Establishing and evaluating a framework for describing variant evidence in clinical genomic reports

Karen Ouyang*, Jackie Tahiliani*, Yuan-Yuan Ho, Michelle Sommargren, Michael Jen, Brandon Bunker, Britt Johnson, Scott Topper, and Keith Nykamp

Invitae, San Francisco, CA

* authors contributed equally

Introduction

With improvements in technology, the cost of DNA sequencing has dropped precipitously. At the same time, the amount of data available for analysis and review has increased significantly. Communicating all this information in a succinct, cohesive manner to clinicians and patients can be extremely challenging and time consuming. To facilitate this process, we use a unique score-based evidence system for variant classification named Sherloc* and a framework for describing the relevant information in a clinical genomics report. We've divided the variant details (VD) into four main sections: (1) a molecular description of the variant, (2) evidence that is directly applicable to the variant (e.g., population data, clinical and functional data), (3) indirect and predictive evidence (e.g., computational analysis), and (4) a final summary. Suggested VD texts were created for different scenarios. Informatics solutions support the automation of these details into the report and metrics are routinely gathered to allow optimization of text language and increase future efficiency.

*For more information on the variant classification methodology at Invitae, see poster #G11: "Sherloc: Evaluation of a score-based implementation of the AMP/ACMG ISV guidelines in a scalable genetic diagnostic laboratory."

What is the variant details system?

The variant details system was developed at Invitae to provide a structured framework for reporting the interpretation (e.g., pathogenic, variant of uncertain significance) of a given variant as well as for describing all evidence used to classify the given variant. To support this structured framework, common text was developed to guide consistency in describing the evidence for variant classification across variants as well as variant interpreters.

Element	What is this?	Details
Element 1	Variant description	Factual description of the observed DNA change; no judgments
Element 2	Direct evidence	Overview: presence/absence in population databases and literature • Description of affected individuals in literature, including segregation data • Presence in clinical databases • Functional data
Element 3	Indirect evidence	In silico predictions Important functional domains
Element 4	Summary	Summarize elements 2-3

Example of variant details write-up

Element 1 →
Element 2 → Clinical information in literature
Element 2 → Functional data
Element 4 →

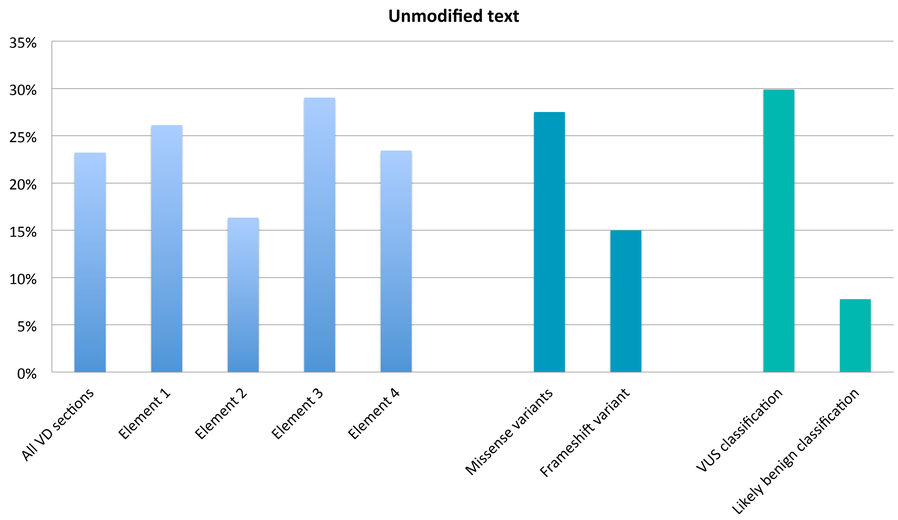
KCNQ1, Exon 6, c.830C>T (p.Ser277Leu), heterozygous, PATHOGENIC

- This sequence change replaces serine with leucine at codon 277 of the KCNQ1 protein (p.Ser277Leu). The leucine residue is highly conserved and there is a large physicochemical difference between serine and leucine.
- This sequence change has been reported in the literature and is not currently found in any individuals from the population databases (rs199472730, no frequency). It has been shown to segregate with long QT syndrome in multiple families (PMID: 21895724, 21241880, 12442276) and has been identified in several unrelated individuals with isolated long QT syndrome (PMID: 19716085).
- Experimental studies have shown that this missense change leads to a non-functional KCNQ1 protein and acts in a dominant-negative manner to reduce the activity and reduces the surface localization of normal KCNQ1 protein (PMID: 21241800, 21895724).
- This sequence change is absent from population databases, has been shown to segregate with long QT syndrome in several families, and leads to a deleterious effect on protein activity. For these reasons, this sequence change has been classified as Pathogenic.

Results

For all suggested text used across all VD sections, 23.2% remained unmodified. Indirect and predictive evidence (29% unmodified) was the least modified VD section; evidence that directly applies to the variant (16% unmodified) was the most frequently modified.

When data was analyzed by variant type, text describing missense variants (27.5% unmodified) was the least modified and frameshift variants (15.0% unmodified) were the most modified variant type. A variant classification of VUS (29.9% unmodified) demonstrated the least text modification, while a classification of likely benign (7.7% unmodified) was the most modified.



Of the four elements, Element 2 had the lowest unmodified text percentage. After review of the content in Element 2, specific pieces of evidence were parsed out to improve automation of this step. Specifically, population frequency details were highlighted as a fully automatable section and therefore changes to Element 2 structure were necessary. The following changes were implemented to improve automation for Element 2:

Element	Bullets	Details
Element 2	2A	ExAC population data, including the highest subpopulation frequency
	2B	Affected individuals from the literature (including PMIDs), segregation data, clinical databases if present (including ClinVar variation ID)
	2C	Functional data, if available (including PMIDs)

Example of modified variant details write-up

Element 2A →
Element 2B →
Element 2C →

KCNQ1, Exon 6, c.830C>T (p.Ser277Leu), heterozygous, PATHOGENIC

- This sequence change replaces serine with leucine at codon 277 of the KCNQ1 protein (p.Ser277Leu). The leucine residue is highly conserved and there is a large physicochemical difference between serine and leucine.
- This variant is not present in population databases (rs199472730, no frequency).
- This variant has been shown to segregate with long QT syndrome in multiple families (PMID: 21895724, 21241880, 12442276) and has been identified in several unrelated individuals with isolated long QT syndrome (PMID: 19716085). ClinVar contains an entry for this variant (Variation ID: 53116).
- Experimental studies have shown that this missense change leads to a non-functional KCNQ1 protein and acts in a dominant-negative manner to reduce the activity and reduces the surface localization of normal KCNQ1 protein (PMID: 21241800, 21895724).
- This sequence change is absent from population databases, has been shown to segregate with long QT syndrome in several families, and leads to a deleterious effect on protein activity. For these reasons, this sequence change has been classified as Pathogenic.

Methods and materials

We collected usage data and analyzed the overall frequency of suggested text modifications by each VD section, variant type (i.e., missense, frameshift, splicing, nonsense), and classification (i.e., pathogenic, likely pathogenic, variant of uncertain significance [VUS], likely benign).

Conclusions

The data indicate that conveying variant analysis results is still generally reliant on operator customization. This customization is a required aspect of variant interpretation as there are pieces of evidence that are not amenable to automation. In order to refine the variant interpretation process, considerable improvements can be incorporated by targeting evidence that can be described with common language and informatic support.

Our analysis provides evidence that improving the efficiency of reporting variant information is a compelling need in genomics. When these findings are incorporated into informatics planning, further elaboration of the framework may decrease the frequency of modified text across various areas. This optimizes the reporting process of each variant as the analysis and description of data can then be reproducible across users.

Further improvements

This initial evaluation of the data highlights areas where automation may improve the efficiency and consistency of the description of variants. Further improvements within the other elements will ensure that variant interpreters can focus on the interpretation of the information available.

Specifically, further development of Element 1 for more complicated variant types (e.g., frameshift, splicing) may reduce modification of this general description of the molecular event.

References and acknowledgements

- Clinical Genomics Group, Invitae
- Genetics Development Group, Invitae
- Poster G11 - Sherloc: Evaluation of a score-based implementation of the AMP/ACMG ISV guidelines in a scalable genetic diagnostic laboratory