Accessing clinical-grade genomic classification data through the ClinGen Data Platform*

Karen P. Dalton[1], Heidi L. Rehm[2,3], Matt W. Wright[1], Mark E. Mandell[1], Kilannin Krysiak[4], Lawrence Babb[3], Kevin Riehle[5], Tristan Nelson[6], Alex H. Wagner[7,8]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA; [2]Massachusetts General Hospital, Boston, MA; [3]Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; [4]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO; [5]Baylor College of Medicine, Houston, TX; [6]Geisinger, Danville, PA; [7]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH; [8]Departments of Pediatrics & Biomedical Informatics, The Ohio State University College of Medicine, Columbus, OH.

The Clinical Genome Resource (ClinGen) serves as an authoritative resource on the clinical relevance of genes and variants. In order to support our curation activities and to disseminate our findings to the community, we have developed a Data Platform of informatics resources backed by standardized data models. In this workshop we demonstrate our publicly available resources including curation interfaces, (Variant Curation Interface, CIViC), supporting infrastructure (Allele Registry, Genegraph), and data models (SEPIO, GA4GH VRS, VA).

Keywords: Clinical Genomics; ClinGen; GA4GH; Data Standards; Variant Interpretation

1. Introduction

Genome-guided precision medicine requires evaluating the clinical significance of genomic variation through the aggregation and standardized evaluation of disparate lines of functional, clinical, and observational evidence. The process by which evidence is combined and turned into a formal classification of significance is guided by professional organization or consortia-driven recommendations, such as the 2015 ACMG/AMP guidelines[1] for Mendelian disease variants, the 2017 AMP/ASCO/CAP guidelines[2] for somatic cancer variants, and the recently published 2022 ClinGen/CGC/VICC guidelines[3] for cancer variant oncogenicity. The application of these guidelines requires carefully controlled curation interfaces and expert vetting of evidence to ensure reproducible and high-quality assertions of clinical significance.

To address this need, the NIH-funded Clinical Genome Resource (ClinGen) was founded in 2013 to serve as a central authority for defining the clinical relevance of genes and variants for use in precision medicine and research. The ClinGen Data Platform represents the coordinated activities of the ClinGen data tools that drive the generation and dissemination of carefully curated, high-quality assertions of clinical relevance in public databases and precision medicine pipelines (clinicalgenome.org/working-groups/data-platform). The Data Platform enables the clinical knowledge journey: the interfaces used to curate clinical significance classifications, the

frameworks for structuring and normalizing them, and the tools for exchanging and widely disseminating this clinical knowledge for use in clinical systems.

2. Workshop Topics and Presenters

2.1. Introduction - The Clinical Genome Resource

Presented by: Heidi Rehm (Broad Institute of MIT and Harvard & Massachusetts General Hospital)

This workshop describes the Clinical Genome Resource (ClinGen) and how ClinGen standardizes and supports the classification of the clinical significance of genes and variants. ClinGen activities include development of standardized frameworks for gene and variant classification, provision of the needed software structures to support this work, and crowd-sourcing the sharing of gene and variant classifications and underlying curated evidence through ClinGen's website (clinicalgenome.org), GenCC (Gene Curation Coalition) and ClinVar (NCBI supported). Conflicting classifications are resolved through interlaboratory efforts for both ClinVar and GenCC entries, and a subset of variants are reviewed and classified through the consensus-driven application of ClinGen's expert panels. This session will also examine forward-looking approaches needed to scale the classification of variants, including example patient cases with variants for use throughout each portion of the workshop. This will entail a review of evidence types used in variant classifications and discussion of how sharing this data according to harmonized data models enables more scalable approaches to variant classification.

2.2. Generating clinical-grade genomic knowledge

2.2.1. Clinical variant knowledge from Variant Curation Expert Panels

Presented by: Matt Wright, Karen Dalton, Mark Mandell (Stanford University)

The ClinGen Variant Curation Interface (VCI)[4] is a global, open-source cloud-native, variant classification platform for supporting the application of evidence-based criteria and classification of variants based on the ACMG/AMP variant classification guidelines. Publicly accessible via https://curation.clinicalgenome.org, the VCI is among a suite of tools developed by ClinGen and supports an FDA-recognized human variant curation process. It enables collaboration and peer review across ClinGen Expert Panels, and supports users in identifying, annotating, and sharing relevant evidence while making variant pathogenicity assertions. Navigation workflows support users by providing guidance to comprehensively apply the ACMG/AMP evidence criteria and document provenance for asserting variant classifications both within ClinGen expert panels and the wider genomics community.

At this part of the data journey from patient genomic data to clinically relevant interpretation of variants, data is ingested from a variety of community resources and, after complete curation, is exported to other resources within the ClinGen ecosystem and also exported with classified variants into ClinVar and the Evidence Repository. We will discuss the use of defined ontologies and data structures to produce consensus interpretations from defined methodologies at scale. The semi-structured workflow in combination with the evaluation by expert panel members moves determinations of variant pathogenicity away from the prior methods of relying on subjective

judgment by a single individual toward structured review of evidence to reach expert consensus, thereby increasing the confidence in the data created.

2.2.2. Somatic cancer clinical variant knowledge from Somatic Cancer Variant Curation Expert Panels

Presented by: Kilannin Krysiak (Washington University in St. Louis), Alex Wagner (Nationwide Children's Hospital and the Ohio State University)

The crowd-sourced, public domain Clinical Interpretations of Variants in Cancer (CIViC) knowledgebase[5] is a cancer variant knowledgebase funded by the NCI Informatics Technology for Cancer Research program that collaborates closely with ClinGen and captures literature-derived evidence for the clinical assessment of genomic variants in cancers through an open evidence curation interface[6]. ClinGen Somatic Cancer Variant Curation Expert Panels (SC-VCEPs) capture evidence in CIViC using concepts from established terminologies for cancer types, therapies, histopathologies, and genes, alongside CIViC-defined structured data fields and human-readable text. The CIViC curation interface supports a rigorous evidence curation protocol[7], which is used and expanded upon by SC-VCEPs in domain-specific (e.g. tumor type and/or gene specific) curation activities. CIViC content is freely available without registration via the web interface, text downloads or API access, and its content is released under a public domain (CC0) declaration.

We will cover the fundamental data types curated in the CIViC interface, and how these apply to professional society guidelines to guide clinical interpretation of tumor variants. A hands-on exercise using Python-based Jupyter notebooks will demonstrate the use of the GraphQL API and the CIViCpy[8] SDK for accessing and applying curated content in clinical and research workflows.

2.3. Standardizing exchange and dissemination of clinical-grade genomic knowledge

2.3.1. Overview of the ClinGen Genomic Knowledge Model and the Variant Annotation framework

Presented by: Larry Babb (Broad Institute of MIT and Harvard), Alex Wagner (Nationwide Children's Hospital and the Ohio State University)

Throughout our infrastructure ClinGen has an ongoing commitment to make genomic knowledge findable, accessible, interoperable and reusable (FAIR) and has devoted consistent data engineering resources over the past 6 years to deliver on that commitment. ClinGen is an ideal platform for evolving these genomic knowledge standards with its consortium comprised of several separate software engineering teams all dedicated to an integrated ecosystem for supporting the collection and curation of evidence, the standardization of variation and other fundamental related genomic concepts, and the dissemination of fully qualified evidence-based genomic knowledge from expert groups. We will be discussing the SEPIO framework, the ClinGen Genomic Knowledge Model, and the application of the Variant Annotation framework[9] that is the foundation for the ongoing standards work being done with the Global Alliance for Genomics and Health (GA4GH)[10,11] within the Genomic Knowledge Standards working group.

We will also examine the GA4GH Genomic Knowledge statement design for representing provenance-based evidence, the assessment of that evidence based on an associated method and the final classification of the knowledge being addressed. ClinGen is leveraging this design to

represent gene and variation based knowledge for Gene Validity, Dosage Sensitivity, Variant Pathogenicity and Clinical Actionability. We will walk through exercises related to Variant Pathogenicity and Therapeutic Response statements to illustrate challenges addressed by this framework and the benefits of standardized, clinical-grade, interoperable and reusable genomic knowledge content. We will then cover the application of this framework to the previously described variation curation platforms, and how it relates to downstream resources such as the Evidence Repository and LDH. A hands-on exercise will be presented for querying (and generating) compliant data with community-developed software tools.

2.3.2. Tools for variant registration and evidence association

Presented by: Kevin Riehle (Baylor College of Medicine)

This session will describe the ClinGen Allele Registry (CAR - https://reg.clinicalgenome.org)[12] which provides a canonicalization service resulting in >2.5B canonical allele identifiers (CA IDs) representing alleles that have equivalent representations across genome builds and transcripts. The Linked Data Hub (LDH: https://ldh.clinicalgenome.org), provides a structured environment that leverages excerpted data from external sources (e.g. molecular consequence, BRCA Exchange, CIViC, ClinVar, population allele frequency, etc.) with links to other core documents (e.g. variants, genes, etc.) that results in aggregation of knowledge for a given query. We will provide an overview and demonstration of the CAR and LDH as it relates to supporting curation efforts in ClinGen and how the functionality can be applied to other projects and consortia.

We will also showcase the incorporation of GA4GH-modeled ClinVar data into the LDH and how this process can be leveraged to support additional resources that maintain SEPIO and non-SEPIO structured documents. Combining the registration service (CAR) with supporting evidence (LDH) provides for downstream tool integration to support curation (e.g., Variant Curation Interface), deduplication, provenance, and other types of applications.

2.3.3. Tools for knowledge dissemination

Presented by: Tristan Nelson (Geisinger)

ClinGen has applied the models developed within the SEPIO Framework and GA4GH Variant Representation and Annotation standards to the variant assessments in ClinVar, as well as Gene Dosage and Gene Validity curations. Through our Genegraph service, we make available a form of ClinVar that represents submissions on a given variant by individual submitters (SCV), as this view of the data allows a fine-grained assessment of the professional assessments made regarding the clinical relevance of a variant, which can then be filtered based on several factors, including the purpose of the assessment and the reputation of the source. We represented the ClinGen Gene Dosage and Validity data in the same formats; demonstrating the utility and flexibility of these models in the context of diverse and highly clinically relevant datasets. We investigate some of the ways these datasets can be explored to produce clinical insights.

3. Conclusion

This workshop will introduce the methods and tools used to support the lifecycle of consuming, generating, and classifying clinical genomic knowledge. We will describe the Variant Curation

Expert Panel evaluation process for constitutional and somatic cancer variant curation, and how these data are disseminated for reuse and expert evaluation between systems through modern data normalization and community-driven data exchange standards.

4. Acknowledgments

National Human Genome Research Institute (NHGRI) awards supported AHW (R35HG011949) and KD, HR, MW, MM, LB, KR, TN (U24HG009649, U24HG006834, U24HG009650). KK is supported by the NIH National Cancer Institute award U24CA237719.

Bibliography

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
2. 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).
3. 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.* (2022) doi:10.1016/j.gim.2022.01.001..
4. Preston, C. G. *et al.* ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. *Genome Med.* **14**, 6 (2022).
5. Krysiak, K. *et al.* A community approach to the cancer-variant-interpretation bottleneck. *Nat Cancer* **3**, 522–525 (2022).
6. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
7. Danos, A. M. *et al.* Standard operating procedure for curation and clinical interpretation of variants in cancer. *Genome Med.* **11**, 76 (2019).
8. Wagner, A. H. *et al.* CIViCpy: A Python software development and analysis toolkit for the CIViC knowledgebase. *JCO Clin. Cancer Inform.* **4**, 245–253 (2020).
9. Brush, M. H., Shefchek, K. & Haendel, M. SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence. in *ICBO/BioCreative* (pdfs.semanticscholar.org, 2016).
10. Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**, 100029 (2021).
11. Wagner, A. H. *et al.* The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification. *Cell Genomics* **1**, 100027 (2021).
12. Pawliczek, P. *et al.* ClinGen Allele Registry links information about genetic variants. *Hum. Mutat.* **39**, 1690–1701 (2018).