STANDARDIZING AND DEMOCRATIZING ACCESS TO CANCER MOLECULAR DIAGNOSTIC TEST DATA FROM PATIENTS TO DRIVE TRANSLATIONAL RESEARCH

SUBHA MADHAVAN[1] Ph.D., DEBORAH RITTER[2] Ph.D., CHRISTINE MICHEEL[3] Ph.D., SHRUTI RAO[1] M.S., ANGSHUMOY ROY[2] M.D, Ph.D., DMITRIY SONKIN[4] Ph.D., MATTHEW MCCOY[1] Ph.D., MALACHI GRIFFITH[5] Ph.D., OBI L GRIFFITH[5] Ph.D., PETER MCGARVEY[1] Ph.D., SHASHIKANT KULKARNI[2] Ph.D. ON BEHALF OF THE CLINGEN SOMATIC WORKING GROUP

1 Innovation Center for Biomedical Informatics, Georgetown University, Washington D.C.; 2. Baylor College of Medicine and Texas Children's Hospital, Houston, TX.; 3. Vanderbilt University School of Medicine, Nashville, TN.; 4. National Cancer Institute, Rockville, MD.; 5. The McDonnell Genome Institute, Washington University, St. Louis, MO.

Abstract

In the last 3-5 years, there has been a rapid increase in clinical use of next generation sequencing (NGS) based cancer molecular diagnostic (MolDx) testing to develop better treatment plans with targeted therapies. To truly achieve precision oncology, it is critical to catalog cancer sequence variants from MolDx testing for their clinical relevance along with treatment information and patient outcomes, and to do so in a way that supports large-scale data aggregation and new hypothesis generation. Through the NIH-funded Clinical Genome Resource (ClinGen), in collaboration with NLM's ClinVar database and >50 academic and industry based cancer research organizations, a Minimal Variant Level Data (MVLD) framework to standardize reporting and interpretation of drug associated alterations was developed. Methodological and technology development to standardize and map MolDx data to the MVLD standard are presented here. Also described is a novel community engagement effort through disease-focused taskforces to provide usecases for technology development.

Introduction

ClinGen

To address these needs of capturing, standardizing and sharing clinically relevant variants, the Clinical Genome Resource, ClinGen [1] collaborative was established in 2012 and has been developing interconnected community resources to improve our understanding of genomic variation and enhance its use in clinical care. ClinGen represents a strong partnership among public, academic, and private institutions that relies on collaboration between the NIH and academic and commercial laboratories operating in both the research and clinical realms. ClinGen is also engaging numerous entities, including professional societies, to ensure that the resources that are produced meet community expectations. The Somatic working group (Somatic WG) is a clinical domain working group within the ClinGen

consortium and was established in 2015 to address standardization and sharing of cancer MolDx test results described here.

The Standard

In order to standardize the collection of clinically relevant somatic data, the Somatic WG of the ClinGen created a framework of consensus data elements titled "Minimum Variant Level Data (MVLD) [2]. MVLD was developed with input from multiple stakeholders ranging from database engineers to researchers and somatic clinical laboratory directors, as well as input from multiple current databases that collect cancer variant data. Briefly, MVLD consists of three sections: allele descriptive, allele interpretive and somatic interpretive. The allele descriptive section contains data elements that describe the genome position, gene, chromosome, genomic location, reference transcript and protein. The allele interpretive contains data elements describing the somatic classification (confirmed somatic, confirmed germline or unknown), the DNA and protein substitution, the variant type and consequence and PubMed identifiers associated with interpretation. The somatic interpretive section contains the most clinically relevant data, and is the section that required the most discussion and consensus-building among working group members. The somatic interpretive section contains a description of the cancer type (suggested ontologies such as NCI Thesaurus or Oncotree, and newly added Disease Ontology), the Biomarker Class (Diagnostic, Prognostic, Predictive), the Therapeutic Context (associated drugs), Effect (Resistant, Responsive, Not-Responsive, Sensitive, Reduced-Sensitivity), Level of Evidence (a tiered system similar to the recent AMP/CAP/ASCO guidelines [3] and Sub-Level of Evidence (reporting of trials, metadata analysis, preclinical data or inferential data). Readers are referred to the publication for a more detailed description of these data elements. Since the publication of MVLD, recent guidelines on somatic variant interpretation have been published through a joint effort of the Association for Molecular Pathology (AMP), College of American Pathologists (CAP) and the American Society of Clinical Oncology (ASCO)[3]. We intend to fully harmonize MVLD elements with these guidelines; mapping any specific criteria to the current version of MVLD and revising MVLD to accommodate new elements. There are distinct areas of agreement between MVLD and AMP/CAP/ASCO guidelines, such as using HUGO-approved nomenclature and HGVS formatting

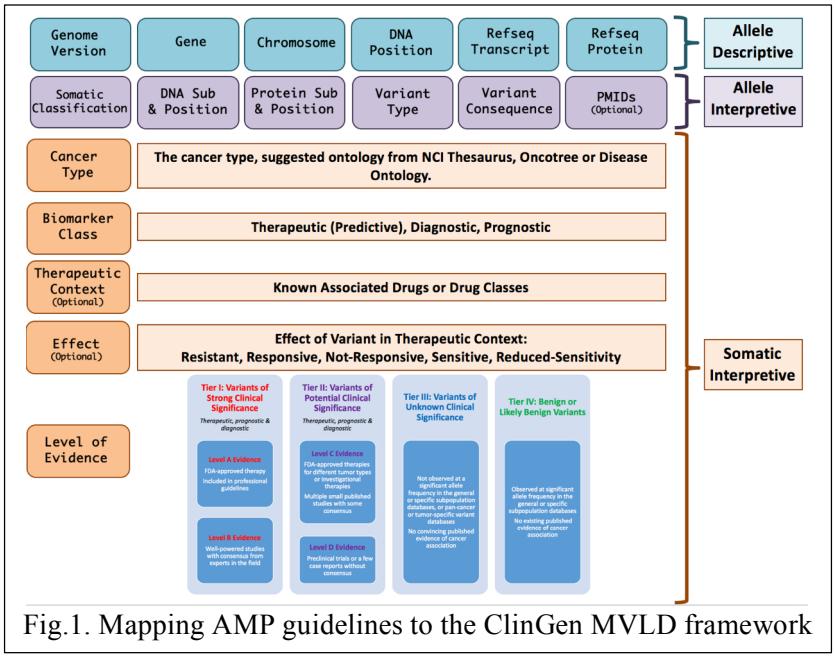

Fig.1. Mapping AMP guidelines to the ClinGen MVLD framework

for variants. However, there are also sizable and nuanced differences that need resolution to sync the guidelines with the MVLD data structure.

One area of immediate critical harmonization needed is in the Somatic Interpretive Level of Evidence and Sub-Level of Evidence in MVLD, which was drawn from the Cancer Driver Log (CanDL) [4]. The AMP/CAP/ASCO guidelines contain classification for uncertain (Tier III) and benign (Tier IV) variants, while MVLD was not initially designed to incorporate these types of variants. However, the necessity and relevance of uncertain or benign variants is apparent in that they too can aid clinical diagnosis. The AMP/CAP/ASCO guidelines Tier I Level A and MVLD Tier 1 are the same, but AMP/CAP/ASCO further provides Level B to sustain interpretations that derive from well-established studies that are not yet FDA or NCCN approved. Similarly, there are numerous nuanced differences between AMP/CAP/ASCO Tier II Level C and D and MVLD Tier 2, 3 and 4. The Sub-Level of Evidence in MVLD is incorporated in AMP/CAP/ASCO at various Tiers as well. Instead of partially modifying the MVLD Level of Evidence and Sub-Level of Evidence, we propose to absorb the Sub-Level of Evidence element into the Level of Evidence and to fully adopt the classification system proposed by AMP/CAP/ASCO into the Somatic Interpretive Level of Evidence shown in **Figure 1**.

Methods

Variant Curation SOP and expert review

Our variant curation and interpretation process leverages the strengths of ClinGen Somatic WG, the consortium of multi-disciplinary experts in somatic variants in cancers, CIViC [5], a cancer variant knowledgebase and crowdsourced curation system and ClinVar [6], an NCBI submission-driven database for variants. ClinGen brings/develops organized clinical and biomedical expertise, best practices and SOPs, CIViC provides a curation interface and interpretation portal, and ClinVar allows widespread dissemination of the expert-curated content and provides patient-level observations of variants in clinical settings back into ClinGen/CIViC.

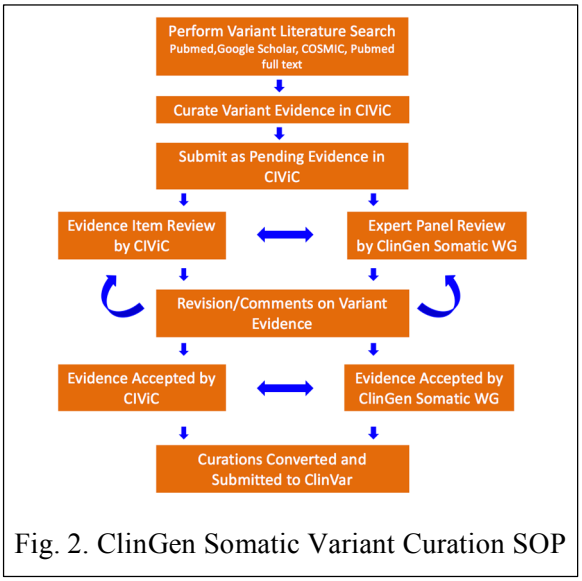

Fig. 2. ClinGen Somatic Variant Curation SOP

The ClinGen Somatic variant curation and expert review process (**Figure 2**). New submissions or revisions are made through data entry pages in CIViC that support dynamic form adjustments, live type-ahead suggestions, ontology look-ups, and warnings for merge conflicts.

Discussion pages track the complete history of comments and revisions. Curators and editors have the option to "follow" any entry (gene, variant, evidence) to receive notifications of comments, proposed changes or additions. Curators can also communicate with others in the CIViC community directly through site mentions, updates and messages. All curated entries can be "flagged" for problems or revisions can be proposed. Flagging allows for easy marking of content needing immediate review or

can flag entries which require more caution with use as diagnostic markers, while revisions are tracked and displayed with detailed GitHub-style diffs and comments. Curators can create detailed profiles so that their efforts are recognized by awarding badges for curation activity milestones to encourage and recognize participation. Curators can also join formal curation organizations within the CIViC community, for example the ClinGen Somatic Working Group exists as a CIViC organization, currently with 12 active members.

Results

Key methodological and technology development results in mapping MolDx data to MVLD are described below.

MolDx2MVLD mapping tools

To complement the crowdsourced expert variant curation process, members of the ClinGen WG are designing and implementing tools to support mapping of clinical MolDx data to standards and automated importation of this data into research databases, for example, G-DOC [7] (Georgetown Database of Cancer) and SEER [8] to drive new hypothesis generation for translational research. The primary goal of this tool is to enable broad sharing of de-identified MolDx data from clinical laboratories for novel hypothesis generation and evidence collection for clinical actionability.

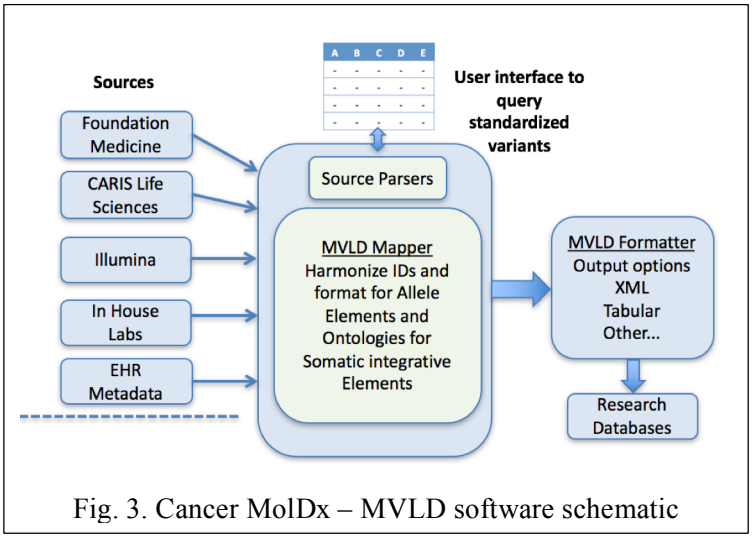

Fig. 3. Cancer MolDx – MVLD software schematic

The general components of the tool to map MolDx data to the MVLD system are outlined in **Figure 3**. The four main components of the tool are: 1) ETL (Extract, Transform, Load) tools to parse individual sources, extract and format information required for MVLD descriptive and interpretive elements; 2) MVLD Mapper to map the extracted information to the MVLD standard, harmonize the elements to standard identifiers and ontologies used in public data repositories, identify missing data elements, and attempt to fill in missing values if possible; 3) a simple QA/QC interface for checking results and correcting or adding missing values; and 4) MVLD formatter to output the information in various formats, e.g., xml, tabular, or others to interface with EHRs or other databases depending on user needs.

ETL tools

MolDx laboratory sources use different internal formats and standards to identify a variant and describe the results of their tests. For example, some labs use only gene names and protein changes as a

description; others provide transcript identifiers and the specific DNA change in the transcript and/or the exact chromosomal location. To date we have not seen a report using a complete HGVS (Human Genome Variation Society) formatted sequenceID+variation name, though some labs provide the information to create the description. There are also differences in how data is distributed, with some labs providing results in tabular formats while others provide XML. We will create parsers and logic for each source to extract and perform initial transformations. **Figure 4** shows an example of an XML excerpt from the Foundation Medicine report on a patient with breast carcinoma. This patient's molecular testing identified a variant E545K in the PIK3CA gene with potential benefit from mTOR inhibitors such as Everolimus or Temsirolimus. The figure shows mapping of XML output from the lab to elements in the MVLD standard. Similar mapping is being conducted for all 18 data elements in MVLD to various commercial labs. Lab formats to integrate are prioritized by our stakeholder community based on most widely used labs.

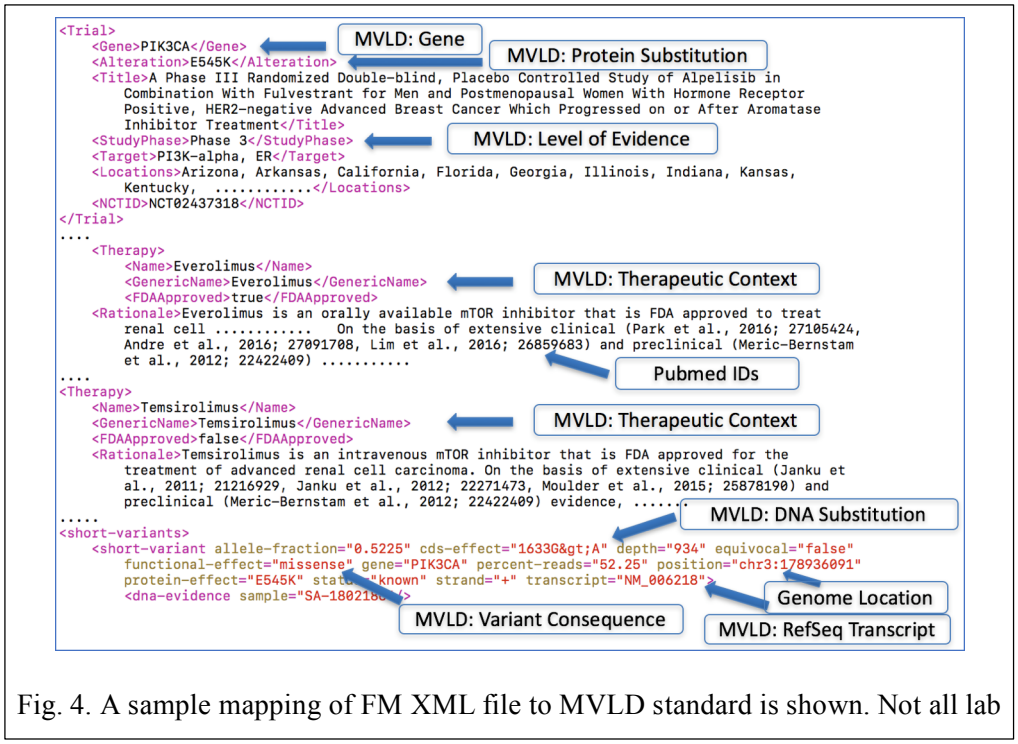

Fig. 4. A sample mapping of FM XML file to MVLD standard is shown. Not all lab

MVLD mapper

This is the core component of the system. We expect variations in how the descriptive and interpretive properties in MVLD will be expressed, and the mapper attempts to harmonize these representations to the most informative common representation using variant-specific APIs connected to international variation databases and ontology servers. For allele properties, we use NCBI's E-utilities including their new variation API services that allow users to compare and return all equivalent alleles using multiple NCBI identifiers including a canonical identifier [9] . The tool first checks the ClinVar database to see if a representation of the variation exists or the test is registered with the Genetic Testing Registry, in which case ClinVar may contain all the variants tested [10]. ClinGen has released an Allele Registry with APIs that also attempts to link equivalent variant alleles to a canonical representation but is not NCBI centric in that ENSEMBL IDs and even EXAC alleles are supported and novel alleles can be submitted [11].

MVLD Interpretive elements like cancer type can be standardized using APIs for terminology servers. The WG identifies terminology standards used in key data fields in lab reports, such as disease and drug names, ICD codes, or the 10-digit national drug code and map them to MVLD-recommended standards. If any term is unknown, the MolDx processor attempts to automatically map to terminologies in the NCI Thesaurus (NCIt) using LexEVS Terminology Server APIs or BioPortal APIs [12,13]. In all cases, the original value is stored along with the selected mappings to defined terminologies. This allows the data to be in a uniform format that follows standards and ontologies, while allowing for more integrated search functions within systems like ClinVar.

MVLD Formatter

The formatter module provides multiple output options for target research databases. The Initial output will be delimited tables or XML mainly for consumption by institutional databases (EHRs) that want to store the MVLD standardized data. We will work with the community to define and build in additional XML or other formats (e.g. JSON) from labs.

Discussion and Future Direction

Community Engagement

Many standards and technology frameworks fail due to lack of community engagement and adoption. To avoid this, the ClinGen clinical domain working group engaged both strategic leaders and tactical implementers from over 60 cancer centers, industry partners and federal agencies. These include active participation from organizations such as Georgetown University Lombardi Cancer Center, Baylor College of Medicine, Vanderbilt University Medical Center, Washington University School of Medicine, Moffitt Cancer Center, Illumina, Molecular Match, NCI, NHGRI and FDA. An initial survey of participating organizations identified major challenges in somatic variant assessment, clinical interpretation pipelines and open tools for variant curation and expert review. The survey results also indicated the use of a tiered system of variants for clinical actionability (FDA-approved/NCC guidelines, clinical trials data, pre-clinical data, mechanistic/pathway level evidence). These results motivated the efforts to develop the MVLD to help standardize how clinical labs report MolDx data to patients, clinicians and regulatory agencies. We also engaged members from AMP (Association of Molecular Pathologists) CAP (College of American Pathologists) somatic practice guideline committees to help drive adoption of MVLD within their professional societies and members. ClinGen Somatic WG is also actively working with Global Alliance for Genomic Health (GA4GH)'s Variant Interpretation for Cancer Consortium (VICC). The VICC seeks to integrate global efforts for the clinical interpretation of cancer variants. The ClinGen Somatic WG engages various experts in the cancer research and care communities through taskforces. These taskforces are self-organized expert groups in a particular cancer type, gene or a pathway. Three such taskforces have been launched during Summer of 2017 and are focused on Pediatric, Pancreatic and non-small cell lung cancer somatic testing with other taskforces being routinely formed. The taskforces are charged with prioritizing variants for curation, using the

technology framework described to standardize and share somatic variants, related clinical evidence and provide direction for new software features.

Data Harmonization

A major goal of the ClinGen Somatic Cancer Working Group is to review and harmonize existing guidelines and guideline efforts related to curation, interpretation, and reporting of somatic alterations in cancer. The working group has formed a task force that will bring together representatives from ClinGen, ACMG, AMP, and CAP, among other relevant organizations. This harmonization task force will seek to build upon recently published work such as the AMP/ASCO/CAP guideline (https://www.ncbi.nlm.nih.gov/pubmed/27993330) and the ClinGen Somatic Cancer Working Group's minimum variant-level data for curation of somatic alterations in cancer (https://www.ncbi.nlm.nih.gov/pubmed/27814769). Further, this task force will work to make its guideline compatible with other related guidelines, such as the ACMG/AMP guideline for interpretation of germline variants (https://www.ncbi.nlm.nih.gov/pubmed/25741868), and an ongoing effort within ACMG on the interpretation of copy number variants in neoplastic diseases. This community-driven task force will review and include work such as My Cancer Genome's efforts to describe and standardize curation practices in the somatic cancer space. My Cancer Genome's curation framework is being developed based on the evidence-based framework recently published by ClinGen for gene-disease relationships in Mendelian disorders (https://www.ncbi.nlm.nih.gov/pubmed/28552198). Finally, the task force will review and include efforts of somatic cancer knowledgebases to map terminologies and levels of evidence schemes across knowledgebases (e.g., CIViC, OncoKB, PMKB, JAX-CKB, CGI, PCT, CanDL, etc.).

REFERENCES

1. Rehm HL, Berg JS, Brooks LD, et al: ClinGen--the Clinical Genome Resource. N Engl J Med 372:2235-42, 2015
2. Ritter DI RS, Roy A, Rao S, Landrum MJ, Sonkin D, Shekar M, Davis CF, Hart R, Micheel C, Weaver M, Allen EV, Parsons DW, McLeod HL, Watson MS, Plon SE, Kulkarni S, Madhavan S: Somatic Cancer Variant Curation and Harmonization through Consensus Minimum Variant Level Data. Genome Medicine, 2016
3. Li MM, Datto M, Duncavage EJ, 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
4. Damodaran S, Miya J, Kautto E, et al: Cancer Driver Log (CanDL): Catalog of Potentially Actionable Cancer Mutations. J Mol Diagn 17:554-9, 2015
5. Griffith M, Spies NC, Krysiak K, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 49:170-174, 2017
6. Landrum MJ, Lee JM, Benson M, et al: ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862-8, 2016
7. Bhuvaneshwar K, Belouali A, Singh V, et al: G-DOC Plus - an integrative bioinformatics platform for precision medicine. BMC Bioinformatics 17:193, 2016

8. Altekruse SF, Rosenfeld GE, Carrick DM, et al: SEER cancer registry biospecimen research: yesterday and tomorrow. Cancer Epidemiol Biomarkers Prev 23:2681-7, 2014
9. NCBI. Services for variation data processing, 2017
10. NCBI. New Web Services for Comparing and Grouping Sequence Variants, 2017
11. Patel RY, Shah N, Jackson AR, et al: ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. Genome Med 9:3, 2017
12. Noy NF, Shah NH, Whetzel PL, et al: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 37:W170-3, 2009
13. Wiki, N. LexEVS Servers and APIs Summary, 2014