Received: 26 October 2016 | Revised: 25 January 2017 | Accepted: 4 February 2017

DOI: 10.1002/humu.23193

RESEARCH ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Predicting phenotype from genotype: Improving accuracy through more robust experimental and computational modeling

Jonathan Gallion[1]* | Amanda Koire[1]* | Panagiotis Katsonis[2] | Anne-Marie Schoenegge[3] | Michel Bouvier[3] | Olivier Lichtarge[1,2]

[1]Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas

[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

[3]Department of Biochemistry, Institute for Research in Immunology and Cancer, Université de Montreal, Quebec, Canada

Correspondence
Olivier Lichtarge, 1 Baylor Plaza, Houston, TX 77030.
Email: lichtarge@bcm.edu

*Jonathan Gallion and Amanda Koire contributed equally to this study.

Contract grant sponsors: National Institutes of Health (NIH) (GM066099 and GM079656); National Science Foundation (NSF) (DBI-1356569); DARPA (N66001-15-C-4042).

Communicated by William S. Oetting

Abstract

Computational prediction yields efficient and scalable initial assessments of how variants of unknown significance may affect human health. However, when discrepancies between these predictions and direct experimental measurements of functional impact arise, inaccurate computational predictions are frequently assumed as the source. Here, we present a methodological analysis indicating that shortcomings in both computational and biological data can contribute to these disagreements. We demonstrate that incomplete assaying of multifunctional proteins can affect the strength of correlations between prediction and experiments; a variant's full impact on function is better quantified by considering multiple assays that probe an ensemble of protein functions. Additionally, many variants predictions are sensitive to protein alignment construction and can be customized to maximize relevance of predictions to a specific experimental question. We conclude that inconsistencies between computation and experiment can often be attributed to the fact that they do not test identical hypotheses. Aligning the design of the computational input with the design of the experimental output will require cooperation between computational and biological scientists, but will also lead to improved estimations of computational prediction accuracy and a better understanding of the genotype–phenotype relationship.

KEYWORDS

functional effect of mutations, genotype–phenotype relationship, in silico prediction, SNV, variant impact prediction, VUS

1 | INTRODUCTION

In order to achieve a vision of personalized medicine in which whole-exome sequencing data is used to both explain and to predict phenotypes in a comprehensive manner, it is necessary to know with a reasonable level of accuracy the effects of each variant at the protein and organismal levels. Given that the latest publication from the 1000 Genomes Project reported 84.7 million unique single-nucleotide variations (SNVs) in a cohort of just over 2,500 individuals (1000 Genomes Project Consortium et al., 2015) and dbSNP144 claims over half a billion submissions (Sherry et al., 2001), individually assaying every mutation in the population for its effect on protein function would prove a herculean task. Instead, computational inference has taken a leading role as a more practical alternative. Since the debut of the classic SIFT algorithm for mutation impact prediction 15 years ago (Ng & Henikoff, 2001), dozens of tools have been developed that use protein structure (Dehouck et al., 2009; Schymkowitz et al., 2005; Zhou & Zhou, 2002), machine learning (Adzhubei et al., 2010; Yates, Filippis, Kelley, & Sternberg, 2014), or evolutionary considerations (Katsonis & Lichtarge, 2014; Stone & Sidow, 2005; Tavtigian et al., 2006) to predict the impact of SNVs and serve as a reasonable stand-in for time-consuming experimental assays. Already, these tools appear in clinical reports related to Mendelian disorders as well as some complex diseases like hearing loss and intellectual disability (Rabbani, Tekin, & Mahdieh, 2014) and gather thousands of citations in studies that aim to bridge the gap between genotype and phenotype.

Still, the ability of computational tools to produce accurate and meaningful predictions at the protein level is not yet sufficient (Sun

et al., 2016). For example, one recent study (Miosge et al., 2015) expands upon previous work (Hicks, Wheeler, Plon, & Kimmel, 2011) to point out that some of the most popular predictors of mutational impact have weak specificity, not just in vivo where incomplete penetrance and epigenetic modification may explain disagreement (Lehner, 2013) but also in vitro. Their conclusions suggest that diagnosis via clinical genome sequencing necessitates direct experimental measurement of putative impactful mutations. An alternative explanation for their results, touched upon briefly in Itan and Casanova (2015), could be that genotype/phenotype prediction discrepancies are not always due to inaccurate computational prediction, but rather inadequate experimental testing to detect the true functional impact of the mutations.

Currently, the most useful approach to assess the performance of these computational predictions objectively is through independent evaluations. One such benchmark is the Critical Assessment of Genome Interpretation (CAGI), which uses new and unpublished experimental data as a gold standard in order to quantify the accuracy of blinded SNV impact predictions on a variety of statistical measures. From these types of comparisons between computational prediction and experimental validation, it is clear that not all computational approaches are equally accurate when applied to a new problem, which may lead to the impression among users that all computational tools are unreliable. However, even top-performing methods are often unable to resolve the genotype to phenotype relationship completely, that is, some of the experimental data refute some of the predictions. In these cases, it becomes unclear whether this residual mismatch results from algorithmic design limitations, experimental design limitations, or both. Understanding the driving forces behind these discrepancies is critical to not only ensuring a robust and accurate venue to benchmark tools but also to predicting and understanding human disease.

Using Evolutionary Action (EA) (Katsonis & Lichtarge, 2014), the consistently top-performing variant impact prediction method for multiple CAGI challenges (2011, 2013, 2015), we explore several sources of potential discordance between predicted and experimentally assessed variant impact: data quality, protein multifunctionality, and input alignment design. We probed the genotype–phenotype relationship for several proteins and found evidence that testing and integrating multiple, diverse parameters of functional impact can improve correlation with in silico predictions. We conclude that the success of predicting phenotype from genotype depends heavily on the computational model and the experimental model testing the same initial biological question in a comprehensive fashion. Moreover, a lack of concordance between predicted and assayed phenotype should lead to the re-examination of both in order to search for possible computational and experimental limitations.

2 | MATERIALS AND METHODS

2.1 | Impact prediction using EA

EA calculations are described at length in the original publication of the method (Katsonis & Lichtarge, 2014). In summary, the action $\Delta\phi$ of an SNV is calculated as the product of the evolutionary gradient $\partial f/\partial ri$ and the perturbation magnitude of the substitution, $\Delta ri,X\rightarrow Y$. The evolutionary gradient is measured by importance ranks of the evolutionary trace (ET) method and the perturbation magnitude by amino acid substitution odds. Computing the evolutionary gradient of a position in a protein using ET involves retrieving homologous sequences, aligning sequences with MUSCLE (Edgar, 2004), accounting for resulting gaps in the query sequence, and maximizing spatial clustering among top-ranked residues and their rank information. The perturbation magnitude of the substitution is derived using the BLOSUM amino acid substitution log odds methodology computed over bins that account for the evolutionary gradient, solvent accessibility, and secondary structure of the substituted position. We normalized both terms and their product to become percentile scores for each protein, with higher action scores indicating a larger predicted impact. For the evaluation of SUMO ligase SNVs in particular, potential gain-of-function variants were identified as substitutions resulting in an amino acid that was more common in the alignment than the query amino acid at that site; these variants were scored as -1*EA.

2.2 | Characterization of *ADRB2* mutants

Twenty-six mutations were designed within *ADRB2* (MIM# 109690) to interrogate the functionality of eight structural positions within the GPCR transmembrane region. The functional effect of each mutation was measured using five assays. Three of these assays measured interaction between *ADRB2* and three downstream binding partners: G_αi, G_αs, and β-arrestin, whereas the two remaining assays measured more global downstream cellular phenotypes: receptor endocytosis and cAMP concentration. The receptor was stimulated with varying concentrations of isoproterenol to generate a dose response curve for each assay. These dose response curves were reduced to five representative quantitative values: EC50, maximal response, minimal response, ligand-induced response (max–min), and Log (T/Ka). Combined with cell surface expression for each assay, this resulted in six measures for each assay and therefore 29 phenotypic measures total for each mutant (β-arrestin response was missing minimal measurements). Total deviance of a variant from wild type was calculated as $\mathbf{\Sigma}(|i_{\mathbf{wt}} - i_{\mathbf{mut}}|)$ for all 29 measures. Each measure was first standardized to be between 0 and 1 so as to contribute equally to this sum despite the scale differences within the original values for each measure.

2.3 | Testing sensitivity of *ADRB2* EA predictions to alignment input

To test the sensitivity of *ADRB2* variant predictions to alignment input, we used the differential ET approach described in Lichtarge, Yamamoto, & Cohen (1997), modified to consider the EA scores of the variants rather than the ET scores of the residues. The FASTA sequence for *ADRB2* was obtained from Uniprot and blasted against the Uniref90 database of human proteins (Suzek et al., 2015) using the method described in Wilkins, Erdin, Lua, & Lichtarge (2012). The results were limited to E-value cutoffs of 0 and 0.5 but no thresholds

for sequence identity. Additionally, blast results were required to be at least 75% the length of the *ADRB2* query sequence. The resulting homolog list was aligned using MUSCLE (Edgar, 2004) and the neighbor-joining tree was extracted with PFAAT (Caffrey et al., 2007) using percentage identity of all columns for all sequences. In reference to the initial query sequence, this tree was pruned into a smaller subdivision encompassing only close homologs to the query sequence. Each alignment was used to generate EA scores specific to that level of evolutionary divergence, for all 19 possible substitutions at every sequence position of *ADRB2*. The above methodology was repeated for additional proteins of interest.

2.4 | Quantifying and mapping EA prediction differences onto *ADRB2*

Comparing EA scores for all of Class A GPCRs to the narrow *ADRB2* alignment, we calculated the EA score differences for all 19 possible amino acid substitutions at each residue and produced a distribution representing these differences. Our narrow alignment, based on sequence similarity, consisted of genes *ADRB1* (MIM# 109630), *ADRB2*, *ADRB3* (MIM# 109691), *HTR4* (MIM# 602164), *DRD1* (MIM# 126449), and *DRD5* (MIM# 126453) from several species, in agreement with prior analyses of GPCR phylogeny (Lin, Sassano, Roth, & Shoichet, 2013). We averaged all 19 substitutions for each residue; residues with an average EA greater than or less than 1 standard deviation from the mean of the distribution were grouped as either more important to *ADRB2*, or to Class A receptors, based on which alignment produced the higher EA score. These residues were mapped onto the 2RH1 crystal structure of *ADRB2* and shown as spheres using PyMOL. Sphere size was scaled based on the equation $\text{Size} = 2 * (\frac{\Delta\text{EAsubstitution}}{\Delta\text{EAmax}})$ where ΔEAmax was the largest change in either the red or blue group. In this way, the largest change in EA was given a sphere size of 2 and each smaller change was scaled linearly. Significance of structural clustering was calculated using a chi-square test measuring the number of residues in the two groups within 4 Å compared with the expected number derived from the adjacency matrix of the 2RH1 structure.

2.5 | Human protein blast search

Following the method outlined in Wilkins et al. (2012), we analyzed all human proteins within the Uniprot_Sprot database and further used the FASTA sequences available from the same location to perform a customized Blastp (Altschul, Gish, Miller, Myers, & Lipman, 1990) search for each protein using the NCBI Blastp tool and the Uniref90 database (Suzek et al., 2015) as a search space. We limited the results to all homologs at least 75% in length compared with the query sequence, and having an E-value between 0 and 0.5. No restrictions were placed on sequence identity or on the maximal number of proteins allowed in an individual blast result.

2.6 | Quantifying computational and experimental biological collaboration

We searched research articles on Scopus for abstracts that explicitly mention one of six commonly used variant impact predictors: SIFT (Ng & Henikoff, 2001), Polyphen (Ramensky, Bork, & Sunyaev, 2002), Polyphen2 (Adzhubei et al., 2010), MutationTaster (Schwarz, Rodelsperger, Schuelke, & Seelow, 2010), Provean (Choi & Chan, 2015), and PANTHER (Thomas et al., 2003). We then compared these numbers with those produced when an additional filter was added to require that one of the tool designers be attributed as an author.

3 | RESULTS

3.1 | Evaluating computational prediction accuracy requires robust, high-quality experimental data

In order to gauge the accuracy of variant impact predictions, the experimental measurements of the functional effects of these variants should themselves be reliable, precise, and performed on a sufficiently large scale to be meaningful. These conditions are not trivial and may require repetitive measurement with a given assay, testing the same single phenotype with multiple different assays, or gathering information on multiple phenotypic outcomes of the same mutation. As a result, in most practical instances, it is challenging to evaluate whether the data that are available are acceptably robust and complete, and to assess whether the experimental design choices affect the relationship with computational prediction.

To quantify the extent to which increasing robustness of experimental testing improves concordance with phenotypic prediction, we first look at the value of repeated measures of the same phenotypic assay. In a recent study of *UBE2I* (SUMO ligase) (MIM# 601661), 682 SNVs were evaluated experimentally using a yeast-based complementation assay in which thousands of barcoded *UBE2I* clones representing nearly 2,000 combinations of amino acid changes were pooled and transformed into a yeast strain carrying a mutant, temperature-sensitive homolog of *UBE2I* (CAGI, 2015). Half of the experimental replicates were grown at a permissive temperature, where there should be no selection for or against variant function, whereas the other half were grown at a restrictive temperature where growth was dependent on the human protein function. Protein fitness in the presence of the variant was represented as the ability of the human protein to rescue SUMO ligase function, and was computed as the ratio of growth between these two conditions. In this experiment, some amino acid changes were represented by multiple, independent clones with unique barcodes, and the resulting data were split into two groups based on whether the fitness score for the variant was an average across at least three independent clones ("high-accuracy" subset) or fewer than three clones. We found that the correlation with EA prediction scores increased substantially ($R^2 = 0.5$ to $R^2 = 0.9$) when the measure of experimental function was an average across a larger number of independent clones (Fig. 1A and B). Even among the "high-accuracy" variant subset, variability in data quality affected concordance between experimental and computational estimations of impact; the power of EA predictions to prioritize the variants increased exponentially as the standard error of the mean for experimental fitness decreased (Fig. 1C). These examples highlight that even very basic improvements in data quality, in this instance

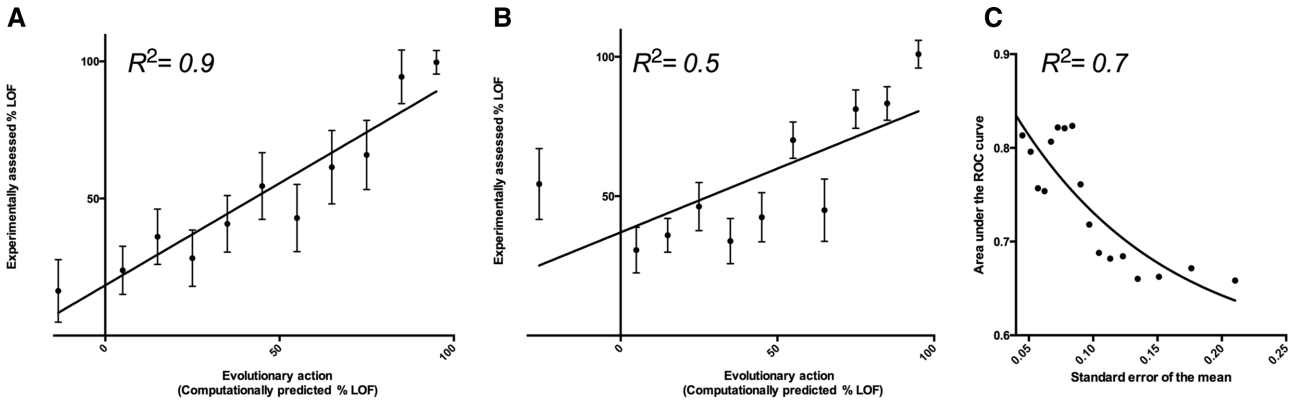

FIGURE 1 Influence of experimental data quality on correlation to computational predictions. **A**: "High-accuracy" subset of 219 SNVs in SUMO ligase, from the 2015 CAGI challenge. "High-accuracy" variants are defined as those that had at least three independent barcoded clones in the complementation screen, providing internal experimental replicates. Variants were binned in deciles according to their Evolutionary Action (EA), and the mean and standard error of the mean was plotted for each bin. Gain-of-function variants were grouped into a single bin and plotted at their average EA value. **B**: Remaining 463 SNVs in SUMO ligase. Variants were binned in deciles according to their Evolutionary Action (EA), and the mean and standard error of the mean were plotted for each bin. **C**: Relationship between experimental standard error of the mean and prioritization accuracy of EA for the "high-accuracy" subset of 219 SNVs in SUMO ligase. Variants were ordered by their standard error of the mean, and the area under the receiver operating curve (AUROC) was calculated for sliding windows with a window size of 50 variants and a slide of 10 variants, using variants with at least a 50% decrease in fitness as a positive set. The AUROC and average standard error for each window was plotted and fitted with an exponential decay function

using at least triplicates, can lead to a greater agreement between phenotype prediction and experimental assessment. However, the complexity of protein biology further necessitates moving beyond simplistic definitions of robustness to consider not just how many replicates or independent runs create a robust dataset, but additionally how many assays accurately characterize the breadth of protein functionality.

3.2 | For many proteins, a single assay may result in an incomplete characterization of overall impact

For any protein, multiple assays may be necessary to get a complete picture of how a variant has affected function. Even for a protein with a single well-defined role, there are many individual phenotypes that may be affected distinctly: expression, localization, binding, or enzymatic activity, for some examples. In addition, over one-quarter of proteins are likely to be multifunctional (Pritykin, Ghersi, & Singh, 2015), and these proteins are of particular interest because they are significantly more likely to be involved in human disease (Pritykin et al., 2015). To emphasize the importance of assaying all relevant functions that could be affected by a variant, we repurposed recent variant assessment studies for two multifunctional proteins: prokineticin receptor 2 (*PROKR2*, MIM# 607123) and dopamine D2 receptor (*DRD2*, MIM# 126450) (Peterson et al., 2015; Sbai et al., 2014). In the analysis of 21 coding SNVs of *PROKR2*, initial phenotypic testing for cell surface expression indicated normal expression for almost half of the variants. After more extensive testing of five other phenotypes, including three signaling pathways, over 95% of the variants were determined to affect function (Fig. 2A). Similarly, in the assessment of 41 coding SNVs from *DRD2*, all of which were computationally predicted by EA to equally and significantly affect function, 32% of the variants could only be detected as impactful in the G-protein signaling assay, whereas 10% of the variants could only be detected as impactful in the β-arrestin signaling assay (Fig. 2B). The few remaining variants in both studies that did not display abnormal behavior on any assay may yet affect other, untested functions. These data show that selective assaying can underestimate a variant's impact on function.

3.3 | Integrative approaches to experimental impact improve concordance with computational prediction

In order to test whether a more holistic, integrative phenotype characterization would increase agreement with computational predictions of variant impact, we next repurposed data collected for a different study by the authors, which analyzed the functional effects of 26 SNVs on *ADRB2* signaling (Schoenegge, submitted). Protein fitness perturbation was computed using six measures of performance across five functional impact assays that were chosen to reflect the primary expected functions of the receptor while additionally capturing the complexity of receptor functionality. When the phenotypic impact of each SNV was calculated to consider total deviance from wild-type function across all assays and measures, the relationship between the computationally predicted impact and experimentally measured phenotypic impact was extremely strong ($R^2 = 0.73$, $P < 0.0001$) (Fig. 3A). We also found that the ability of computational prediction to distinguish functional effects of variants was highly accurate even for variants affecting the same protein residue (Fig. 3B). These variants were correctly prioritized 87% of the time, and those that were not had significantly smaller differences both in computationally predicted impact ($P = 0.008$) and experimentally measured impact ($P = 0.03$), indicating that errors were both infrequent and of small magnitude. However, when the

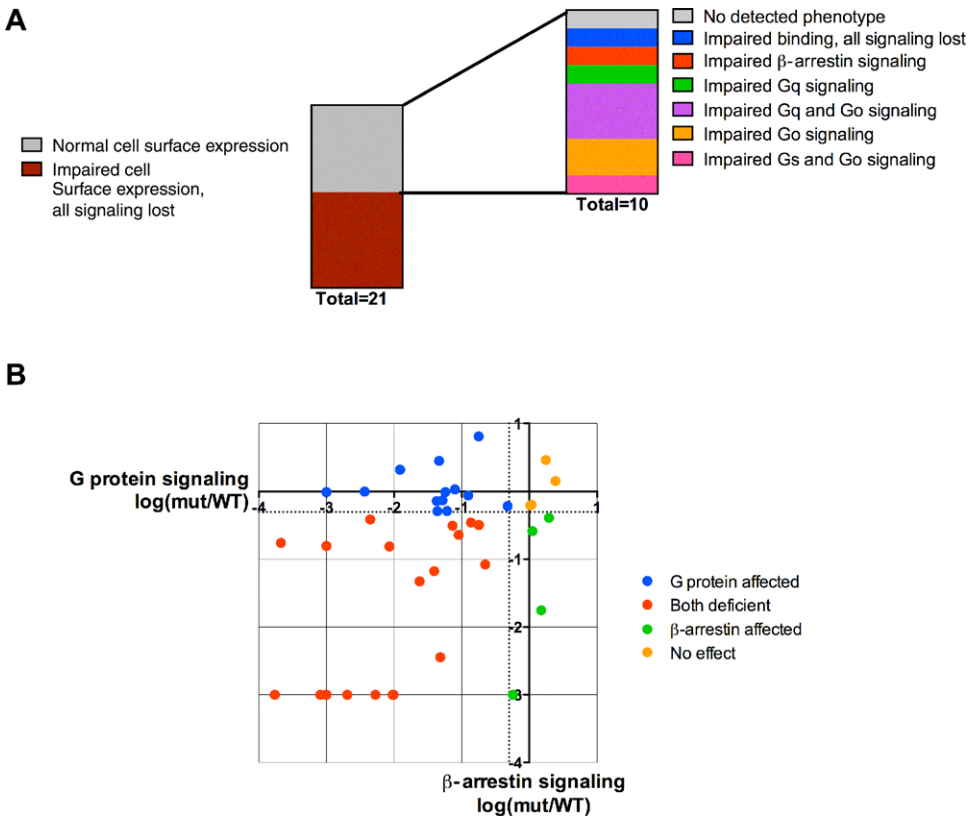

FIGURE 2 Diversity in functional effect caused by different variants in a protein. **A**: *PROKR2* mutational data adapted from Sbai et al. (2014), testing functional effects of 21 total variants. Initial assessment of cell surface expression left 10 variants with wild-type phenotype; five additional assays revealed a spectrum of affected function for the remaining variants. **B**: *DRD2* mutational data adapted from Peterson et al. (2015). All SNVs with high predicted impact (defined as top 20% of all possible DRD2 variants, EA>80) were assessed for phenotypic impact. Axes represent log(mutant function/WT function) values for G protein (x axis) and *β*-arrestin (y axis) signaling. Signaling was considered to be affected if there was at least a 50% decrease in signaling function for the mutant *DRD2* compared with wild type (dotted lines)

correlation between computationally predicted impact and measured phenotypic effect of the *ADRB2* variants was calculated considering only one phenotypic measure at a time, every single measure performed worse than the integrative approach, and many individual measures in fact performed extremely poorly indeed (Fig. 3C). The integrative approach was able to account for over 10% more variability than the best performing individual measure, and over 30% more variability than the median performing individual measure. In addition, the average performance of the integrative approach increased as a greater number of phenotypic measures was incorporated, showing a clear advantage to approaching phenotype from a holistic perspective and making experimental models more robust through multiple assays and measures. The importance of this integrative approach is highlighted further when examining the performance of any given phenotypic measure across different functional assays (Fig. 3D). No one phenotypic measure produces either the best or the worst results in all assays, and in some cases a measure that is relatively uninformative in one assay has a strong relationship with predicted impact in another. These data show that to choose only one assay, or one phenotypic measure, could easily lead to the impression of genotype/phenotype discrepancies when the predicted phenotypic effect was simply not detected by the experimental setup. Furthermore, given the differing effects of individual variants, it would be challenging to preemptively predict the single most relevant assay to utilize. A thorough characterization compensates for this uncertainty.

Next, we assessed whether the value of an integrative approach to experimental impact can be seen even when multifunctional signaling is not a major concern for a protein and when experimental reporting is not on a continuous scale. To do so, we examined the relationship between EA and experimental assessment for the 28 coding SNVs in *MLH1* (MIM# 120436) reported by Raevaara et al. (2005). The variants of *MLH1*, a DNA mismatch repair gene involved in hereditary colon cancer, were assessed using five separate assays of function (expression, localization to nucleus, repair efficiency, localization with binding partner, and interaction with binding partner). For each variant, the study authors reported each category as normal or abnormal and then provided an "overall interpretation" of impact. We found that prioritization of the variants based on their EA scores was most predictive of the "overall interpretation" (AUC = 0.91, $P = 0.0004$) (Fig. 4). While the assay measuring the primary function of *MLH1*, mismatch repair, was the best performing individual assay (AUC = 0.86), an integrative approach to defining experimental impact was still able to increase concordance with computational prediction. These data show that the principle of improving experimental robustness through multiple assays is likely to apply broadly, and may benefit the interpretation of other disease genes like *MLH1*.

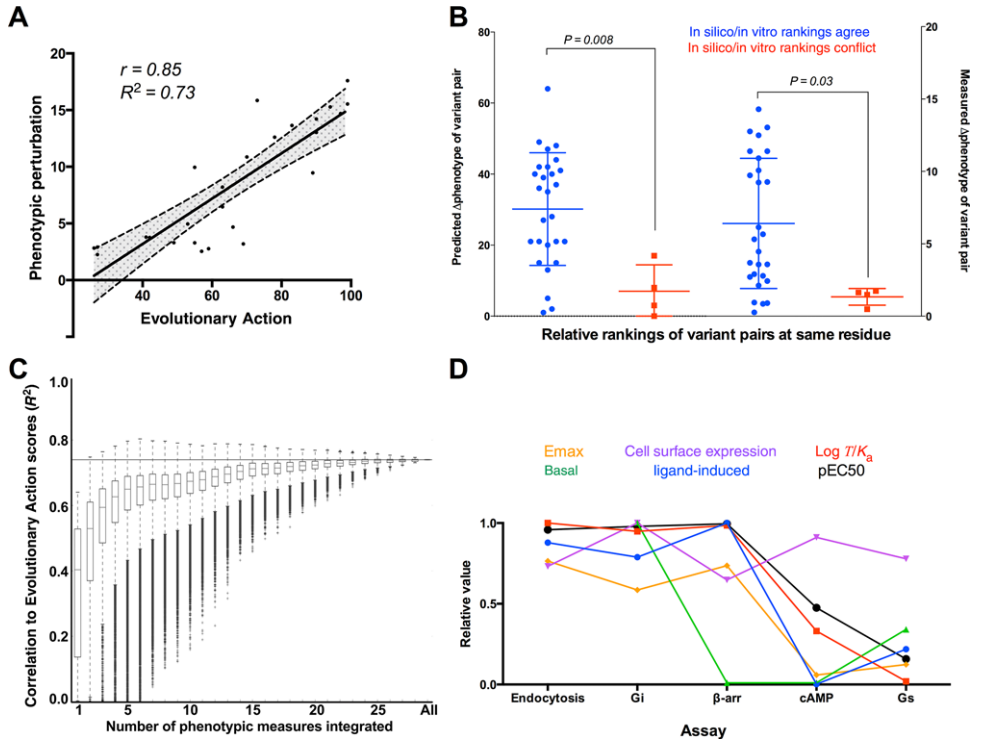

FIGURE 3 Holistic characterization of SNV perturbation of *ADRB2* function. **A**: Correlation between predicted SNV impact (Evolutionary Action [EA]) and overall experimentally assessed impact for 26 variants in *ADRB2*, quantified using 29 measures across five biological assays. **B**: Ability to correctly rank the impact of amino acid substitutions at the same residue. Left y-axis represents ΔEA between two mutations at same residue, right y-axis represents the measured biological difference between mutations. Substitution pairs at the same residue are shown in blue if in silico and in vitro approaches agree on their relative impact ranking, and in red if they do not. **C**: Correlation (R^2) between EA scores and 29 phenotypic measures for *ADRB2* variants. Each phenotypic measure individually (1), all phenotypic measures integrated as a whole (All), and different combinations of the 29 measures (2–28) are represented on the x-axis. For combinations of six measures through combinations of 23 measures, a random sample of 300,000 possibilities was assessed. The performance of all phenotypic measures integrated as a whole is also represented as a solid line at y = 0.73. **D**: Relative performance of phenotypic measures is inconsistent across assays. For each assay, performance of each phenotypic measure (EC50, maximal response, etc.) is shown as a relative value compared with best performing measure for the assay

3.4 | Computational input design affects impact prediction scores for many variants, and must reflect the same underlying hypothesis as the experimental design

Holistic experimental phenotyping improves correlation in the examples above largely because evolutionary impact prediction methods are designed to consider the "global" impact of a variant on a protein. Within these tools, variant predictions are derived from analyses of the protein's evolutionary lineage using alignments of homologous proteins. We hypothesized that these "global predictions" likely stem from a global input and that alterations to this alignment may result in variations in the output predictions, with more divergent alignments resulting in greater differences in impact predictions. If impact predictions varied depending on the specificity of the input, we further hypothesized that input selection could be used to tailor predictions to specific biological questions, and that incorrect selection of protein alignments could result in discordance between computation and biology.

To quantify the variance in EA predictions resulting from deviations in the alignment input, we first generated EA scores for *ADRB2* from a broad, class-specific alignment (Class A GPCRs), and a narrow alignment composed only of very close homologs. Using these two alignments, we observed how the evolutionary time scale of the alignment affected EA scores of *ADRB2* variants. For every possible amino acid change in *ADRB2*, we compared the predictions produced using different alignments as inputs (Fig. 5A). We then averaged all 19 substitutions for each residue; residues with an average EA greater than or less than 1 standard deviation from the mean of the distribution were grouped as either more important to *ADRB2* or to Class A receptors, respectively, whereas residues with an average EA within 1 standard deviation were considered "robust." Residues within functional motifs such as the DRY, NPXXY, and PIF domain showed minimal variation between the two alignments, in agreement with prior biological characterization indicating their conserved functionality throughout Class A GPCRs (Katritch, Cherezov, & Stevens, 2013). Mapping additional robust residues onto the *ADRB2* structure demonstrates that many of these predictions are located at inward facing core residues.

To test whether the deviations (ΔEA >±1STD) resulting from different evolutionary time scales of the alignments reflected genuine differences in biology, rather than noise or a nonrobust methodology, we separated substitutions into two groups based on whether the substitution was more impactful in the narrow (*ADRB2*) alignment or the broad (Class A) alignment and then mapped these positions onto the *ADRB2* structure (Fig. 6). Of the 110 residues with an EA

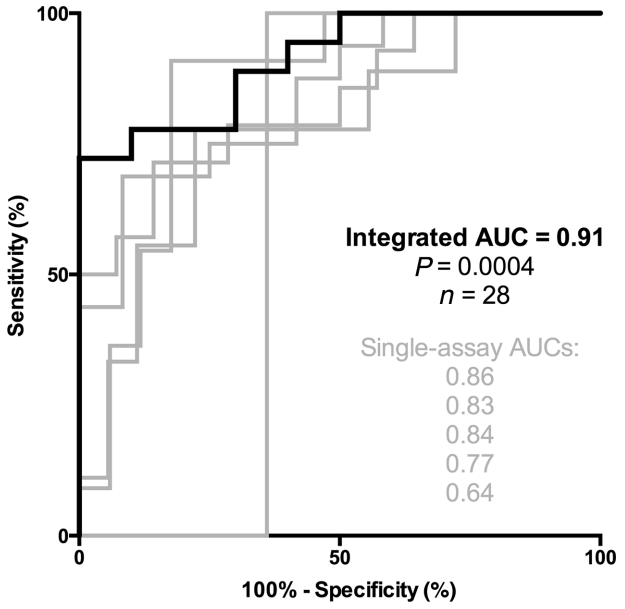

FIGURE 4 Assessment of *MLH1* variants using multiple assays. *MLH1* mutational data were adapted from Raevaara et al. (2005), in which 28 SNVs were measured for performance on five separate assays of protein function. The ability of EA to rank the functional impact of the variants within each individual assay is displayed in gray, whereas the ability of EA to prioritize "overall" variant effect is displayed in black

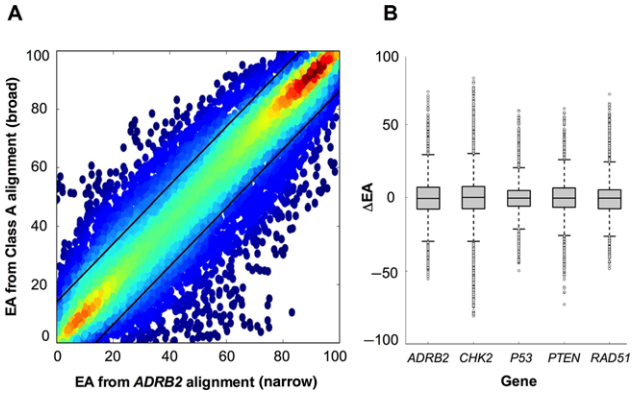

FIGURE 5 Relationship between scope of input alignment and EA predictions. **A**: Identification of alignment-sensitive variant predictions. Alignment input was varied to reflect narrow and broad definitions of *ADRB2* homology and fluctuations in the resulting EA predictions were quantified. The x-axis represents EA scores from a very narrow alignment using only close homologs to *ADRB2*, whereas the y-axis represents EA scores from a broad alignment of *ADRB2* homologs spanning all of Class A GPCRs. Color represents point density, with red representing the highest possible density. **B**: Application of narrow versus broad alignment analysis to identify input-sensitive variants in other multifunctional proteins related to disease: CHK2 (MIM# 113705), TP53 (MIM#191170), PTEN (MIM#601728), and RAD51 (MIM# 179617). ΔEA reflects the EA score produced using the broad (Class A) alignment subtracted from the EA score produced using the narrow (*ADRB2*) alignment

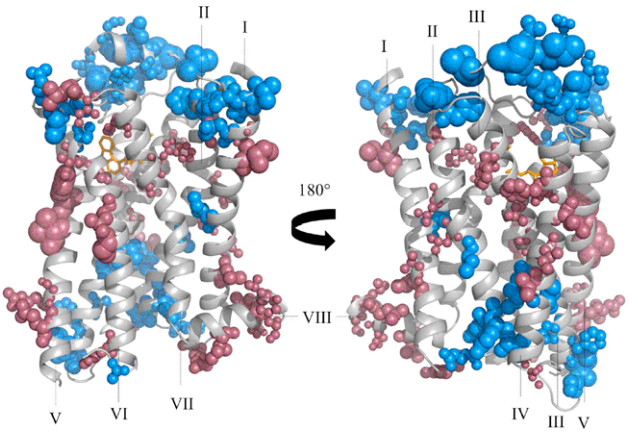

FIGURE 6 Structural locations of residues sensitive to scope of alignment input. Positions with ΔEA > 1 standard deviation when comparing broad to narrow alignment inputs are shown as spheres where size denotes the magnitude of the ΔEA difference. Red spheres = *ADRB2* EA>Class A EA and Blue spheres = Class A EA >*ADRB2* EA. Partial inverse agonist Carazolol shown as orange sticks. Roman numerals identify the GPCR helices

difference larger than 1 standard deviation, mutations at 60 residues were predicted to be more impactful to *ADRB2* alignment, and 50 residues were predicted to be more impactful in the full Class-A alignment. Of these, 12 residues and five residues had an EA difference larger than 2 standard deviations (*ADRB2* and Class A, respectively). Both groups (ΔEA >±1STD) demonstrated significant structural clustering ($P < 0.00001$), and these clusters were found at sites of particular relevance to GPCRs. Residues identified as higher scoring and of greater importance when using the narrow *ADRB2* alignment clustered near the predicted dimerization interfaces of close homologs (Lee, O'Dowd, Rajaram, Nguyen, & George, 2003; Milligan, 2004), as well as the ligand-binding pocket, with several residues (F290, Y308) previously shown to be in direct contact with the ligand of *ADRB2* (Plazinska, Plazinski, & Jozwiak, 2015; Rosenbaum et al., 2007). On the other hand, residues that were more important to the rest of the Class A GPCRs clustered significantly ($P < 0.00001$) around the extracellular loop structure, which directly binds ligand in more distant Class A GPCRs like Rhodopsin but which evolved to interact indirectly and be less essential in *ADRB2* (Wheatley et al., 2011). These data show with *ADRB2* that the majority of EA predictions are robust to input variability reflecting the conserved functionality of essential biological residues throughout protein homologs. However, by using alignments tailored to a specific biological question, EA further demonstrates the sensitivity to capture functionalities unique to that protein in question. While not all variants could benefit from a customized alignment to improve accuracy, 2,112 of the possible 7,828 substitutions, including many substitutions to residues involved in specific functions like ligand binding, demonstrate some sensitivity to choices in the computational input (ΔEA >±1STD), with 417 having significant alterations in EA score (ΔEA >±2STD).

To test whether other proteins are similarly sensitive to the evolutionary time scale of the alignment used, we conducted the same analysis on four other multifunctional, disease-related proteins (Fig. 5B).

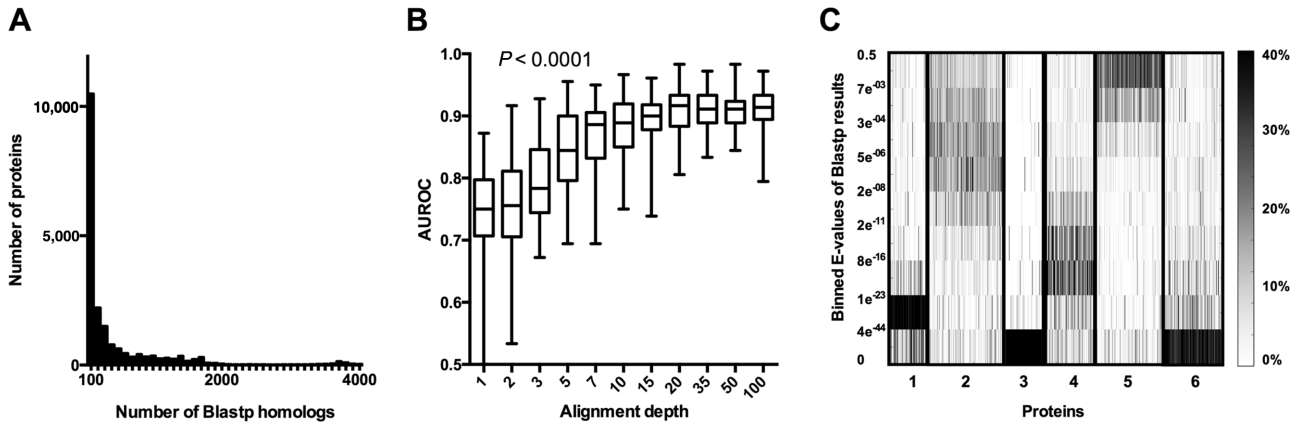

FIGURE 7 Variable evolutionary history of human proteins. **A**: Number of Blastp homologs, from Blastp search with E-value between 0 and 0.5, for all human proteins. **B**: Relationship between alignment depth and variant prioritization accuracy for a set of 28 SNVs in *MLH1*. For each alignment depth, 50 alignments were created by randomly selecting without replacement from the set of *MLH1* homologs with E-values between 0 and 0.5. Variant EA scores were derived from each alignment and variants rankings were assessed using the experimental "overall interpretation of impact" as a gold standard and quantified using the area under the receiver operating curve (AUROC). For each box plot, the center line indicates the median, the box indicates quartiles, and the whiskers indicate the range. Alignment depths were compared statistically using a Kruskal-Wallis test. **C**: E-value distributions of homologous proteins from 20,000 protein blasts. Bins were chosen such that the occupancy of each bin is equal when considering all homologs produced by Blastp searches of human proteins. The y-axis represents the binned E-value distribution of the Blastp search for each human protein (x-axis). Shade represents the percent of all homologs from the protein blast result that fall into that E-value bin. Proteins were organized into six clusters using non-negative matrix factorization

Similar to the β2 adrenergic receptor, the majority of the predictions in these proteins were consistent in both the specific and broad alignments, whereas some variants showed a marked variance in EA predictions. These data show that this phenomenon is generalizable and that customized alignments can help refine results for many proteins. A broad alignment emphasizes global characteristics representing conserved biological function across an entire class of proteins, whereas a customized narrow alignment highlights features unique to a specific protein that have diverged from more distant homologs. A similar analysis investigating rare oncogenic kinase mutations further concluded that the evolutionary depth of the mutated residue was a strong predictor of the mutation's "oncogenic effect" (ManChon, Talevich, Katiyar, Rasheed, & Kanaan, 2014). In the same way that experimental assays must be optimized to the specific biological question of interest, we find that computational experimental design also has a profound effect on the interpretation of predictions for many variants and requires careful consideration.

3.5 | Robustness of homolog availability and distribution varies by protein and may affect computational prediction confidence

Having demonstrated that evolution-based impact predictions are sensitive to intentional changes in the depth and diversity of the homolog tree used for the protein alignment, we next hypothesized that there may be unintentional biases in generic predictions stemming from inherent variability in the availability and evolutionary divergence of homologs for individual proteins. To test this hypothesis, we first acquired, for each human protein, all available homologs from the Uniref90 database (Suzek et al., 2015) using NCBI Blastp and quantified how both the number of homologs and the evolutionary distribution of the homologs varied among proteins. The overall distribution of homolog availability for human proteins fits a power law decay (Fig. 7A) with 50% of all proteins having fewer than 100 homologs and 5,234 proteins having fewer than 20 homologs, indicating that they are below or nearing the lower bounds of appropriate alignment depth recommended in order to produce reliable results (Mihalek et al., 2004). To test how limited homolog availability may affect impact prediction, we randomly sampled the original alignment to generate different alignments of between 1 and 100 homologs for *MLH1*, calculated EA scores for each alignment, and compared the EA scores of the same set of 28 variants explored earlier with an experimental gold standard (Fig. 7B). Prediction accuracy varied significantly ($P < 0.0001$) depending on the number of sequences in the alignment (depth), with increasing alignment depth rapidly leading both to higher average accuracy scores and smaller variance in the accuracy scores until finally reaching a plateau around 20 homologs. These data indicate that the large variability in available data when creating alignments of human proteins is likely to prove highly relevant when attempting to translate computational predictions to experimentally validated results.

Because the current pool of reference sequences is biased to extensively studied species (Zhou et al., 2014), the number of homologs is not necessarily identical to the coverage of the phylogenetic tree or the span of evolutionary history captured by the sequences. To investigate the depth and distribution of evolutionary information captured for each protein, we considered the representation of E-values produced by each protein blast search (Fig. 7C). For each human protein, we converted the number of homologs within an E-value range into a percentage of all homologs from that blast search. In this way, we calculated the balance between highly diverged (larger E-value) and highly similar (smaller E-values) homolog matches. We found that the evolutionary distribution of available homologs varied by protein, and used

non-negative matrix factorization of the binned E-value distributions to group proteins with similar E-value distributions for ease of visualization (Fig. 7C). Blast homolog results for proteins within groups 1, 3, and 6 were composed predominantly of highly similar, very significant matches (E-value $< 1e^{-23}$) with few homologs exhibiting more extensive divergence from the query. For proteins in group 4, the majority of homologs fell into intermediate E-values ($1e^{-23} <$ E-value $< 2e^{-08}$), whereas proteins in groups 2 and 5 had disproportionately more distantly related homologs (E-value $> 3e^{-08}$). These data indicate that human proteins have broad diversity in both the depth and breadth of the evolutionary information present in their available homologs, which can create challenges in computing impact predictions from alignments that sample from these homologs. Differences in the quality and specificity of the alignment between proteins may unknowingly influence impact predictions.

4 | DISCUSSION

The usage of computational methods to identify and predict harmful variants in both a clinical and laboratory setting is dependent on their ability to accurately and reliably represent biology. Studies demonstrating disagreements between prediction and validation have assumed inaccurate computational prediction to be the driver, and suggested that clinicians might be better served by forgoing computational analysis entirely and proceeding directly to testing in vitro function. Here, we presented a methodological analysis exploring several possible reasons beyond inaccurate prediction for why these disagreements may occur, and conclude that disagreements are likely to occur when the computational tool and validation assays are not testing the same underlying hypothesis. We find that improved agreement between prediction and validation can be achieved when experimental testing reflects a holistic definition of function that is attuned to the computational methodology. Furthermore, many variant predictions depend heavily on an appropriate alignment choice and require a set of sequences that, by virtue of their evolutionary span, reflect the functional features that best match the biological question and the subsequent experimental validation. Therefore, an increased collaboration between computational biologists and experimentalists would improve the biological relevance of the predictions.

A call for more extensive experimental validation poses a large task both for biologists and for gold standard prediction competitions that use biological assays to benchmark the ability of computational tools to score SNV impact on protein function. Yet, as we have shown here, both the quality and the thoroughness of biological characterization can profoundly influence the final conclusions. Mutations with no effect in certain assays may result in constitutive activation or loss of function in other tests. Consequently, variants predicted to be impactful are sometimes initially deemed as "false positives" based on existing literature (Bromberg, Overton, Vaisse, Leibel, & Rost, 2009), only to be confirmed years later after additional testing (He & Tao, 2014). Moreover, the same experimental assay may reveal variant pathogenicity only in a particular environment; in a study of 10 *BRCA1* (MIM# 113705) variants assayed for transactivation activity in both yeast and mammalian cells, 20% of the variants could only be verified as loss-of-function causative variants, consistent with observed pedigrees, when using the mammalian system (Vallon-Christersson et al., 2001). An exhaustive assessment of the impact is especially crucial for proteins known to be multifunctional; though with estimates of multifunctionality ranging from 25% to 65% of human proteins (Ekman, Bjorklund, Frey-Skott, & Elofsson, 2005; Pritykin et al., 2015), this group is hardly a minority. Beyond testing more functions, higher reproducibility for experimental determination of impact also has a tangible benefit; we described above how averaging experimental replicates improved correlation between prediction and validation, but other prediction methods have also found that when multiple studies agree on the experimental interpretation of a variant, agreement with prediction also rises markedly (Bromberg et al., 2009).

However, even best efforts to maximize experimental rigor are unlikely to resolve these issues entirely. In reality, there are many ways in which a variant's behavior may be altered by environmental conditions: cell line variations, in vitro versus in vivo context, temperature, and cell composition, on top of variations within the spatiotemporal factors in these systems. These considerations create a formidable task for experimentalists to design a rigorous panel of tests that is not prohibitively cost and time-intensive. Furthermore, while computational methods predict global impact, often clinical and biological relevance may be directly linked to a specific functionality. Although it is often not possible to know a priori which tests will be most efficient in this regard, it is necessary to find a balance between rigor and practical value.

Moreover, there are challenges to validation that would not be easily resolved by additional or repetitive testing. Even well-studied proteins may have currently unknown functions, limiting a satisfactory reconciliation of predicted and assayed variant impact. In addition, one recent study indicates that current state-of-the-art experimental methodologies used for measuring fitness effects do not adequately reproduce constraints found in nature and can cause many impactful mutations to falsely appear benign in the laboratory setting (Rockah-Shmuel et al., 2015); more realistic mutation conditions dramatically increased the fraction of mutations assessed experimentally to be deleterious. These possibilities highlight how the positive predictive value of computational methods may be underestimated by current approaches to validation, and can perhaps explain why studies like Miosge et al. (2015) found the false-positive rate for computational prediction methods to be two to five times higher than the false-negative rate, which was generally quite low (3%–5%). For these reasons, bypassing computational prediction and proceeding directly to testing would not necessarily improve clinical diagnosis, as inconsistencies between prediction and testing can be driven by the experimental setup as well as by prediction errors.

Of course, prediction errors should not be ignored as a potential source of disagreement. We, like others (Adebali, Reznik, Ory, & Zhulin, 2016; Hicks et al., 2011), find that alignment choice plays an important role for evolutionary-based prediction methods. In *ADRB2*, we found that nearly half of potential variants were sensitive to computational input. However, changes to the variant prediction scores did

not reflect a nonrobust methodology, but rather a different underlying hypothesis. By using a narrow alignment for close homologs of *ADRB2*, we identified biologically relevant clusters of residues within the structure that were predicted to be more or less important to *ADRB2*-like GPCRs than the rest of the Class A GPCRs, and even more precision could be gained by using advanced tools for alignment subdivision in order to predict positions essential for even smaller alignments. By tailoring the computational approach to the specific biological question in this way, predictions can be customized to better reflect variant effect in a specific functional context. Yet while method developers can and do create bespoke predictions, this level of knowledgeable refinement may be out of reach for an average user who has downloaded one of the many accessible predictor tools for personal use (Katsonis et al., 2014). There is a great deal of opportunity for collaboration between computational biologists and experimental biologists in this area that remains to be harnessed; across more than 1,000 research articles with a reference to any of six popular prediction methods in their abstract, less than 2% of articles appeared to involve a direct collaboration between study authors and the tool developers. Ideally, communication across disciplines would increase to the point that when a user finds that default alignments or parameters for a tool produce results inconsistent with experimentation, both groups would begin the process of troubleshooting together.

Naturally, just as there are some challenges in experimental design that cannot be addressed by additional testing, there are computational challenges to prediction that cannot be addressed by alignment optimization. In particular, nonuniform homolog availability across proteins as well as the evolutionary age of a functional region in a protein can affect prediction confidence. Limited evolutionary information can lead to less certain or even inaccurate predictions in these cases, with variant impact on very newly emergent or divergent functional properties most likely to be missed by standard approaches to prediction (Ng & Henikoff, 2006). In addition, the vast majority of variant impact prediction tools predict the effect of each variant independently, and none are currently able to consider the full genomic context of the variant and factor in the magnitude and direction of epistatic modulation. These modifications may not be too far in the future—EA and other methods have demonstrated strong predictive ability when considering the impact of multiple concurrent variants within a protein (CAGI, 2016), so the incorporation of genetic or even multiomic information from outside the protein in question (Raimondi, Gazzo, Rooman, Lenaerts, Vranken, 2016) is a logical next step. Furthermore, given the practical and biological limitations to comprehensively assaying a protein, computational prediction may need to take on the challenge of developing advanced methods to predict a variant's impact on a specific protein function, thereby increasing applicability when guiding precise biological investigation.

The ultimate goal of computational prediction is to integrate extensive variant information into a single measure that accurately represents the presentation of a complex, multigenic phenotype. Although this type of modeling is still a work in progress, one notable success in yeast came when the protein alignments were optimized through collaboration with tool developers before variant impact prediction, 20 different growth conditions were considered when testing functionality, and the group of genes responsible for the phenotype was well-delineated (Jelier, Semple, Garcia-Verdugo, & Lehner, 2011). This study demonstrates a successful union of biological investigation and computational prediction and further shows how active cooperation between the two fields leads to more accurate estimations of computational prediction accuracy and a better understanding of the relationship between genotypes and phenotypes. An increased adoption of the type of approach will enable an increased personalization of future computational methods.

ACKNOWLEDGMENTS

A.K. is supported by a CPRIT predoctoral fellowship, and the Baylor Research Advocates for Student Scientists (BRASS).

DISCLOSURE STATEMENT

The authors have no conflict of interest to declare.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., ... Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Adebali, O., Reznik, A. O., Ory, D. S., & Zhulin, I. B. (2016). Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genetics in Medicine*, *18*, 1029–1036.

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Bromberg, Y., Overton, J., Vaisse, C., Leibel, R. L., & Rost, B. (2009). In silico mutagenesis a case study of the melanocortin 4 receptor. *The FASEB Journal*, *23*(9), 3059–3069.

Caffrey, D. R., Dana, P. H., Mathur, V., Ocano, M., Hong, E. J., Wang, Y. E., ... Huang, E. S. (2007). PFAAT version 2.0: A tool for editing, annotating, and analyzing multiple sequence alignments. *BMC Bioinformatics*, *8*, 381.

CAGI (Critical Assessment of Genome Interpretation). (2016). Human SUMO ligase (UBE2I): Predict the effects of missense mutations on competitive growth in a high-throughput yeast complementation assay. Retrieved from https://genomeinterpretation.org/content/4-SUMO_ligase

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745–2747.

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., & Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, *25*, 2537–2543.

Edgar, RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Ekman, D., Bjorklund, A. K., Frey-Skott, J., & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *Journal of Molecular Biology*, *348*(1), 231–243.

He, S., & Tao, Y. X. (2014). Defect in MAPK signaling as a cause for monogenic obesity caused by inactivating mutations in the melanocortin-4 receptor gene. *International Journal of Biological Sciences*, *10*(10), 1128–1137.

Hicks, S., Wheeler, D. A., Plon, S. E., & Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutation*, *32*(6), 661–668.

Itan, Y., & Casanova, J. L. (2015). Can the impact of human genetic variations be predicted? *Proceedings of the National Academy of Sciences*, *112*(37), 11426–11427.

Jelier, R., Semple, J. I., Garcia-Verdugo, R., & Lehner, B. (2011). Predicting phenotypic variation in yeast from individual genome sequences. *Nature Genetics*, *43*(12), 1270–1274.

Katritch, V., Cherezov, V., & Stevens, R. C. (2013). Structure-function of the G-protein-coupled receptor superfamily. *Annual Review of Pharmacology and Toxicology*, *53*, 531–556.

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Science*, *23*(12), 1650–1666.

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Research*, *24*(12), 2050–2058.

Lee, S. P., O'Dowd, B. F., Rajaram, R. D., Nguyen, T., & George, S. R. (2003). D2 dopamine receptor homodimerization is mediated by multiple sites of interaction, including an intermolecular interaction involving transmembrane domain 4. *Biochemistry*, *24*(37), 11023–11031.

Lehner, B. (2013). Genotype to phenotype: Lessons from model organisms for human genetics. *Nature Reviews Genetics*, *14*(3), 168–178.

Lichtarge, O., Yamamoto, K. R., & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *Journal of Molecular Biology*, *274*(3), 325–337.

Lin, H., Sassano, M. F., Roth, B. L., & Shoichet, B. K. (2013). A pharmacological organization of G protein coupled receptors. *Nature Methods*, *10*(2), 140–146.

ManChon, U., Talevich, E., Katiyar, S., Rasheed, K., & Kanaan, N. (2014). Predicton and prioritization of rare oncogenic mutations in the cancer kinome using novel features and multiple classifiers. *PLOS Computational Biology*, *10*(4), e1003545.

Milligan, G. (2004). G protein-coupled receptor dimerization: Function and ligand pharmacology. *Molecular Pharmacology*, *66*(1), 1–7.

Mihalek, I, Res, I, Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol.*, *336*(5), 1265–82.

Miosge, L. A., Field, M. A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., ... Whittle, B., et al. (2015). Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(37), E5189–E5198.

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, *11*(5), 863–874.

Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, *7*, 61–80.

Peterson, S. M., Pack, T. F., Wilkins, A. D., Urs, N. M., Urban, D. J., Bass, C. E., … Caron, M. G. (2015). Elucidation of G-protein and β-arrestin functional selectivity at the dopamine D2 receptor. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(22), 7097–7102.

Plazinska, A., Plazinski, W., & Jozwiak, K. (2015). Agonist binding by the β2-adrengergic receptor: An effect of receptor conformation on ligand association-dissociation characteristics. *European Biophysics Journal*, *44*, 149–163.

Pritykin, Y., Ghersi, D., & Singh, M. (2015). Genome-wide detection and analysis of multifunctional genes. *PLOS Computational Biology*, *11*(10), e1004467.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*(1), 5–15.

Raevaara, T. E., Korhonen, M. K., Lohi, H., Hampel, H., Lynch, E., Lonnqvist, K. E., ... Nystrom. M. (2005). Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. *Gastroenterology*, *129*(2), 537–549.

Raimondi, D., Gazzo, A. M., Rooman, M., Lenaerts, T., & Vranken, W. F. (2016). Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. *Bioinformatics*, *32*(12), 1797–1804.

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: Server and survey. *Nucleic Acids Research*, *30*(17):3894–3900.

Rockah-Shmuel, L, Toth-Petroczy, A, Tawfik, DS. (2015). Systemic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput Biol.*, *11*(8), e1004421. PMCID: PMC4537296.

Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kolbilka, T. S., ... Kobilka, B. K. (2007). GPCR Engineering yields high-resolution structural insights into 2-Adrenergic receptor function. *Science*, *318*(5854), 1266–1273.

Sbai, O., Monnier, C., Dode, C., Pin, J. P., Hardelin, J. P., & Rondard, P. (2014). Biased signaling through G-protein-coupled PROKR2 receptors harboring missense mutations. *The FASEB Journal*, *28*(8), 3734–3744.

Schwarz, J. M., Rodelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–576.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, *33*, W382–W388.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Stone, E. A., & Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research*, *15*(7), 978–986.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt Consortium. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932.

Sun, S., Yang, F., Tan, G., Costanzo, M., Oughtred, R., Hirschman, J., ... Roth, F. P. (2016). An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*, *26*(5), 670–680.

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., ... Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *Journal of Medical Genetics*, *43*(4), 295–305.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141.

Wheatley, M., Wootten, D., Conner, M. T., Simms, J., Kendrick, R., Logan, R. T., ... Warwell, J. (2011). Lifting the lid on GPCRs: The role of extracellular loops. *British Journal of Pharmacology*, *165*(6), 1688–1703.

Vallon-Christersson, J., Cayanan, C., Haraldsson, K., Loman, N., Bergthorsson, J. T., Brondum-Nielsen, K., ... Monteiro, A. N. A. (2001). Functional analysis of BRCA1 C-terminal missense mutations identified in breast

and ovarian cancer families. *Human Molecular Genetics*, *10*(4), 353–360.

Wilkins, A. D., Erdin, S., Lua, R., & Lichtarge, O. (2012). Evolutionary trace for prediction and redesign of protein functional sites. *Methods in Molecular Biology*, *819*, 29–42.

Yates, C. M., Filippis, I., Kelley, L. A., & Sternberg, M. J. (2014). SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology*, *426*(14), 2692–2701.

Zhou, C., Mao, F., Yin, Y., Huang, J., Gogarten, J. P., & Xu, Y. (2014). AST: An automated sequence-sampling method for improving the taxonomic diversity of gene phylogenetic trees. *PLoS One*, *9*(6), e98844.

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, *11*, 2714–2726.

How to cite this article: Gallion J, Koire A, Katsonis P, Schoenegge A, Bouvier M, Lichtarge O. Predicting phenotype from genotype: improving accuracy through more robust experimental and computational modeling. *Human Mutation*. 2017;38:569–580. https://doi.org/10.1002/humu.23193