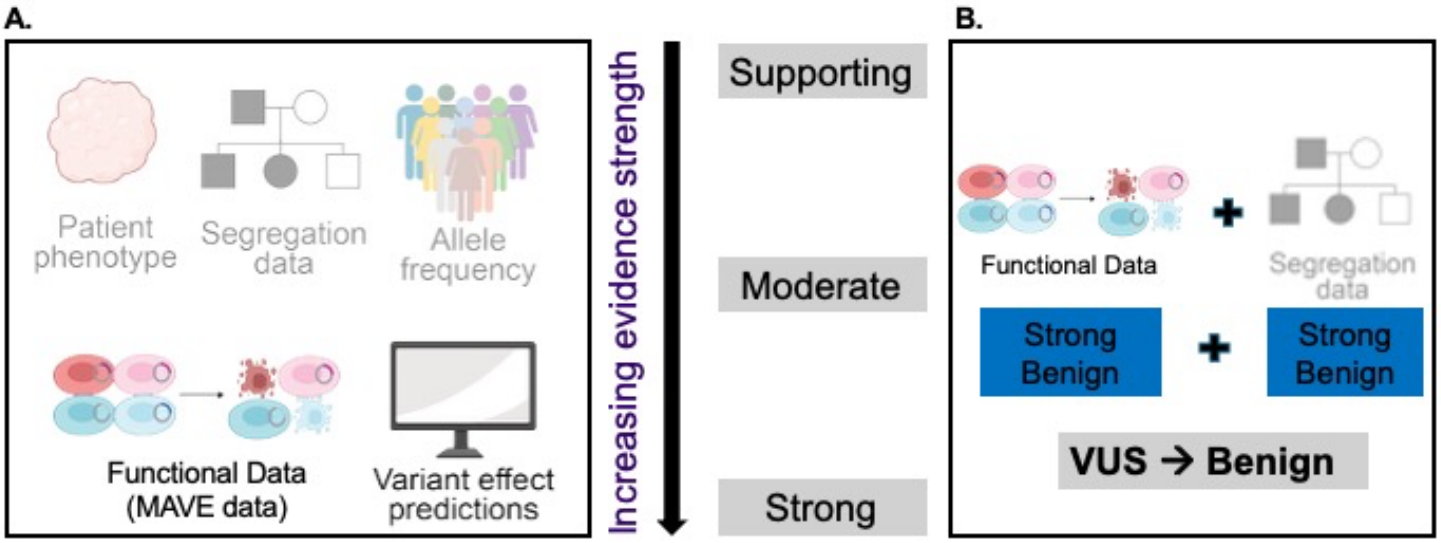


Figure 1.2 : Combination of evidence types and their respective strengths leads to VUS reclassification as outlined in Richards et al. Genet. Med. (2015)

A Different types of data can be used as evidence for variant classification, and each evidence type is assigned a specific weight: supporting, moderate, strong, or very strong, according to the ACMG/AMP guidelines. **B** Evidence codes can be combined to reach a final variant classification. In the example shown, two strong lines of evidence, functional data and segregation data, combine to support a benign classification. For a complete list of allowable evidence code combinations, see Richards et al. (Genet. Med., 2015).

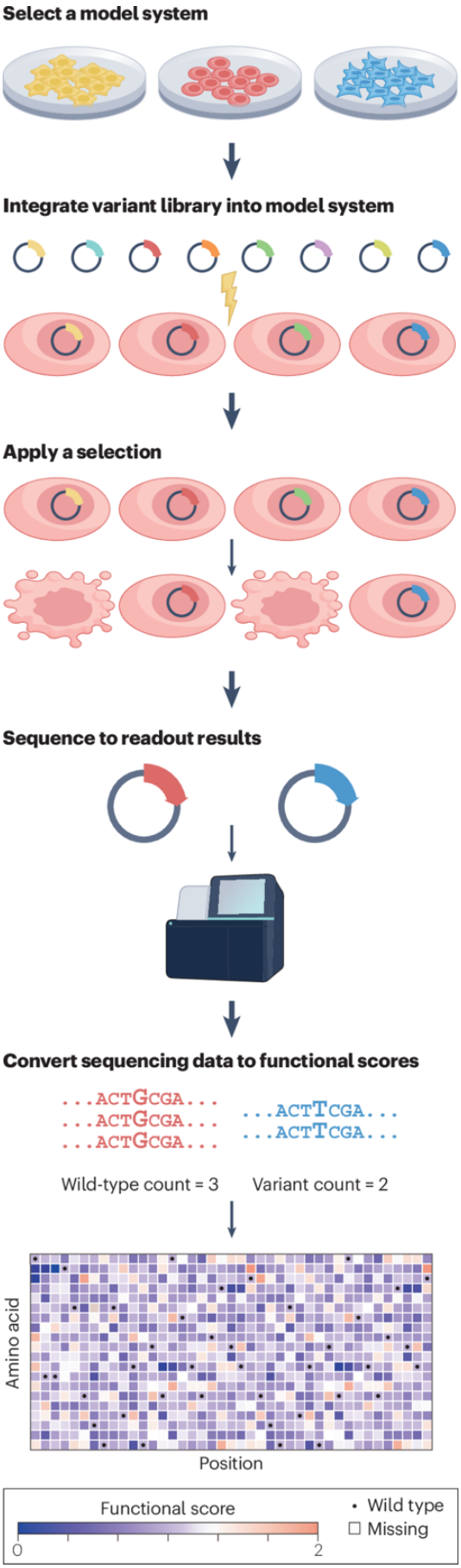


Figure 1.3 Adapted from McEwen et al. Nat Rev Genet (2025) Overview of multiplexed assays of variant effect workflows

After an appropriate model system is selected, a variant library is introduced into cells followed by phenotypic selection with a functional assay. Cells are then sequenced both before and after selection, and variant functional scores are computed to quantify changes in variant frequencies, reflecting their functional impact.

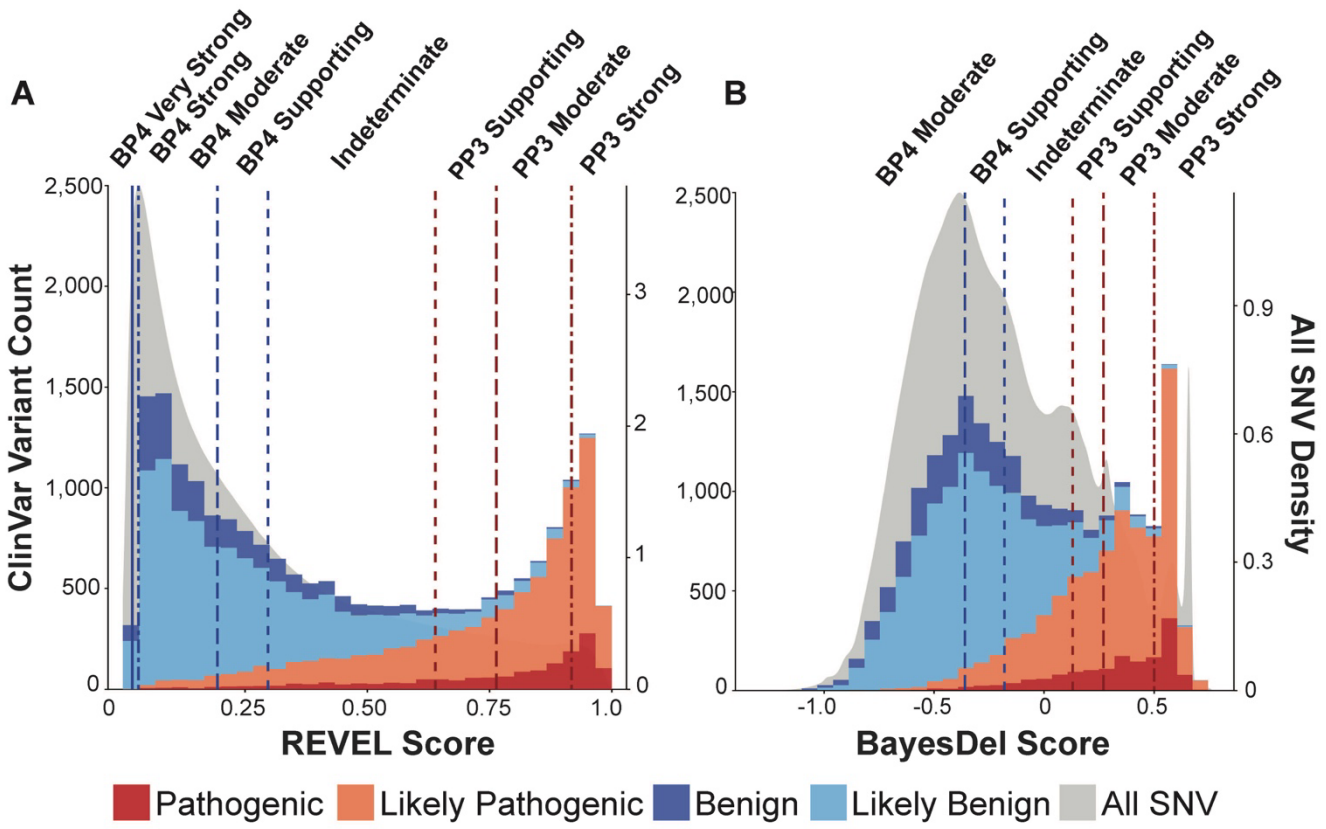


Figure 1.4 Figure 1A Adapted from Tejura et al. AJHG (2024)

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 and B)

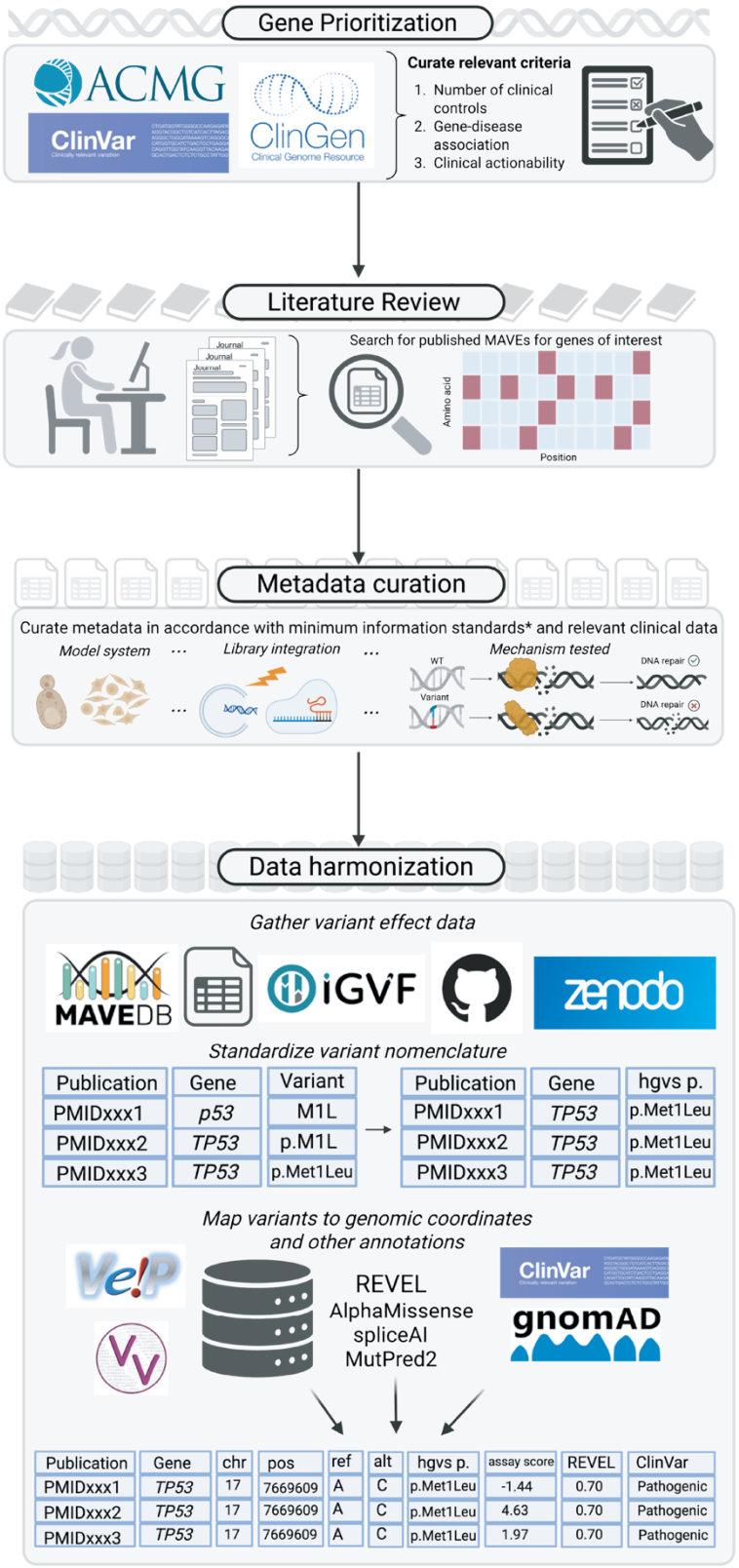


Figure 1.5 Data curation and harmonization workflow

Gene prioritization was performed by selecting clinically actionable genes with sufficient control variants. A comprehensive literature search identified published MAVE datasets for these genes, followed by metadata curation according to the Minimum Information Standards for MAVEs (Claussnitzer et al. *Genome Biol*. 2024). All functional datasets were harmonized into standardized variant nomenclature and mapped to variant effect predictors (REVEL, AlphaMissense, and MutPred2), ClinVar assertions, gnomAD population frequencies, and SpliceAI scores. The resulting integrated dataset provided a unified framework for combining functional and predictive evidence to support VUS reclassification analyses.

1.7 Tables

Table 1: Adapted from Tavtigian et al. *Human Mutat* (2020) Point values for ACMG/AMP strength of evidence categories

Evidence Strength	**Pathogenic (Points)**	**Benign (Points)**
Indeterminate	0	0
Supporting	1	–1
Moderate	2	–2
Strong	4	–4
Very Strong	8	–8

Table 2: Adapted from Tavtigian et al. *Human Mutat* (2020) Point based variant classification categories

Category	**Point Range**
Pathogenic	≥ 10
Likely Pathogenic	6 – 9
Uncertain	0 – 5
Likely Benign	–1 – –6
Benign	≤ –7

1.8 References

1. Lionel, A. C. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* **20**, 435–443 (2018).
2. Liu, Z., Zhu, L., Roberts, R. & Tong, W. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: Where are we? *Trends Genet.* **35**, 852–867 (2019).
3. Mani, S., Lalani, S. R. & Pammi, M. Genomics and multiomics in the age of precision medicine. *Pediatr. Res.* **97**, 1399–1410 (2025).
4. Human genomic variation. *Genome.gov* https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation?utm_source=chatgpt.com.
5. McEwen, A. E., Tejura, M., Fayer, S., Starita, L. M. & Fowler, D. M. Multiplexed assays of variant effect for clinical variant interpretation. *Nat. Rev. Genet.* (2025) doi:10.1038/s41576-025-00870-x.
6. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
7. Hoffman-Andrews, L. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J. Law Biosci.* **4**, 648–657 (2017).
8. Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).

9. 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).
10. Christensen, K. D., Dukhovny, D., Siebert, U. & Green, R. C. Assessing the costs and cost-effectiveness of genomic sequencing. *J. Pers. Med.* **5**, 470–486 (2015).
11. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).
12. Tavtigian, S. V., Harrison, S. M., Boucher, K. M. & Biesecker, L. G. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum. Mutat.* **41**, 1734–1737 (2020).
13. Starita, L. M. *et al.* Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
14. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
15. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
16. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation Mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
17. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
18. Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).

19. Findlay, G. M. Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Hum. Mol. Genet.* **30**, R187–R197 (2021).
20. 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).
21. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
22. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
23. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
24. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
25. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**, 1125–1132 (2006).
26. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543 (2009).
27. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
28. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382-8 (2005).

29. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).

30. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* **44**, 575–588 (2019).

31. Nielsen, J. T. & Mulder, F. A. A. Quality and bias of protein disorder predictors. *Sci. Rep.* **9**, 5137 (2019).

32. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).

33. 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).

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

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

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

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

38. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* (2021) doi:10.1101/2021.07.09.450648.

39. Vihinen, M. Guidelines for reporting and using prediction tools for genetic variation analysis. *Hum. Mutat.* **34**, 275–282 (2013).

40. 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).
41. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
42. 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).
43. 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).
44. Zeiberg, D. *et al.* Gene-based calibration of high-throughput functional assays for clinical variant classification. *bioRxivorg* (2025) doi:10.1101/2025.04.29.651326.
45. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
46. Claussnitzer, M. *et al.* Minimum information and guidelines for reporting a multiplexed assay of variant effect. *Genome Biol.* **25**, 100 (2024).
47. Gersing, S. *et al.* A comprehensive map of human glucokinase variant activity. *Genome Biol.* **24**, 97 (2023).
48. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

Chapter 2: Multiplexed assays of variant effect for clinical variant interpretation

This chapter is adapted with minimal modifications from:

*McEwen, A.E, **Tejura, M**, Fayer, S, Starita LM, Fowler DM. Multiplexed assays of variant effect for clinical variant interpretation. Nat Rev Genet (2025).*

2.1 Abstract

The rapid expansion of clinical genetic testing has markedly improved the detection of genetic variants. However, most variants lack the evidence needed to classify them as pathogenic or benign, resulting in the accumulation of variants of uncertain significance that cannot be used to diagnose or guide treatment of disease. Moreover, targeted therapy for cancer treatment increasingly depends on correctly identifying oncogenic driver mutations, but the oncogenicity of many variants identified in tumours remains unclear. To address these challenges, efforts to classify variants are increasingly using multiplexed assays of variant effect (MAVEs), which are massively scaled experiments that can generate functional data for thousands of variants simultaneously. The rise of MAVEs is accompanied by better guidance on the use of MAVE data for classifying germline variants to aid their clinical implementation. Here, we overview MAVE technologies from their inception to their increased use in the clinic, including their roles in uncovering mechanisms for variant pathogenicity and guiding targeted therapy and drug development.

2.2 Introduction

Genetic information can provide an accurate diagnosis, reveal variant pathomechanism, inform drug development and guide treatment. However, this genetic information is often unusable because the impact of most human variants on molecular and cellular phenotypes or disease is unknown, and such variants of uncertain significance (VUS) cannot be used for diagnosis or clinical decision-making. Addressing this knowledge gap is important because VUS comprise the majority of unique genetic variants found through genetic testing thus far and are reported in ~20–40% of individuals[1,2,3] (Fig. 1 and Supplementary information). VUS also disproportionately affect individuals from ancestry groups that have been historically under-represented in genetics research[4,5,6].

VUS arise when there is insufficient evidence to classify a variant as either pathogenic or benign, which is often based on variant frequencies among populations, case–control studies, patterns of variant segregation within families, computational predictions and functional data[7]. In the case of functional data, variant effects are measured in experimental assays that, historically, were time-intensive and labour-intensive and performed after a patient with VUS presented to the clinic. However, these limitations resulted in the VUS discovered by clinical testing rapidly outpacing the functional data that could be generated. To address these challenges of scale and efficiency in classifying VUS, the same high-throughput DNA sequencing technologies that led to the massive increase in VUS also enabled the development of multiplexed assays of variant effect (MAVEs). In contrast to the traditional assays, MAVEs simultaneously measure the effects of thousands (or even tens of thousands) of variants in a single experiment by testing the functions of a large library of variants introduced into a model system. Recent advances in technology and expert guidance have increased the clinical utility of MAVE data. More specifically, new editing methods and model systems are enabling more appropriate phenotypes

to be measured, and the decreasing cost of next-generation sequencing and DNA synthesis has made these experiments much more affordable. Consequently, MAVEs are now estimated to have measured the effects of more than 10 million variants[8], which are available to a clinical community that is becoming more comfortable with using MAVE data to classify variants. Reflecting this increased use, experts have now published guidance for evaluating and using functional data[9,10], and several gene-specific expert groups have incorporated MAVE data sets into their own guidelines for variant classification[11]. Moreover, translation of the first MAVE data into clinical evidence has demonstrated its value in resolving VUS. In previously published work, more than 1,500 variants in clinically relevant genes (such as those on the ACMG list of reportable secondary findings) have been re-evaluated in the context of MAVE data, and over half of these variants were reclassified. Thus, we are in an exciting moment in genetic medicine in which comprehensive functional data are available for a handful of clinically relevant genes, with more data being generated rapidly[12,13]. In this Review, we describe how MAVEs are designed, including some of the more recent technological advances. Then, we discuss the evolution of MAVEs over three eras, with a focus on the increasing guidance for clinicians and variant review scientists to interpret MAVE data. Finally, we discuss how MAVE data reveal the molecular mechanisms of variant pathogenicity, guide the development of new drugs and enable the personalization of therapies. Although we discuss the clinical implications of assay design choices, we do not review MAVE assays and technologies in depth, as recently done elsewhere[12,14,15].

2.3 MAVE technologies and design

MAVEs measure the effects of thousands of variants in a single experiment by introducing a large library of variants into a model system and then tracking each variant through an assay that selects for one or more functions (Fig. 2). MAVEs have been performed in various model systems, including phage[16,17,18], bacteria[19,20,21], yeast[22,23,24], immortalized mammalian cells and stem cells[25,26,27], and they comprise coding variant-focused assays, such as deep mutational scanning, and non-coding variant-focused assays, such as massively parallel reporter assays (MPRAs). Additionally, they can involve transgenic elements or endogenous genome editing with homology-directed repair (HDR), base or prime editors.

After a model system is chosen, the variant library is introduced into the model system such that each cell expresses a single variant. One common approach uses complementary DNA encoding open reading frames, which are introduced either transiently (for example, from plasmids) or from the genome following integration by a retroviral vector or recombinase-based landing pad[28,29,30,31,32,33]. Such approaches may leave the original gene intact or, in cases in which the original gene is knocked down or out, replace its function[29,31]. Another approach involves directly editing a variant into the endogenous gene, which has been made increasingly precise by advances in CRISPR–Cas9 technology that can alter the endogenous gene while preserving the surrounding regulatory elements[34,35,36]. For example, saturation genome editing (SGE) uses Cas9 and HDR to introduce a variant-bearing repair template[37]. Newer technologies include base editing, which combines Cas9 nickase with a deaminase enzyme to make specific base changes[38], and prime editing, which uses a Cas9 nickase fused to a reverse transcriptase along with a prime editing guide RNA to integrate variants[39].

After introducing variants, diverse methods can be used to assay variant function. One commonly used assay is based on cell growth, in which variants are classified on the basis of how they affect cell growth or survival. For example, variants that confer drug resistance can be identified by measuring variant frequencies with high-throughput DNA sequencing before and at fixed intervals throughout a period of cell growth in the presence of a chosen drug. Increased variant frequencies over time indicate that a variant had a positive effect on cell growth and thus likely conferred drug resistance. Another class of variant function assays involve fluorescence-based or barcode-based reporters, in which variants are classified on the basis of whether they activate the reporter. Examples of such assays include variant abundance by massively parallel sequencing (VAMP-seq), in which a fluorescent reporter reads out target protein abundance, and MPRAs, in which a transcribed RNA barcode reporter reads out enhancer or promoter function[32,40]. Beyond growth and reporter assays, a plethora of other MAVEs have been described that can capture diverse aspects of variant function, including assays that measure variant effects on specific protein functions[23,32,41], cell transcriptomes[42,43,44] and cell morphology[45].

Choices regarding model system, method of variant introduction and assay have consequences on the clinical usability of the data and often involve tradeoffs (Table 1). For example, phage, bacterial and yeast assays are attractive owing to low cost but do not always capture human-specific features such as protein post-translational modifications or protein–protein interactions[46,47]. Furthermore, exogenous expression approaches relying on complementary DNAs are scalable but generally cannot detect splicing variants, which contrasts with the features of endogenous editing approaches[32,37]. Some assays measure a particular molecular property, such as transcription or protein abundance, whereas other assays capture

many molecular functions but provide less insight into mechanism (for example, cell survival assays)[14]. However, prior to use for variant interpretation, the performance of all assays must ultimately be assessed by benchmarking against variants with well-established clinical effects. Ideally, these clinical control variants are used to select an appropriate assay system for their gene of interest[9,10].

2.4 Three eras of germline variant classification

The translation of functional data, derived from both traditional assays and MAVEs, has co-evolved with the associated clinical guidelines through three distinct eras (Fig. 3). In the 'early era', with one remarkable exception[48], only traditional, low-throughput assays were available and there was little guidance on how to use functional data. In the 'middle era', from 2015 to 2019, MAVE data were available for a handful of clinically relevant genes and there was increased guidance. The current era began in 2020 with the accumulation of MAVE data and the articulation of guidelines, which has now enabled systematic use of functional data in clinical variant classification.

2.4.1 Early era: low-throughput assays and little guidance

As molecular cloning and clinical sequencing became more common, researchers began analysing genetic variants found in patients by expressing them in bacteria, yeast, mammalian cells in culture, transgenic mice and other model systems. Many of these early experiments were originally performed to investigate protein function or uncover the molecular mechanisms underlying genetic variants and were only later adapted for assessing variant pathogenicity[49,50,51]. Although many functional assays were performed for clinically relevant genes, the clinical

genetics community was slow to incorporate this type of evidence into variant classification. This sentiment was reflected in the first two versions of the recommendations for standards for interpretation of sequence variation published by the American College of Medical Genetics (ACMG), which offered no guidance beyond suggesting that laboratories 'might establish collaborative relationships with research facilities to determine genotype phenotype correlations, such as through functional studies'[52,53].

Most early assays tested only a handful of variants, usually long after the individual presented to the clinic, but one study published in 2003 individually tested the effect of 2,314 missense variants of p53 (encoded by TP53) on transcriptional activation of eight different promoters[48]. Although the original goal was to better understand the molecular function of p53, it was later observed that the most common mutations found in tumours showed a substantial loss of transcriptional activity[54,55,56]. This data set was remarkable at the time for its scale and inclusion of many variants that had not yet been observed in humans, which provided data that could be consulted as new variants were discovered by genetic testing. Nearly two decades later, current guidelines for classifying newly discovered variants in TP53 recommend using functional evidence from this data set[57], highlighting the utility and longevity of comprehensive, prospective measurement of variant effects.

2.4.2 Middle era: advent of MAVEs and limited guidance

The first MAVEs were developed in 2009 and initially focused on investigating gene expression, sequence–structure–function relationships, enzyme function and molecular evolution[16,18,22,23,24,41,58,59,60]. Multiple advances in MAVE technology also took place during this era, including the development of generalizable assays that could be deployed for many genes to accelerate data generation. These advances in MAVE technology coincided with the first official

recommendations for incorporating functional evidence into clinical variant classification, which were published in the 2015 version of the ACMG/AMP Standards and Guidelines for the Interpretation of Sequence Variants[7]. The guidelines stated that 'well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product' and 'well-established in vitro or in vivo functional studies show[ing] no damaging effect on protein function or splicing' were strong evidence for variant pathogenicity (PS3) and benignity (BS3), respectively. However, 'well-established' was not defined nor was guidance provided for adjusting evidence strength.

Shortly after the release of the ACMG/AMP guidelines, several MAVEs in clinically relevant genes were performed with the explicit goal of investigating the functions human variants. For example, a 2015 study measured the effects of nearly 2,000 variants affecting the BRCA1 RING domain on two molecular functions of BRCA1 commonly disrupted by missense mutations observed in the clinic. One of these BRCA1 functions, BARD1 binding activity, was assayed using a yeast two-hybrid system, and the other, E3 ubiquitin ligase activity, was assayed using phage display[61]. Neither assay perfectly distinguished the benign from the pathogenic control variants, but combining data from both assays improved predictive power. In 2016, a MAVE of 10,000 variants of PPARG — a nuclear hormone receptor whose dysfunction is associated with autosomal dominant lipodystrophy and increased risk of type 2 diabetes — was performed by measuring the effects of variants on cell surface expression of a downstream target[62]. In support of the clinical utility of the data, the functional score assigned to each variant was inversely correlated with type 2 diabetes risk, and predicted pathogenic variants were enriched in exomes from a cohort with a clinical history suggestive of monogenic diabetes or insulin resistance. In 2018, a MAVE was performed in yeast that measured lipid phosphatase

activity of 7,000 variants of PTEN — a tumour suppressor with dual-specificity lipid and protein phosphatase that antagonizes the PI3K/AKT/mTOR signalling pathway whose dysfunction is associated with autosomal dominant hereditary cancer predisposition syndromes and neurodevelopmental disorders, such as autism spectrum disorder. Again supporting the clinical utility of this MAVE, the functional scores assigned to each variant were able to differentiate the control pathogenic variants obtained from ClinVar from the putative benign variants present in gnomAD with high specificity and moderate sensitivity[33]. For both the 2016 and 2018 MAVEs, the authors did not reclassify any VUS because of the paucity of controls available at the time and the unclear guidance surrounding functional evidence, but their strong performance suggested that MAVE data could be a powerful tool for clinical variant classification[63].

Many of the first MAVEs in clinically relevant genes, including the examples mentioned earlier, measured specific molecular phenotypes related to the function of the target protein, which required a bespoke MAVE for each gene. Further into this middle era, studies began to develop and implement more generalizable MAVEs, such as complementation assays, SGE and VAMP-seq. Complementation assays often entail introducing a human gene — and its many variants — into a yeast strain in which the orthologous gene has been removed; the variants are classified on the basis of whether they restore (or complement) the function and promote cell survival. Although yeast and humans are separated by ~1 billion years of evolution, many molecular components are sufficiently conserved such that yeast complementation assays can reasonably distinguish pathogenic and benign variation in the human genome. For example, such a MAVE for CALM1, CALM2 and CALM3, all of which encode the calmodulin protein involved in cell signalling[64], performed well in distinguishing pathogenic from benign variants[29]. Yeast complementation assays have already been applied to several other genes, including TPK1

(ref. [29]), HMBS[65], OTC[66], MTHFR[67], MSH2 (ref. [68]), CBS[69], GDI1 (ref. [70]) and GCK[71], which could in principle be used for the nearly 200 human genes known to rescue the loss of their yeast orthologue. Notably, similar complementation assays have been performed in immortalized human cells for MSH2 (ref. [72]), NDUFAF6 (ref. [73]) and TP53 (refs. [30,31]).

Another generalizable MAVE, SGE, can be used in mammalian cells. More specifically, Cas9-mediated HDR is used to edit an essential gene at its endogenous locus and the subsequent effect on cell survival and growth is used to classify the variant. The first large SGE experiment was performed on nearly 4,000 BRCA1 variants and, although this assay was not performed in cell lines of breast nor ovarian cancer origin, functional scores assigned to variants could distinguish between benign and pathogenic variants with high sensitivity and specificity[74]. SGE has since been used to target DDX3X[75], VHL[76], RAD51C[77] and BAP1 (ref. [78]) in the haploid HAP1 cell line, in which nearly 2,000 other genes are essential and thus could be targeted in the future[79]. SGE has also been used in non-haploid cell lines to assay certain genes, including CARD11 (ref. [80]), BRCA2 (refs. [81,82]) and TP53 (ref. [83]).

VAMP-seq is a generalizable assay that was developed on the basis of the observation that ~75% of disease-causing missense variants reduce protein abundance[84,85]. In VAMP-seq, a fluorescent reporter is fused to the target protein expressed in a human cell line, which is used to sort cells into bins on the basis of the effect of variants on protein abundance as measured by fluorescence[32]. VAMP-seq was initially used to quantify the abundance of more than 4,000 variants in PTEN and showed high positive predictive value by identifying 80% of pathogenic variants on the basis of the effects on protein abundance. These data were used to reclassify the p.I135K variant from likely pathogenic to pathogenic, marking the first time MAVE evidence was used within the ACMG/AMP framework for variant classification. VAMP-seq has been

applied to several genes, including TMPT32, CYP2C9 (ref. [86]), CYP2C19 (ref. [87]), VKOR[88], NUDT15 (ref. [89]), PRKN[90], KCNE1 (ref. [91]), KCNH2 (ref. [92]), SCN5A[93], ASPA[94], OCT1 (ref. [95]) and F9 (ref. [96]).

In summary, technological advances that enabled more generalizable MAVEs led to increasing data sets for clinically relevant genes during this era. However, these data sets were not widely used for variant interpretation because guidance remained unclear.

2.4.3 Current era: wide use of MAVEs and clear guidance

In 2020, the ClinGenSequence Variant Interpretation (SVI) working group released updated guidelines for application of functional evidence to the ACMG/AMP framework[10] and the Brotman Baty Institute Mutational Scanning Working Group (BBI-MSWG) released additional guidance specifically for MAVE data[9]. The guidance from these groups provided the foundation for MAVE data to be systematically included in clinical variant classification workflows by describing how to assess the clinical validity of functional assays and assign an appropriate strength of evidence. To determine evidence strength (a process called calibration), the ClinGen SVI developed OddsPath, a calculation that quantifies how well the assay distinguishes between pathogenic and benign clinical control variants[97]. These updated guidelines clarified how to include functional data in variant classification and were a perfect match for MAVE data sets that contain thousands of variant effect measurements, including those for many control variants (Fig. 4).

To systematically assess the clinical utility of MAVE data, a 2021 study used previously published data to reclassify VUS in three clinically actionable genes: BRCA1, TP53 and PTEN[98]. The functional data from each assay were first calibrated according to the SVI guidelines, which revealed that the number of clinical control variants limited evidence strength.

For BRCA1, a single high-quality SGE functional data set that accurately discriminated between known benign and pathogenic variants was used[74], and the assay generated strong evidence for pathogenicity or benignity. Combining the functional data with other lines of evidence resulted in the reclassification of 49% of BRCA1 VUS. The BRCA1 data set was straightforward to calibrate because the SGE assay accurately discriminated between known benign and pathogenic variants. Moreover, because many known benign and pathogenic variants existed, the SVI framework could be applied to determine the strength of evidence for pathogenicity and benignity supported by the data set. By contrast, none of the multiple MAVEs for TP53 — each testing different disease-related functions — could accurately discriminate between benign and pathogenic variants. However, combining data from multiple assays to better predict the probability of variant pathogenicity (see the MAVEs and AI section discussed subsequently) generated strong evidence for pathogenicity and moderate evidence for benignity, ultimately leading to 69% of TP53 VUS being reclassified. In the case of PTEN, two MAVEs existed but could not be combined using a model nor calibrated using the SVI guidelines because of the scarcity of control benign variants owing to missense PTEN mutations being so detrimental and thus rare in humans[32,33]. Nonetheless, 15% of VUS were reclassified using the functional data and rules articulated by a panel of experts developing guidelines for interpreting PTEN variants[99]. Thus, this 2021 study of BRCA1, TP53 and PTEN highlighted the power of using MAVE data to reclassify VUS but also exposed the challenges of integrating multiple data sets without sufficient control variants. Since the 2021 study, many other MAVE data sets have been generated, calibrated or used to reclassify VUS (Table 2). For example, MAVE-derived MSH2 functional data provided strong and moderate evidence for pathogenic and benign classifications, respectively, which helped to reclassify 532 of 682 (78%) VUS identified by a commercial

laboratory[72,100]. Across many previously published studies, a total 937 of 1,711 (55%) VUS have been reclassified with the addition of MAVE data. Notably, many of these genes are on the ACMG secondary findings list of genes associated with diseases considered to be medically actionable, indicating that established interventions informed by MAVEs can prevent or substantially reduce morbidity and mortality101. Such ACMG secondary findings genes with reclassified VUS include BRCA1 (refs. [74,98,102]), CALM1, CALM2, CALM3 (refs. [29,103]), MSH2 (refs. [72,100]), OTC66, PTEN[32,33,98] and TP53 (refs. [31,48,98,104]), and calibrated assays exist for other ACMG secondary findings genes, including BRCA2 (ref. [81]), KCNH2 (ref. [105]) and VHL[76]. These studies add to the growing body of evidence that MAVEs are a powerful tool for clinical variant classification and have the potential to positively affect patient care (Fig. 5 and Supplementary Table 1).

Most clinically focused MAVEs thus far have assayed variants that cause a loss of function, which is the most common and easy-to-detect molecular mechanism of disease, but some MAVEs have been developed to detect a wider spectrum of disease mechanisms. For example, MAVEs of TP53 have been performed both in the absence of wild-type TP53 to detect loss-of-function variants and in the presence of wild-type TP53 to detect dominant-negative variants, as both types of TP53 variants are associated with cancer predisposition[30,31,104]. By contrast, a MAVE of SCN5A using sodium influx toxicity as a readout successfully distinguished loss-of-function variants associated with Brugada syndrome from gain-of-function variants associated with long QT syndrome, demonstrating how different molecular mechanisms can produce distinct clinical phenotypes[106]. Similarly, a yeast complementation assay of GCK differentiated between loss-of-function variants associated with maturity-onset diabetes of the young and gain-of-function variants associated with hyperinsulinaemic hypoglycaemia[71]. A

study of 200 KRAS and TP53 variants demonstrated that single-cell RNA-sequencing could simultaneously identify both dominant-negative and loss-of-function variants in TP53 and gain-of-function variants in KRAS[43,44], highlighting how newer MAVE approaches can comprehensively capture diverse disease mechanisms. Reflecting interest and usage, a large community and critical infrastructure has recently established itself around MAVEs. For example, the database of record for MAVE data includes ~2,000 data sets comprising ~7 million MAVE-derived variant effect measurements[107]. The Atlas of Variant Effects Alliance is an international community focused on deploying MAVEs to understand the human genome at nucleotide resolution[8] and the NHGRI-funded Impact of Genomic Variation on Function Consortium seeks to understand how genomic variation affects genome function[108]. Both organizations have worked to markedly expand the amount of MAVE data available, set standards and disseminate MAVE data widely[8,108,109]. Thus, an ecosystem of technologies and tools has arisen to for MAVEs, including those to benchmark and distribute the resulting functional data.

2.5 MAVEs for non-coding germline variants

About 98% of the human genome is non-coding and includes functional elements crucial for regulating gene expression. Genome-wide association studies have revealed that these non-coding regions contain a substantial amount of genetic variation associated with common diseases[110,111], highlighting the need for MAVEs that extend to non-coding variants. Indeed, several MAVEs of non-coding variants have been performed to better understand the functions of promoters41,112, enhancers[26,113,114,115,116,117,118], silencers[119], RNA stability or untranslated regions[115,120,121], RNA splicing[122,123,124,125,126,127], RNA editing[128,129] and RNA localization[130].

A type of MAVE known as MPRAs is emerging as a powerful tool to investigate non-coding variants that are identified during clinical genetic testing. For example, a study of 111 probands with negative exome sequencing in a rare disease cohort identified a median of 14 de novo and 15.5 inherited rare variants within 100 kb of known transcription start sites, ~4.5% of which substantially affected regulatory activity when assessed with an MPRA[131]. Moreover, an MPRA-like assay in mice that assesses splicing was used to follow up on de novo exonic variants identified in individuals with autism spectrum disorder and revealed that 6.3% of these variants alter the splicing of many paralogues of human genes known to be associated with autism spectrum disorder[132]. MPRAs were also used to investigate more than 30,000 variants spanning the regulatory elements of 20 genes, including the promoters of TERT and LDLR, which successfully identified non-coding variants known to be pathogenic, suggesting that MPRAs may be a valuable data source for clinical variant interpretation[133].

Currently, little guidance exists for using functional data to interpret non-coding variants. The 2022 recommendations for clinical interpretation of non-coding variants suggest that MPRAs can be used to generate evidence for variant pathogenicity or benignity if they are evaluated following the ClinGen SVI guidelines for functional data, with some additional considerations in mind[134]. Considerations include that MPRAs will not detect variants that rely on distant DNA elements or secondary structure and that the MPRAs use a disease-relevant model system. Moreover, calculating evidence strength is not feasible for most MPRA data sets owing to the lack of control benign and pathogenic variants in non-coding regions, which currently hampers their clinical utility.

2.6 MAVEs for tumour variants and therapy resistance

Tumour genome sequencing is becoming increasingly common and, as with germline sequencing, is revealing massive numbers of genetic variants that could have prognostic or therapeutic implications. Variants found in cancers are classified in terms of both oncogenicity (oncogenic, likely oncogenic, uncertain, likely benign and benign) and clinical significance (I, strong clinical significance; II, potential clinical significance; III, unknown clinical significance; IV, benign or likely benign)[135,136]. Similarly, when interpreting germline variants, the guidelines for classifying the oncogenicity of somatic variants in cancer suggest that functional data can generate strong evidence and recommend evaluating assays according to the ClinGen SVI guidance for the interpretation of functional data.

Patients with actionable variants are more likely to be enrolled in a genotype-matched clinical trial, but determining the functional and therapeutic relevance of tumour variants remains challenging[137]. One of the earliest uses of MAVEs in molecular oncology used an immortalized human breast epithelial cell (MCF10A) xenograft model, from which cells were transformed with a variant library of PIK3CA, which encodes the catalytic subunit of a lipid kinase that regulates cell proliferation, survival and metabolism. PIK3CA gain-of-function mutations are associated with several human cancers, including breast, colorectal, endometrial and ovarian cancers. Cells expressing PIK3CA variants observed in mutational hotspots of tumours were injected into mice and the variant frequencies in the resulting tumour were quantified[138]. This experiment found that oncogenic variants were over-represented in the tumour compared with pre-injected cells, showing that this approach can be used to measure variant oncogenicity. Several low-frequency variants that were classified as oncogenic using this method now fulfil biomarker criteria for a treatment that uses the PIK3CA inhibitor alpelisib[139]. MAPK1 is another

example of an oncogene that was previously thought to only be affected by variants at a single mutational hotspot. However, a cell-growth-based MAVE of nearly 7,000 MAPK1 missense variants identified dozens of novel activating mutations and revealed a subset of variants that were sensitive to ERK-directed kinase inhibitors138, and many of these variants are now classified as oncogenic[140]. Similar studies have identified additional driver variants in oncogenes such as MP[141], MET[142], Ras[143] and EGFR[144,145].

Kinase inhibitor therapy is currently used to treat multiple types of cancers, but successful treatment is often limited by acquired resistance. In 2014, a MAVE was used to prospectively identify vemurafenib resistance variants in BRAF V600E, a constitutively active BRAF variant that aberrantly activates the MAPK/ERK signalling pathway often found in malignancies (including melanoma, papillary thyroid cancer, colorectal cancer and non-small-cell lung cancer). Measurement of vemurafinib resistance of nearly 5,000 variants surrounding the drug binding site identified several resistance-conferring variants outside the binding site, a previously undescribed resistance mechanism[27]. Similar MAVEs have been deployed to identify resistance variants in EGFR[145,146,147,148], BCR-ABL[149], SRC[150], KRAS G12C151, CDK4, CDK6 and ERK2 (ref. [152]).

2.7 MAVEs for drug development and personalization

MAVEs that illuminate sequence–function relationships for DNA, RNA and proteins at high resolution are also valuable for implicating novel variants in common disease, uncovering unexpected disease mechanisms and revealing mechanisms of protein function. These insights can shed light on new variant–disease links, bring new understanding to the spectrum of variant

effects in known disease-linked genes, guide treatment decisions and catalyse the development of novel therapeutics.

2.7.1 Drug target identification and personalization

The comprehensive nature of MAVEs can help uncover the mechanisms by which certain variants cause disease, thereby empowering the personalization of therapies. Therapeutically relevant mechanisms that MAVEs can measure include protein stability, enzymatic activity, cellular localization, effects on the global transcriptome, splicing and gene expression. For example, a MAVE for CARD11 revealed a previously unknown exon-skipping mechanism that leads to dominant-negative CARD11 activity and causes CADINS[80]. SGE and other splice-aware MAVEs are able to identify variants that are pathogenic owing to exon-skipping events that could be targeted by splice-switching antisense oligonucleotides as a potential therapeutic direction[153]. Indeed, a splice-switching ASO was recently developed for a TP53 variant identified as splice-altering by SGE[83].

The ability to quantify pathomechanisms of variants can also identify those amenable to corrector drugs that stabilize partially unfolded variant proteins. For example, corrector drugs have been designed to stabilize and improve the function of CFTR to treat cystic fibrosis, and a MAVE was deployed to measure which CFTR variants were ameliorated by two of these drugs[154]. In another example, a MAVE used yeast complementation to identify the variants of cystathionine beta-synthase (CBS), which can cause tall stature, brittle bones and other phenotypes that can be ameliorated with vitamin B6 treatment[69]. Finally, a MAVE of MC4R, a G-protein-coupled receptor implicated in obesity, revealed variants amenable to a corrector therapy[155].

2.7.2 Developing and refining drugs, vaccines and cell therapies

MAVEs can empower the development or refinement of drugs ranging from small molecules to proteins to cells. For example, a MAVE was used to overcome challenges in identifying allosteric pockets on protein surfaces — which can be used to control disease-associated protein activity — by measuring the effects of ~26,000 variants on the folding and interaction partner binding of the oncogenic kinase KRAS[156,157,158,159]. One of these sites is occupied by previously developed allosteric KRAS inhibitors that are currently used to treat cancer, highlighting how knowledge of allosteric sites can be leveraged to create drugs[160]. By generating unbiased, comprehensive maps of allosteric sites in clinically important proteins, MAVE data can improve drug development and enable drugs to be developed for previously undruggable targets.

MAVEs can also guide the engineering and use of therapeutic monoclonal antibodies, an important class of drugs[161,162,163]. Such antibodies were rapidly developed against SARS-CoV-2 at the onset of the COVID-19 pandemic[164], specifically targeting the receptor-binding domain (RBD) within the spike protein of the virus. However, the same epitopes being targeted by therapeutic monoclonal antibodies were rapidly mutating and gaining resistance to the human immune system as SARS-CoV-2 spread through the human population[165]. To address this challenge, a MAVE was developed to leverage yeast display to assess the ability of thousands of variants in the spike RBD to bind to the human ACE2 receptor, which ultimately showed that the virus was highly mutable and many RBD variants could actually enhance ACE2 binding[166]. Additionally, the authors repeated the RBD MAVEs in the presence of antibodies from multiple sources and identified spike RBD mutations that abrogated binding and rendered the therapeutic antibodies ineffective[167,168,169]. Using these aggregated data sets, the authors generated a publicly

available antibody escape calculator that could be consulted to interpret the effects of new variants detected through surveillance sequencing[170]. The ACE2 binding and antibody escape data sets remain critical for assessing the monoclonal antibodies that are still used to treat immunosuppressed individuals and make important contributions to strain choice for vaccine updates[171]. Generating MAVE data for viral proteins represents a new paradigm for safely studying the evolution of viruses currently circulating in the human population[172,173,174,175] and those with pandemic potential[176,177,178,179].

Base-editing screens have been used to identify variants that affect cell fate and cell state in blood cells, which has led to insights into haematopoiesis[180] and promising avenues towards improved chimeric antigen receptor (CAR)-T immunotherapy[181,182]. These studies took advantage of advanced adenine base editors fused[183] to a Cas9 engineered to have a more flexible protospacer adjacent motif site[184] to increase editing efficiency and the density of edited variants at candidate loci. One study also included a cytosine base editor to increase the number of possible variants[181,185]. Another study identified splice variants that would eliminate CD33 expression in donor haematopoietic stem cells that could help develop a CAR-T immunotherapy against CD33-expressing acute myeloid leukaemia cells[186]. The other two studies targeted candidate genes that could potentially tune cytotoxic T cells to affect CAR-T function[181,182].

2.8 MAVEs and AI

MAVE data and artificial intelligence (AI) models have intersected in various ways, driven by the comprehensive, large-scale nature of the functional data. The recent advent of protein language models, which can rival the accuracy of the best MAVEs, has substantially contributed to this intersection and created a new set of opportunities[187,188,189]. AI models have

been used to empower MAVE studies by imputing missing data and combining multiple MAVE data sets; reciprocally, MAVE data have enabled benchmarking and training AI models (Fig. 6).

2.8.1 Improving MAVE data sets with AI

Increasingly, multiple MAVE data sets are available for clinically relevant targets. In some cases, multiple data sets are collected because one experimental assay cannot capture every important disease-related mechanism. For example, both loss-of-function and dominant-negative variants in TP53 can be pathogenic190. To capture this complexity, MAVEs of TP53 were conducted in cells that either had a wild-type copy of TP53 or were TP53 null — and also in the presence of nutlin-3, a key negative regulator of p53 stability[31]. In addition, TP53 loss of function was examined with an etoposide MAVE in the TP53 null cells. Although the loss-of-function MAVEs could broadly segregate nonsense and synonymous variants, the dynamic range of each assay was limited, making it difficult to partition missense variants on functional impact. Consequently, the authors built a generalized linear model to combine all three MAVEs with cancer mutational signatures from the COSMIC database[191] and yielded accurate predictions for the frequency of somatic TP53 mutations found in tumour samples.

MAVE data sets have also been combined to increase their ability to discriminate between known pathogenic and benign variants. One study combined four TP53 MAVEs with a naive Bayes classifier trained on ClinVar variants. This classifier improved the ability to discriminate clinically pathogenic and benign variants over any single assay and was calibrated for variant interpretation following ACMG recommendations[98]. In another study, multiple F9 MAVEs were combined with a random forest classifier to yield more accurate functional classifications and calibration for variant interpretation[96]. In both cases, model-based integration of the multiple data sets yielded improved separation of ClinVar pathogenic and benign variants.

Importantly, these models include only MAVE-derived functional scores as features and do not use features related to protein structure and conservation used by computational variant effect predictors. As these models merely transform functional data, they can be included in clinical variant interpretation workflows along with computational variant effect predictors, enabling the reclassification of VUS.

2.8.2 MAVE data sets for benchmarking and training AI models

The comprehensive nature of MAVE data is well suited to contribute to evaluating and creating computational models that predict variant effects. For example, the Critical Assessment of Genome Interpretation (CAGI) challenges teams to make specific sets of variant effect predictions that are then benchmarked against unpublished data[192], such as in 2016, when CAGI challenged teams to predict missense variant effects in UBE2I and then benchmarked predictions on a MAVE of UBE2I (ref. [29]). Subsequent cycles of CAGI have included various MAVE data sets[192]. Large collections of MAVE data sets have also been used to retrospectively assess variant effect predictor performance. For example, one study collected 31 previously published MAVE data sets on missense variant effects to compare the performance of 46 different variant effect predictors, enabling comparison of MAVE data and variant effect predictor performance on ClinVar pathogenic and benign variants[193]. Another study created ProteinGym to help assess AI models that predict variant effects by compiling several sources of variant effect data, including ~250 MAVE data sets, that can be compared with predictions[194]. The MAVE data in ProteinGym were used to benchmark a wide variety of variant effect predictors and, subsequently, newly published predictors have also used the data for benchmarking[187]. Models to predict non-coding variant effects have also been benchmarked using data derived from MPRAs, as was done to compare the Enformer model with other non-coding variant effect predictors[195].

MAVE data have also been used to train or refine variant effect predictors. For example, a study used MAVE-derived missense variant effect data from nine different proteins to train a variant effect predictor whose performance was slightly better than other contemporary predictors when predicting variant functional effects but inferior at predicting clinical variant effects[196]. Another study used MPRA data to train MpraNet, a model that can predict the functional consequences of non-coding variants. More recently, missense variant effect data from five proteins were used to train a protein language model, CPT-1, which performs well at predicting both clinical and functional variant effects[197]. MAVE-derived missense variant effect data from 30 proteins were also used to fine-tune the ESM1v protein language model to predict variant effects[198]. Finally, ultra-large-scale MAVE data sets have begun to be generated and used for model training, such as the ~250,000 variant effect measurements from a protease sensitivity assay that were used to train ThermoMPNN, a deep neural network that can predict variant stability effects[199,200].

2.9 Conclusions and future perspectives

In conclusion, MAVEs are having a major impact in the clinic, from improving clinical variant interpretation to leveraging variant pathomechanisms to develop therapeutic strategies. The main strengths of MAVE data — scale, comprehensiveness and uniformity — are well suited to a future in which increasing genome sequencing and genetic testing will continue to reveal more variants in patients. High-quality MAVE data can have considerable clinical impact on germline variant classification and, because MAVE data are unbiased with respect to ancestry, it could reduce disparities in VUS burden across populations[201]. Thus, MAVE data are already in high demand from clinicians and disease biologists[202], as comprehensive atlases of variant effects give prospective access to variant effect information for nearly all variants in

clinically relevant genes. Indeed, efforts to organize and scale up data production within the Atlas of Variant Effects Alliance and the NHGRI Impact of Genomic Variation on Function Consortium, along with the efforts of many individual laboratories, have markedly increased the availability of MAVE data in recent years[8,108].

MAVEs have substantial potential to grow from certain technological advancements. Although current technologies can produce clinically useful data, MAVEs are still labour-intensive and expensive. Thus, technologies that enable longer and cheaper DNA synthesis, base editing and prime editing, and cheaper long-read DNA sequencing all have the capacity to transform MAVEs by providing lower cost, higher scale approaches. Indeed, the recent appearance of ultra-large-scale MAVE data sets points to a future in which MAVE data are available for all human genes[200,203]. MAVEs are also still limited by the range of model systems and phenotypes that can be measured, which are limitations that particularly affect the study of genes and non-coding regions that must be studied in specific cellular contexts or measured by specialized phenotypes. To address these challenges, approaches that move MAVEs into induced pluripotent and human embryonic stem cell types[204] and multicellular model organisms such as organoids[205,206], along with single-cell imaging-based and omics-based phenotyping, could be transformative.

Another area for growth regards the lack of sufficient well-established benign and pathogenic control variants. The lack of control variants has hampered or precluded evaluation of assay sensitivity, specificity and predictive value for many genes. In addition to merely evaluating assay results, control variants are important for deploying AI to combine multiple MAVE data sets and for disentangling the relationship between the molecular and cellular phenotypes measured by MAVEs and different disease outcomes. We hope that a virtuous cycle

between clinicians and researchers executing MAVEs will result in additional control variants being classified. Case–control data from large cohort studies and biobanks are another source of control variant information that could help.

A final area for growth regards the difficulty in discovering and using MAVE data. Although MaveDB is a first step towards organizing and distributing MAVE data, it is designed for data generators and data scientists and thus lacks features needed by clinicians[207]. Additionally, a recent survey of clinical laboratory geneticists indicated that most were not comfortable with MAVE data regardless of years of experience, highlighting the need for clinician-focused infrastructure and training[208,209]. We foresee that the Atlas of Variant Effects Alliance, working in partnership with ClinGen, GA4GH and other entities, can foster both formal and information interactions between researchers and clinicians. The goals of these interactions should include development and refinement of standards, link data in MaveDB to the places that clinicians find variant information and provide needed training. Thus, having settled the questions of whether MAVE data can be generated and whether it can be clinically useful, the remaining work to do is to produce the data needed for a comprehensive atlas of variant effects, build the infrastructure required for clinical use of the data and create a community of clinicians and researchers working to translate the data into the clinic.

2.10 Figures

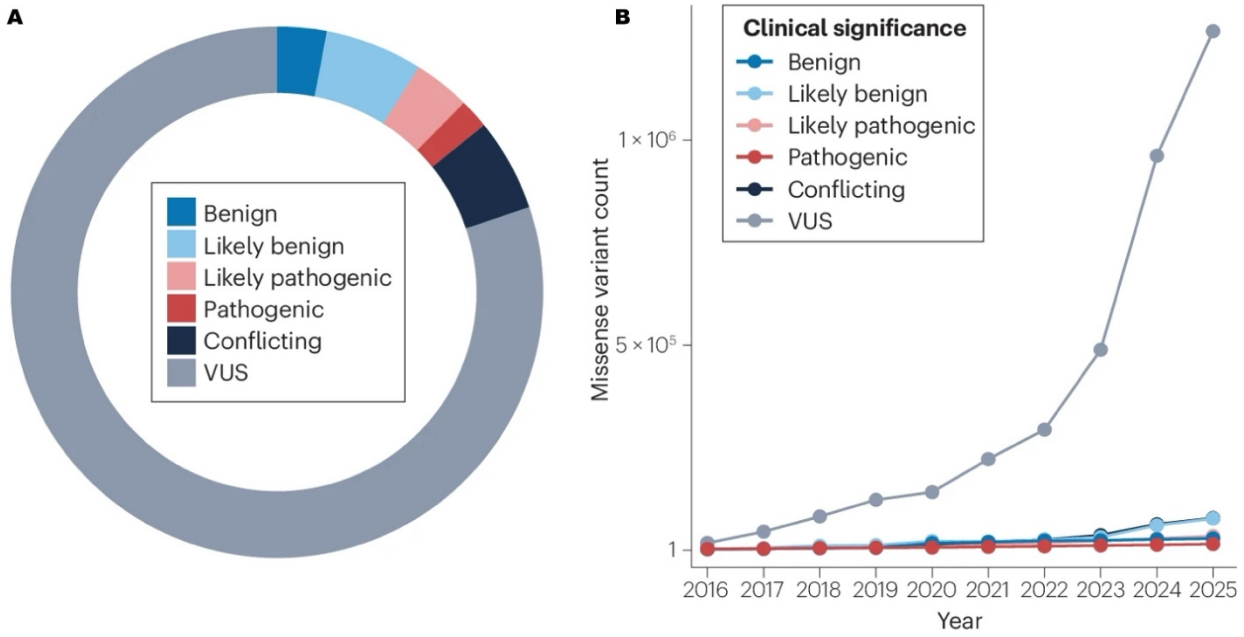


Figure 2.1: Variants of uncertain significance are an exponentially growing clinical problem

Single-nucleotide missense variants coloured by ClinVar classifications: benign (dark blue) = 28,616; likely benign (light blue) = 77,190; variants of uncertain significance (VUS) (grey) = 1,265,600; likely pathogenic (light red) = 34,378; pathogenic (dark red) = 14,985; conflicting interpretations (black) = 782,856. ClinVar data downloaded on 12 December 2024. b, Missense variants in ClinVar from 2015 to 2024, shown by clinical significance. See Supplementary information for more details on how ClinVar data were processed.

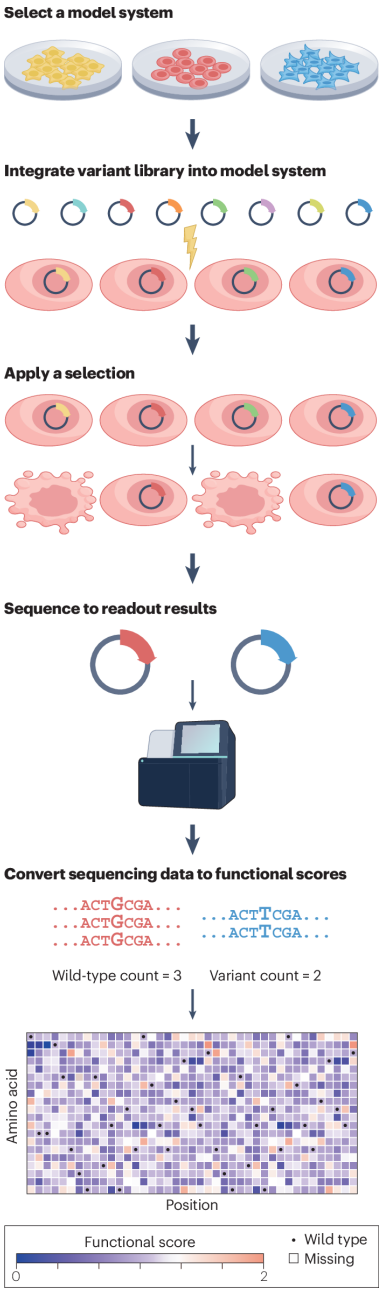


Figure 2.2 Overview of multiplexed assays of variant effect workflows.
After an appropriate model system is selected, a variant library is introduced into cells followed by phenotypic selection with a functional assay. Cells are then sequenced both before and after selection, and variant functional scores are computed to quantify changes in variant frequencies, reflecting their functional impact.

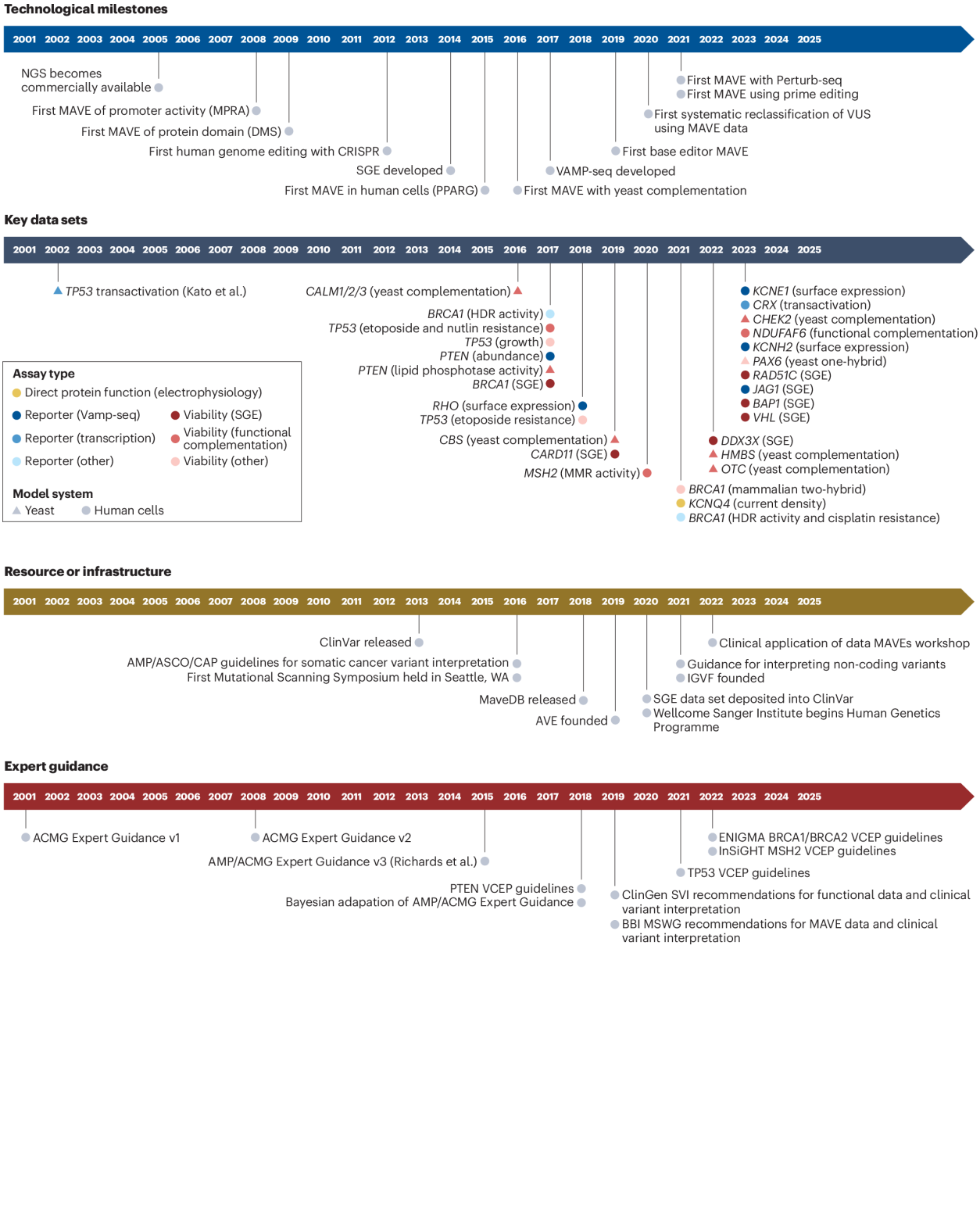

Technological milestones
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
NGS becomes commercially available
First MAVE of promoter activity (MPRA)
First MAVE of protein domain (DMS)
First human genome editing with CRISPR
SGE developed
First MAVE in human cells (PPARG)
VAMP-seq developed
First MAVE with yeast complementation
First base editor MAVE
First systematic reclassification of VUS using MAVE data
First MAVE with Perturb-seq
First MAVE using prime editing
Key data sets
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
TP53 transactivation (Kato et al.)
CALM1/2/3 (yeast complementation)
BRCA1 (HDR activity)
TP53 (etoposide and nutlin resistance)
TP53 (growth)
PTEN (abundance)
PTEN (lipid phosphotase activity)
BRCA1 (SGE)
RHO (surface expression)
TP53 (etoposide resistance)
CBS (yeast complementation)
CARD11 (SGE)
MSH2 (MMR activity)
KCNE1 (surface expression)
CRX (transactivation)
CHEK2 (yeast complementation)
NDUFAF6 (functional complementation)
KCNH2 (surface expression)
PAX6 (yeast one-hybrid)
RAD51C (SGE)
JAG1 (SGE)
BAP1 (SGE)
VHL (SGE)
DDX3X (SGE)
HMBS (yeast complementation)
OTC (yeast complementation)
BRCA1 (mammalian two-hybrid)
KCNQ4 (current density)
BRCA1 (HDR activity and cisplatin resistance)
Assay type
Direct protein function (electrophysiology)
Reporter (Vamp-seq)
Reporter (transcription)
Reporter (other)
Viability (SGE)
Viability (functional complementation)
Viability (other)
Model system
Yeast
Human cells
Resource or infrastructure
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
ClinVar released
AMP/ASCO/CAP guidelines for somatic cancer variant interpretation
First Mutational Scanning Symposium held in Seattle, WA
MaveDB released
AVE founded
SGE data set deposited into ClinVar
Wellcome Sanger Institute begins Human Genetics Programme
Guidance for interpreting non-coding variants
IGVF founded
Clinical application of data MAVEs workshop
Expert guidance
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
ACMG Expert Guidance v1
ACMG Expert Guidance v2
AMP/ACMG Expert Guidance v3 (Richards et al.)
PTEN VCEP guidelines
Bayesian adapation of AMP/ACMG Expert Guidance
ClinGen SVI recommendations for functional data and clinical variant interpretation
BBI MSWG recommendations for MAVE data and clinical variant interpretation
TP53 VCEP guidelines
ENIGMA BRCA1/BRCA2 VCEP guidelines
InSiGHT MSH2 VCEP guidelines

Figure 2.3 Timeline of key events in the use of multiplexed assay of variant effect data in clinical contexts.

This timeline depicts critical milestones in the application of multiplexed assay of variant effect (MAVE) data for clinical variant interpretation. It highlights key technological breakthroughs, including the development of MAVE assays, the establishment of resources and infrastructure to support education and data dissemination and the creation of expert guidance for variant interpretation. Also featured are major data sets that have been clinically calibrated, endorsed by expert panels or used to reclassify variants of uncertain significance, illustrating the progressive integration of MAVE data into clinical genetics and precision medicine. ACMG, American College of Medical Genetics; AMP, Association for Molecular Pathology; ASCO, American Society of Clinical Oncology; AVE, Atlas of Variant Effects Alliance; BBI-MSWG, Brotman Baty Institute-Mutational Scanning Working Group; CAP, College of American Pathologists; DMS, deep mutational scanning; HDR, homology-directed repair; IGVF, Impact of Genomic Variation on Function; MMR, mismatch repair; MPRA, massively parallel reporter assay; NGS, next-generation sequencing; SGE, saturation genome editing; SVI, sequence variant interpretation; VAMP-seq, variant abundance by massively parallel sequencing; VCEP, variant curation expert panel; VUS, variants of uncertain significance.

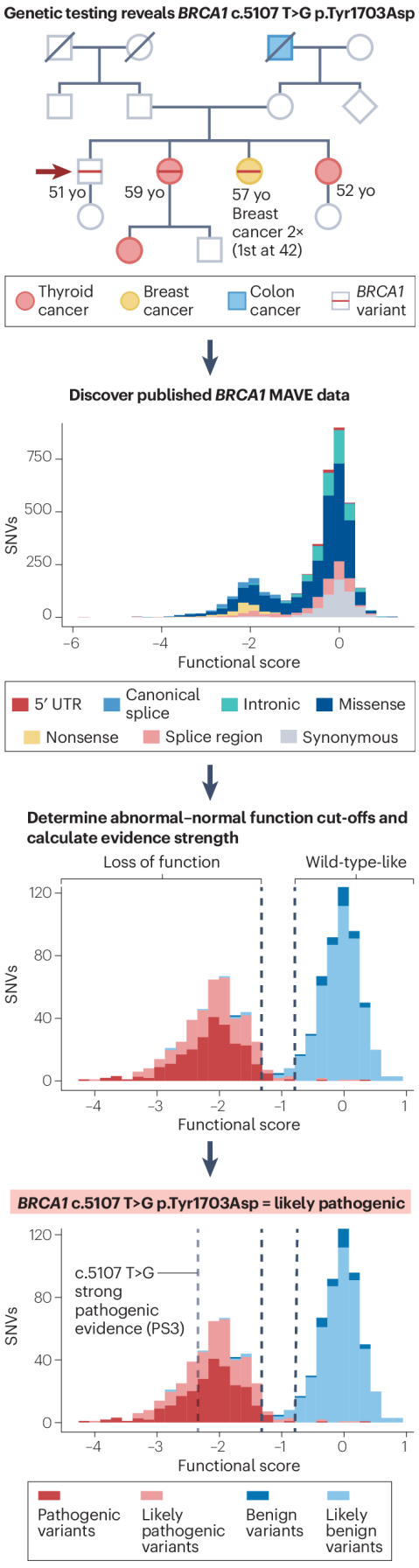


Figure 2.4 An example of multiplexed assays of variant effect data in action.

A BRCA1 variant of uncertain significance (c.5107 T>G, p.Tyr1703Asp, ClinVar Accession: SCV001244102.2) was identified in a 51-year-old (yo) male of European ancestry with a family history of multiple cancers. This variant was also identified in a 57-year-old sister, who had breast cancer twice, first at age 42, and a 59-year-old sister without history of cancer, who had a prophylactic mastectomy and oophorectomy. The effect of this variant was measured in a saturation genome editing multiplexed assay of variant effect (MAVE)[37]. Analysis of the MAVE data according to ClinGen Sequence Variant Interpretation guidelines[98] revealed that these MAVE data provide strong evidence for both benignness and pathogenicity. The patient's c.5107 T>G variant had a functional score in the loss-of-function range, comprising strong evidence of pathogenicity. With the addition of this functional evidence, this variant was reclassified as likely pathogenic. SNV, single-nucleotide variant; UTR, untranslated region.

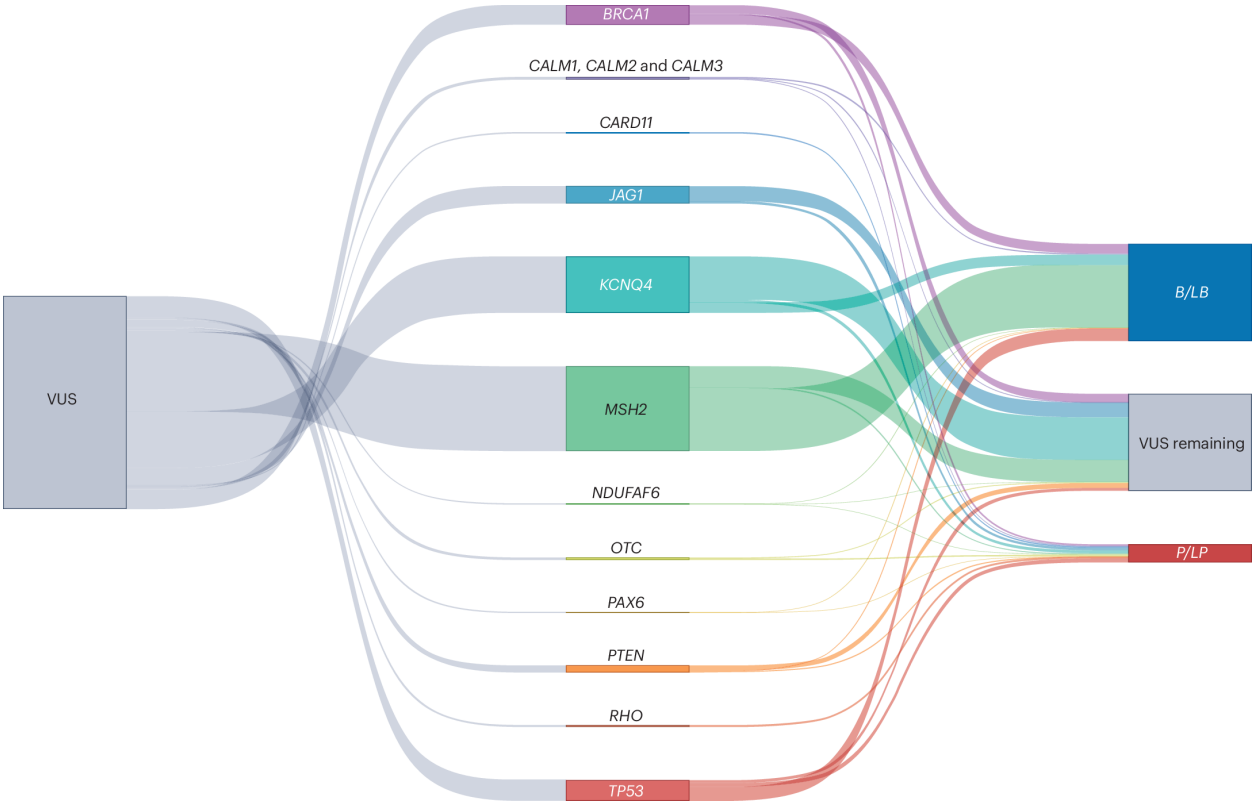


Figure 2.5 Reclassification of variants of uncertain significance with multiplexed assays of variant effect data.

This Sankey diagram depicts the previously published reclassification of variants of uncertain significance (VUS) into more definitive categories with the addition of multiplexed assays of variant effect data. A total of 1,711 VUS across 12 genes were reclassified in 12 published studies[29,66,73,80,98,100,102,103,105,213,214,215]. The data and R code used to make the Sankey diagram can be found within the Supplementary information and at https://github.com/FowlerLab/NRG_MAVEs_in_clinic.

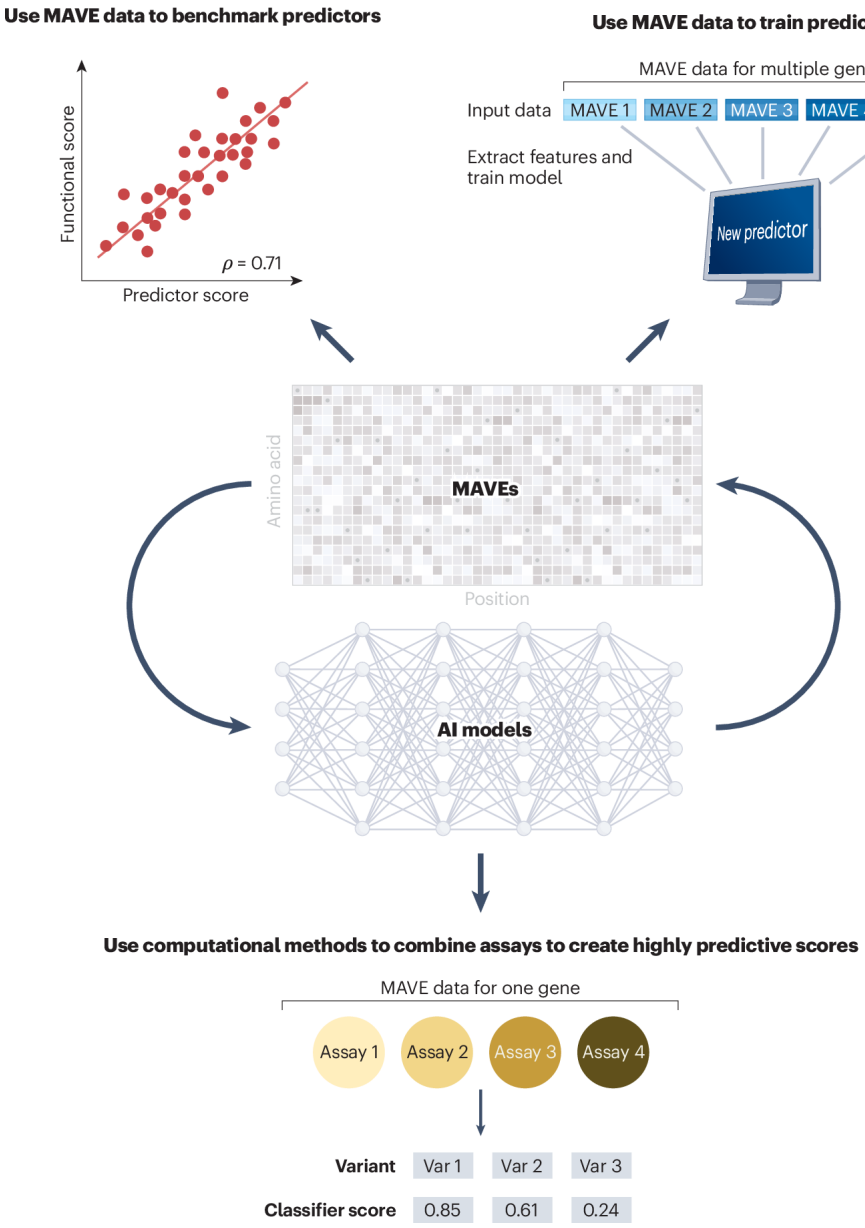


Figure 2.6 Multiplexed assays of variant effect data and artificial intelligence.

Multiplexed assays of variant effect (MAVE) data synergize with artificial intelligence (AI) models in various ways. For example, MAVE data can be used to benchmark variant effect predictors. Similarly, MAVE data sets can be used to train new variant effect prediction models. Finally, multiple MAVE data sets for the same gene can be combined to create a combined score with optimal performance.

2.11 Tables

2.11.1 Table 1: Clinical implications of assay design choices

Design choice	**Advantages**	**Limitations**
Model system		
Phage	Large library size Inexpensive to culture Can measure binding in vitro	Lack of eukaryotic protein-folding environment No post-translational modifications Limited protein size
Bacteria	Large library size Inexpensive to culture Extremely easy transgenesis and editing	Lack of eukaryotic protein folding environment No endogenous protein–protein or signaling interactions No post-translational modifications
Yeast	Large library size Inexpensive to culture Extremely easy transgenesis and editing Many conserved proteins–protein interactions	Lack of human protein folding environment Many protein–protein and signaling interactions missing or altered Some post-translational modifications missing or altered
Human cancer cell lines	Easy to culture Easy transgenesis and editing Preserves generic human protein folding environment Most protein–protein and signaling interactions and post-translational modifications	Some protein–protein interactions missing Many pathways disrupted or altered Generally lacks cell-type specific proteins or structures

Stem cell-derived specialized cells	Recapitulates specific human cell types	Expensive and difficult to culture Transgenesis is limited and editing is difficult
Library integration		
Exogenous expression	Allows for barcoding and protein tagging Allows for introduction of inducible expression, selectable markers and reporters Generally lower expense and effort than endogenous approaches	Splicing effects generally not detected Truncating variants do not undergo NMD Expression level and timing not endogenous
Transient expression (plasmid)	Does not require cell line engineering Ideal for MAVEs such as MPRAs where multiple variants can be assayed per cell	Difficult to titrate to one variant per cell Duration of expression cannot be controlled
Viral integration	Does not require upfront cell line engineering	Template switching precludes barcoded designs Expression level depends on integration site
Landing pad	Integration at single locus One variant per cell Enables barcoded designs	Required upfront cell line engineering Lower expression than other methods
Endogenous expression	Endogenous regulatory control avoids overexpression or expression timing artifacts Can detect canonical and cryptic splice variants Truncating variants are physiologic	May be limited by low editing efficiency Some methods require introduction of DNA breaks Harder to capture phenotypes other than growth Generally higher cost and effort

Haploid HDR editing	Editing rates are higher, and editing is simple (for example, no heterozygous versus homozygous edits) Especially amenable for the ~2,000 genes essential in well-established haploid human cell lines	Cannot detect dominant-negative variants Best results limited to essential genes Cell line choices are limited
Diploid HDR editing	Expanded choice of cell lines	Cell line engineering required to limit editing to single allele High false negative rate: loss of function variants may have functional copy at second allele
Base editing and prime editing	Multiple genomic regions can be assessed in a single experiment Lower cost and higher throughput than HDR-based editing assays	Sequencing pegRNA or sgRNA may not reflect actual edit made in cells Works best with positive selection
Phenotype selection		
Cell viability	Integrates multiple mechanisms of protein function Facile and scalable	Provides less mechanistic insight
Endogenous copy is knocked out (functional complementation)	Can be performed in yeast or mammalian cells Removes effect of endogenous or wild-type copies of target gene	Some disease-related genes do not have yeast orthologues Will not detect dominant-negative variants
Endogenous copy is edited (haploid SGE)	Many genes are essential in haploid human cell lines	Some disease-related genes are not essential in suitable cell lines Will not detect dominant-negative variants

Endogenous copy is intact	Will detect dominant-negative and gain-of-function variants	Will not detect loss-of-function variants
Reporter	Measures specific molecular or cellular phenotype Can provide mechanistic insight	Does not detect variants that affect other phenotypes
Protein abundance (VAMPseq)	Loss of protein stability is a common disease mechanism Generalizable to many proteins	Not all pathogenic variants affect protein stability Unsuitable for some proteins
Transactivation (MPRA)	Essential function for promoters and enhancers Generalizable and scalable	Only detects the ability to activate a single promoter or enhancer Capturing cell type-specific effects is challenging
Direct protein function	Measures specific protein function Provide the most granular mechanistic insight.	Does not detect variants that affect other phenotypes
Protein interaction	Many important interactions known	Not all pathogenic variants impact interactions
Enzymatic activity	Essential function for many proteins Integrates many aspects of function	Many proteins are not enzymes

HDR, homology-directed repair; MAVE, multiplexed assay of variant effect; MPRA, massively parallel reporter assay; NMD, nonsense-mediated mRNA decay; pegRNA, prime editing guide RNA; SGE, saturation genome editing; sgRNA, single guide RNA; VAMP-seq, variant abundance by massively parallel sequencing.

2.11.2 Table 2: Multiplexed assays of variant effect data sets that have been clinically calibrated or used to reclassify variants of uncertain significance

Gene (HGNC ID)	Assay name (Assay type)	Phenotypic read-out	Molecular mechanism assessed	Model system	Library type	Splice aware?	Evidence strength (Benign, pathogenic)	VUS reclassified
BAP1 (950)(Waters et al. 2024)	SGE (Cell viability)	Growth	Overall function	Immortalized human cells	Endogenous locus	Yes	Moderate, Moderate	NA
BRCA1 (1100)(Findlay et al. 2018),(Fayer et al. 2021)	SGE (Cell viability)	Growth	SGE	Immortalized human cells	Endogenous locus	Yes	Strong, Strong	54 of 110 (49%)
BRCA1 (1100)(Adamovich et al. 2022)	Other (Reporter)	Fluorescence	HDR	Immortalized human cells	Exogenous expression	No	Moderate, Moderate	56 of 66 (84%)
BRCA1 (1100)(Clark et al. 2022)	Other (Reporter)	Fluorescence	BARD1 binding	Immortalized human cells	Exogenous expression	No	Variant Specific	NA
BRCA2 (1101)(Sahu et al. 2023)	SGE (Cell viability)	Growth	Overall function	Mouse primary cells	Endogenous locus	Yes	Moderate, Moderate	NA
CALM1 (1442), ***CALM2*** (1445), ***CALM3*** (1449)(Weile et al. 2017),(Floyd et al. 2023)	Functional complementation (Cell viability)	Growth	Overall function	Yeast	Exogenous expression	No	Variant Specific	8 of 14 (57%)
CARD11 (16393)(Meitlis et al. 2020)	SGE (Cell viability)	Growth	Overall function	Immortalized human cells	Endogenous locus	Yes	Moderate, Moderate	6 of 6 (100%
CBS (1550)(Sun et al. 2020)	Functional complementation (Cell viability)	Growth	Overall function	Yeast	Exogenous expression	No	Variant Specific	NA
CHEK2 (16627)(Gebbia et al. 2024)	Functional complementation (Cell viability)	Growth	Overall function	Yeast	Exogenous expression	No	Variant Specific	NA
CRX (2383)(Shepherdson et al. 2024)	Other (Reporter)	Fluorescence	Transactivation	Immortalized human cells	Exogenous expression	No	None, Moderate	NA
DDX3X (2745)(Radford et al. 2023)	SGE (Cell viability)	Growth	Overall function	Immortalized human Ccells	Endogenous locus	Yes	Strong, Strong	NA
HMBS (4982)(van Loggerenberg et al. 2023)	Functional complementation (Cell viability)	Growth	Overall function	Yeast	Exogenous expression	No	Variant Specific	NA
JAG1 (6188)(Gilbert et al. 2024)	VAMPSeq (Reporter)	Fluorescence	Cell surface expression	immortalized mammalian cells	Exogenous expression	No	None, Strong	23 of 140 (16%)
KCNE1 (6240)(Muhammad et al. 2024)	VAMPSeq (Reporter)	Fluorescence	Cell surface expression	Immortalized human cells	Exogenous expression	No	Moderate, Moderate	NA

KCNH2 (6251)(Zheng et al. 2022)	VAMPSeq (Reporter)	Fluorescence	Cell surface expression	Immortalized human cells	Exogenous expression	No	Moderate, Strong	NA
KCNQ4 (6296)(O'Neill et al. 2024)	Other (Direct function)	Automated patch clamping	Ion channel function	Immortalized human cells	Exogenous expression	No	Strong, Strong	112 of 456 (25%)
MSH2 (7325)(Jia et al. 2021),(Scott et al. 2022)	Functional complementation (Cell viability)	Growth	MMR activity	Immortalized human cells	Exogenous expression	No	Strong, Strong	507 of 682 (74%)
NDUFAF6 (28625)(Sung et al. 2024)	Functional complementation (Cell viability)	Growth	Overall function	Immortalized human cells	Exogenous expression	No	Moderate, Moderate	3 of 5 (60%)
OTC (8512)(Lo et al. 2023)	Functional complementation (Cell viability)	Growth	Overall function	Yeast	Exogenous expression	No	Not Calibrated	10 of 16 (63%)
PAX6 (8620)(McDonnell et al. 2024)	Other (Cell viability)	Growth	Transactivation	Yeast	Exogenous expression	No	Moderate, Moderate	3 of 3 (100%)
PTEN (9588)(Mighell, Evans-Dutson, and O'Roak 2018),(Fayer et al. 2021)	Other (Cell viability)	Growth	Lipid phosphatase activity	Yeast	Exogenous expression	No	Moderate, None	NA
PTEN (9588)(Matreyek et al. 2018),(Fayer et al. 2021)	VAMPSeq (Reporter)	Fluorescence	Protein abundance	Immortalized human cells	Exogenous expression	No	Moderate, None	NA
PTEN (9588)(Mighell, Evans-Dutson, and O'Roak 2018; Matreyek et al. 2018),(Fayer et al. 2021)	Multiple assays combined	Fluorescence, Growth	Protein abundance, lipid phosphatase activity	Immortalized human cells, Yeast	Exogenous expression	No	Per VCEP recommendations	7 of 48 (15%)
RAD51C (9820)(Olvera-León et al. 2024)	SGE (Cell viability)	Growth	Overall function	Immortalized human cells	Endogenous locus	Yes	Moderate, Supporting	NA
RHO (10012)(Wan et al. 2019)	VAMPSeq (Reporter)	Fluorescence	Cell surface expression	Immortalized human cells	Exogenous expression	No	Not calibrated	4 of 15 (27%)
TP53 (11998)(Kato et al. 2003),(Fortuno, Pesaran, et al. 2021)	Other (Reporter)	Fluorescence	Transactivation	Yeast	Exogenous expression	No	Strong, Strong	NA
TP53 (11998)(Giacomelli et al. 2018),(Fayer et al. 2021)	Other (Cell viability)	Growth	Nutlin resistance	Immortalized human cells	Exogenous expression	No	Moderate, Moderate	NA
TP53 (11998)(Giacomelli et al. 2018),(Fayer et al. 2021)	Other (Cell viability)	Growth	Etoposide resistance	Immortalized human cells	Exogenous expression	No	Moderate, Moderate	NA
TP53 (11998)(Giacomelli et al. 2018; Boettcher et al. 2019),(Fayer et al. 2021)	Multiple assays combined	Cell viability	Etoposide resistance, Nutlin resistance	Immortalized human cells	Exogenous expression	No	Moderate, Strong	115 of 166 (69%)
VHL (12687)(Buckley et al. 2024)	SGE (Cell viability)	Growth	Overall function	Immortalized human cells	Endogenous locus	Yes	Moderate, Strong	NA

Genes that are on the secondary findings list of American College of Medical Genetics101 are bolded. HDR, homology-directed repair; MMR, mismatch repair; NA, not available; SGE, saturation genome editing; VUS, variants of uncertain significance.

2.12 References

1. Chen, E. *et al.* Rates and classification of variants of uncertain significance in hereditary disease genetic testing. *JAMA Netw. Open* **6**, e2339571 (2023).
2. Rehm, H. L. *et al.* The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. *Genet. Med.* **25**, 100947 (2023).
3. LaDuca, H. *et al.* 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.* **22**, 407–415 (2020).
4. Ndugga-Kabuye, M. K. & Issaka, R. B. Inequities in multi-gene hereditary cancer testing: lower diagnostic yield and higher VUS rate in individuals who identify as Hispanic, African or Asian and Pacific Islander as compared to European. *Fam. Cancer* **18**, 465–469 (2019).
5. Roberts, M. E. *et al.* Ancestry-specific hereditary cancer panel yields: Moving toward more personalized risk assessment. *J. Genet. Couns.* **29**, 598–606 (2020).
6. Kwon, D. H.-M., Borno, H. T., Cheng, H. H., Zhou, A. Y. & Small, E. J. Ethnic disparities among men with prostate cancer undergoing germline testing. *Urol. Oncol.* **38**, 80.e1–80.e7 (2020).
7. 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).

8. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).
9. Gelman, H. *et al.* Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med.* **11**, 85 (2019).
10. 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).
11. Criteria Specification Registry. https://cspec.genome.network/cspec/ui/svi/.
12. Findlay, G. M. Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Hum. Mol. Genet.* **30**, R187–R197 (2021).
13. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.* **137**, 665–678 (2018).
14. Tabet, D., Parikh, V., Mali, P., Roth, F. P. & Claussnitzer, M. Scalable functional assays for the interpretation of human genetic variation. *Annu. Rev. Genet.* **56**, 441–465 (2022).
15. Geck, R. C., Boyle, G., Amorosi, C. J., Fowler, D. M. & Dunham, M. J. Measuring pharmacogene variant function at scale using multiplexed assays. *Annu. Rev. Pharmacol. Toxicol.* **62**, 531–550 (2022).
16. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
17. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* **6**, 1782–1790 (2010).
18. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E1263–72 (2013).

19. Adkar, B. V. *et al.* Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* **20**, 371–381 (2012).
20. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13067–13072 (2013).
21. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
22. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7896–7901 (2011).
23. Dutta, S. *et al.* Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. *J. Mol. Biol.* **398**, 747–762 (2010).
24. Traxlmayr, M. W. *et al.* Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *J. Mol. Biol.* **423**, 397–412 (2012).
25. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31**, 1956–1978 (2014).
26. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
27. Wagenaar, T. R. *et al.* Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res.* **27**, 124–133 (2014).
28. Chan, K. K. *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
29. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).

30. Kotler, E. *et al.* A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. *Mol. Cell* **71**, 178–190.e8 (2018).
31. Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).
32. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
33. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation Mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
34. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
35. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
36. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
37. 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).
38. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
39. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

40. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
41. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
42. Ozturk, K. *et al.* Interface-guided phenotyping of coding variants in the transcription factor RUNX1. *Cell Rep.* **43**, 114436 (2024).
43. Ursu, O. *et al.* Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat. Biotechnol.* **40**, 896–905 (2022).
44. Xu, H. *et al.* Single cell sequencing as a general variant interpretation assay. *Genomics* (2023).
45. Hasle, N. *et al.* High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Mol. Syst. Biol.* **16**, e9442 (2020).
46. Kato, S. *et al.* 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 (2003).
47. Monteiro, A. N., August, A. & Hanafusa, H. Evidence for a transcriptional activation function of BRCA1 C-terminal region. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13595–13599 (1996).
48. Monteiro, A. N., August, A. & Hanafusa, H. Common BRCA1 variants and transcriptional activation. *Am. J. Hum. Genet.* **61**, 761–762 (1997).
49. Vallon-Christersson, J. *et al.* Functional analysis of BRCA1 C-terminal missense mutations identified in breast and ovarian cancer families. *Hum. Mol. Genet.* **10**, 353–360 (2001).
50. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008).

51. Kazazian, H. H., Boehm, C. D. & Seltzer, W. K. ACMG recommendations for standards for interpretation of sequence variations. *Genet. Med.* **2**, 302–303 (2000).
52. Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008 (2010).
53. Soussi, T., Kato, S., Levy, P. P. & Ishioka, C. Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations. *Hum. Mutat.* **25**, 6–17 (2005).
54. Iacopetta, B. *et al.* Functional categories of TP53 mutation in colorectal cancer: results of an International Collaborative Study. *Ann. Oncol.* **17**, 842–847 (2006).
55. Fortuno, C. *et al.* Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum. Mutat.* **42**, 223–236 (2021).
56. McLaughlin, R. N., Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).
57. Gajula, K. S. *et al.* High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase. *Nucleic Acids Res.* **42**, 9964–9975 (2014).
58. Doolan, K. M. & Colby, D. W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *J. Mol. Biol.* **427**, 328–340 (2015).
59. Starita, L. M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
60. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).

61. Starita, L. M. *et al.* Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

62. van Loggerenberg, W. *et al.* Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *Am. J. Hum. Genet.* **110**, 1769–1786 (2023).

63. Lo, R. S. *et al.* The functional impact of 1,570 individual amino acid substitutions in human OTC. *Am. J. Hum. Genet.* **110**, 863–879 (2023).

64. Weile, J. *et al.* Shifting landscapes of human MTHFR missense-variant effects. *Am. J. Hum. Genet.* **108**, 1283–1300 (2021).

65. Ollodart, A. R. *et al.* Multiplexing mutation rate assessment: determining pathogenicity of Msh2 variants in Saccharomyces cerevisiae. *Genetics* **218**, (2021).

66. Sun, S. *et al.* A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med.* **12**, 13 (2020).

67. Silverstein, R. A. *et al.* A systematic genotype-phenotype map for missense variants in the human intellectual disability-associated gene*GDI1*. *bioRxiv* (2021) doi:10.1101/2021.10.06.463360.

68. Gersing, S. *et al.* A comprehensive map of human glucokinase variant activity. *Genome Biol.* **24**, 97 (2023).

69. Jia, X. *et al.* Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108**, 163–175 (2021).

70. Sung, A. Y. *et al.* Systematic analysis of NDUFAF6 in complex I assembly and mitochondrial disease. *Nat. Metab.* **6**, 1128–1142 (2024).

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

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

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

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

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

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

77. Meitlis, I. *et al.* Multiplexed functional assessment of genetic variants in CARD11. *Am. J. Hum. Genet.* **107**, 1029–1043 (2020).

78. Sahu, S. *et al.* Saturation genome editing of 11 codons and exon 13 of BRCA2 coupled with chemotherapeutic drug response accurately determines pathogenicity of variants. *PLoS Genet.* **19**, e1010940 (2023).

79. Huang, H. *et al.* Saturation genome editing-based functional evaluation and clinical classification of BRCA2 single nucleotide variants. *Genomics* (2023).

80. Funk, J. *et al.* Functional diversity of the TP53 mutome revealed by saturating CRISPR mutagenesis. (2023) doi:10.1101/2023.03.10.531074.

81. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).

82. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein destabilization as a common factor in diverse inherited disorders. *J. Mol. Evol.* **82**, 11–16 (2016).

83. Amorosi, C. J. *et al.* Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).
84. Boyle, G. E. *et al.* Deep mutational scanning of CYP2C19 in human cells reveals a substrate specificity-abundance tradeoff. *Genetics* **228**, (2024).
85. Chiasson, M. A. *et al.* Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**, (2020).
86. Suiter, C. C. *et al.* Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5394–5401 (2020).
87. Clausen, L. *et al.* A mutational atlas for Parkin proteostasis. *Nat. Commun.* **15**, 1541 (2024).
88. Muhammad, A. *et al.* High-throughput functional mapping of variants in an arrhythmia gene, KCNE1, reveals novel biology. *Genome Med.* **16**, 73 (2024).
89. O'Neill, M. J. *et al.* Multiplexed assays of variant effect and automated patch clamping improve KCNH2-LQTS variant classification and cardiac event risk stratification. *Circulation* **150**, 1869–1881 (2024).
90. Kozek, K. A. *et al.* High-throughput discovery of trafficking-deficient variants in the cardiac potassium channel KV11.1. *Heart Rhythm* **17**, 2180–2189 (2020).
91. Grønbæk-Thygesen, M. *et al.* Deep mutational scanning reveals a correlation between degradation and toxicity of thousands of aspartoacylase variants. *Nat. Commun.* **15**, 4026 (2024).
92. Yee, S. W. *et al.* The full spectrum of SLC22 OCT1 mutations illuminates the bridge between drug transporter biophysics and pharmacogenomics. *Mol. Cell* **84**, 1932–1947.e10 (2024).
93. Popp, N. A. *et al.* Multiplex, multimodal mapping of variant effects in secreted proteins. *Genomics* (2024).

94. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).
95. 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).
96. Scott, A. *et al.* Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. *Genome Biol.* **23**, 266 (2022).
97. Miller, D. T. *et al.* 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 (2023).
98. Adamovich, A. I. *et al.* The functional impact of BRCA1 BRCT domain variants using multiplexed DNA double-strand break repair assays. *Am. J. Hum. Genet.* **109**, 618–630 (2022).
99. Floyd, B. J. *et al.* Proactive variant effect mapping aids diagnosis in pediatric cardiac arrest. *Circ. Genom. Precis. Med.* **16**, e003792 (2023).
100. Boettcher, S. *et al.* A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science* **365**, 599–604 (2019).
101. Zheng, H. *et al.* Proactive functional classification of all possible missense single-nucleotide variants in KCNQ4. *Genome Res.* **32**, 1573–1584 (2022).
102. Glazer, A. M. *et al.* Deep mutational scan of an SCN5A voltage sensor. *Circ. Genom. Precis. Med.* **13**, e002786 (2020).
103. Rubin, A. F. *et al.* MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. *Genomics* (2021).

104. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).

105. Claussnitzer, M. *et al.* Minimum information and guidelines for reporting a multiplexed assay of variant effect. *Genome Biol.* **25**, 100 (2024).

106. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–10 (2015).

107. French, J. D. & Edwards, S. L. The role of noncoding variants in heritable disease. *Trends Genet.* **36**, 880–891 (2020).

108. Mattioli, K. *et al.* High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res.* **29**, 344–355 (2019).

109. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

110. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).

111. Zhao, W. *et al.* Massively parallel functional annotation of 3' untranslated regions. *Nat. Biotechnol.* **32**, 387–391 (2014).

112. Birnbaum, R. Y. *et al.* Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genet.* **10**, e1004592 (2014).

113. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).

114. Wong, M. S., Kinney, J. B. & Krainer, A. R. Quantitative activity profile and context dependence of all human 5' splice sites. *Mol. Cell* **71**, 1012–1026.e3 (2018).

115. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).

116. Oikonomou, P., Goodarzi, H. & Tavazoie, S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Rep.* **7**, 281–292 (2014).

117. Sample, P. J. *et al.* Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).

118. Mueller, W. F., Larsen, L. S. Z., Garibaldi, A., Hatfield, G. W. & Hertel, K. J. The silent sway of splicing by synonymous substitutions. *J. Biol. Chem.* **290**, 27700–27711 (2015).

119. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).

120. Ke, S. *et al.* Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).

121. Chong, R. *et al.* A multiplexed assay for Exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Mol. Cell* **73**, 183–194.e8 (2019).

122. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).

123. Braun, S. *et al.* Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* **9**, 3315 (2018).

124. Safra, M., Nir, R., Farouq, D., Vainberg Slutskin, I. & Schwartz, S. TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res.* **27**, 393–406 (2017).

125. Liu, X. *et al.* Learning cis-regulatory principles of ADAR-based RNA editing from CRISPR-mediated mutagenesis. *Nat. Commun.* **12**, 2165 (2021).

126. Shukla, C. J. *et al.* High-throughput identification of RNA nuclear enrichment sequences. *EMBO J.* **37**, (2018).

127. McQuerry, J. A. *et al.* Massively parallel identification of functionally consequential noncoding genetic variants in undiagnosed rare disease patients. *Sci. Rep.* **12**, 7576 (2022).

128. Rhine, C. L. *et al.* Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. *PLoS Genet.* **18**, e1009884 (2022).

129. Ellingford, J. M. *et al.* Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* **14**, 73 (2022).

130. Li, M. M. *et al.* Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).

131. Horak, P. *et al.* Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): Joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for Cancer Consortium (VICC). *Genet. Med.* **24**, 986–998 (2022).

132. Johnson, A. *et al.* Clinical use of Precision Oncology Decision Support. *JCO Precis. Oncol.* **2017**, 1–12 (2017).

133. Dogruluk, T. *et al.* Identification of variant-specific functions of PIK3CA by rapid phenotyping of rare mutations. *Cancer Res.* **75**, 5341–5354 (2015).

134. Narayan, P. *et al.* FDA approval summary: Alpelisib plus fulvestrant for patients with HR-positive, HER2-negative, PIK3CA-mutated, advanced or metastatic breast cancer. *Clin. Cancer Res.* **27**, 1842–1849 (2021).

135. de Bruijn, I. *et al.* Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).

136. Bridgford, J. L. *et al.* Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood* **135**, 287–292 (2020).

137. Estevam, G. O. *et al.* Conserved regulatory motifs in the juxtamembrane domain and kinase N-lobe revealed through deep mutational scanning of the MET receptor tyrosine kinase domain. *Elife* **12**, (2024).

138. Bandaru, P. *et al.* Deconstruction of the Ras switching cycle through saturation mutagenesis. *Elife* **6**, (2017).

139. Cantor, A. J., Shah, N. H. & Kuriyan, J. Deep mutational analysis reveals functional trade-offs in the sequences of EGFR autophosphorylation sites. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7303–E7312 (2018).

140. Belli, O., Karava, K., Farouni, R. & Platt, R. J. Multimodal scanning of genetic variants with base and prime editing. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02439-1.

141. Kohsaka, S. *et al.* A method of high-throughput functional evaluation of EGFR gene variants of unknown significance in cancer. *Sci. Transl. Med.* **9**, (2017).

142. Kim, Y., Oh, H.-C., Lee, S. & Kim, H. H. Saturation profiling of drug-resistant genetic variants using prime editing. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02465-z.

143. Chardon, F. M. *et al.* A multiplex, prime editing framework for identifying drug resistance variants at scale. *bioRxiv* (2023) doi:10.1101/2023.07.27.550902.

144. Ma, L. *et al.* CRISPR-Cas9-mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11751–11756 (2017).

145. Chakraborty, S. *et al.* Profiling of drug resistance in Src kinase at scale uncovers a regulatory network coupling autoinhibition and catalytic domain dynamics. *Cell Chem. Biol.* **31**, 207–220.e11 (2024).

146. Awad, M. M. *et al.* Acquired resistance to KRASG12C inhibition in cancer. *N. Engl. J. Med.* **384**, 2382–2393 (2021).

147. Persky, N. S. *et al.* Defining the landscape of ATP-competitive inhibitor resistance residues in protein kinases. *Nat. Struct. Mol. Biol.* **27**, 92–104 (2020).

148. Kim, J. *et al.* A framework for individualized splice-switching oligonucleotide therapy. *Nature* **619**, 828–836 (2023).

149. McKee, A. G. *et al.* General trends in the effects of VX-661 and VX-445 on the plasma membrane expression of clinical CFTR variants. *Cell Chem. Biol.* **30**, 632–642.e5 (2023).

150. Howard, C. J. *et al.* High resolution deep mutational scanning of the melanocortin-4 receptor enables target characterization for drug discovery. *Synthetic Biology* (2024).

151. Mathy, C. J. P. *et al.* A complete allosteric map of a GTPase switch in its native cellular network. *Cell Syst.* **14**, 237–246.e7 (2023).

152. Tack, D. S. *et al.* The genotype-phenotype landscape of an allosteric protein. *Mol. Syst. Biol.* **17**, e10179 (2021).

153. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

154. Leander, M., Yuan, Y., Meger, A., Cui, Q. & Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25445–25454 (2020).

155. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).

156. Mason, D. M. *et al.* High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. *Nucleic Acids Res.* **46**, 7436–7449 (2018).

157. Wollacott, A. M. *et al.* Structural prediction of antibody-APRIL complexes by computational docking constrained by antigen saturation mutagenesis library data. *J. Mol. Recognit.* **32**, e2778 (2019).

158. Koenig, P., Sanowar, S., Lee, C. V. & Fuh, G. Tuning the specificity of a Two-in-One Fab against three angiogenic antigens by fully utilizing the information of deep mutational scanning. *MAbs* **9**, 959–967 (2017).

159. Renn, A., Fu, Y., Hu, X., Hall, M. D. & Simeonov, A. Fruitful neutralizing antibody pipeline brings hope to defeat SARS-CoV-2. *Trends Pharmacol. Sci.* **41**, 815–829 (2020).

160. Mengist, H. M. *et al.* Mutations of SARS-CoV-2 spike protein: Implications on immune evasion and vaccine-induced immunity. *Semin. Immunol.* **55**, 101533 (2021).

161. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020).

162. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021).

163. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep. Med.* **2**, 100255 (2021).

164. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021).

165. Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evol.* **8**, veac021 (2022).

166. Dadonaite, B. *et al.* Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature* **631**, 617–626 (2024).

167. Li, Y., Arcos, S., Sabsay, K. R., Te Velthuis, A. J. W. & Lauring, A. S. Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza A virus PB1 protein. *J. Virol.* **97**, e0132923 (2023).

168. Kikawa, C. *et al.* The effect of single mutations in Zika virus envelope on escape from broadly neutralizing antibodies. *J. Virol.* **97**, e0141423 (2023).

169. Radford, C. E. *et al.* Mapping the neutralizing specificity of human anti-HIV serum by deep mutational scanning. *Cell Host Microbe* **31**, 1200–1215.e9 (2023).

170. Lei, R. *et al.* Mutational fitness landscape of human influenza H3N2 neuraminidase. *Cell Rep.* **42**, 111951 (2023).

171. Soh, Y. S., Moncla, L. H., Eguia, R., Bedford, T. & Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *Elife* **8**, (2019).

172. Dadonaite, B. *et al.* Deep mutational scanning of H5 hemagglutinin to inform influenza virus surveillance. *PLoS Biol.* **22**, e3002916 (2024).

173. Larsen, B. B. *et al.* Functional and antigenic landscape of the Nipah virus receptor binding protein. *bioRxivorg* (2024) doi:10.1101/2024.04.17.589977.

174. Carr, C. R. *et al.* Deep mutational scanning reveals functional constraints and antibody-escape potential of Lassa virus glycoprotein complex. *Immunity* **57**, 2061–2076.e11 (2024).

175. Martin-Rufino, J. D. *et al.* Massively parallel base editing to map variant effects in human hematopoiesis. *Cell* **186**, 2456–2474.e24 (2023).

176. Schmidt, R. *et al.* Base-editing mutagenesis maps alleles to tune human T cell functions. *Nature* **625**, 805–812 (2024).

177. Walsh, Z. H. *et al.* Mapping variant effects on anti-tumor hallmarks of primary human T cells with base-editing screens. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02235-x.

178. Richter, M. F. *et al.* Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).

179. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).

180. Thuronyi, B. W. *et al.* Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* **37**, 1070–1079 (2019).

181. Kim, M. Y. *et al.* Genetic inactivation of CD33 in hematopoietic stem cells to enable CAR T cell immunotherapy for acute myeloid leukemia. *Cell* **173**, 1439–1453.e19 (2018).

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

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

184. Orenbuch, R. *et al.* Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. *medRxiv* (2023) doi:10.1101/2023.11.27.23299062.

185. Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).

186. Critical Assessment of Genome Interpretation Consortium. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. *Genome Biol.* **25**, 53 (2024).

187. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).

188. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *Synthetic Biology* (2023).

189. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

190. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative missense variant effect prediction using large-scale Mutagenesis data. *Cell Syst.* **6**, 116–124.e3 (2018).

191. Jagota, M. *et al.* Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biol.* **24**, 182 (2023).

192. Lafita, A. *et al.* Fine-tuning protein Language Models with Deep Mutational Scanning improves variant effect prediction. *arXiv [q-bio.GN]* (2024).

193. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2314853121 (2024).

194. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).

195. Dawood, M. *et al.* Using multiplexed functional data to reduce variant classification inequities in underrepresented populations. *Genome Med.* **16**, 143 (2024).

196. Kuang, D. *et al.* MaveRegistry: a collaboration platform for multiplexed assays of variant effect. *Bioinformatics* **37**, 3382–3383 (2021).

197. Beltran, A., Jiang, X. ’er, Shen, Y. & Lehner, B. Site-saturation mutagenesis of 500 human protein domains. *Nature* **637**, 885–894 (2025).

198. Kreimer, A. *et al.* Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* **13**, 1504 (2022).

199. Capauto, D. *et al.* Characterization of enhancer activity in early human neurodevelopment using Massively Parallel Reporter Assay (MPRA) and forebrain organoids. *Sci. Rep.* **14**, 3936 (2024).

200. Deng, C. *et al.* Massively parallel characterization of regulatory elements in the developing human cortex. *Science* **384**, eadh0559 (2024).

201. Allen, S. *et al.* Workshop report: the clinical application of data from multiplex assays of variant effect (MAVEs), 12 July 2023. *Eur. J. Hum. Genet.* **32**, 593–600 (2024).

202. Villani, R. M. *et al.* Consultation informs strategies to improve functional evidence use in variant classification. *medRxiv* (2024) doi:10.1101/2024.12.04.24318523.

203. Park, M. S. *et al.* Insights on improving accessibility and usability of functional data to unlock its potential for variant interpretation. *Genetic and Genomic Medicine* (2025).

204. Clark, K. A. *et al.* Comprehensive evaluation and efficient classification of BRCA1 RING domain missense substitutions. *Am. J. Hum. Genet.* **109**, 1153–1174 (2022).

205. Gebbia, M. *et al.* A missense variant effect map for the human tumor-suppressor protein CHK2. *Am. J. Hum. Genet.* **111**, 2675–2692 (2024).

206. Shepherdson, J. L. *et al.* Mutational scanning of CRX classifies clinical variants and reveals biochemical properties of the transcriptional effector domain. *Genome Res.* **34**, 1540–1552 (2024).

207. Gilbert, M. A. *et al.* Functional characterization of 2,832 JAG1 variants supports reclassification for Alagille syndrome and improves guidance for clinical variant interpretation. *Am. J. Hum. Genet.* **111**, 1656–1672 (2024).

208. McDonnell, A. F. *et al.* Deep mutational scanning quantifies DNA binding and predicts clinical outcomes of PAX6 variants. *Mol. Syst. Biol.* **20**, 825–844 (2024).

209. Wan, A., Place, E., Pierce, E. A. & Comander, J. Characterizing variants of unknown significance in rhodopsin: A functional genomics approach. *Hum. Mutat.* **40**, 1127–1144 (2019).

210. Fortuno, C. *et al.* An updated quantitative model to classify missense variants in the TP53 gene: A novel multifactorial strategy. *Hum. Mutat.* **42**, 1351–1361 (2021).

Chapter 3: Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes

This chapter is adapted with minimal modifications from:

Tejura M, *Fayer S, McEwen AE, Flynn J, Starita LM, Fowler DM. Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. Am J Hum Genet. 2024;111(9):2031-2043.*

3.1 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.

3.2 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.

3.3 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**).

$x_{incorrect}$

$= number\ of\ incorrectly\ predicted\ variants\ in\ the\ evidence\ strength\ interval$

$n = minimum\ number\ of\ control\ variants\ needed\ in\ the\ evidence\ strength\ interval$

$p_{incorrect}$

$= probability\ of\ an\ incorrectly\ predicted\ variant\ (e.g. incorrect\ prediction\ tolerance)$

$$P(x_{incorrect} \geq 1) = 1 - P(x = 0)$$

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

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

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**).

$x_{correct} = number\ of\ correctly\ predicted\ variants\ in\ the\ evidence\ strength\ interval$

$n = minimum\ number\ of\ control\ variants\ needed\ in\ the\ evidence\ strength\ interval$

$(n - x_{correct}) =$

$number\ of\ incorrectly\ predicted\ variants\ in\ evidence\ strength\ interval$

$p_{correct} = probability\ of\ a\ correctly\ predicted\ variant\ or\ (1 - incorrect\ prediction\ tolerance)$

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

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

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

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.

3.4 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.

3.5 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.

Declaration of Interests The authors declare no competing interests.

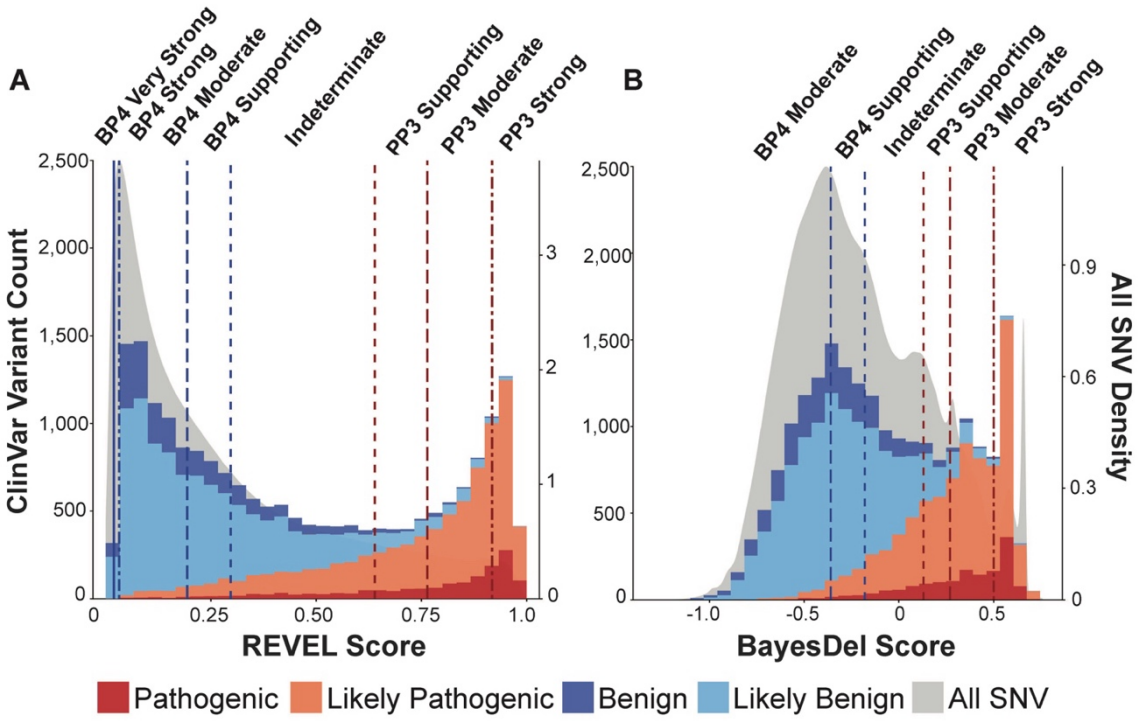
Acknowledgments We would like to thank members of the Fowler and Starita labs, Dr. Andrew B. Stergachis, and Dr. Brian H. Shirts for helpful feedback. This work was supported by the NHGRI Center for Multiplexed Assessment of Phenotype RM1HG010461 (DMF, LS, AEM, SF, MT), Center for Actionable Variant 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).

Web Resources OMIM, http://www.omim.org

3.6 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.

3.7 Figures



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

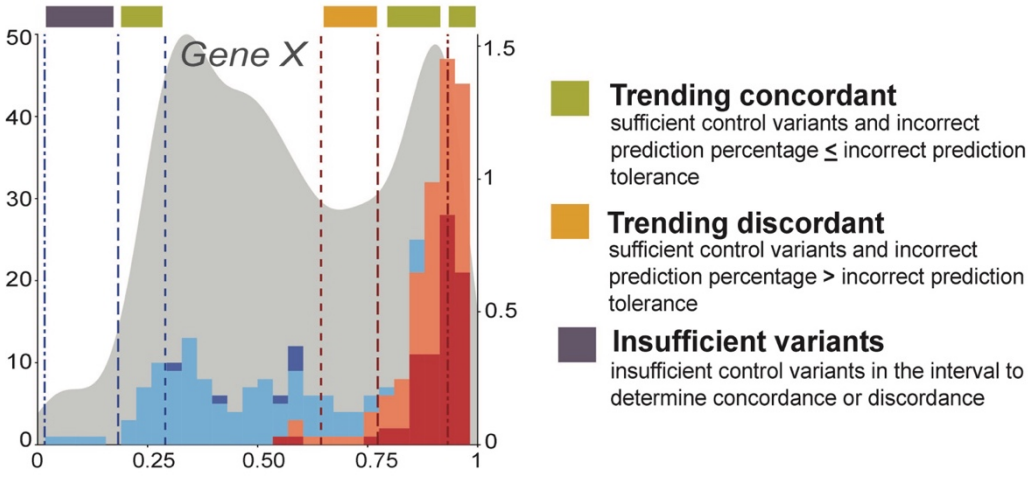


Figure 3.1 Gene-specific analysis workflow

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).

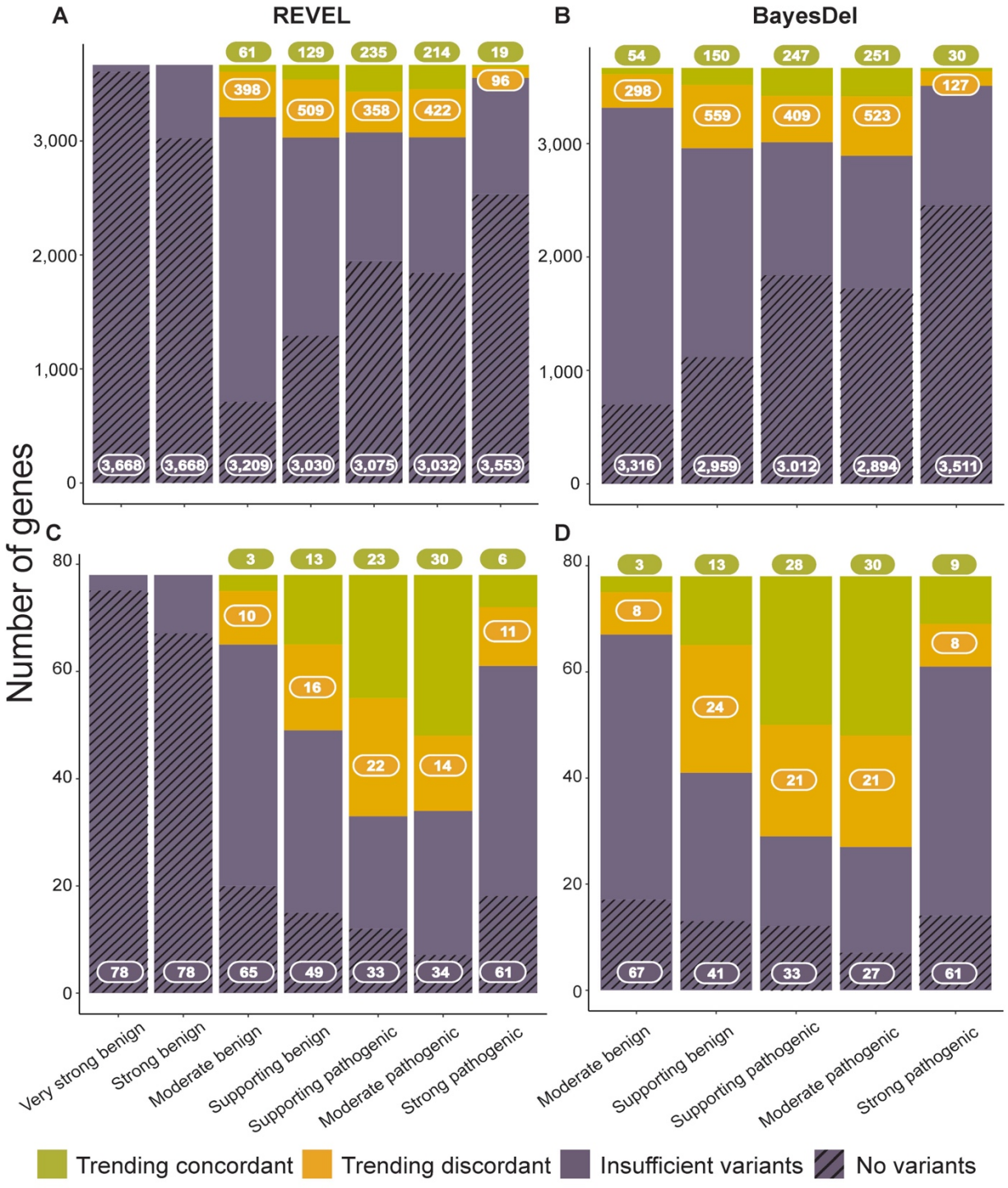

A
REVEL
B
BayesDel
C
D
Number of genes
Very strong benign
Strong benign
Moderate benign
Supporting benign
Supporting pathogenic
Moderate pathogenic
Strong pathogenic
Trending concordant
Trending discordant
Insufficient variants
No variants

Figure 3.2 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.

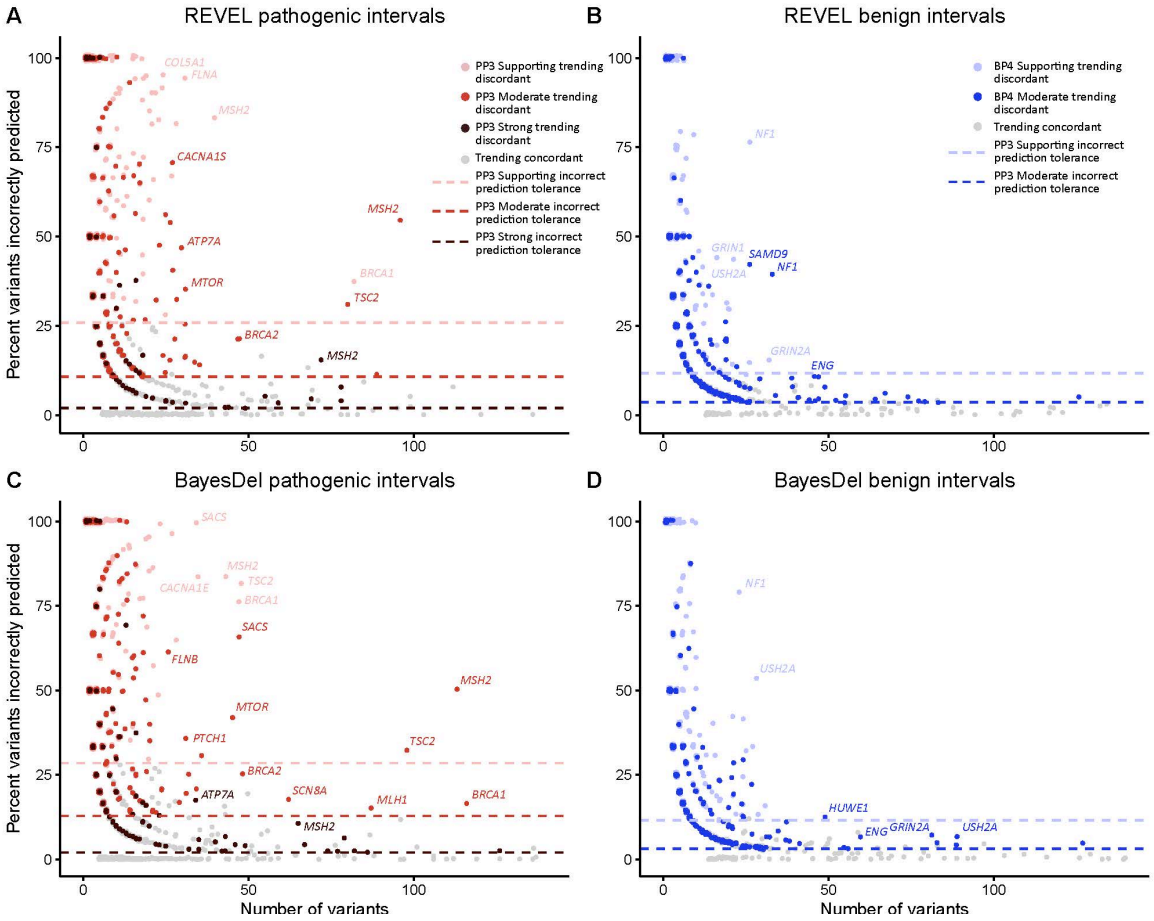


Figure 3.3 Some trending discordant intervals arise from genes with many control variants

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

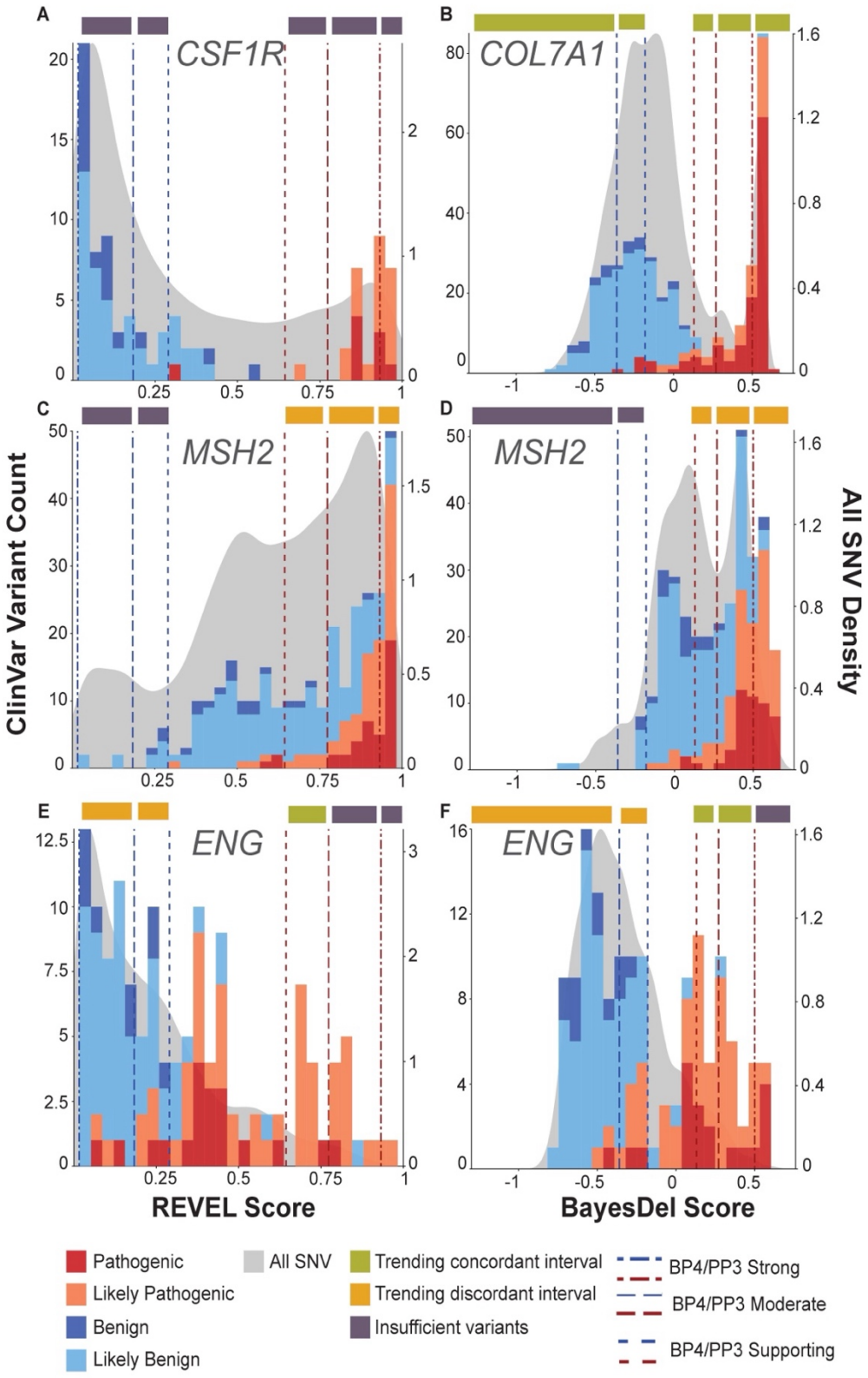

A
CSF1R
B
COL7A1
C
MSH2
D
MSH2
E
ENG
F
ENG
ClinVar Variant Count
All SNV Density
REVEL Score
BayesDel Score
Pathogenic
Likely Pathogenic
Benign
Likely Benign
All SNV
Trending concordant interval
Trending discordant interval
Insufficient variants
BP4/PP3 Strong
BP4/PP3 Moderate
BP4/PP3 Supporting

Figure 3.4 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.

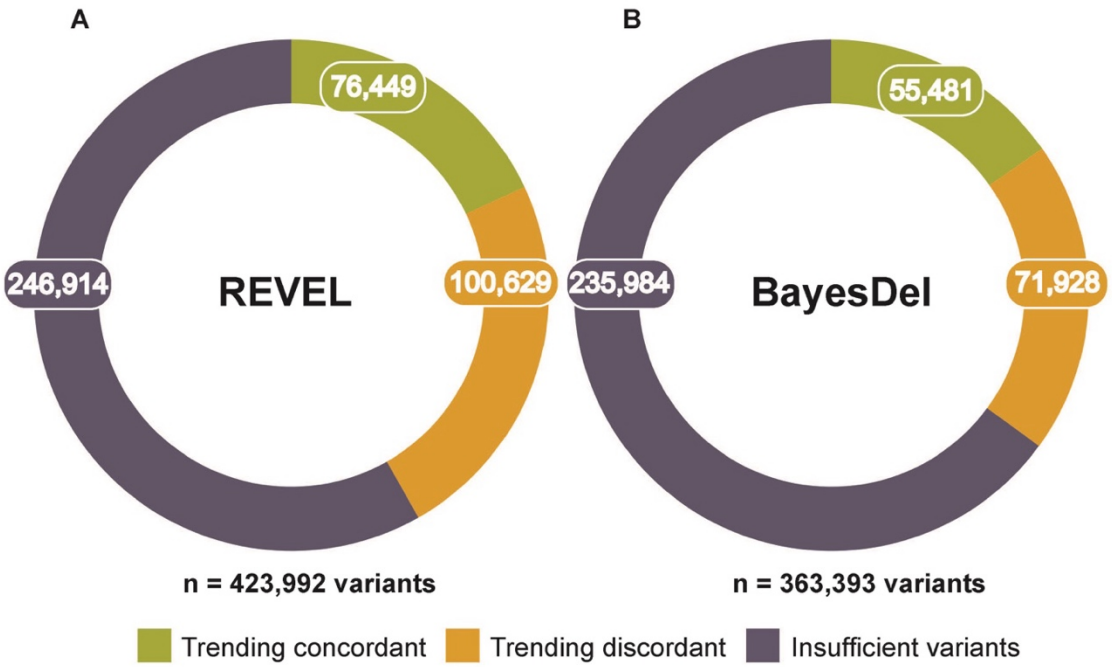


Figure 3.5 Many ClinVar variants of uncertain significance are in trending discordant intervals

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).

3.8 Tables

3.8.1 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%

3.8.2 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

3.8.3 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

3.9 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: A scalable approach to resolving variants of uncertain significance

This chapter is adapted with minimal modifications from:

***Tejura M**, Chen Y, McEwen AE, Stewart R, Sverchkov Y, Laval F, et al. A scalable approach to resolving variants of uncertain significance. bioRxiv. 2026 Feb 15.*

4.1 Abstract

Over 90% of missense variants across ~4,000 disease-associated genes are variants of uncertain significance (VUS). Experimental variant effect measurements provide critical evidence about pathogenicity and inform disease biology, but most variants lack data and clinical translation has been limited. The Impact of Genomic Variation on Function Consortium generated experimental data for 62,215 variants across ten genes using multiplexed assays and 1,407 variants across 163 genes using arrayed assays, curated 193,139 additional community-generated variant effect measurements across 30 additional genes, and developed automated calibration methods for translating experimental data and variant effect predictions into clinical evidence. To reduce current VUS, we developed a scalable workflow using only experimental and predictive evidence, enabling reclassification of 75% of the 16,115 VUS in these genes as pathogenic or benign with <1% error. To minimize future VUS, we analyzed >90,000 unobserved variants; 62% had enough evidence to be "preclassified" as pathogenic or benign. We validated our data, evidence and classifications using All of Us and created interactive resources to enable clinical use of the calibrated data. Thus, for 40 genes, representing 1% of the clinical genome, we resolve most existing VUS and future variants, illustrating how systematic use of scalable evidence can empower genomic medicine.

4.2 Main

Genetic testing has identified millions of unique single nucleotide variants (SNVs) in genes associated with disease[1,2]. However, most of these variants have never been functionally characterized, and over 90% of missense and near-intronic variants are classified as variants of uncertain significance (VUS)[3]. VUS cause confusion and distress, and cannot be used in diagnosis or treatment, therefore representing a major barrier to genomic medicine. Moreover,

individuals from ancestries historically underrepresented in genomic medicine and research bear a disproportionately large proportion of VUS, perpetuating inequities[4–8]. In 2020, the NHGRI predicted that by 2030 the term 'VUS' would become obsolete and that individuals from ancestrally diverse backgrounds would benefit equitably from advances in human genomics[9]. Making this prediction a reality requires understanding whether variants affect gene and protein function and knowing how variants impact function, including by altering splicing, disrupting protein folding, destabilizing complexes, perturbing molecular interactions, or rewiring cellular networks.

Experimental assays and computational variant effect predictions are the most scalable sources of knowledge about how variants affect gene and protein function. In particular, Multiplexed Assays of Variant Effect (MAVEs) leverage high-throughput DNA sequencing to enable measurement of the impact of nearly all single nucleotide or amino acid substitutions[10–12]. MAVEs for *BRCA1*[13], *MSH2*[14], *TP53* [15,16] and other clinically relevant genes have demonstrated the power of experimental data to reclassify VUS and reveal mechanisms of disease[3]. MAVE-derived experimental data is comprehensive, and therefore benefits individuals with VUS regardless of ancestry[17]. In parallel, variant effect predictors can estimate the impact of most variants in the genome, and advances in machine learning have increased their accuracy[18–23]. However, use of experimental and predictive data in the clinic has been limited; experimental data is lacking for most genes[24], both data sources are difficult to translate into clinical evidence[25], and there is limited dissemination infrastructure[26].

We hypothesized that systematic generation and application of experimental and predictive evidence would resolve VUS at scale[27–29]. To enable us to test this hypothesis, as part of the NHGRI's Impact of Genomic Variation on Function (IGVF) Consortium[30], we:

Established production-scale MAVEs to systematically characterize all possible missense and near-intronic variants in clinically relevant genes, generating high-quality experimental data for 62,215 variants across 10 genes[12,13,31–36] (Fig. 1a,b). These data accurately discriminate known pathogenic and benign variants and demonstrate disease association for loss-of-function variants in a population biobank.

Developed high-throughput arrayed phenotypic assays that reveal how 1,407 variants in 163 genes impact function, including protein abundance, protein-protein interactions, and subcellular localization. These results established protein mislocalization as a hallmark of pathogenic variants and prioritized phenotypes for future MAVEs[37] (Fig. 1a,b).

Integrated our IGVF-generated experimental data with data from 68 published assays across an additional 30 clinically relevant genes and scores from top-performing variant effect predictors[18,20,21,38,39] (Fig. 1c,d). This integrated variant effect dataset contains variant effect measurements and predictions for 255,354 variants and enables systematic variant classification.

Developed automated calibration methods to translate experimental and predictive data into standardized evidence compatible with established guidelines for clinical variant classification[33,40–42](Fig. 1e). These new methods increase the accuracy of experimental and predictive evidence[25,43,44] and enable the classification of more variants (Fig. 1f).

Enabled discovery of our calibrated experimental evidence through the IGVF catalog[41] and MaveMD, a clinically-focused interface for MaveDB[26,38,45] (Fig. 1g). Additionally, we present PredictMD, a platform which enables discovery of calibrated predictive evidence. These resources facilitate exploration of the evidence we provide, enabling variant classification by clinicians.

Leveraging our experimental data and calibrations, we establish the first scalable workflow for resolving VUS, capable of providing definitive classifications for 75% (12,135 of 16,115) of VUS in the 40 clinically relevant genes we study. To minimize future VUS, we preclassify 90,530 previously unobserved missense variants in these genes, with 62% receiving sufficient evidence to be classified as pathogenic or benign once they are observed. Our workflow results in <1% of known pathogenic or benign variants receiving discordant classifications with ClinVar, and our pathogenic evidence and classifications are strongly associated with disease in a population biobank. Thus, comprehensive functional characterization of genetic variants can simultaneously advance our understanding of gene and protein biology while enabling clinical variant classification at scale, helping make VUS obsolete and enabling equitable precision healthcare.

4.2.1 Production-scale MAVEs provide variant effect measurements for 62,215 variants across 10 genes

We applied two complementary MAVEs to systematically measure variant effects across 10 clinically relevant genes. Variant Abundance by Massively Parallel sequencing (VAMP-seq) measures steady-state protein abundance for all possible amino acid substitutions by fusing variants to GFP, followed by fluorescence-activated cell sorting and sequencing[12] (Fig. 2a). VAMP-seq is effective for identifying pathogenic missense variants because a primary mechanism of missense pathogenicity for most proteins is reduced steady-state abundance due to protein destabilization and degradation[46–49]. Saturation Genome Editing (SGE) measures the effect of SNVs in essential genes on cellular fitness in haploid human cells, and thus detects the effects on RNA expression, splicing, and protein function in a single assay[11,13] (Fig. 2b).

We applied VAMP-seq to three proteins with distinct protein folds and disease associations. For *G6PD*, which encodes glucose-6-phosphate dehydrogenase and where pathogenic loss-of-function variants are associated with increased risk of hemolytic anemia, we measured the abundance of 10,675 variants[35]. For *TSC2*, which encodes the tumor suppressor tuberous sclerosis complex subunit 2, we focused on the tuberin and RapGAP domains where pathogenic missense variants cluster, measuring the abundance of 9,253 variants[34]. For *F9*, which encodes factor IX and where pathogenic variants cause hemophilia B, we deployed a modified VAMP-seq protocol using antibody labeling to make 45,035 measurements of the secretion, folding, and post-translational modifications of 9,007 variants[32]. Together, these VAMP-seq datasets represent 64,963 measurements across 27,097 missense, 1,287 synonymous, and 551 nonsense variants (Fig. 2c,d; Supplementary Data 1). Abundance score distributions show clear separation of synonymous variants, which reflect normal abundance, from nonsense and low abundance missense variants (Fig. 2c).

We used SGE to measure the effects of SNVs on cellular fitness for seven genes, five of which encode proteins in the homology-directed DNA break repair pathway: *BARD1* (8,851 SNVs)[36], *PALB2* (8,514), *BRCA2* N-terminus (795), *RAD51D* (4,358), and *XRCC2* (3,743) (Fig. 2e,f). In these genes, pathogenic variants are established or suspected risk factors for hereditary breast and/or ovarian cancer[50,51], leading to VUS accumulation due to the inclusion of these genes on cancer risk panel tests. We also measured the effects of 4,790 SNVs in *CTCF*, required for genome organization, where pathogenic *de novo* variants cause neurodevelopmental delay and autism[52]; and 2,229 SNVs in *SFPQ*, where variants are suspected to cause intellectual disability[53]. Together, these SGE datasets represent 33,280 SNVs across coding exons, exon-

intron junctions, and UTRs (Fig. 2e). Unlike VAMP-seq, SGE's single-nucleotide resolution enables assessment of splice variants and loss or gain of start and stop codons (Fig. 2f).
All datasets were uniformly processed and met rigorous IGVF quality control standards for variant coverage and replicate correlation (IGVF standards). Together, VAMP-seq and SGE queried 48,402 codon substitutions at 15,222 nucleotide positions in 10 genes, with VAMP-seq covering more variants and SGE more positions (Fig. 2g). VAMP-seq saturates amino acid variants, permitting deep exploration of sequence-structure-function relationships whereas SGE saturates single nucleotide variants, capturing splicing and covering more genomic space (Fig. 2d,f,g). VAMP-seq shows high precision for identifying pathogenic variants (range = 0.897 - 0.965), while SGE demonstrates both high precision (range = 0.934 - 0.969) and recall (range = 0.942 - 0.974) in discriminating loss-of-function pathogenic variants from benign variants (Fig. 2h). We validated these experimental measurements in the All of Us biobank, where participants carrying variants identified as loss-of-function in *TSC2, BARD1*, *PALB2* and *RAD51D* showed elevated risk for cancer phenotypes (Fig. 2i, Supplementary Data 2).

4.2.2 Curating experimental data from the community to generate an integrated variant effect dataset

To augment our experimental data, we assembled 68 community-generated large-scale datasets encompassing 295,058 variant effect measurements and 31,858 composite scores for 193,139 unique variants of 30 additional genes associated with rare diseases, cancer predisposition, metabolic diseases and cardiovascular phenotypes (see Supplementary Data 3 and references therein). We curated these community data, adding clinically relevant metadata, mapping all variants to genomic coordinates, standardizing variant nomenclature, and integrating them with

our newly generated IGVF datasets to create a harmonized experimental dataset that includes more than 400,000 variant effect measurements for 255,354 variants of 40 unique genes[38]. The 40 genes in the dataset are components of 7,885 registered genetic tests[54] and nearly half (18) are listed as ACMG Secondary Findings genes[55], indicating clinical actionability (Fig. 3a). Experimental data in the set was generated using multiple model systems across three major modalities: reporter assays measuring protein expression or activity, cell-fitness assays, and direct protein-function assays evaluating biochemical or biophysical properties (Fig. 3b). VAMP-seq and SGE were the most common assays, and 25% of all variant effect measurements were generated by the IGVF Consortium (Fig. 3c).

We annotated variants with scores from AlphaMissense[18] MutPred2[21] and REVEL[20]. AlphaMissense and REVEL had accessible scores for non-synonymous SNVs, annotating ~37% of our variants and MutPred2 had scores for multinucleotide codon changes, annotating ~77% of our variants. In addition, we annotated variants with ClinVar clinical significance and presence in gnomAD. This integrated variant effect dataset encompasses 17,676 VUS, 12,075 control variants with known pathogenic or benign classifications, 27,157 variants observed in gnomAD and 98,585 possible SNVs that have not been observed in population screening[56] or genetic testing[1] (Fig. 3d, Supplementary Data 1).

4.2.3 Transforming experimental and predictive data into clinical evidence with improved calibration methods

The first step in using experimental or predictive data for clinical variant classification is calibration, which translates the data into evidence conforming to American College of Medical

Genetics/Association for Molecular Pathology (ACMG/AMP) variant classification guidelines[57,58]. In this framework, experimental data yields 'functional evidence' and data from variant effect predictors yields 'computational evidence'. Calibration is based on the ability of the data to discriminate pathogenic from benign control variants and results in each measurement or prediction having a point value ranging from -8, the strongest benign evidence, to +8, the strongest pathogenic evidence. Points from different sources of evidence, including allele frequency, segregation, and patient phenotype as well as experimental and predictive data, are combined to generate a final classification (Benign ≤-6, Likely Benign -5– -1, VUS 0–5, Likely Pathogenic 6–9, Pathogenic ≥10)[59]. However, current calibration methods have shortcomings. For experimental data, the ClinGen-accepted OddsPath method relies on study-specific thresholds that define broad functional categories (e.g., loss of function or functionally normal) without leveraging the full distribution of scores[25]. For predictive data, the accepted 'genome-wide' method aggregates variants across disease-associated genes, masking gene-specific predictor behavior and leading to errors in evidence assignment[43,44,60]. We developed new approaches to address these shortcomings.

For calibrating experimental data, we created ExCALIBR (Experimental score CALIBRator). ExCALIBR calculates gene-specific priors from gnomAD frequencies; fits mixture models to assay score distributions of gnomAD, synonymous, pathogenic/likely pathogenic (PLP) and benign/likely benign (BLB) variants; and calculates a posterior probability of pathogenicity for each variant that scales evidence strength based on these distributions[33] (Fig. 4a). We applied ExCALIBR to 37 of the 40 genes in our dataset, excluding *F9* and *TP53* where classifier models were used to integrate multiple datasets[32,61] and *SFPQ* which remains a candidate disease gene.

ExCALIBR assigned at least 1 point of pathogenic or benign evidence to variants of 35 of 37 genes (Extended Data Fig. 1, Supplementary Data 4).

For the *MSH2* dataset[14], ExCALIBR expands the indeterminate range, better reflecting the mixture of pathogenic and benign control variants compared to the accepted threshold-based approach[62] (Fig. 4b). Moreover, evidence strength increases continuously as variant scores become more extreme, rather than being uniformly strong beyond a fixed threshold (Fig. 4b). ExCALIBR evidence assignments agreed with ClinVar PLP and BLB control variants for 97.9% of variants compared to functional annotations, which agreed for 93.6% of variants (Fig. 4c). However, the gain in accuracy comes at the cost of reduced coverage; the proportion of VUS assigned to the indeterminate score range increased from 6.8% to 20.3% (Fig. 4c). Nonetheless, because ExCALIBR takes advantage of control variants beyond ClinVar, it was able to calibrate an additional 19 datasets contributing 78,022 variant effect measurements for 48,707 variants that could not be calibrated using the established OddsPath approach. 14 of 17 gene-disease pairs had significant associations between pathogenic evidence and disease phenotypes in the All of Us biobank (Extended Data Fig. 2a, Supplementary Data 2). Overall, our calibrated experimental data provides benign evidence for 122,708 variant effect measurements and pathogenic evidence for 74,652 variant effect measurements representing 166,562 unique variants across 35 genes.

For predictive data, we developed a gene-specific calibration method to address the pitfalls of the accepted genome-wide method that we previously identified[60]. Eight of our 40 genes had enough control variants for gene-specific calibration: *BRCA1*, *BRCA2*, *F9*, *JAG1*, *MSH2*, *SCN5A*, *TP53*, *TSC2*[42] (Supplementary Data 4). For each of these genes, we computed gene-specific priors[40,42] and applied post-hoc calibration via a data-adaptive framework, fitting calibration models to

assign posterior probabilities and evidence strength for REVEL, MutPred2 and AlphaMissense (Fig. 4d, Extended Data Fig. 3)[42].

For *MSH2*, REVEL scores are higher overall, causing the accepted genome-wide calibration method to inappropriately assign pathogenic evidence to some variants and no evidence to others (Fig. 4e). Our gene-specific calibration aligned the evidence thresholds with the *MSH2* control variant distributions, preventing the misassignment of pathogenic evidence to BLB variants and reducing the number of variants receiving no evidence (Fig. 4e). Our gene-specific calibration also yielded odds ratios for cancer association with variants assigned benign evidence in the All of Us Biobank closer to 1 (0.98, 95% CI 0.90-1.07) than the genome-wide method (1.61, 95% CI 0.69-3.78), supporting the superiority of our approach (Extended Data Fig. 2b, Supplementary Data 2). When considering all eight genes calibrated using our method, BLB variants misassigned to the indeterminate category was reduced by 41.7% for REVEL, and 30.4% (17,846) more variants were assigned pathogenic or benign evidence[43,44]. Thus, our new calibration methods reduce evidence misassignment, improve the number of variants receiving evidence and move further toward automated calibration by removing subjective assignment of thresholds for both data sources.

4.2.4 Variant classification workflow using experimental and predictive evidence is highly accurate

We hypothesized that systematically generated, curated, and calibrated experimental and predictive evidence could resolve most VUS in the absence of other evidence. Therefore, we developed a scalable workflow using experimental and predictive evidence to classify variants according to the points-based ACMG/AMP variant classification framework[59]. In particular, we:

Applied calibrated experimental evidence from single assays, except for *F9* and *TP53* where models were used to integrate data from multiple assays[32,61], removing variants that had conflicting scores from multiple assays.

Applied predictive evidence, calibrated with our new gene-specific method where available, and with the genome-wide method for the remainder[43,44].

Removed the start-loss variants, variants predicted to affect splicing and variants used in training REVEL and MutPred2 when appropriate.

Combined experimental and predictive evidence to generate classifications for the remaining 94% (239,836 of 255,354) of variants.

The accepted variant classification workflow involves combining all available evidence including population frequency, segregation, and clinical presentation[57,63–65]. To determine the accuracy of our scalable workflow combining only experimental and predictive evidence we assessed its ability to recapitulate established BLB and PLP classifications from ClinVar that were made with all available evidence[1]. 69% of variants (7,085/10,228) remained consistent with original ClinVar BLB or PLP classifications, and 30% of variants reverted to VUS due to the absence of other evidence (Fig. 5a). Critically, our workflow did not result in definitive (i.e. PLP or BLB) classifications that disagreed with ClinVar. The rate of discordance was very low, ranging from 0.66% (67 of 10,228 variants) for classifications using experimental evidence plus predictive evidence from REVEL to 0.69% using evidence from MutPred2 (63 of 9,099 variants) and 1.14% (117 of 10,250 variants) for those employing AlphaMissense evidence (Fig. 5b,c, Extended Data Fig. 4a-d, Supplementary Data 5). Of the 67 discordant variants, 62 were PLP to BLB and 5 were BLB to PLP (Fig. 5b). The three genes with the most discordant variants were

TP53 (10), *VHL* (8) and *F9* (3) (Extended Data Fig. 5). Thus, our scalable workflow using only experimental and predictive evidence was largely concordant with ClinVar classifications that incorporated all available evidence, was robust across predictors and led to a remarkably low rate of discordance among definitive classifications.

We further evaluated the accuracy of our scalable workflow using the ClinGen Evidence Repository, which contains variants annotated with all evidence used (Fig. 5d). To avoid circularity, we removed all experimental and predictive evidence that contributed to the original classifications; 194 variants across 11 genes retained sufficient evidence to remain BLB (19) or PLP (175). The resulting ClinGen Evidence Repository classifications that did not depend on experimental or predictive evidence were highly concordant with our classifications using only experimental and predictive evidence, with discordance rates ranged from 2.23% (REVEL) to 2.86% (MutPred2), further validating the low error rate of our scalable workflow (Fig. 5e,f, Extended Data Fig. 4e-h, Supplementary Note 1 and Supplementary Data 5).

Finally, we compared variant classifications made using evidence strength computed with our new calibration methods to classifications using evidence strength computed using the currently accepted approaches[25,43,44]. Using accepted calibrations 6,269 (60%) of ClinVar BLB and PLP variant classifications were consistent with the original classification, 39% reverted to VUS and 99 (0.95%) were discordant (Extended Data Fig. 6a-f, Supplementary Data 6). 81 (45.76%) of ClinGen Evidence Repository variants were consistent and 3 (1.69%) were discordant (Extended Data Fig. 6g-l). Calibrations produced with accepted methods performed well when compared to ClinVar classifications, but our new calibrations produced fewer discordant classifications

(0.66% vs 0.95%) and fewer VUS reversions (30% vs 39%). Overall, our improved calibration methods enhanced both the accuracy and classification power of experimental and predictive evidence for clinical variant classification.

4.2.5 Experimental and predictive evidence alone can reclassify 75% of existing VUS and preclassify 62% of unobserved variants

Our scalable workflow using only experimental evidence and predictive evidence from REVEL yielded a discordance rate of 0.66%, less than the uncertainty thresholds acceptable for pathogenic and benign classifications[66]. This accuracy enables, for the first time, systematic variant classification at scale using standardized experimental and predictive evidence. We applied experimental and predictive evidence from REVEL to three categories of variants across 40 genes: VUS identified through genetic testing in ClinVar[1], variants detected in human populations via exome or genome sequencing and in gnomAD[56] and all possible missense SNVs not yet observed by population or clinical sequencing (Fig. 3d).

For the 16,115 ClinVar VUS in the 40 genes in our study, more definitive classifications would be immediately useful to improve the efficacy of genetic testing. Of these VUS, 12,135 (75%) received sufficient evidence to be classified as BLB or PLP using REVEL as the predictor; 11,184 (69%) reached BLB and 951 (6%) PLP (Fig. 6a, Supplementary Data 5). This outcome is consistent both with the expectation that most VUS are benign[4,67–69] and with our use of the points-based classification scheme[59], which reduces the evidence required to reach likely benign compared to previous variant classification framework[57]. Variants receiving pathogenic evidence (combined points $\geq$ 1) had significant disease associations: 10 cancer predisposition genes

showed elevated odds ratios for cancer phenotypes, *SCN5A*, *KCNQ4* and *GCK* variants showed significant associations with long QT syndrome, hearing loss and maturity-onset diabetes of the young, respectively (Fig. 6b, Supplementary Data 2). VUS resolution rates using MutPred2 and AlphaMissense (78% and 79% respectively) were comparable (Extended Data Fig. 7). 82% of variants with sufficient evidence to be classified as PLP had experimental and predictive evidence, whereas 18% had only experimental evidence. 7,223 of 11,184 BLB variants (65%) had evidence from both sources, 3039 (27%) had only experimental evidence and 922 (8%) had only predictive evidence, suggesting that experimental data drove more reclassifications (Fig. 6c).

Of the 16,115 ClinVar VUS, 3,980 (25%) did not receive enough evidence (0 to 5 total points) to be reclassified. 531 (13%) of the unresolved VUS had concordant evidence from both sources, but the evidence was too weak to change the classification. Another 2,154 (54%) of the unresolved VUS were missing one or more sources of evidence: 466 (12%) had only experimental evidence, 970 (24%) had only predictive evidence and 718 (18%) had no evidence from either source. The remaining 33% of the unresolved VUS (Fig. 6c) had discordant experimental and predictive evidence, which most often resulted in 0 total points (no evidence). Notably, 698 (17%) of the unresolved VUS accrued 4 or 5 points of pathogenic evidence, suggesting they could reach likely pathogenic classification with additional evidence. The unresolved VUS were distributed unevenly across genes based on the strength of experimental and predictive evidence available for each gene (Extended Data Fig. 8). Thus, experimental and predictive evidence alone can reclassify the majority of existing VUS, addressing a substantial portion of variants already discovered through clinical genetic testing.

We also used our scalable workflow to assess the 24,992 variants in our 40 genes in gnomAD. 1,208 (5%) gnomAD variants received sufficient evidence to be classified as PLP, a modest but statistically significant reduction relative to the 6% of ClinVar VUS reclassified as PLP in ClinVar ($p = 2.4e-6$, chi-squared test, Fig. 6d, Extended Data Fig. 9, Supplementary Data 5). 16,409 (66%) gnomAD variants received sufficient evidence for BLB classification, a modest but significantly lower percentage than for ClinVar VUS (69%, $p = 3.3e-15$, chi-squared test). Thus, ClinVar VUS and variants in gnomAD yielded a similar fraction of variants that could be classified as PLP or BLB.

Under the existing variant classification framework, VUS are appearing rapidly[3]. Moreover, all possible SNVs compatible with life will eventually be observed in the human population[70]. Ideally, all newly discovered variants would receive definitive classifications at the time of detection rather than becoming VUS that require reevaluation. To avoid future VUS, we introduce the concept of variant preclassification, where experimental and predictive evidence are systematically generated, calibrated and applied to variants that have not yet been observed in the human population.

We therefore applied our scalable workflow to the 90,530 possible SNVs that have not yet been observed in ClinVar or gnomAD (Fig. 6e, Extended Data Fig. 9). Of these, 56,341 (62%) received sufficient evidence to be immediately classified as BLB or PLP. While the overall classification rate of 62% is lower than for ClinVar VUS (75%) or gnomAD variants (71%), a significantly higher proportion were classified as PLP: 8% (7,568) compared to 6% in ClinVar (p

< 2.e-16, chi-squared test) and 5% in gnomAD ($p < 2e-16$). This result is consistent with the findings that variants absent from population databases are more likely to be deleterious[71–76]. We predict that, by providing preclassifications for most variants, application of our scalable workflow to unobserved variants will prevent most new VUS in these genes.

4.2.6 Identifying protein properties disrupted by variants guides future MAVEs and provides mechanistic insights into pathogenicity

To complement MAVEs and provide broader gene coverage, we developed high-throughput arrayed assays measuring three core protein properties often affected by variants[37,77,78]. Here, variant impacts on protein abundance (DUAL-IPA), subcellular localization (Variant Painting), and protein-protein interactions (sqY2H) were measured for 791 (91 genes), 794 (116 genes) and 502 (74 genes) variants, respectively (Fig. 7a, Supplementary Data 7). Altogether, 1,407 variants across 163 genes were experimentally profiled. These assays distinguished between pathogenic and benign variants: 32% of ClinVar PLP variants were abnormal compared to only 9% of BLB variants ($p = 2 \times 10^{-9}$, two-sided Fisher's exact test) (Fig. 7b). Each assay independently discriminated pathogenic from benign variants (DUAL-IPA: $p = 2 \times 10^{-7}$; Variant Painting: $p = 9 \times 10^{-10}$; sqY2H: $p = 0.04$ Mann-Whitney test) (Fig. 7c-e).

Among the 30 genes profiled across all three assays, 67% exhibited detectable phenotypes for pathogenic variants in at least one assay. Among them, 40% exhibited phenotypes in only one assay, 13% in two assays, and 13% in all three (Fig. 7f,g). This phenotypic orthogonality demonstrates that different genes require different assays to detect pathogenic variants, providing a roadmap for prioritizing future MAVE targets and illustrating that pathogenic variants in different genes act via distinct mechanisms.

Integrating arrayed assay data with MAVE results provided insight into variant pathomechanism. For example, we applied a modified version of VAMP-seq to factor IX (*F9*) revealing that 39.5% of missense variants caused reduced secretion of the protein from cells[32] (Fig. 2d). This assay could not distinguish degradation of misfolded protein from intracellular retention. To resolve this ambiguity, we leveraged Variant Painting, a high-content cellular imaging assay that characterizes the localization pattern of missense variants and changes to cellular morphology[37]. Variant Painting-derived phenotype scores, which summarize variant effects across thousands of image-derived features, correlated strongly with VAMP-Seq secretion measurements (Fig 7h,i). Individual feature analysis revealed that 14 poorly secreted variants showed increased co-localization with Golgi markers, suggesting retention in the secretory pathway (Fig 7j). In contrast, the signal peptide variant p.Cys28Arg showed loss of staining that appeared cytoplasmic, consistent with translocation failure and rapid degradation (Fig. 7k). These mechanistic distinctions may have therapeutic implications since total loss of factor IX expression can lead to the development of inhibitory anti-factor IX antibodies in response to exogenous factor IX therapy and worse clinical outcomes[79]. Moreover, variants retained in the secretory pathway represent ‘rescuable’ targets amenable to pharmacological stabilization, while degradation-prone variants may require alternative strategies.

4.2.7 An ecosystem of resources to disseminate and visualize experimental and predictive data and evidence

Despite the value of experimental and predictive evidence for variant classification, clinical implementation is challenged by several usability barriers[26,80,81]. Experimental data are scattered across supplemental tables and repositories with inconsistent formats, requiring specialized

expertise to evaluate quality and clinical relevance. Predictive data is available from multiple sources[82–84], but the appropriateness of predictive evidence for particular genes and variants is reliant on expert guidance[85,86]. Moreover, no resource allows discovery, assessment, and use of experimental and predictive evidence simultaneously. To address these challenges, we developed MaveMD, to enable the direct clinical use of curated experimental evidence. We also created a dedicated resource for IGVF coding variant data that provides centralized access to the data, analyses and calibrated evidence presented here.

MaveMD (mavemd.mavedb.org) is a clinical interface that transforms experimental data into clinically interpretable evidence[38]. MaveMD implements flexible variant search using clinical notation, interactive histograms contextualizing variant scores within complete functional distributions, and overlays of ClinVar classifications to reveal concordance between experimental data and clinical classifications. Users can easily compare evidence assignments from multiple calibration methods, including ClinGen-specified OddsPath[25] and ExCALIBR, our improved approach[33]. Standardized "assay facts" labels distill complex metadata into essential criteria, including assay type, molecular phenotype, model system, and variant consequences detected by the assay.

We developed a dedicated resource for IGVF coding variant data (igvf.mavedb.org) that contains all experimental data, calibrations, and analysis tools. We built on the MaveMD framework to present the initial version of PredictMD, which includes calibrated variant effect predictions from three state-of-the-art predictors (REVEL, AlphaMissense, MutPred2) comparing three calibration methods, enabling systematic comparison of variant effect predictions. Together,

MaveMD and PredictMD directly address requirements identified in clinical user surveys by enabling detailed, variant-wise comparison of predictions and experimental evidence for the first time, and are designed to substantially reduce technical barriers to routine integration of functional data into clinical practice.

4.3 Discussion

We integrated expertise across experimental genomics, computational biology and clinical genetics to address the VUS challenge. Systematically generated experimental data, combined with calibrated predictions, can resolve 12,135 (~75%) of VUS across 40 clinically relevant genes. Moreover, we introduce the concept of preclassification, showing that experimental and predictive evidence would enable definitive classification of 62% of the ~90,000 variants in these 40 genes when they are eventually observed. By establishing production-scale multiplexed and arrayed experimental platforms, developing automated calibration methods, and creating clinically-oriented data resources, we provide both immediate clinical utility and a generalizable framework for eliminating most VUS.

Our production-scale MAVEs generated approximately 25% of extant clinically validated variant effect data. We, along with the wider community[29,87], established quality standards[39,88] and demonstrated the feasibility of comprehensive variant characterization across clinically relevant genes. We show that experimental data is required to reach a PLP classification for most variants and that experimental data generally provides more power to classify variants as benign than predictive data. Thus, experimental data for additional genes would directly improve the utility of clinical testing, but new MAVE technologies that increase scale, reduce costs and capture a wider array of phenotypes are needed[89]. This need is highlighted by the gene-specific variation in

variant classification accuracy we observed and by our arrayed assays showing that pathogenic variants disrupt multiple functions necessitating diverse assays. Our emerging MAVE platforms dramatically expand scale and mechanistic scope: LABEL-seq measures protein abundance, interactions, and pathway activity at scale[90], imaging-based assays capture multimodal variant effects and new editing approaches enable assays in physiologically relevant cellular and diploid contexts[37,91–93].

Our automated calibration approaches represent a fundamental shift from expert-defined thresholds to empirical, unbiased evidence assignment. To calibrate experimental data, we fit mixture models to control variant distributions and calculate gene-specific priors. This approach eliminates subjective priors and thresholds, enabling systematic calibration across datasets[33]. To calibrate predictive data, we developed a gene-specific approach, ameliorating the inappropriate evidence generated for some genes by the previous genome-wide approach[42]. We also developed a promising domain-aggregation approach that groups domains based on similar predictor behavior to expand the number of genes that can be calibrated while increasing calibration power over the genome-wide approach[42]. These automated calibration approaches can keep pace with rapid experimental data production and new predictors while reducing error and improving classification power, although these new approaches should be further vetted by the expert community, including ClinGen Expert Panels.

Our scalable variant classification workflow makes numerous simplifying assumptions, including using only experimental and predictive evidence, discarding variants with conflicting experimental evidence instead of adjudicating these conflicts, ignoring mode of inheritance, and

not stratifying ClinVar assertions by mechanism of pathogenicity (e.g., loss-of-function versus gain-of-function) when evaluating concordance. Remarkably, this simplified approach achieves an 0.66% variant discordance rate with ClinVar and 2.23% with ClinGen Evidence Repository control variants. However, our approach does have drawbacks. Using only experimental and predictive evidence led to ~30% of control variants reverting to VUS. 4.9% (693/13,050) of our BLB reclassifications relied on weak, single-source evidence (-1 point). In the current ACMG/AMP points-based classification system, such weak evidence is sufficient for a likely benign classification, but many of these classifications will be incorrect. For example, an experiment or predictor generating such weak evidence would miss ~40% of pathogenic variants (Extended Data Fig. 10 and Supplementary Note 2). Thus, while these weak, single-source likely benign classifications follow current guidelines, they should be used with caution and ideally augmented with other sources of evidence.

Making the term ‘VUS’ obsolete will require expanding the approach demonstrated here in several dimensions. Experimental data must be produced for thousands of clinically relevant genes, which will require new technologies and expanded data production capacity. Calibration methods must be refined to preserve accuracy with the smaller numbers of control variants available for most genes. New data generated by the community must be curated, harmonized and incorporated into infrastructure that connects experimental and predictive evidence to clinical workflows. Other evidence types must be systematically curated, calibrated and incorporated into a scalable framework. We demonstrate that this investment, if made, could eliminate most current and future VUS.

4.4 Methods

Multiplexed Assays of Variant Effect

Variant Abundance by Massively Parallel Sequencing (VAMP-seq)

VAMP-seq was performed as described in Briar et al. 2026 and Geck et al 2025[34,35]. IGVF standards documents for VAMP-seq include links to protocols.io and quality control metrics. For factor IX, we performed a modified version of VAMP-seq for secreted proteins called MultiSTEP, as described in Popp et al. 2025[32]. IGVF standards documents for MultiSTEP include links to protocols.io and detailed quality control metrics.

Saturation Genome Editing (SGE)

SGE was performed as in Woo et al. 2025[36] with the exception that experiments for *RAD51D* included an extra timepoint at day 17. IGVF standards documents for SGE include links to protocols.io and detailed quality control metrics.

For both Vamp-seq and SGE assays, raw functional assay data were deposited in the IGVF Data Portal (portal.igvf.org), following FAIR principles and consortium data standards, accession numbers are in Supplementary Table 1.

Arrayed Phenotypic Assays

For all arrayed phenotypic assays, raw functional assay data were deposited in the IGVF Data Portal (portal.igvf.org), following FAIR principles and consortium data standards, accession numbers are in Supplementary Data 7.

Protein Abundance by DUAL-IPA

To measure variant impact on protein abundance, we developed DUAL-Impact on Protein Abundance (DUAL-IPA), advancing the DUAL-FLUO system[78]. GFP-fused variants are co-expressed alongside a housekeeping mCherry control from a single plasmid in human cells. Flow

cytometry quantifies both green and red fluorescences across thousands of cells per variant, normalizing GFP signal to mCherry within each cell to account for cell-to-cell expression variation. The comprehensive experimental protocol is available at: dx.doi.org/10.17504/protocols.io.36wgqpn55vk5/v1

Protein-protein interactions

To profile the impact of variants on protein-protein interactions, we adapted the systematic yeast two-hybrid (Y2H) screening pipeline we previously described[94]. Semi-quantitative Y2H (sqY2H) employs 3-Amino-1,2,4-triazole (3-AT) titration to identify optimal stringency where wild-type interactions approach the detection limit, enabling detection of variant-induced weakening. Automated image analysis provides objective, continuous measurement of reporter activation. To ensure data quality, interaction profiles generated by sqY2H were subsequently validated using the mammalian NanoLuc Two-Hybrid (mN2H) protocol we developed[95]. Step-by-step experimental procedures for the semi-quantitative Y2H (sqY2H) and the mN2H validation are available at dx.doi.org/10.17504/protocols.io.bp2l6eknzgqe/v1

[78]

dx.doi.org/10.17504/protocols.io.rm7vzem8rvx1/v1[78]

[78]

To visualize the impact of genetic variants on protein mislocalization, we adapted the previously established Cell Painting V3[37,96] to create Variant Painting, which relies on the expression of mNeonGreen (mNG)-tagged open reading frames. Changes in protein mislocalization were determined computationally and visually by comparing protein localization of the mutant allele to the corresponding WT reference. [78]

dx.doi.org/10.17504/protocols.io.36wgqnwkxgk5/v1

Image Analysis: We used CellProfiler bioimage analysis software (version 4.2.6)[97] to process the images using classical algorithms following a prior protocol[96]. Illumination variations in images were corrected using the CellProfiler CorrectIlluminationApply module. We segmented nuclei and cells using CellProfiler's IdentifyPrimaryObjects module with Minimum Cross-Entropy thresholding and IdentifySecondaryObjects module with Otsu three-class watershed thresholding, respectively, and measured feature categories including fluorescence intensity, texture, granularity, density, and location (see http://cellprofiler-manual.s3.amazonaws.com/CellProfiler-4.2.4/index.html for more details) across all segmented compartments and all imaged channels. We obtained ~3,200 feature measurements from each of about 800 cells per well. We parallelized our image processing workflow using Distributed-CellProfiler [https://doi.org/10.1371/journal.pbio.2005970]. The image analysis pipelines we used are available in the Cell Painting Gallery[98]. (https://cellpainting-gallery.s3.amazonaws.com/index.html#cpg0020-varchamp/).

The overall methods for processing single-cell profiles to produce population-level profiles were developed previously and are standard practice in image-based profiling[99]. The imaging wells with low signal-to-noise ratios (calculated by taking the ratio between 99th-percentile and 50th percentile pixel intensity of the imaging plate) in the GFP (protein localization) channel were removed from downstream analyses. A small number of new steps were incorporated to handle data artefacts unique to single-cell data. First, missing values were handled by filtering out any feature with more than 100 missing values, and then filtering out any cell that still had missing

values. Next, we computed plate-level statistics (median and median absolute deviation [MAD]) for each feature and removed features with extremely low variance, defined as having a robust coefficient of variation (MAD divided by the absolute median) below 1e-3 in any plate. Data were then MAD-normalized by plate, which involved subtracting the plate-matched median from each feature value and dividing the result by the plate-matched MAD. After normalization, we filtered out CellProfiler features with extreme values (the 99th percentile is greater than 100) caused by technical artefacts and additionally clipped extreme feature values to -100 or 100. We performed feature selection to remove uninformative and redundant features with Pycytominer using the "variance_threshold" and "correlation_threshold" with default settings. Finally, downstream analysis revealed that occasionally small cell debris was incorrectly identified as cells or many cells were incorrectly identified as one very large cell. These segmentation errors were identified by computing the ratio between the nuclei area and the cell area and filtering to only include cells with an area ratio between 0.1 and 0.4.

Protein Mislocalization Analysis: The objective of this analysis was to computationally detect whether a variant allele has a different protein localization pattern than its matched reference allele. We did this by training a set of binary XGBoost classifiers to discriminate between single-cell protein profiles from variant alleles and reference alleles from the same gene. Because the reference and variant alleles were screened on the same plate in separate wells with four replicates each of the two batches, we used a 4-fold cross-validation strategy where samples were stratified by plate. Variant impact was quantified as the model's AUROC score, with the rationale that variants with greater subcellular localization perturbations would have higher

performing classifiers. We refer to the model's AUROC scores as the Variant Painting-derived phenotype scores in the main text.

Plate layout impacts image-based profiles and therefore single-cell profiles from two different wells can usually be partially distinguished even if the perturbation is the same. Therefore, we needed an estimate of the well position effect to robustly infer which variant-reference classifier scores were greater than expected due to technical effects alone. To estimate the well position effect, we trained an additional set of binary XGBoost classifiers between pairwise combinations of repeated controls on each plate. This resulted in a group of binary classifiers between two wells with different positions on the plate but the same allele per batch. We defined the upper limit of the well position effect as the 95th percentile of the AUROC scores from the repeated control allele thresholds. We defined initial mislocalization hits as variants with variant-reference AUROC scores that surpassed these thresholds.

Finally, we applied various QA/QC filters to call the final mislocalization hits. First, we filtered out any variant-reference pair where there was a class imbalance of more than 3-fold. When there are large differences in cell count, image-based profiles will be easily distinguishable simply due to the differences in cell shape at different densities, and so we couldn't be confident that hits with large cell count differences actually represented differences in protein localization. Final hits were defined as variants that surpassed the well-position threshold in both biological replicate batches.

Experimental data curation and integration workflow

Gene prioritization and literature review

Experimental data was collected from the published literature through January 2025. Experimental data was prioritized based on the disease association of the gene studied, potential clinical utility of experimental evidence and number of variants experimentally assessed. Disease association was defined using the GenCC (The Gene Curation Coalition) database (downloaded March 28th, 2025)[2] at moderate or higher disease association. Clinical utility was defined by inclusion in the ACMG Secondary Findings list[55] or by expert review of gene-disease associations. For each gene, we tallied control variants in ClinVar (pathogenic, likely pathogenic, benign, or likely benign), and genes with more than 20 control variants were prioritized for inclusion. Experimental data curation focused on datasets generated using multiplexed assays of variant effect (MAVEs) as well as experimental assays that tested a large number of single variants, prioritizing datasets with published and accessible raw or composite scores.

Metadata curation and harmonization

Metadata associated with each assay were curated according to the minimum information standards for MAVE datasets[39] providing a framework for harmonizing key fields and controlled vocabularies. Specifically, we curated information such as the variant library generation method, cellular or model system used, variant library integration method, and phenotype measured. We further extended this schema to include additional variables specific to our data model, such as the biological mechanism tested, alignment with gene-disease mechanism, and ability to detect splicing and nonsense mediated decay. In total, we curated 68 datasets of which 6 were meta-analyses, consisting of classifier models trained on more than one experimental dataset for a given gene or composite scores derived from the original datasets (Supplementary Data 3).

After curating assay-level metadata, we collected and standardized all variant identifiers across datasets using custom scripts. Raw data were retrieved from multiple sources, including supplementary files, GitHub repositories, Zenodo archives, and other public databases (variant score files). Minor manual corrections were made to ensure consistent parsing and are documented in Supplementary Data 3, column `Additional Notes`

Each experimental dataset targeted a specific gene or protein. For each dataset we identified and mapped the corresponding reference transcript using custom scripts to align author-provided protein sequences to reference protein sequences, when not provided by the authors. 11 datasets had a single amino acid discrepancy from the translated reference transcript; these are documented in Supplementary Data 3, column `Transcript ID Additional Details`. Mapping to a reference transcript was critical for subsequent annotations as ClinVar, gnomAD, and computational predictors store genomic coordinates.

Experimental datasets fell into two broad categories:

1. Nucleotide-level assays (*e.g.,* SGE) that introduce variants directly into the endogenous locus, allowing genomic coordinates to be assigned directly.

2. Protein-level assays where variants are expressed as a cDNA (*e.g.,* VAMP-seq) that include single and multinucleotide codon changes or small deletions. These datasets lack genomic coordinates and require mapping to reference transcripts and genomic coordinates.

For protein-level assays, we used transcript identifiers and the standard nuclear genetic code to infer corresponding HGVS annotations (v.21.0.4,[100,101]) for variants. For example, protein variant p.M1V in PTEN corresponds to four possible codon changes encoding valine: ENST00000371953:c.1_3delinsGTT, ENST00000371953:c.1_3delinsGTC, ENST00000371953:c.1_3delinsGTA, ENST00000371953:c.1_3delinsGTG. For each variant in

protein-level assays, all possible codon substitutions were compiled in HGVS format (as shown above) and annotated with Variant Effect Predictor (VEP) v113[82] to add genomic coordinates, variant consequences, and additional sequence-based metadata (https://zenodo.org/records/18637474).

Annotations from VEP[82] were then integrated with standardized variant identifiers and functional scores. Protein-level datasets that already reported nucleotide changes (BRCA2_Hu_2024, JAG1_Gilbert_2024, RHO_Wan_2019, LARGE1_Ma_2024, FKRP_Ma_2024) were retained as originally reported and were not expanded to all possible nucleotide substitutions to preserve the authors' intended variant representations. A small subset of variants unable to be annotated by VEP were annotated using VariantValidator[102,103] (version 3.0.1). Variants that could not be mapped using either VEP or VariantValidator were flagged with " * " (Supplementary Data 1, column `Flag`). This process produced a variant effect dataset with harmonized variant identifiers and metadata across 68 published experimental assays (Supplementary Data 1).

Integration with ClinVar, computational predictors and IGVF experimental data.

The harmonized experimental dataset was annotated with data from multiple external resources, including ClinVar classifications[1] (variant summary files from January 2025 and December 2018 releases), SpliceAI[104] predictions (precomputed SpliceAI tables based on Gencode v24), gnomAD v4 minor allele frequency[56], ClinGen Evidence Repository (version 2.5.0) and computational variant effect predictions from AlphaMissense[18], MutPred2[21], and REVEL[20]. For REVEL and MutPred2, we recorded which variants appeared in the training sets of these predictors (personal communication from authors). IGVF datasets (BARD1_IGVF, BRCA2_IGVF, CTCF_IGVF,PALB2_IGVF, RAD51D_IGVF, SFPQ_unpublished,

XRCC2_IGVF, TSC2_IGVF, G6PD_IGVF, and all F9 datasets) were incorporated and standardized using the same pipeline.

Overall, by combining curated, published datasets with newly generated data, we created an integrated variant effect dataset encompassing 83 experimental datasets across 40 clinically relevant genes, including 15 genes with more than one dataset. This resource comprised 434,166 variant effect measurements for 255,354 unique variants (295,058 measurements for 187,000 variants from published datasets and 31,858 composite scores from published meta-analyses (6,139 variants with composite scores generated from BRCA2_Sahu_2025_SGE); 98,243 newly generated measurements for 62,215 unique variants and 9,007 newly generated composite scores from a meta-analysis (*F9*)).

This integrated variant effect dataset is provided in an expanded format: each row corresponds to a nucleotide-level variant where all variants from protein level assays are mapped to all possible nucleotide substitutions (1,079,133 variants) and the associated protein-level experimental measurements are replicated and applied to each unique nucleotide-level variant (example table below) (Supplementary Data 1).

PMID	Gene	Chromosome	Hg38 Start	Hg38 End	Ref allele	Alt allele	HGVS protein	Experimental score
29785012	*PTEN*	10	87864470	87864472	ATG	GTT	p.Met1Val	1.065143
29785012	*PTEN*	10	87864470	87864472	ATG	GTC	p.Met1Val	1.065143

29785012	*PTEN*	10	87864470	87864472	ATG	GTA	p.Met1Val	1.065143
29785012	*PTEN*	10	87864470	87864470	A	G	p.Met1Val	1.065143

Calibration of experimental and predictive data

Calibration of experimental data

OddsPath calibrations: As part of our metadata curation, we extracted reported functional classifications determined by the authors representing functionally normal or functionally abnormal classes, and for some published datasets (PTEN_Matreyek_2018, select F9_Popp_2025 datasets) and all newly generated IGVF datasets, created new functional classifications using a Gaussian mixture model as in Woo et al 2025 ([36]. Variants classified as indeterminate or belonging to variant classes not relevant to the disease mechanism under study (e.g. hypermorphic) were annotated but removed from calculations. When more than one functionally normal or functionally abnormal interval or functional classification was provided (ex: ‘slow depleting’ and ‘fast depleting’ for functionally abnormal), they were combined to create one class. We calculated the Odds of Pathogenicity (OddsPath) using ClinVar controls. ClinVar controls were defined as variants with clinical significance of Pathogenic, Likely pathogenic, Pathogenic/Likely pathogenic, Benign, Likely benign, or Benign/Likely benign from the variant summary file. To map these variant classifications to protein-level assays, all possible nucleotide-level substitutions for a given amino acid variant were mapped to ClinVar assertions. When multiple nucleotide substitutions mapped to the same amino acid substitution with concordant ClinVar classifications, the less confident classification was retained (e.g. LP would

be chosen instead of P). Amino acid substitutions with discordant nucleotide-level classifications were removed.This resulted in at most one ClinVar assertion per amino acid variant for a given gene and transcript pair (see example below). For *BRCA1*, *TP53*, *MSH2*, and *PTEN*, ClinVar control classifications from 2018 were used to avoid circularity when calibrating or evaluating experimental evidence, as experimental data from these assays may have contributed to ClinVar classifications since their publication[61,62]. For MSH2_Jia_2021 the clinically relevant thresholds from Scott et al. 2022 were used[62]. We followed the framework described by Brnich et al.[25] including the addition of one misclassified variant per functional class (when necessary) to avoid posterior probabilities of 0 or infinity. Separate values for OddsNormal and OddsAbnormal were derived to assign evidence strength as ACMG/AMP points[59]. Not all experimental datasets included author-defined functional classifications or the thresholds were not provided in an accessible format preventing us from calculating OddsPath. In total, we

assigned some evidence (either OddsNormal or OddsAbnormal) to 43 of 83 experimental datasets using the OddsPath calibration method (Supplementary Data 4, OddsPath calibrations) ExCALIBR calibrations: We developed ExCALIBR, a statistical method to calibrate high-throughput experimental data[33]. Briefly, we jointly modeled scores of pathogenic, benign, gnomAD, and synonymous variants as skew normal mixtures. For *BRCA1*, *TP53*, *MSH2*, and *PTEN*, ClinVar control classifications from 2018 were used. We performed 1,000 bootstrap iterations of each dataset, using the 5th and 95th percentiles of the local positive likelihood ratio along with the median estimated prior probability of pathogenicity to define pathogenic and benign evidence strengths expressed in ACMG/AMP points with the exception that our approach enables assignment of all possible evidence points (e.g. ±1 to ±8) (Supplementary Data 4, ExCALIBR calibrations). The prior probability of pathogenicity was estimated using the score

PMID	Gene	Chrom	Hg38 Start	Hg38 End	Reference allele	Alternate allele	Hgvs protein	Experimental score	Clinical Significance
29785012	*PTEN*	10	87894048	87894050	ATG	CGT	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894050	ATG	CGC	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894050	ATG	CGA	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894049	AT	CG	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894050	TG	GA	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894049	T	G	p.Met35Arg	0.343871	Pathogenic

1 Pathogenic control retained: p.Met35Arg (*PTEN*)

PMID	Gene	Chrom	Hg38 Start	Hg38 End	Reference allele	Alternate allele	Hgvs protein	Experimental score	Clinical Significance
29785012	*PTEN*	10	87894048	87894050	ATG	CGT	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894050	ATG	CGC	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894050	ATG	CGA	p.Met35Arg	0.343871	NA
29785012	*PTEN*	10	87894048	87894049	AT	CG	p.Met35Arg	0.343871	Conflicting interpretations
29785012	*PTEN*	10	87894048	87894050	TG	GA	p.Met35Arg	0.343871	Likely Benign
29785012	*PTEN*	10	87894048	87894049	T	G	p.Met35Arg	0.343871	Pathogenic

Variant removed

distribution of gnomAD variants as a population reference[33]. When comparing ExCALIBR calibrations to author-designated functional classes (Fig. 4c), evidence was assigned in an out-of-

bag manner; i.e., each variant's evidence only considered bootstrap models that excluded it during fitting, ensuring unbiased evaluation. Conversely, threshold-based approaches may classify ClinVar controls that were simultaneously used to define the score intervals representing functional classes, potentially inflating performance.

Calibration of computational variant effect predictions

We developed a gene-specific calibration workflow that adaptively selects the optimal calibration strategy for each gene based on score distribution characteristics and data availability. Our strategy is divided into four steps.

Step 1 (gene-specific prior estimation): For each gene, we first estimated the prior probability of pathogenicity using an adapted *DistCurve* algorithm[40,42] leveraging ClinVar control pathogenic variants together with unlabelled variants from gnomAD. This procedure yields a gene-specific prior probability of pathogenicity that was subsequently used both to generate simulated datasets and to define the posterior probability ranges corresponding to ACMG/AMP evidence points.

Step 2 (score distribution modeling): For each gene, we independently modeled the pathogenic and benign variant score distributions using four parametric approaches (*Beta*, *Truncated Normal*, *Truncated Skew-t*, and *Truncated Skew-Cauchy*) and selected the best-fitting model from each approach by maximum log-likelihood.

Step 3 (simulation and method evaluation): Using the fitted distributions and the estimated gene-specific prior, we generated 30 synthetic datasets per gene with known posterior probabilities. For each gene, we evaluate 9 local calibration configurations (varying window size and gnomAD

smoothing fraction) as in Pejaver et al.[43] and 10 alternative post hoc calibration methods[42], each assessed with 1,000 bootstrap iterations. Calibration thresholds were estimated from variants not used in model training (out-of-bag), and we use the 5th percentile of posterior estimate for each threshold to mitigate overestimation of evidence strength.

Step 4 (multi-stage model selection): A single calibration model was selected for each gene through a three-stage filtering procedure: (i) an overestimation risk filter that removed calibration models exhibiting clinically unsafe behavior—defined as ≥2 evidence point error at the 75th percentile or ≥3 points maximum in pathogenic- or benign-supporting regions in at least 25% of simulations; (ii) average misestimation ranking, retaining the three calibration models with the lowest mean ACMG-point error across pathogenic, benign, and indeterminate regions; and (iii) an overestimation frequency filter that selected the calibration model with the lowest combined pathogenic- and benign-fraction, reflecting the fewest ≥1-point overestimations in key evidence regions. The selected optimal calibration model was then applied to the target gene to produce final calibrated posterior probabilities and to define predictor score intervals corresponding to each ACMG/AMP evidence strength.

Scalable workflow for applying experimental and predictive evidence to variant classification

Filtering variants for classification analyses: All analyses described below began with filtering of the integrated variant effect dataset (434,166 variant effect measurements) Supplementary Data 1). We excluded all variants in *SFPQ* because it does not have a definitive gene-disease association, a subset of variants in *CHEK2* deemed inappropriate for clinical analysis by the original authors, and all variants flagged as having unreliable or inconsistent variant annotations (Supplementary Data 1, `Flag` column). Variants receiving conflicting experimental evidence

(i.e. assigned both positive and negative evidence points from two or more assays) were removed.

We excluded *TP53* and *F9* variant effect measurements, using OddsPath calibrations on output from classifier models that integrate these data instead[32,61]. We excluded all splice-altering variants defined as having a SpliceAI score greater than or equal to 0.2 for donor (loss or gain) and acceptor (loss or gain) because although some assays are sensitive to splice variants (e.g. SGE), the REVEL, MutPred2 and AlphaMissense are not trained to detect them. We also exclude start-loss variants in non-SGE assays. Applying these filters resulted in 307,558 variant effect measurements. These filters resulted in a set of 239,836 variants across 39 genes for classification analysis.

Filtering step	Variant effect measurements	Δ Measurements	Unique variants	Δ Variants
Integrated variant effect dataset (Supplementary Data 1)	434,166		255,354	
Replace F9 and TP53 datasets with model calibrations; remove SFPQ	325,151	-109,015	252,755	-2,599
Exclude *CHEK2* variants considered unsuitable for clinical analysis based on author guidance[105]	322,446	-2,705	250,050	-2,705

Remove variants with conflicting experimental evidence	317,479	-4,967	248,161	-1,889
Remove predicted splice variants (Splice AI ≥ 0.2)	307,778	-9,701	239,967	-8,194
Remove start-loss variants in non-SGE assays	307,651	-127	239,907	-60
Remove variants with inconsistent mapping and annotations (Flag column = *)	307,558	-93	239,836	-71

In addition to these filters, when evaluating REVEL and MutPred2, we excluded variants in the respective predictor that were used in training (REVEL, 115 variants pre-filtering; MutPred2, 3,703 variants pre-filtering).

Summing evidence points to generate classifications: To generate variant classifications, we summed experimental and predictor evidence points. The experimental evidence points were generated by either the ExCALIBR or OddsPath calibrations (Supplementary Data 4, OddsPath calibrations, ExCALBR calibrations). The predictive evidence points were generated for REVEL, MutPred2, or AlphaMissense using gene-specific calibrations (Supplementary Data 4, Gene-specific calibrations) or genome-wide calibrations from Bergquist et al. for all other genes[42–44]. The summed evidence points yielded a classification in the points-based ACMG/AMP framework[59] (Supplementary Data 5, Supplementary Data 6). For all downstream analysis, for duplicate variants across assays, we retained the variant classification associated with the largest absolute point value from our scalable workflow.

ClinVar control analysis: ClinVar controls were defined as variants with clinical significance of Pathogenic, Likely pathogenic, Pathogenic/Likely pathogenic or Benign, Likely benign, Benign/Likely benign from the ClinVar variant summary file downloaded from January 2025 and December 2018 releases. ClinVar controls were filtered to retain variants with 1+ stars in ClinVar meaning that some assertion criteria were provided by the submitter. We used controls from ClinVar January 2025 except for *BRCA1*, *TP53*, *MSH2*, and *PTEN* where classifications from 2018 were used to avoid comparison with variants classified with evidence from these datasets. For protein-level assays, at most one ClinVar assertion per amino acid variant for a given gene and transcript pair was retained (OddsPath calibrations). After applying filters mentioned above, we retained 10,228 ClinVar control variants for analysis with REVEL (excluding variants used in REVEL training), 10,250 ClinVar controls for AlphaMissense, and 9,099 ClinVar controls for MutPred2 (excluding variants used in MutPred2 training).

To quantify the performance of our scalable workflow, we assessed the concordance between each variant's classification from our workflow with its original ClinVar classification. True positives (TP) were defined as variants for which our pathogenic or likely pathogenic (PLP) classifications were concordant with the original ClinVar classification PLP. True negatives (TN) were defined as variants for which our benign or likely benign (BLB) classifications were concordant with the original ClinVar classification of BLB. False positives (FP) were defined as variants classified as PLP by our scalable workflow but classified as BLB in ClinVar. False negatives (FN) were defined as variants classified as BLB by our scalable workflow but classified as PLP in ClinVar. We then calculated the following metrics to assess our framework:

$$Sensitivity = \frac{TP}{TP + FN}, Specificity = \frac{TN}{TN + FP}, MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

$$Concordance = \frac{TP + TN}{TP+TN+FP+FN}$$

We deemed classifications of BLB and PLP “determinate” (in contrast to VUS). We therefore computed the fraction of determinate classifications by dividing the number of variants classified as PLP and BLB by the total number of variants classified. We deemed classifications “discordant” if our workflow gave a determinate classification that disagreed with ClinVar (e.g. a variant we classified as BLB that was PLP in ClinVar and vice versa). Thus, we computed the fraction of discordant classifications by dividing the number of BLB->PLP and PLP->BLB classifications by the total number of variants analyzed (Supplementary Data 5, Supplementary Data 6).

ClinGen Evidence Repository analysis: We retrieved all nucleotide variants from the ClinGen Evidence Repository for our genes which included all evidence codes used to classify each variant (downloaded November 10th, 2025, v 2.5.0). We mapped all possible nucleotide substitutions for amino acid variants from protein-level assays to variants found in the ClinGen Evidence Repository, and found no conflicting classifications. 351 variants had determinate (PLP, BLP) classifications in the Repository and experimental data. For each variant, we removed the experimental evidence (BS3, PS3) and/or predictive evidence (PP3, BP4) used in the original classifications and then recalculated the classification without this evidence[57]. 194 variants across 11 genes retained sufficient evidence to remain BLB (19) or PLP (175). Of these, post filtering, we retained 179 variants for analysis with REVEL (excluding variants used in REVEL training), 182 variants for AlphaMissense, and 35 variants for MutPred2 (excluding variants used in MutPred2 training) We then assessed the concordance of our classifications using only experimental or predictive evidence to the expert panel classifications generated from all other sources of evidence using the same performance metrics as the ClinVar control analysis (Supplementary Data 5, Supplementary Data 6).

ClinVar VUS, gnomAD, and unobserved variants analysis:

VUS were defined as variants classified as Uncertain significance in the January 2025 ClinVar variant summary file (Supplementary Data 1). gnomAD variants were defined as variants observed in gnomAD (i.e., with a reported allele frequency). Unobserved variants were defined as single nucleotide variants absent from both ClinVar in January 2025 and gnomAD v4. Post filtering, we retained 16,115 VUS, 24,992 gnomAD variants, and 90,530 unobserved variants for analysis with REVEL. For analysis with AlphaMissense, we retained 16,149 VUS, 25,063 gnomAD variants, and 90,550 unobserved variants. For MutPred2 we retained 15,705 VUS, 23,918 gnomAD variants and 89,463 unobserved variants.

Variant classifications for ClinVar VUS, gnomAD variants, and unobserved variants were determined using the same scalable framework described above. For each nucleotide variant, experimental and predictive evidence points were summed, and the resulting total points were used to assign a classification according to the framework described by Tavtigian et al[59]. (Supplementary Data 5, Supplementary Data 6).

Biobank validation of variant classes

To further validate our scalable workflow, we assessed the relationship between variant classes (defined by experimental data, evidence assigned by calibrations of experimental and predictive data, and classifications) and clinical phenotypes in ~400,000 participants in the All of Us biobank. The validation was performed within the All of Us Workbench using the All of Us Controlled Tier Dataset v8[106] Curated Data Repository (CDR) version 8 Release Notes – User Support.

Gene-specific cohort definition

For each of the 40 clinically-relevant genes we analyzed, we evaluated whether the gene is known to be associated with an autosomal dominant phenotype that can be defined in terms of diagnosis codes (reviewed by a medical genetics physician and a molecular pathologist), measurements, and surveys recorded in All of Us (AoU) and whether there is a sufficient number of participants (at least 40) with whole genome sequencing data in the biobank that fit each definition and have variants in the corresponding gene. We determined that 23 gene-disease pairs did not meet these criteria (Supplementary Table 2), yielding 17 gene-disease pairs for analysis. For each of these 17 gene-disease pairs, we established a case cohort by selecting participants that had specific conditions, measurements, or survey answers that indicated they had the phenotype of interest. For each gene-disease pair, we also established a control cohort by excluding participants that had broad conditions, measurements, or survey answers that indicated they might have the phenotype of interest or a closely related phenotype (Supplementary Table 3, Supplementary Data 8).

Definition of carrier status for a set of variants

We aggregated variants into sets based on four classification schemes: (i) experimental data functional classes, (ii) ExCALIBR-calibrated experimental evidence aggregated by assigned points, (iii) calibrated predictive evidence aggregated by assigned points (iv) the sum of experimental and predictive evidence aggregated by assigned evidence points. For each variant set, we queried the All of Us short-read whole genome sequencing matrix to identify carriers, defined as participants with one or more variants in that set.

Odds ratio estimation

For every set of variants considered (e.g. functional class, variants with ≥2 points of experimental evidence), we estimate the odds ratio of disease phenotype given carrier status by

fitting a logistic regression model to the cohort of cases and controls established for the corresponding gene as described above. When the variant set is defined by computational predictors or assays that do not measure the effect of splicing, we additionally exclude all participants with any variants in the gene that have a SpliceAI score above 0.2 for donor or acceptor gain or loss. The model response variable is case or control status, and explanatory variables are carrier status for the variant set, sex assigned at birth (trichotomized into male, female, and other), age at the time of AoU data release in years, and 16 genetic ancestry PCA components computed by the AoU research program[107]. We report the point estimate, 95% confidence interval, and Wald test p-value for the regression coefficient associated with the carrier status variable, which is the natural log of the odds ratio of interest (Supplementary Data 2, Supplementary Data 8).

4.5 Data Availability

Raw and analyzed SGE and VAMP-seq data, calibrations, arrayed assay data, and Supplementary Data 1 are available from the IGVF Portal, see Supplementary Table 1 for accession numbers.

SGE and VAMP-seq results are available on MaveDB and MaveMD see Supplementary Table 1 for MaveDB urns.

All supplementary data files and large supporting files for analysis and figure generation are also hosted at https://zenodo.org/records/18637474.

4.6 Code Availability

Code supporting this study is publicly available at the following repositories:

SGE analysis pipeline: https://github.com/bbi-lab/sge-pipeline/

VAMP-seq analysis pipeline (CountESS): https://github.com/CountESS-Project/CountESS

Pipelines for arrayed assays: https://github.com/broadinstitute/2025_Pillar_VarChAMP

Biobank analyses https://github.com/Craven-Biostat-Lab/igvf-cvfg-biobank

The integrated variant effect datasets pipeline and variant classification analysis are available at: https://github.com/bbi-lab/IGVF-cvfg-pillar-project, including scripts to generate main and extended data figures.

4.7 Supplementary Notes

Supplementary Note 1

PTEN Variant Misclassification

One misclassified variant was *PTEN* (NM_000314.6:c.500C>A), a *de novo* variant associated with autism spectrum disorder. Experimental evidence supported a benign classification (−2 points) while predictive evidence supported pathogenicity (+1 point), resulting in −1 combined points and an LB classification. This discordance suggests that the experimental assay measuring lipid phosphatase activity[108] may not fully capture the ASD-relevant mechanism for PTEN. Specifically, NM_000314.8(PTEN):c.500C>A (p.Thr167Asn) was identified *de novo* in a 99-month-old white female with autism spectrum disorder from the Simons Simplex Collection (SSC; proband 11390.p1)[109]. The individual presented with macrocephaly, speech delay with regression, social communication deficits, ADHD, pica, sleep disturbances, recurrent otitis media, and gastrointestinal issues. Family history is notable for autism features in the mother, ADHD in a sibling, and an extensive family history of neuropsychiatric conditions. Both parents and the sibling did not have macrocephaly. While functional evidence does not show that this variant is low abundance or lacking lipid phosphatase activity, we have recently learned PTEN

variants associated with neurodevelopmental disorders may disrupt PTEN function through other mechanisms. A recent study which used visual phenotype to assess variant effects on PTEN subcellular localization demonstrated that this variant is nuclearly mislocalized, similar to other PTEN variants known to be associated with neurodevelopmental disorders[91].

Supplementary Note 2

Minimum Sensitivity for BS3_supporting

The likelihood ratio for a normal assay result (LR−) equals (1 − sensitivity) divided by specificity. Setting LR− equal to the BS3_supporting threshold of 0.48 and solving for sensitivity at 90% specificity yields a minimum sensitivity of 56.8%, calculated as 1 − (0.48 × 0.90). An assay with 90% specificity can therefore have sensitivity as low as 57% and still qualify for BS3_supporting, meaning such an assay fails to detect 43% of pathogenic variants.

False Omission Rate at Different Priors

The false omission rate (FOR) equals the product of (1 − sensitivity) and the prior probability, divided by the sum of that product and the product of specificity and (1 − prior). At 57% sensitivity and 90% specificity, the FOR is 4.5% when the prior probability of pathogenicity is 10%, rises to 13.7% at a 25% prior, and reaches 32.3% at a 50% prior. When the prior exceeds the ACMG/AMP assumption of 10%, the proportion of misclassified pathogenic variants increases substantially. At a 50% prior, nearly one in three likely benign classifications based on single-source weak evidence would be incorrect

4.8 Figures

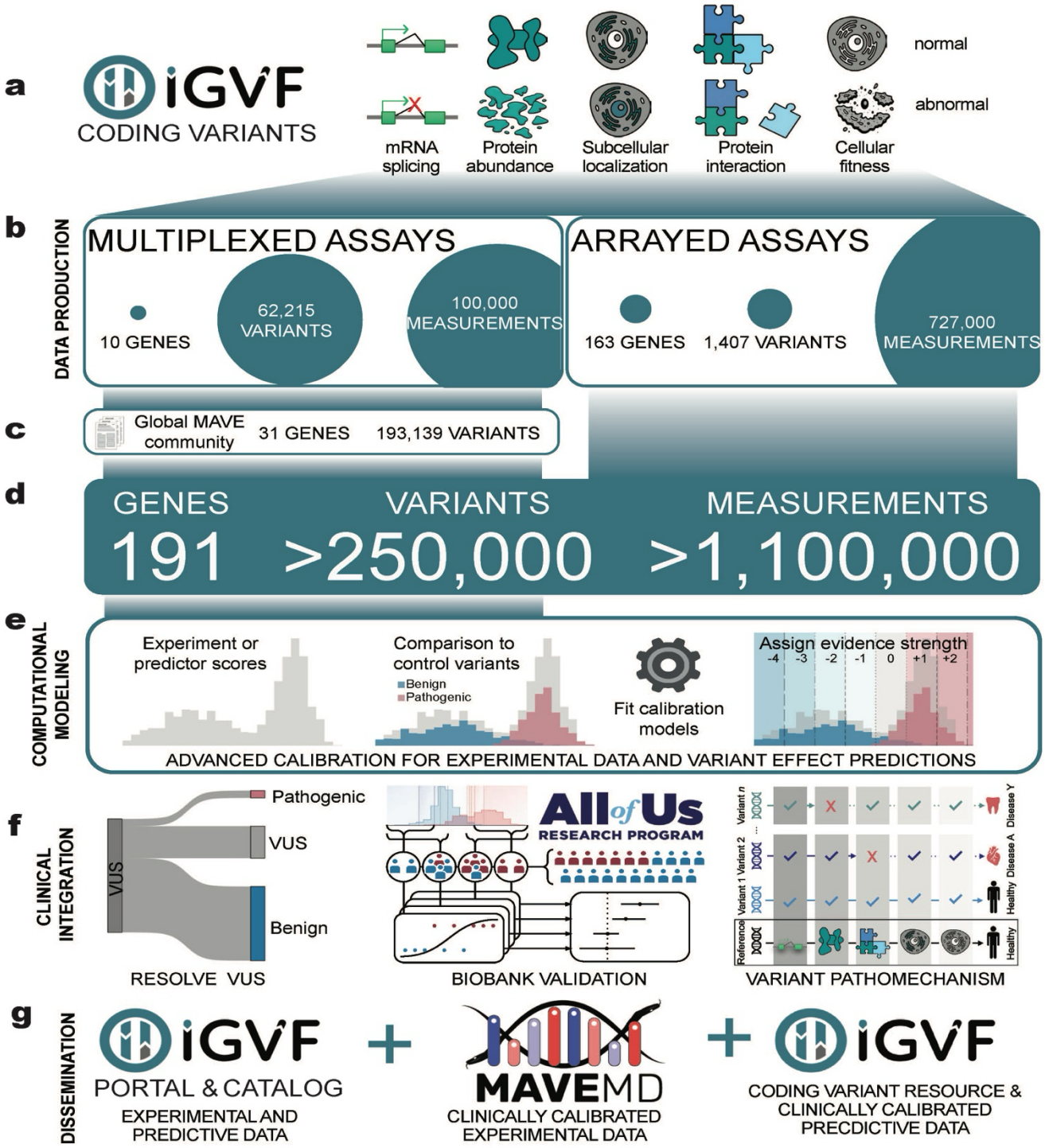

a
iGVF
CODING VARIANTS
normal
abnormal
mRNA splicing
Protein abundance
Subcellular localization
Protein interaction
Cellular fitness
b
DATA PRODUCTION
MULTIPLEXED ASSAYS
10 GENES
62,215 VARIANTS
100,000 MEASUREMENTS
ARRAYED ASSAYS
163 GENES
1,407 VARIANTS
727,000 MEASUREMENTS
c
Global MAVE community
31 GENES
193,139 VARIANTS
d
GENES
191
VARIANTS
>250,000
MEASUREMENTS
>1,100,000
e
COMPUTATIONAL MODELING
Experiment or predictor scores
Comparison to control variants
Benign
Pathogenic
Fit calibration models
Assign evidence strength
-4 -3 -2 -1 0 +1 +2
ADVANCED CALIBRATION FOR EXPERIMENTAL DATA AND VARIANT EFFECT PREDICTIONS
f
CLINICAL INTEGRATION
Pathogenic
VUS
Benign
VUS
RESOLVE VUS
All of Us
RESEARCH PROGRAM
BIOBANK VALIDATION
Variant n
Variant 2
Variant 1
Reference
Disease Y
Disease A
Healthy
Healthy
VARIANT PATHOMECHANISM
g
DISSEMINATION
iGVF
PORTAL & CATALOG
EXPERIMENTAL AND PREDICTIVE DATA
+
MAVEMD
CLINICALLY CALIBRATED EXPERIMENTAL DATA
+
iGVF
CODING VARIANT RESOURCE & CLINICALLY CALIBRATED PRECDICTIVE DATA

Figure 4.1 Integrated framework for experimental and computational characterization of coding variant effects

a. The IGVF Coding Variants Focus Group characterizes variant effects across diverse functional readouts including mRNA splicing, protein abundance, subcellular localization, protein-protein interactions, and cellular fitness.
b. IGVF data production teams generated experimental data using multiplexed assays measuring 60,000 variants across 10 genes (100,000 measurements) and arrayed assays measuring 1,407 variants across 163 genes (600,000 measurements).
c. Integration with curated experimental data from the global MAVE community expands data coverage.
d. The integrated variant effect dataset comprises over 1 million functional measurements across 262,000 variants in 200 clinically relevant genes.
e. IGVF computational modeling teams developed advanced calibration methods for both experimental data and variant effect predictions.
f. The integrated variant effect dataset enables systematic VUS resolution, biobank validation of variant classifications in diverse populations, and mechanistic characterization of variant pathomechanisms.
g. Data is disseminated through three complementary resources: the IGVF Portal (raw experimental and predictive data), IGVF Catalog (browsable experimental and predictive data), and MaveMD (clinically calibrated experimental data for immediate clinical use).

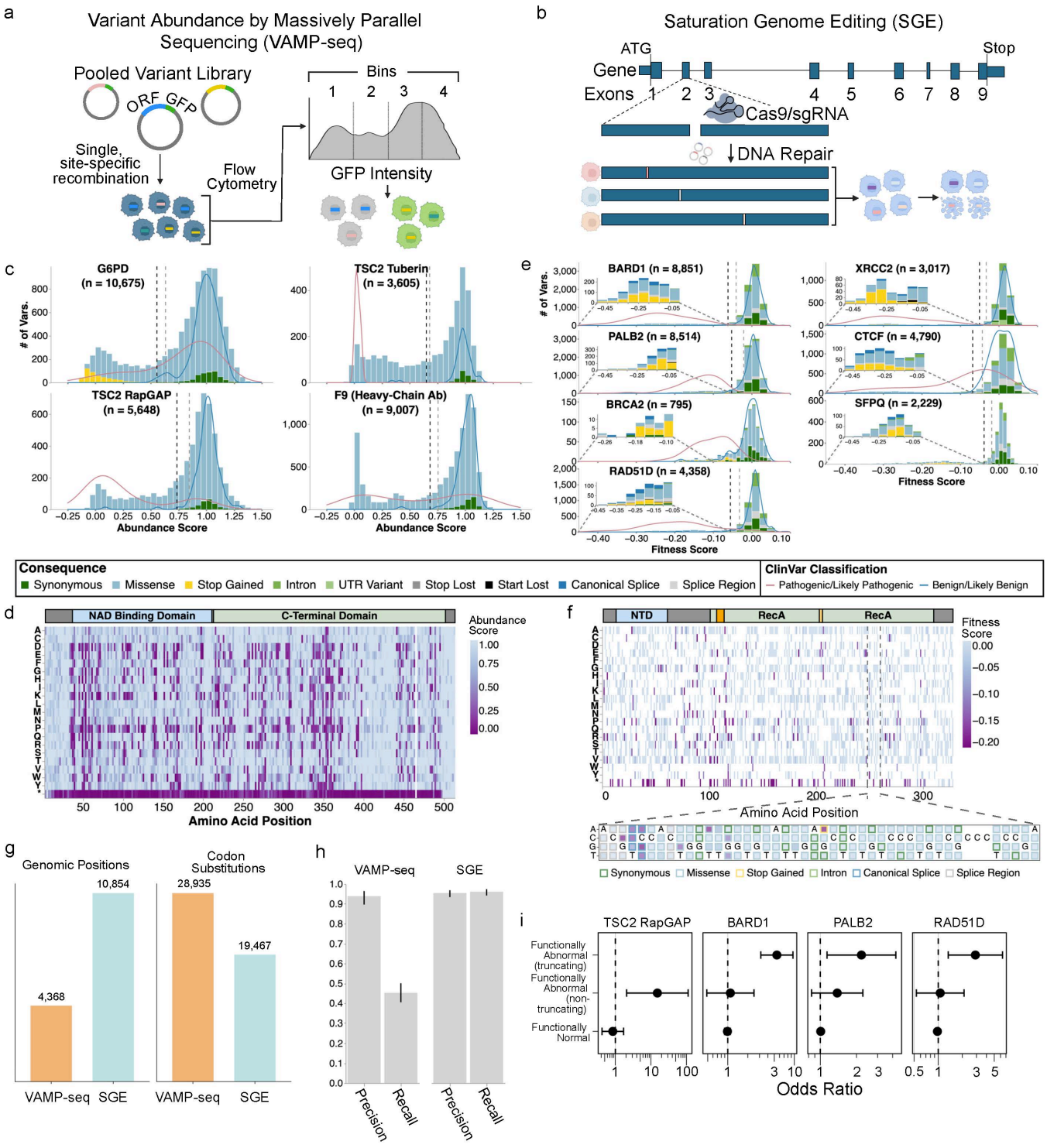

a
Variant Abundance by Massively Parallel Sequencing (VAMP-seq)
Pooled Variant Library
ORF GFP
Single, site-specific recombination
Flow Cytometry
Bins
1 2 3 4
GFP Intensity
b
Saturation Genome Editing (SGE)
ATG
Stop
Gene
Exons 1 2 3 4 5 6 7 8 9
Cas9/sgRNA
DNA Repair
c
G6PD (n = 10,675)
TSC2 Tuberin (n = 3,605)
TSC2 RapGAP (n = 5,648)
F9 (Heavy-Chain Ab) (n = 9,007)
# of Vars.
Abundance Score
e
BARD1 (n = 8,851)
XRCC2 (n = 3,017)
PALB2 (n = 8,514)
CTCF (n = 4,790)
BRCA2 (n = 795)
SFPQ (n = 2,229)
RAD51D (n = 4,358)
Fitness Score
Consequence
Synonymous
Missense
Stop Gained
Intron
UTR Variant
Stop Lost
Start Lost
Canonical Splice
Splice Region
ClinVar Classification
Pathogenic/Likely Pathogenic
Benign/Likely Benign
d
NAD Binding Domain
C-Terminal Domain
Abundance Score
Amino Acid Position
f
NTD
RecA
RecA
Fitness Score
Amino Acid Position
Synonymous
Missense
Stop Gained
Intron
Canonical Splice
Splice Region
g
Genomic Positions
10,854
4,368
VAMP-seq
SGE
Codon Substitutions
28,935
19,467
VAMP-seq
SGE
h
VAMP-seq
SGE
Precision
Recall
Precision
Recall
i
TSC2 RapGAP
BARD1
PALB2
RAD51D
Functionally Abnormal (truncating)
Functionally Abnormal (non-truncating)
Functionally Normal
Odds Ratio

Figure 4.2 IGVF-produced MAVE data for 62,000 variants

a. Schematic of experimental workflow for Variant Abundance by Massively Parallel Sequencing (VAMP-seq)

b. Schematic of experimental workflow for Saturation Genome Editing (SGE)

c. Histograms showing the distribution of VAMP-seq scores for the indicated genes and domains. The molecular consequence of variants are colored as indicated. Black and gray dashed vertical lines indicate thresholds for variants called functionally abnormal or functionally normal respectively. Overlaid density plots show the distribution of pathogenic/likely pathogenic (red) and benign/likely benign (blue) from ClinVar.

d. Heatmap representation of VAMP-seq scores for G6PD. A cartoon of G6PD domains is at the top. Amino acid positions in G6PD are on the X-axis and amino acid and stop-gained (*) substitutions are indicated on the Y-axis. Abundance scores are colored as indicated.

e. Histograms showing the distribution of SGE fitness scores for the indicated genes. The molecular consequence of variants are colored as indicated. Black and gray dashed vertical lines indicate thresholds for variants called functionally abnormal or functionally normal respectively for datasets with at least 1,000 variants scored. Insets highlight functionally abnormal variants or all variants scoring < -0.10 for datasets without thresholds. Overlaid density plots represent the distribution of pathogenic/likely pathogenic (red) and benign/likely benign (blue) from ClinVar for each gene with at least five PLP and BLB variants.

f. Amino acid-level heatmap of SGE fitness scores for missense and stop-gained variants in RAD51D with a cartoon of RAD51D protein domains (top). A nucleotide-level heatmap for the indicated region of *RAD51D* (bottom). At each position, the wild-type base is written and alternate bases are denoted on the Y-axis. Colored borders for each variant indicate the molecular consequence and fitness scores for each variant are filled as indicated.

g. Bar plot comparing the number of nucleotide positions (left) targeted by VAMP-seq (orange) and SGE (teal) across genes and number of unique codon substitutions (right) assayed by VAMP-seq and SGE.

h. Bar plot showing the precision and recall for VAMP-seq (right) and SGE (left) datasets on known pathogenic/likely pathogenic ClinVar variants. Black vertical lines at the end of each bar denote a 95% confidence interval.

i. Odds ratios for occurrence of disease in individuals with functionally abnormal and functionally normal variants in the All of Us biobank for the TSC2 RapGAP domain, BARD1, PALB2 and RAD51D. X-axis denotes the odds ratio with the black vertical dashed line indicating an odds ratio of 1. Y-axis denotes the functional consequence of variants used in the analysis. Points denote the estimated odds ratio and whiskers denote the 95% confidence interval.

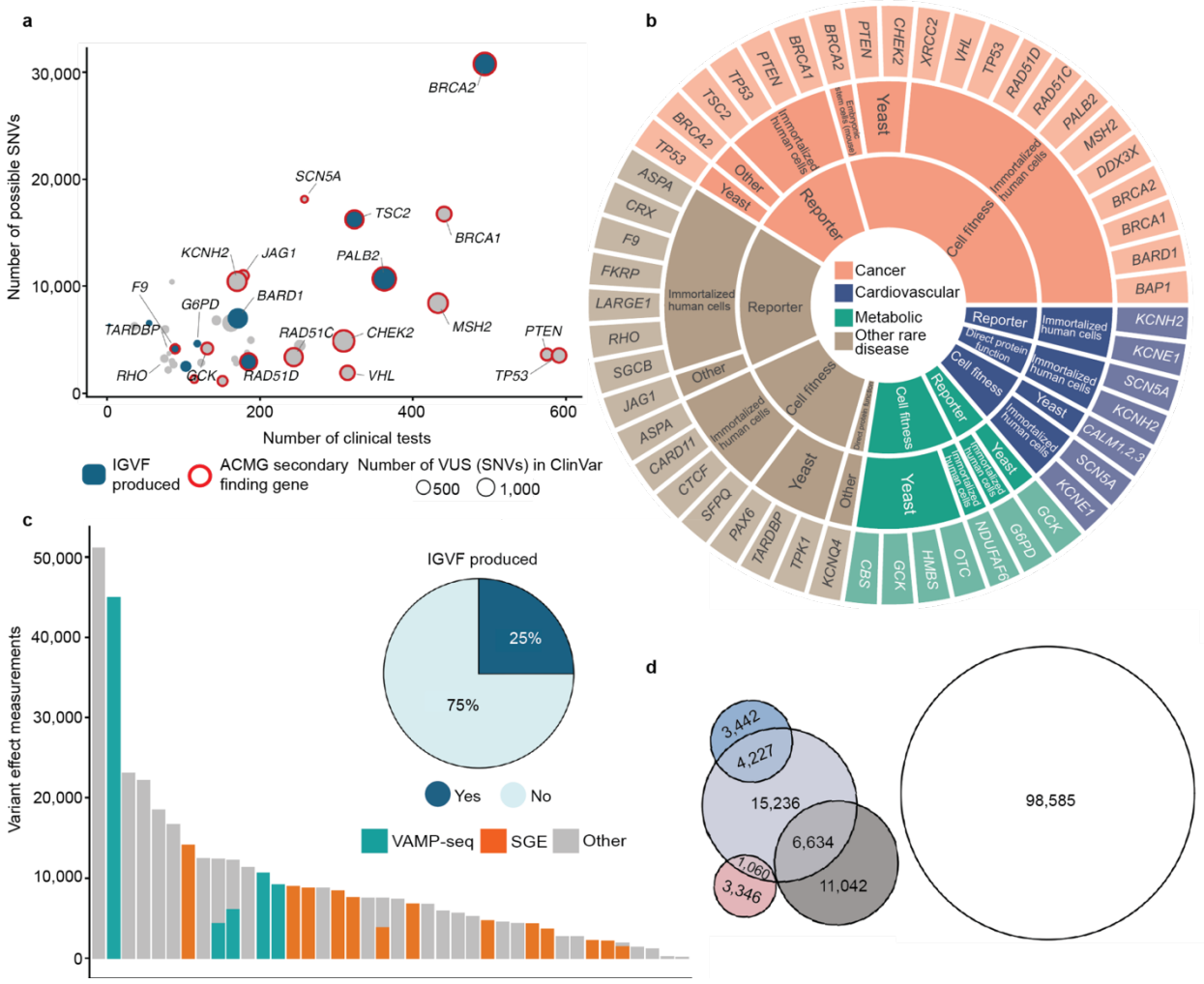


Figure 4.3: Curating an expanded set of experimental data from the MAVE community

a. Each bubble represents a gene in the integrated variant effect dataset. The number of genetic tests from the Genetic Testing Registry is on the x-axis and the y-axis indicates the total number of possible coding single-nucleotide variants (SNVs). Bubble size is proportional to the number of SNV ClinVar VUS. Bubble color denotes whether experimental data was generated by the IGVF consortium and red outlines denote ACMG secondary findings genes.
b. Sunburst plot summarizing the distribution of experimental data across genes (outer ring), experimental modalities (middle ring) and model systems (inner ring). Segments are colored according to clinical association for each gene as indicated (center).
c. Bar plot showing the number of variant effect measurements (y-axis) per gene (x-axis). Bars are colored by assay type (SGE, VAMP-seq, or other functional assays). A pie chart indicates the proportion of variant effect measurements generated by the IGVF consortium.
d. Venn diagram showing the number of variants that have not yet been observed as well as those in gnomAD or ClinVar (pathogenic/likely pathogenic, benign/likely benign, and VUS colored as indicated).

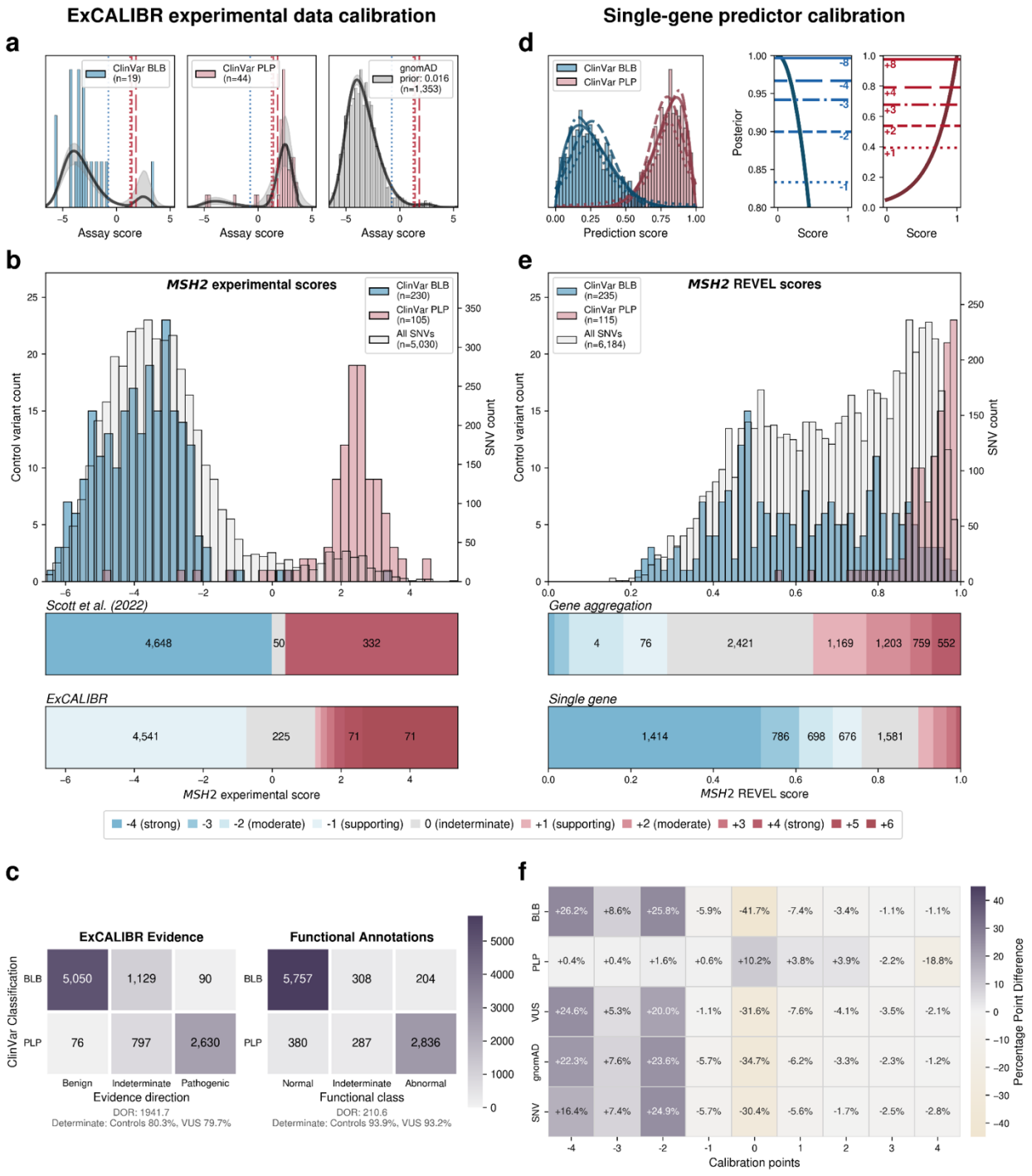


Figure 4.4: Improved gene- and variant-specific calibration methods improve calibration of experimental and predictive data

a. ExCALIBR experimental data calibration strategy illustrated using *MSH2* data[14]. Score distributions of BLB (blue) and PLP (red) variants from ClinVar (December 2018), as well as population variants from gnomAD (gray), are jointly modeled using skew normal mixtures. The prior probability of pathogenicity is estimated using the gnomAD sample, which is used

alongside the local positive likelihood ratio to determine evidence strength thresholds. Evidence strength thresholds are drawn as vertical lines, with red and blue representing pathogenic and benign evidence strength thresholds, respectively (short dashes = 1 point, medium dashes = 2 points, long dashes = 4 points).

b. Comparison of *MSH2* evidence strength thresholds. Top: experimental score distributions for all SNVs (gray), ClinVar BLB (blue) and PLP (red) controls (January 2025). Middle: author-defined evidence thresholds[62]. Bottom: ExCALIBR calibrated thresholds (using December 2018 ClinVar release). Total SNV counts per evidence category are displayed in each bin.

c. Comparison of ExCALIBR evidence assignments (left) and functional annotations (right) for ClinVar PLP and BLB variants for 28 genes with annotated thresholds. For *BRCA1*, *MSH2*, and *PTEN*, ClinVar controls from the December 2018 release were evaluated. Evidence assignments less than zero indicate benign direction; greater than zero indicate pathogenic direction; zero is indeterminate. Diagnostic odds ratios are computed on determinate calls. Determinate assignment rates are shown separately for control variants and VUS.

d. Data-adaptive, single-gene calibration framework. For genes with sufficient control variants, simulation analyses evaluate calibration feasibility and identify the best-fitting gene-specific control distribution. Across 30 simulations and 10 calibration methods, posterior probabilities are estimated. The optimal method is selected based on ACMG/AMP-aligned miscalibration metrics, prioritizing accuracy, robustness, and control of overestimation, and is used to derive calibrated score thresholds and corresponding evidence assignments.

e. Calibration results for *MSH2* REVEL scores. Top: Histogram of REVEL scores for ClinVar PLP (red), BLB (blue), and all possible missense SNVs (gray), with SNV counts shown on the right y-axis. Middle: The number of variants with the indicated evidence assignment for REVEL using the genome-wide[43]or (bottom) for single-gene calibrations.

f. A confusion matrix shows the difference in the percentage of variants assigned evidence points between the gene-specific method and the genome-wide method for all 8 genes[43] for variants with ClinVar classifications, gnomAD variants and all possible SNVs.

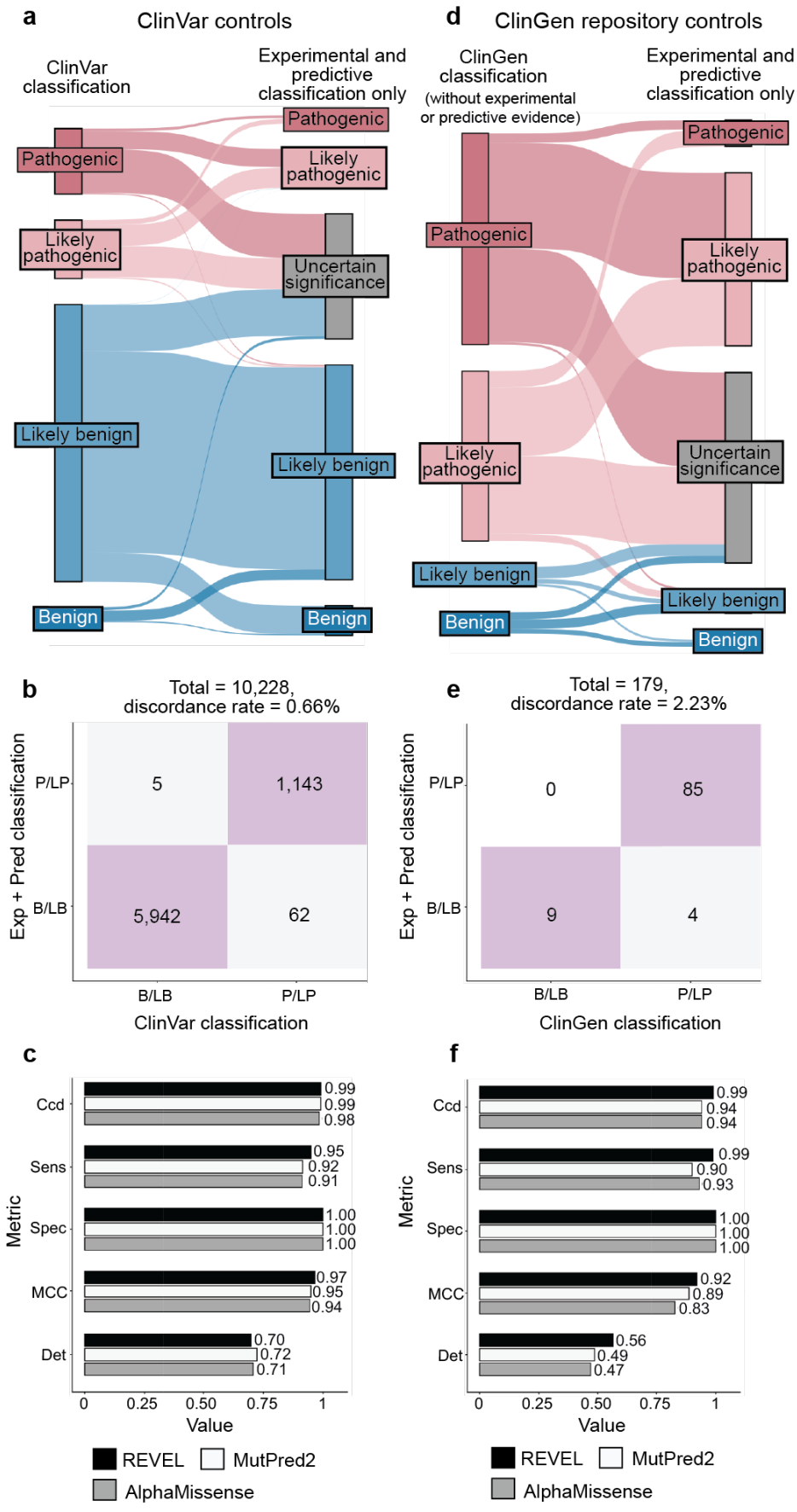
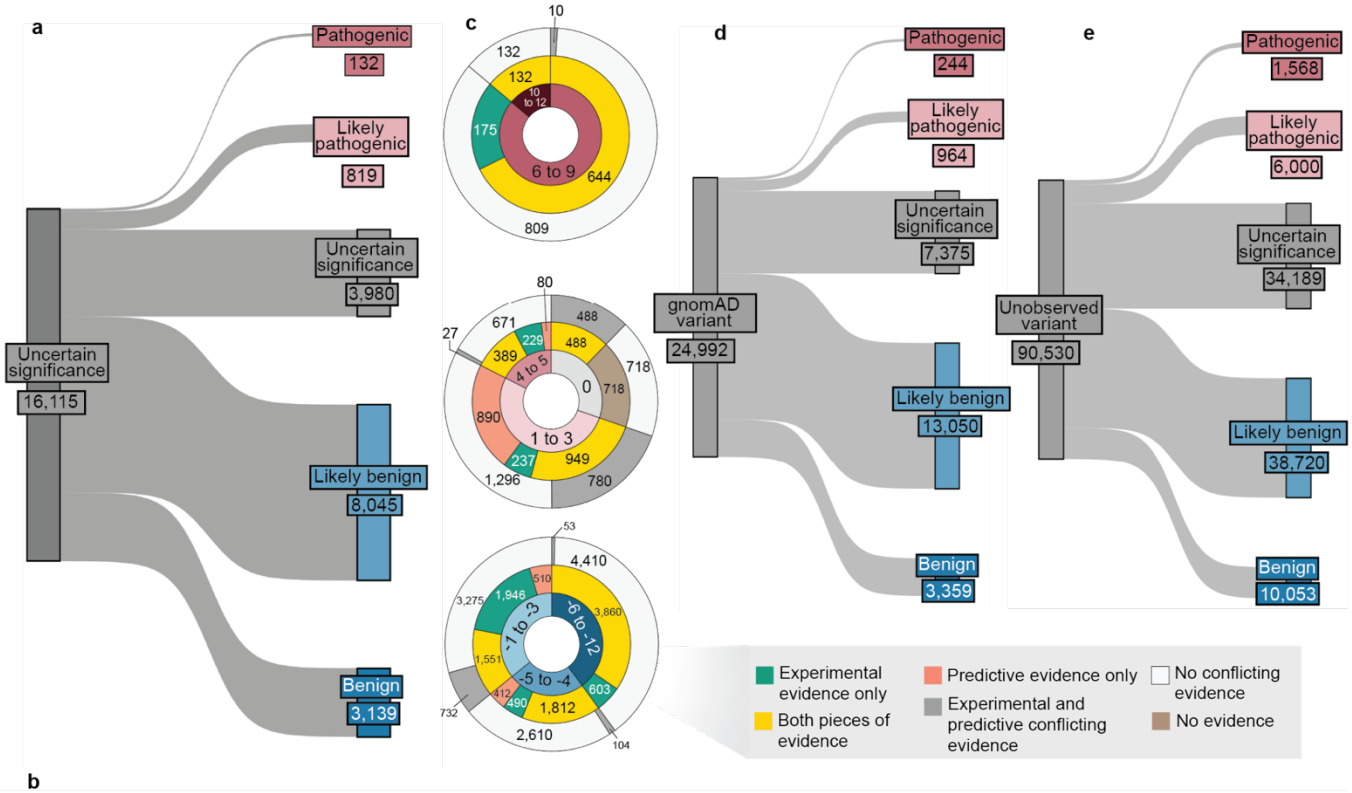

a
ClinVar controls
ClinVar classification
Experimental and predictive classification only
Pathogenic
Likely pathogenic
Likely benign
Benign
Pathogenic
Likely pathogenic
Uncertain significance
Likely benign
Benign
d
ClinGen repository controls
ClinGen classification (without experimental or predictive evidence)
Experimental and predictive classification only
Pathogenic
Likely pathogenic
Likely benign
Benign
Pathogenic
Likely pathogenic
Uncertain significance
Likely benign
Benign
b
Total = 10,228, discordance rate = 0.66%
Exp + Pred classification
P/LP
B/LB
5
1,143
5,942
62
B/LB
P/LP
ClinVar classification
e
Total = 179, discordance rate = 2.23%
Exp + Pred classification
P/LP
B/LB
0
85
9
4
B/LB
P/LP
ClinGen classification
c
Metric
Ccd
Sens
Spec
MCC
Det
0.99
0.99
0.98
0.95
0.92
0.91
1.00
1.00
1.00
0.97
0.95
0.94
0.70
0.72
0.71
0
0.25
0.50
0.75
1
Value
REVEL
MutPred2
AlphaMissense
f
Metric
Ccd
Sens
Spec
MCC
Det
0.99
0.94
0.94
0.99
0.90
0.93
1.00
1.00
1.00
0.92
0.89
0.83
0.56
0.49
0.47
0
0.25
0.50
0.75
1
Value
REVEL
MutPred2
AlphaMissense

Figure 4.5 99.3% of PLP and BLB classifications from experimental and predictive evidence match ClinVar

a. Sankey diagram depicting the relationship of the variant classifications in ClinVar and their classifications produced by our scalable workflow using only experimental and predictive (REVEL) evidence.
b. Confusion matrix comparing original ClinVar classifications to classifications from our workflow, summarizing concordant (lilac) and discordant (grey) variant classifications.
c. Bar plots summarizing the concordance of classifications using only experimental and predictive evidence with ClinVar controls. Predictive evidence from REVEL, AlphaMissense, and MutPred2 as indicated. Ccd = concordance, Sens = sensitivity, Sepc = specificity, MCC = Matthew's correlation coefficient, Det = determinate classifications.
d. Sankey diagram depicting the relationship between variant classifications from the ClinGen Evidence Repository, updated to exclude experimental and predictive evidence (see Methods) to classifications assigned by only experimental and predictive (REVEL) evidence.
e. Confusion matrix comparing updated ClinGen Evidence Repository classifications updated to exclude experimental and predictive evidence to classifications assigned using only experimental and predictive evidence, colored as in (b).
f. Bar plots summarizing the concordance of classifications using only experimental and predictive evidence with updated ClinGen Evidence Repository controls. Predictive evidence from REVEL, AlphaMissense, and MutPred2 as indicated.


a
Uncertain significance
16,115
Pathogenic
132
Likely pathogenic
819
Uncertain significance
3,980
Likely benign
8,045
Benign
3,139
c
10
132
132
10 to 12
175
6 to 9
644
809
80
671
488
488
229
27
389
4 to 5
718
718
0
890
1 to 3
237
949
1,296
780
53
4,410
510
3,275
1,946
-7 to -3
3,860
-8 to -12
1,551
-5 to -4
603
412
490
1,812
732
2,610
104
d
gnomAD variant
24,992
Pathogenic
244
Likely pathogenic
964
Uncertain significance
7,375
Likely benign
13,050
Benign
3,359
e
Unobserved variant
90,530
Pathogenic
1,568
Likely pathogenic
6,000
Uncertain significance
34,189
Likely benign
38,720
Benign
10,053
Experimental evidence only
Both pieces of evidence
Predictive evidence only
Experimental and predictive conflicting evidence
No conflicting evidence
No evidence

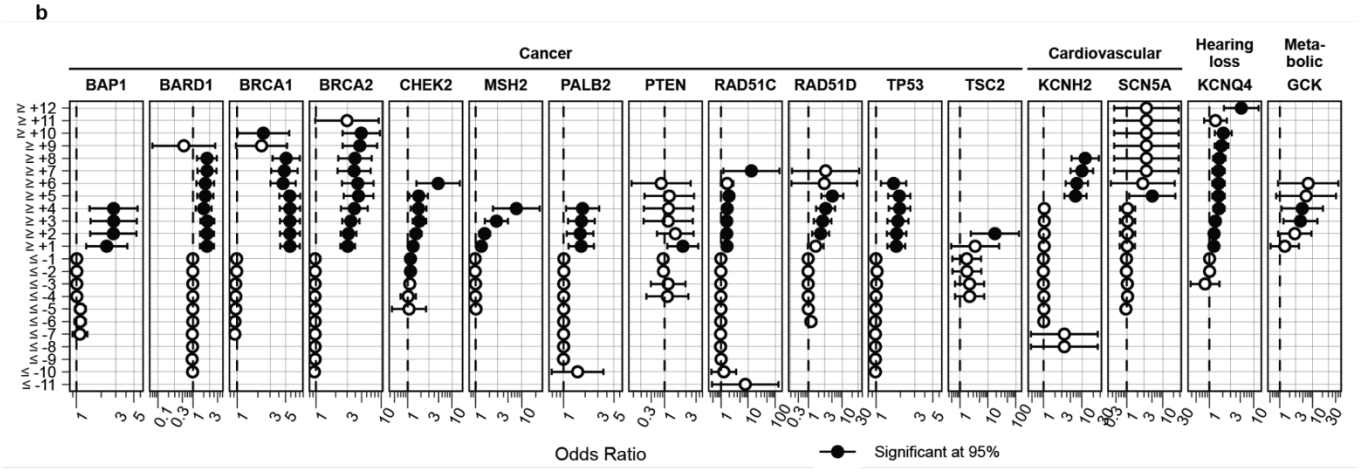

b
Cancer
Cardiovascular
Hearing loss
Meta-bolic
BAP1
BARD1
BRCA1
BRCA2
CHEK2
MSH2
PALB2
PTEN
RAD51C
RAD51D
TP53
TSC2
KCNH2
SCN5A
KCNQ4
GCK
Odds Ratio
Significant at 95%

Figure 4.6 Experimental and predictive evidence can resolve ~75% of VUS and preclassify 62% of unobserved variants

a. Sankey diagram depicting the relationship of ClinVar VUS and classifications using only experimental and predictive evidence.
b. Odds ratios for occurrence of disease in individuals in the All of Us biobank with variants meeting each indicated evidence strength threshold for genes that were found to have significant association with disease at some level of pathogenic evidence (**Supplementary Data 2**). X-axis denotes the odds ratio with the black vertical dashed line indicating an odds ratio of 1. Y-axis denotes the total evidence points for sets variants. Circles denote the estimated odds ratio and whiskers denote the 95% confidence interval.
c. Donut plots summarizing the strength, source, and consistency of evidence across ClinVar VUS shown in (a). The inner ring denotes the cumulative evidence points assigned to variants. The middle ring indicates the source of evidence as indicated. The outer ring shows the number of variants with concordant (evidence in the same direction) versus conflicting (evidence in opposite directions) evidence for each evidence source category.
d. Sankey diagram depicting the relationship of gnomAD variants to classifications using only experimental and predictive evidence.
e. Sankey diagram depicting the relationship of unobserved variants to classifications using only experimental and predictive evidence.

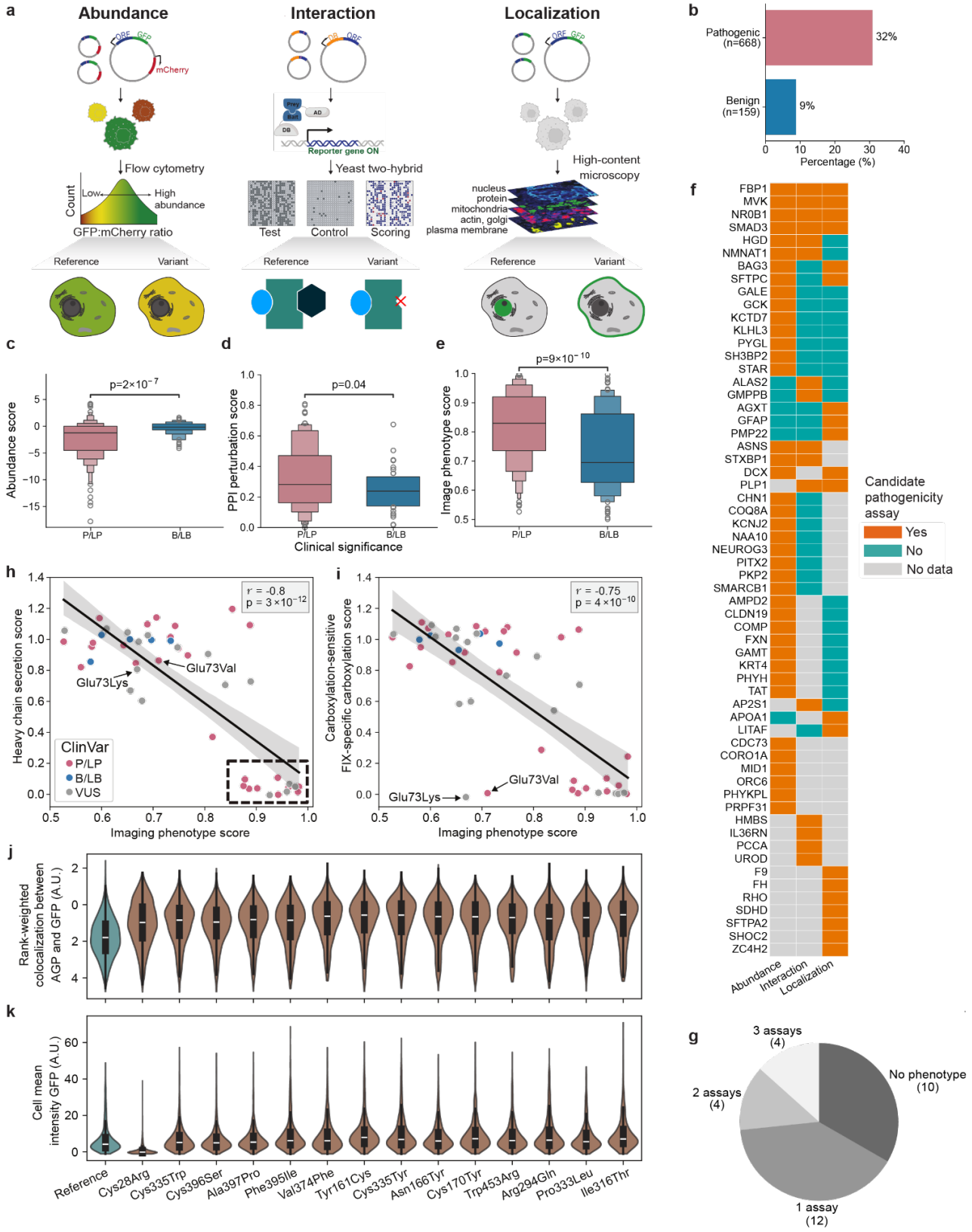

a
Abundance
Interaction
Localization
ORF
GFP
mCherry
Flow cytometry
Count
Low
High abundance
GFP:mCherry ratio
Reporter gene ON
Yeast two-hybrid
Test
Control
Scoring
High-content microscopy
nucleus
protein
mitochondria
actin, golgi
plasma membrane
Reference
Variant
b
Pathogenic (n=668)
32%
Benign (n=159)
9%
Percentage (%)
c
p=2×10−7
Abundance score
P/LP
B/LB
d
p=0.04
PPI perturbation score
Clinical significance
e
p=9×10−10
Image phenotype score
f
Candidate pathogenicity assay
Yes
No
No data
Abundance
Interaction
Localization
g
3 assays (4)
No phenotype (10)
2 assays (4)
1 assay (12)
h
r = -0.8
p = 3×10−12
Heavy chain secretion score
Glu73Val
Glu73Lys
ClinVar
P/LP
B/LB
VUS
Imaging phenotype score
i
r = -0.75
p = 4×10−10
Carboxylation-sensitive FIX-specific carboxylation score
j
Rank-weighted colocalization between AGP and GFP (A.U.)
k
Cell mean intensity GFP (A.U.)
Reference
Cys28Arg
Cys335Trp
Cys396Ser
Ala397Pro
Phe395Ile
Val374Phe
Tyr161Cys
Cys335Tyr
Asn166Tyr
Cys170Tyr
Trp453Arg
Arg294Gln
Pro333Leu
Ile316Thr

Figure 4.7 Profiling protein properties to guide future MAVEs and provide mechanistic insights into pathogenicity

a. Schematic of assays for variant effects on abundance (DUAL-IPA), protein-protein interactions (sqY2H) and subcellular localization (Variant Painting).
b. Percentage of variants scoring as hits in 163 genes identified by any of the three assays for PLP and BLB variants. Statistical significance was assessed using Fisher's two-sided exact test.
c. Distribution of DUAL-IPA abundance scores. Statistical significance was assessed using Mann-Whitney tests.
d. Distribution of PPI perturbation scores. Statistical significance was assessed using Mann-Whitney tests.
e. Distribution of Variant Painting imaging phenotype scores for PLP and BLB variants. Statistical significance was assessed using Mann-Whitney tests.
f. Heatmap showing whether each assay can discriminate PLP from BLB variants and thus serve as a candidate assay for future MAVE development. A gene is assigned a candidate assay if at least one pathogenic variant is identified as a hit in the assay while none of its benign or reference variants are hits. Only genes with at least one candidate assay are shown; other genes are listed in **Supplementary Data 7**.
g. Percentage of genes screened by all three assays that have 3, 2, 1, or no candidate assays.
h. Correlation between imaging phenotype scores and factor IX heavy-chain secretion scores. Boxed variants (n=14) are poorly secreted and show increased Golgi co-localization, indicating secretory pathway retention.
i. Correlation between imaging phenotype scores and factor IX-specific carboxylation score.
j. Distribution of single-cell normalized values for rank-weighted colocalization between AGP staining (actin, Golgi, plasma membrane markers) and GFP-tagged variants
k. Distribution of single-cell normalized values for mean GFP intensity per cell, across the factor IX reference allele and selected variants.

4.9 References

1. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

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

3. McEwen, A. E., Tejura, M., Fayer, S., Starita, L. M. & Fowler, D. M. Multiplexed assays of variant effect for clinical variant interpretation. *Nat. Rev. Genet.* (2025) doi:10.1038/s41576-025-00870-x.

4. Chen, E. *et al.* Rates and classification of variants of uncertain significance in hereditary disease genetic testing. *JAMA Netw. Open* **6**, e2339571 (2023).

5. Gold, J. I. *et al.* Racial and socioeconomic disparities in genetic evaluation and testing in the adult patient population. *Am. J. Hum. Genet.* **113**, 29–40 (2026).

6. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

7. Popejoy, A. B. *et al.* The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).

8. Bianco, B. C. F. & Planello, A. C. Variant classification of hereditary cancer genes is affected by genomic underrepresentation of admixed populations. *Mol. Genet. Genomics* **300**, 86 (2025).

9. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).

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

11. 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).

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

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

14. Jia, X. *et al.* Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108**, 163–175 (2021).

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

16. Boettcher, S. *et al.* A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science* **365**, 599–604 (2019).

17. Dawood, M. *et al.* Using multiplexed functional data to reduce variant classification inequities in underrepresented populations. *Genome Med.* **16**, 143 (2024).

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

19. Orenbuch, R. *et al.* Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. *Res. Sq.* (2024) doi:10.21203/rs.3.rs-3740259/v1.

20. 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).

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

22. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).

23. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023).

24. Rubin, A. F. *et al.* MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. *Genomics* (2021).

25. 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).

26. Park, M. S. *et al.* Insights on improving accessibility and usability of functional data to unlock their potential for variant interpretation. *Am. J. Hum. Genet.* **112**, 1468–1478 (2025).

27. Starita, L. M. *et al.* Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

28. Fowler, D. M. & Rehm, H. L. Will variants of uncertain significance still exist in 2030? *Am. J. Hum. Genet.* **111**, 5–10 (2024).

29. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).

30. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).

31. Chen, Y. *et al.* Multi-objective prioritization of genes for high-throughput functional assays towards improved clinical variant classification. *Pac. Symp. Biocomput.* **28**, 323–334 (2023).

32. Popp, N. A. *et al.* Multiplex, multimodal mapping of variant effects in secreted proteins. *Genomics* (2024).

33. Zeiberg, D. *et al.* Gene-based calibration of high-throughput functional assays for clinical variant classification. *bioRxivorg* (2025) doi:10.1101/2025.04.29.651326.

34. Biar, C. G. *et al.* An integrated, scaled approach to resolve TSC2 variants of uncertain significance. *bioRxiv* 2026.01.16.699909 (2026) doi:10.64898/2026.01.16.699909.

35. Geck, R. C. *et al.* Evidence for G6PD variant classification from multiplexed functional assays. *bioRxivorg* (2025) doi:10.1101/2025.08.11.669723.

36. Woo, I. *et al.* Saturation genome editing of BARD1 resolves VUS and provides insight into BRCA1-BARD1 tumor suppression. *medRxiv* (2025) doi:10.1101/2025.11.03.25339440.

37. Lacoste, J. *et al.* Pervasive mislocalization of pathogenic coding variants underlying human disorders. *Cell* **187**, 6725–6741.e13 (2024).

38. McEwen, A. E. *et al.* MaveMD: A functional data resource for genomic medicine. *Genetic and Genomic Medicine* (2025).

39. Claussnitzer, M. *et al.* Minimum information and guidelines for reporting a multiplexed assay of variant effect. *Genome Biol.* **25**, 100 (2024).

40. Zeiberg, D., Jain, S. & Radivojac, P. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. *Proc. Conf. AAAI Artif. Intell.* **34**, 6729–6736 (2020).

41. Li, D. *et al.* The IGVF catalog-from genetic variation to function. *Nucleic Acids Res.* **54**, D1437–D1445 (2026).

42. Chen, Y. et al. in preparation. Gene- and domain-aware calibration increases the clinical utility of variant effect predictors.

43. 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).

44. Bergquist, T. *et al.* Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. *Genet. Med.* **27**, 101402 (2025).

45. Rubin, A. F. *et al.* MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biol.* **26**, 13 (2025).

46. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).

47. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein destabilization as a common factor in diverse inherited disorders. *J. Mol. Evol.* **82**, 11–16 (2016).

48. Wang, Z. & Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270 (2001).

49. Taipale, M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms 1. *Biochem. Cell Biol.* **97**, 46–57 (2019).

50. Breast Cancer Association Consortium *et al.* Breast cancer risk genes - association analysis in more than 113,000 women. *N. Engl. J. Med.* **384**, 428–439 (2021).

51. Hu, C. *et al.* A population-based study of genes previously implicated in breast cancer. *N. Engl. J. Med.* **384**, 440–451 (2021).

52. Fu, J. M. *et al.* Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).

53. Dawood, M. *et al.* GREGoR: accelerating genomics for rare diseases. *Nature* **647**, 331–342 (2025).

54. Rubinstein, W. S. *et al.* The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.* **41**, D925–35 (2013).

55. Lee, K. *et al.* ACMG SF v3.3 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.* **27**, 101454 (2025).

56. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

57. 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).

58. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).

59. Tavtigian, S. V., Harrison, S. M., Boucher, K. M. & Biesecker, L. G. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum. Mutat.* **41**, 1734–1737 (2020).

60. 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).

61. 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).

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

63. Rivera-Muñoz, E. A. *et al.* ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum. Mutat.* **39**, 1614–1622 (2018).

64. Wu, D. *et al.* 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 (2020).

65. Costa, M., García S, A., León, A. & Pastor, O. The promises and pitfalls of automated variant interpretation: a comprehensive review. *Brief. Bioinform.* **26**, (2025).

66. Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).

67. Mersch, J. *et al.* Prevalence of variant reclassification following hereditary cancer genetic testing. *JAMA* **320**, 1266–1274 (2018).

68. Hahn, E. *et al.* Variant classification changes over time in the clinical molecular diagnostic laboratory setting. *J. Med. Genet.* **61**, 788–793 (2024).

69. Tsai, G. J. *et al.* Outcomes of 92 patient-driven family studies for reclassification of variants of uncertain significance. *Genet. Med.* **21**, 1435–1442 (2019).

70. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-specific variants and the limits of human genetics. *Trends Mol. Med.* **22**, 925–934 (2016).

71. Rapaport, F. *et al.* Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2001248118 (2021).

72. Weghorn, D. *et al.* Applicability of the mutation-selection balance model to population genetics of heterozygous protein-truncating variants in humans. *Mol. Biol. Evol.* **36**, 1701–1710 (2019).

73. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).

74. Gudmundsson, S. *et al.* Exploring penetrance of clinically relevant variants in over 800,000 humans from the Genome Aggregation Database. *Nat. Commun.* **16**, 9623 (2025).

75. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

76. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).

77. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).

78. Fragoza, R. *et al.* Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat. Commun.* **10**, 4141 (2019).

79. Santoro, C. *et al.* Inhibitors in hemophilia B. *Semin. Thromb. Hemost.* **44**, 578–589 (2018).

80. Allen, S. *et al.* Workshop report: the clinical application of data from multiplex assays of variant effect (MAVEs), 12 July 2023. *Eur. J. Hum. Genet.* **32**, 593–600 (2024).

81. Villani, R. M. *et al.* Consultation informs strategies for improving the use of functional evidence in variant classification. *Am. J. Hum. Genet.* **112**, 1489–1495 (2025).

82. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

83. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).

84. Zhou, H. *et al.* FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res.* **51**, D1300–D1311 (2023).

85. Parsons, M. T. *et al.* Evidence-based recommendations for gene-specific ACMG/AMP variant classification from the ClinGen ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel. *Am. J. Hum. Genet.* **111**, 2044–2058 (2024).

86. Fortuno, C. *et al.* A quantitative, Bayesian-informed approach to gene-specific variant classification: Updated Expert Panel recommendations improve classification of TP53 germline variants for Li-Fraumeni syndrome. *Genome Med.* **17**, 128 (2025).

87. Fowler, D. *et al.* Atlas of Variant Effects 2030 Roadmap: resolving human variants of uncertain significance. Preprint at https://doi.org/10.5281/ZENODO.15420413 (2025).

88. Arbesfeld, J. A. *et al.* Mapping MAVE data for use in human genomics applications. *Genome Biol.* **26**, 179 (2025).

89. Sinnott-Armstrong, N. *et al.* Understanding genetic variants in context. *Elife* **13**, (2024).

90. Simon, J. J., Fowler, D. M. & Maly, D. J. Multiplexed profiling of intracellular protein abundance, activity, interactions and druggability with LABEL-seq. *Nat. Methods* **21**, 2094–2106 (2024).

91. Pendyala, S. *et al.* Image-based, pooled phenotyping reveals multidimensional, disease-specific variant effects. *bioRxivorg* (2025) doi:10.1101/2025.07.03.663081.

92. Fayer, S. *et al.* Editing stem cell genomes at scale to measure variant effects in diverse cell and genetic contexts. *medRxiv* (2025) doi:10.1101/2025.11.12.25340127.

93. 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).

94. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

95. Choi, S. G. *et al.* Maximizing binary interactome mapping with a minimal number of assays. *Nat. Commun.* **10**, 3907 (2019).

96. Cimini, B. A. *et al.* Optimizing the Cell Painting assay for image-based profiling. *Nat. Protoc.* **18**, 1981–2013 (2023).

97. Stirling, D. R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).

98. Weisbart, E. *et al.* Cell Painting Gallery: an open resource for image-based profiling. *Nat. Methods* **21**, 1775–1777 (2024).

99. Serrano, E. *et al.* Reproducible image-based profiling with Pycytominer. *Nat. Methods* **22**, 677–680 (2025).

100. Hart, R. K. *et al.* HGVS Nomenclature 2024: improvements to community engagement, usability, and computability. *Genome Med.* **16**, 149 (2024).

101. den Dunnen, J. T. *et al.* HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* **37**, 564–569 (2016).

102. Freeman, P. J. *et al.* Standardizing variant naming in literature with VariantValidator to increase diagnostic rates. *Nat. Genet.* **56**, 2284–2286 (2024).

103. Freeman, P. J., Hart, R. K., Gretton, L. J., Brookes, A. J. & Dalgleish, R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. *Hum. Mutat.* **39**, 61–68 (2018).

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

105. Gebbia, M. *et al.* A missense variant effect map for the human tumor-suppressor protein CHK2. *Am. J. Hum. Genet.* **111**, 2675–2692 (2024).

106. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

107. All of Us Genomic Quality Report. *User Support* https://support.researchallofus.org/hc/en-us/articles/29390274413716-All-of-Us-Genomic-Quality-Report.

108. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation Mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).

109. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).

Acknowledgements

This work was primarily supported by the following National Institute of Health (NIH) National Human Genome Research Institute (NHGRI) Impact of Genomic Variation on Function (IGVF) consortium center awards: the Center for Actionable Variant Analysis (CAVA) to D.M.F. and L.M.S. (UM1HG011969), the Radivojac Center to P.R. (U01HG012022), the Craven Center to M.C. (U01HG012039), the Variant Characterization Across the Mendelian Proteome (VarChAMP) Center to M.V. (UM1HG011989) and the Data Analysis and Coordinating Center (DACC) to B.C.H. (U24HG012012). Additional financial support was provided by an NHGRI Advancing Genome Medicine Research consortium award (R01HG013025, A.E.M., A.F.R., D.M.F., A.B.S. and L.M.S.), an NHGRI GREGoR consortium award (U01HG011744, support to L.M.S., S.C. and S.B.), NHGRI grants (R01HG013350, V.P.; P50HG004233, M.V.), NIH National Institute of Mental Health (NIMH) grants (R56MH128365, L.M.I.; R21MH128827,

L.M.I.), an NIH National Heart, Lungs and Blood Institute (NHLBI) grant (R01HL152066, J.M.J. and D.M.F.), the NIH Intramural Research Program, a Clinical and Translational Science Awards (CTSA) grant from the National Center for Advancing Translational Sciences (UL1TR004419), awards from the NIH Office of Research Infrastructure (S10OD026880; S10OD030463), the generous support of the Brotman and Baty families through the Brotman Baty Institute for Precision Medicine, a Chan Zuckerberg Initiative Foundation grant (CZIF2024-010284, D.M.F. and L.M.S.), a Simons Foundation Autism Research Initiative (SFARI) grant (896503, L.M.I. and M.V.), the Australian Government (A.F.R.), a RUNX1 Foundation/Alex's Lemonade Stand for Childhood Cancer Early Career award (21-25037, A.E.M.), a Brotman Baty Institute Catalytic Collaborations grant (CC28, A.E.M.), a Medical Genetics Postdoctoral Fellowship (T32GM007454, S.F.), Belgian American Educational Foundation (BAEF) Doctoral Research Fellowships (F.L., M.T.), Wallonia-Brussels International (WBI)-World Excellence Fellowships (F.L., G.C.), a Fonds de la Recherche Scientifique (FRS-FNRS)-Télévie grant (FC31747–7459421F, F.L.), a Herman-van Beneden Prize (F.L.), a Josée and Jean Schmets Prize (F.L.), the Léon Fredericq Foundation (F.L.), a FRS-FNRS-Fund for Research Training in Industry and Agriculture (FRIA) grant (FC31543–1E00419F, G.C.), a University of Toronto Open Fellowship (C.R.), a Cecil Yip Doctoral Research Award (C.R.) and Dana-Farber Cancer Institute Center for Cancer Systems Biology (CCSB) Deborah F. Allinger Fellowships (A.Y., L.L.). M.V. is a Chercheur Qualifié Honoraire from FRS-FNRS (Wallonia-Brussels Federation, Belgium). We gratefully acknowledge All of Us participants for their contributions and thank the NIH All of Us Research Program for making available the participant data. We acknowledge support from the Minerva computational and data resources and staff expertise provided by the Scientific Computing and Data group at the

Icahn School of Medicine at Mount Sinai, and from the Explorer Cluster, supported by Northeastern University's Research Computing team. Support for title page creation and format was provided by AuthorArranger, a tool developed at the NIH National Cancer Institute (NCI). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this document are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. We thank past and current members of our groups for helpful discussions and guidance.

Conflicts of interest

Robert D. Steiner serves as a consultant for Amgen, Astellas and Mirum, and has received research support from Regeneron. Frederick P. Roth holds equity in Ranomics, and is a shareholder and scientific advisor for SeqWell and Constantiam Biosciences. Shantanu Singh and Anne E. Carpenter serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C: Recursion, SyzOnc, Quiver Bioscience, S.S.: Waypoint Bio, Dewpoint Therapeutics, Deepcell) and receive honoraria for occasional scientific visits to pharmaceutical and biotechnology companies. Douglas M. Fowler is a member of the Alloz Bio scientific advisory board.

Author Contributions

This project was conceptualized by M.Te., Y.C., A.E.M., R.St., Y.S., F.L., D.Z., S.F., J.S., M.V., M.Ta., A.E.C., M.A.C., M.C., V.P., A.F.R., P.R., D.M.F. and L.M.S. The original manuscript was drafted by M.Te., Y.C., A.E.M., R.St., Y.S., F.L., I.W., R.Sh., S.F., P.R., D.M.F. and L.M.S. The manuscript was reviewed by all authors and edited by M.Te., Y.C., A.E.M., R.St., Y.S., F.L., I.W., R.Sh., S.F., P.G., M.B., S.P., J.M., L.M.I., S.S., B.A.C., M.Ta., A.E.C., M.A.C., V.P., A.F.R., P.R., D.M.F. and L.M.S. MAVE data was generated by S.C., Z.R.W. and M.W.S., with

contributions from I.W., S.B., L.R.S., O.S., A.J.V., C.W., M.K.W., S.P., D.H., A.X., A.H., under the supervision of D.M.F. and L.M.S. Experimental data from the community was curated and integrated by M.Te., P.G., and A.E.M., with feedback from D.M.F. and L.M.S. Experimental data calibration was conducted by R.St. and D.Z., with contributions from S.J., and feedback from V.P. and P.R. Predictive data calibration was led by Y.C. and S.F., with contributions from M.B. and S.J., and feedback from V.P., P.R., D.M.F. and L.M.S. Biobank analyses were performed by Y.S., with feedback from R.D.S. and M.C. Clinical reclassification of VUS and unobserved variants was conducted by M.Te., A.E.M., Y.C., R.St., P.G., S.F., M.B., under the supervision of V.P., P.R., D.M.F. and L.M.S. Protein abundance data was generated by Z.Y. and F.L, with feedback from M.A.C. Interactomic data was generated by F.L., G.C. and A.Y., with help from K.S.F., and supervision from M.A.C. Imaging data was generated by C.R. and T.T., with help from F.L., and feedback from M.Ta. Sequence confirmation of DNA clones used in arrayed phenotypic assays was performed by A.G.C., M.G., W.v.L. and K.M.F., under the supervision of F.P.R. Analysis of the protein abundance dataset was led by R.Sh., with assistance from M.Ti., J.D.E., F.L. and G.C., and feedback from M.A.C. and A.E.C. Analysis of the interactomic dataset was led by M.Ti., with assistance from G.C., J.D.E. and F.L., and feedback from A.E.C., L.L. and M.A.C. Image processing and analysis pipelines were developed by Z.S.C., M.H., A.K.H. and J.D.E., and the imaging dataset was analyzed by R.Sh., E.A.M. and E.W., with feedback from S.S., M.Ta., B.A.C. and A.E.C. The specific analysis on factor IX was conducted by R.Sh., with feedback from A.E.C., J.M.J. and D.M.F. MaveDB development was led by B.J.C. and D.R. MaveMD was developed by A.E.M., J.S., E.Y.D. and S.G. PredictMD was built by J.S. The CVFG website was designed and launched by J.S.; development of these four platforms received feedback from D.M.F. and L.M.S. The Data Analysis and Coordinating

Center was managed by B.C.H., I.G., K.L. and S.D. Data submission to DACC was handled by N.S. and T.H. Figures were assembled by M.Te., Y.C., R.St., Y.S., F.L., I.W., R.Sh. and C.R., with feedback from M.A.C., A.F.R., P.R., D.M.F. and L.M.S. The multi-institutional IGVF Coding Variant Focus Group was supervised by L.M.I., S.S., B.A.C., F.P.R., R.G.J., M.V., M.Ta., A.E.C., M.A.C., M.C., V.P., A.F.R., P.R., D.M.F. and L.M.S., and overall project administration was handled by J.S., S.H. and L.M. Funding was acquired by J.M.J., A.B.S., L.M.I., F.P.R., M.V., M.Ta., A.E.C., M.A.C., M.C., A.F.R., P.R., D.M.F. and L.M.S.

Chapter 5: Conclusions and Future Directions

The rise of next-generation sequencing has increased our ability to detect genetic variation; however, our ability to interpret the clinical significance of those variants has not kept pace. This gap poses one of the most fundamental challenges in genomic medicine. In this thesis, I aimed to address this challenge by developing a scalable framework that integrates and applies two of the most abundant forms of evidence for variant classification: multiplexed assays of variant effect (MAVEs) and computational variant effect predictors (VEPs), leading to the resolution of approximately 75% of VUS in 40 clinically relevant genes.

A major contribution of this thesis is the creation of a harmonized resource that maps computational predictions, clinical assertions, population data, and assay metadata across 40 clinically relevant genes in 74 curated and 9 newly generated functional datasets. Because functional datasets have historically been published with inconsistent practices, they have been difficult to use systematically. The dataset developed here required extensive curation, where our curation team and I extracted approximately 200 layers of metadata for functional assays, mapped variants to relevant transcripts and genomic coordinates, and integrated multiple layers of annotation including ClinVar[1], gnomAD[2], SpliceAI[3], and three major VEPs (REVEL, MutPred2, and AlphaMissense)[4–6]. This effort to create a coherent, standardized resource laid the groundwork for subsequent analyses and reflects a core strength of this work.

For functional datasets, the OddsPath statistic represented an early quantitative solution but was limited by its dependency on large pathogenic and benign control sets, which many genes lack, and assay-wide evidence strength given to all variants regardless of how deleterious they score[7]. The skew-normal mixture model developed within the IGVF Coding Variant Focus Group addressed several of these limitations by generating variant-specific evidence strengths

based on modeled distributions of pathogenic, benign, synonymous, and population variants[8]. Previous calibrations for computational predictors included genome-wide thresholds defined by the ClinGen SVI[9], where I showed that genome-wide thresholds can both under-estimate and over-estimate predictive strength as they do not capture the heterogeneity in score distributions between genes[10]. However, these thresholds represented an important conceptual advance, enabling VEPs to contribute stronger evidence for variant classification. As was done with functional data, new gene-specific and domain-clustered calibration approaches for predictors were developed within the consortium to address the limitations mentioned above[11] .

Using the harmonized dataset I constructed, I evaluated the impact of applying calibrated functional and predictive data on VUS. My analysis showed that systematic application of functional and predictive data resolves approximately 75% of existing VUS across 40 clinically relevant genes. Equally important, the framework provides provisional classifications for approximately 62% of previously unobserved variants, offering a starting point for future clinical sequencing results to be interpreted. These findings illustrate the power of combining two scalable datatypes in variant classification.

Despite these advances, several limitations remain. Even well-designed MAVEs may show ambiguous results in poorly understood pathways, and the availability of pathogenic and benign controls remains uneven across genes. Computational predictors, although improving, are still constrained by circularity risks[11], and the limitations of reference databases such as ClinVar and gnomAD[12–14]. In addition, the curation and harmonization process to transform functional datasets into a uniform resource is time-intensive and laborious.

Looking forward, several directions emerge. Expanding functional assays to additional clinically relevant genes, particularly those with high VUS burdens, will be essential.

Additionally, integrating features from next-generation protein language models[15] can improve predictive accuracy, especially for genes lacking extensive evolutionary data. Improved calibration frameworks that incorporate uncertainty measurements (e.g. standard error) can refine evidence strength assignments. As functional and predictive datasets continue to grow, there will be a need for automated systems that update calibrations and interpretations as ClinVar annotations change. Finally, advances in AI can offer a promising route to reduce the manual effort required to identify, annotate, and standardize functional datasets, enabling faster integration of new evidence into clinical frameworks.

Beyond research, a core component of this thesis is its translational impact. The harmonized datasets and calibration methods will be disseminated through MaveMD[16], a clinician-facing platform. By making these resources accessible to clinical geneticists and variant scientists, this work will contribute to improving the scalability of variant interpretation in the real-world. In conclusion, this thesis demonstrates the combined power of functional genomics and computational modeling. When calibrated correctly, both functional datasets and variant effect predictions can meaningfully reduce the burden of VUS. As sequencing individuals in a clinical setting continues to grow, approaches like those developed here will be essential for ensuring that genomic data can be accurately interpreted and at scale. This work contributes both conceptual frameworks and practical tools toward that goal, helping to move genomic medicine closer to fulfilling its promise.

5.1 References

1. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
2. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
3. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
4. 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).
5. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).
6. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
7. 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).
8. Zeiberg, D. *et al.* Gene-based calibration of high-throughput functional assays for clinical variant classification. *bioRxivorg* (2025) doi:10.1101/2025.04.29.651326.
9. 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).
10. 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).

11. . Chen, Y. et al. in preparation. Gene- and domain-aware calibration increases the clinical utility of variant effect predictors. bioRxiv (2026) doi:/10.64898/2026.02.17.706269.
12. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
13. Han, A. L. *et al.* Diverse ancestral representation improves genetic intolerance metrics. *Nat. Commun.* **16**, 2648 (2025).
14. Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* **17**, (2016).
15. Sharo, A. G., Zou, Y., Adhikari, A. N. & Brenner, S. E. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med.* **15**, 51 (2023).
16. 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).
17. McEwen, A. E. *et al.* MaveMD: A functional data resource for genomic medicine. *Genetic and Genomic Medicine* (2025).