INVITAE

Integrating functional modeling to enhance clinical variant interpretation in Mendelian disease

Challenges of interpreting missense variants

Research over the past decade has illuminated the dramatic extent of variation in the human genome. Any two unrelated individuals have millions of DNA sequence changes that differ between them, including a large proportion that alter protein coding sequences with single amino acid changes (i.e., missense variants).[1] Amid this natural variation, individuals may carry certain sequence changes that cause or predispose them to monogenic hereditary disease. Identifying and correctly classifying these variants is central to clinical genetic testing.

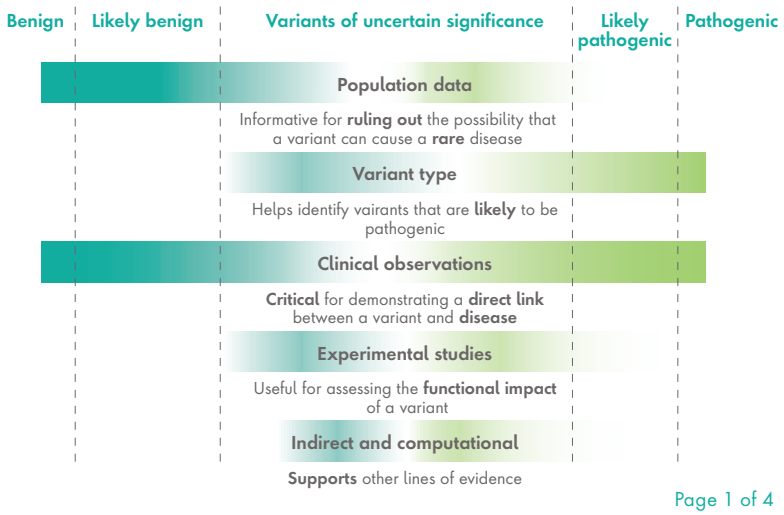
The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have developed useful guidelines for evaluating evidence for the clinical interpretation of DNA sequence variants.[2] Invitae has also pioneered a semi-quantitative, score-based approach for clinically classifying DNA variants. This approach, known as Sherloc, was modeled after the framework recommended by ACMG/AMP.[2,3] However, Sherloc also includes many refinements that provide important granular context for when and how diverse types of evidence should be used and combined into robust and consistent variant classification. Importantly, the points-based framework has proven to be a powerful tool for evaluating and assigning relative weights (i.e., scores) to individual lines of evidence and criteria.

Importantly, prospective evaluation of 624 likely pathogenic or pathogenic and likely benign or benign classifications from Sherloc, compared with highly confident variant classifications submitted by diagnostic laboratories to ClinVar for these variants, demonstrates that clinical interpretations made with Sherloc are at least 99% confident. Importantly, this exceeds recommendations provided by the ACMG/AMP working group that likely pathogenic and likely benign classifications should be more than 90% confident.[2]

Within the Sherloc framework, five categories of evidence are considered for each sequence variant: the frequency of the variant in the general population (population data), the expected effect of the variant on the mRNA or protein (variant type), the clinical presentation of the individual carrying the variant and whether the variant segregates with disease among family members (clinical observations), the observed impact of the variant on RNA or protein function in vitro or in cell culture (experimental studies), and inferences made about the variant based on known biological properties of the RNA and protein (indirect and computational).

Five Sherloc evidence categories that determine pathogenicity

Each category of evidence provides different information about a variant and its likelihood of being pathogenic or benign. The shading of the bars within each category illustrates the level to which data within that category are impactful for classifying a variant as pathogenic or benign. For example, the population frequency of a variant can only be used to rule out the possibility that a variant causes disease; on its own, it does not define a variant as pathogenic. Conversely, the type of variant (e.g., a frameshift variant) can help in identifying variants that are most likely to be pathogenic, but is much less useful for classifying a variant as benign. For more details about the evidence categories and data used for variant classification within the Sherloc framework, see Nykamp *et al.* 2017.



Clinical significance is relatively easy to interpret for two types of sequence variants: those expected to result in a complete loss of protein expression or function (e.g., nonsense, frameshift, and consensus splice-site variants) in genes where protein loss-of-function is known to cause disease, and those that do not directly impact the protein sequence (e.g., synonymous variants). In contrast, it is more difficult to interpret the clinical significance of variants that result in amino acid changes in the protein (e.g., missense variants and in-frame insertions or deletions). This is because they are the most abundant changes in the genome and while they can impact protein function and lead to disease, they can also lead to amino acid changes that have no discernible impact on the protein's function. Within Invitae's test population, missense variation contributes significantly to uncertain results—roughly 28% of all individuals tested have a missense variant of uncertain significance (VUS), compared with 5% of individuals who have a non-missense VUS.

Clearly, in order to provide more definitive results to more patients and decrease the uncertainty in clinical genetic testing, we need to better understand how missense variants throughout the genome affect protein function.

Need for methods with better predictive values

Our ability to readily interpret DNA variants has dramatically improved with the rapid expansion of public genomic resources such as the Genome Aggregation Database (gnomAD). This gives us much better visibility to the population frequency of each variant and helps define which variants are too common to be the primary cause of a rare disease.[4] In addition, with a steady increase in the volume, scale, and quality of clinical genetic testing, the ability to correlate a rare variant with disease has also improved. Yet, while substantial progress has been made in systematically evaluating the clinical consequences of DNA variants overall, understanding the functional consequences of missense variants has lagged behind.

Two categories of methods are currently available for evaluating the functional impact of missense variants: experimental methods and computational methods. Experimental methods include in vitro and in vivo assays designed to test the functional impact of variants on protein function. Well-established and validated functional assays that reflect the full biological function of the protein, such as cellular signaling or substrate breakdown by an enzyme, provide the strongest evidence that a missense variant may or may not disrupt protein function. Assays that test just one aspect of the protein, such as protein stability or cellular localization, provide much weaker evidence as these may not be the primary mechanisms by which changes in the protein function leads to disease.[2] Although experimental methods can be extremely powerful for interpreting missense variants, in vitro and in vivo assays have remained substantially limited in clinical utility and number due to the costs and variable performance associated with these methods.[5,6]

Computational methods include in silico algorithms developed to predict the functional impact of rare or novel missense variation throughout the genome. These often leverage protein sequence, structural, and functional data published in the literature or available in public databases. Historically, computational methods have provided relatively low or inconsistent predictive values (i.e., accuracy) among diverse classes of proteins.[7] For example, the accuracy (i.e., Matthews correlation coefficient*) of commonly used tools such as SIFT, PolyPhen2, REVEL, and MetaSVM has been generally limited to 75% to 85%.[8-10] A major limitation, and likely contributor to the overall low accuracy of these tools, is that they attempt to generate a single algorithm for all genes, relying on generalized biological properties that are evaluated across all residues in all proteins, such as amino acid conservation or physicochemical properties. However, this approach cannot account for protein-specific and domain-level characteristics (i.e., molecular stability, domain interactions, allosteric networks, epistatic effects, and conformational dynamics) that contribute to the unique properties and activities of each protein. As a result, the accuracy of these general algorithms is highly variable across different classes of proteins, ranging from 60% to 95% depending on the gene and algorithm tested.[11]

* The Matthews correlation coefficient is an established analytic metric of balanced accuracy, which accounts for accuracy in both positive (i.e., pathogenic) and negative (i.e., benign) calls.

Due to the the relatively low and highly variable accuracy of published missense predictors, the ACMG/AMP guidelines for the interpretation of sequence variants recommend that these computational methods be used only as supporting evidence for variant classification.[2] Unfortunately, this limits the potential utility of these methods for resolving the uncertainty of novel or rare missense variants since these variants have limited clinical evidence.[7]

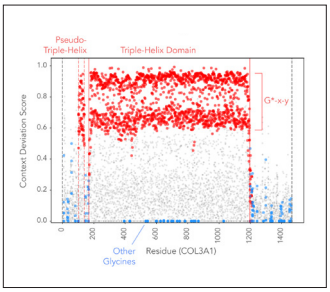
Invitae's functional modeling platform

To directly address these concerns and develop computational methods that both more accurately classify variants and greatly improve our understanding of how specific missense variants affect protein function, Invitae has invested in a functional modeling platform (FMP). Importantly, the FMP leverages recent improvements in our understanding of genomic variation and the biophysical properties of proteins with a machine learning approach that generates algorithms with >95% accuracy for all genes analyzed and tested. Given the higher performance of these gene-specific algorithms, compared to those recommended by the ACMG/AMP guidelines,[2,7] predictions from the FMP can be given higher weight within our Sherloc variant classification system.

With data generated from the most recently updated FMP models, we empirically assessed the prospective performance (i.e., accuracy) of these models and, using these performance metrics, carefully determined the appropriate weight of the functional modeling algorithms relative to the performance of other lines of evidence used by Sherloc. As with other types of functional evidence, we found that evidence from the FMP would not be sufficient, by itself, to result in a likely pathogenic or likely benign variant classification; however, the FMP evidence contributes to the overall Sherloc score and, in combination with clinical evidence, results in many more definitive classifications than would be possible without this evidence.

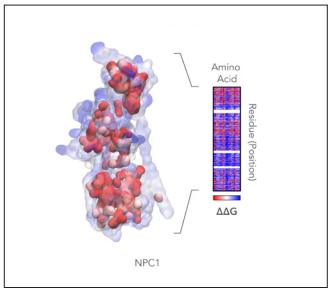
Invitae's functional models (>95% accuracy)

Sequence analyses
"Gene-specific engines"



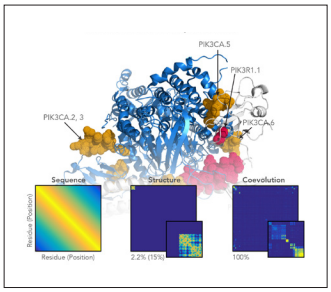
696 genes covered by this model

Biophysical analyses
"Molecular stability engines"



117 genes covered by this model

Spatial relationships
"Multi-dimensional hotspots"



416 genes covered by this model

Cellular assays
"Deep mutational assays"

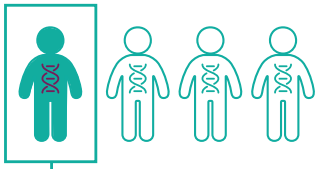
2 genes covered by this model

Four types of functional models were trained and tested for performance using data available in July 2019. Importantly, only gene models with >95% accuracy are used for variant classification. Therefore, some genes may have predictions from all four models, while others will have no predictions from any model. While this limits the number of genes for which we can make predictions, it ensures that the confidence of the predictions provided by these models is high.

Importantly, integration of the functional modeling platform into the Sherloc framework will offer a more precise variant classification in more than 1 in 4 patients tested at Invitae (28%). When analyzing the prospective performance of variants with classifications informed by the FMP data, we found that the accuracy was greater than 99% (n = 728).

Altogether, the incorporation of FMP data into Sherloc substantially increases the clinical value of Invitae's genetic tests, while maintaining the same overall confidence of our variant classifications.

FMP informs missense VUS, which are found in more than 1 in 4 (28%) of all patients tested (N=689,321).



FMP evidence increased the certainty of the Sherloc score for 12% of all patients tested (43% of patients with a missense VUS result).

Invitae is committed to providing healthcare providers and their patients with the most accurate and up-to-date information they need to make healthcare decisions. Invitae's variant classification system, Sherloc, and the newly developed FMP, are clear examples of that commitment. We will continue to leverage innovations in both experimental and computational technologies along with the entire variant classification approach to bring the best of science and technology to healthcare providers and the individuals under their clinical care.

References

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
2. 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*. 2015;17(5):405-424.
3. Nykamp K, *et al.* Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105-1117.
4. Kobayashi Y, *et al.* Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med*. 2017;9(1):13.
5. Majithia AR, *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat Genet*. 2016;48(12):1570-1575.
6. Findlay GM, *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. 2018;562(7726):217-222.
7. Ghosh R, *et al.* Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol*. 2017;18(1):225.
8. Sim NL, *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40(web server issue):W452-W457.
9. Adzhubei IA, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249.
10. Ioannidis NM, *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. 2016;99(4):877-885.
11. Tian Y, *et al.* REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Sci Rep*. 2019;9:12752.

HEADQUARTERS | 1400 16th Street, San Francisco, CA 94103 | ONLINE | www.invitae.com | CONTACT | www.invitae.com/contact