original reports abstract

My Cancer Genome: Coevolution of Precision Oncology and a Molecular Oncology Knowledgebase

Marilyn E. Holt, PhD[1]; Kathleen F. Mittendorf, PhD[1]; Michele LeNoue-Newton, PhD[1]; Neha M. Jain, PhD[1]; Ingrid Anderson, PhD[1]; Christine M. Lovly, PhD[1]; Travis Osterman, MS[1]; Christine Micheel, PhD[1]; and Mia Levy, PhD[2]

PURPOSE The My Cancer Genome (MCG) knowledgebase and resulting website were launched in 2011 with the purpose of guiding clinicians in the application of genomic testing results for treatment of patients with cancer. Both knowledgebase and website were originally developed using a wiki-style approach that relied on manual evidence curation and synthesis of that evidence into cancer-related biomarker, disease, and pathway pages on the website that summarized the literature for a clinical audience. This approach required significant time investment for each page, which limited website scalability as the field advanced. To address this challenge, we designed and used an assertion-based data model that allows the knowledgebase and website to expand with the field of precision oncology.

METHODS Assertions, or computationally accessible cause and effect statements, are both manually curated from primary sources and imported from external databases and stored in a knowledge management system. To generate pages for the MCG website, reusable templates transform assertions into reconfigurable text and visualizations that form the building blocks for automatically updating disease, biomarker, drug, and clinical trial pages.

RESULTS Combining text and graph templates with assertions in our knowledgebase allows generation of web pages that automatically update with our knowledgebase. Automated page generation empowers rapid scaling of the website as assertions with new biomarkers and drugs are added to the knowledgebase. This process has generated more than 9,100 clinical trial pages, 18,100 gene and alteration pages, 900 disease pages, and 2,700 drug pages to date.

CONCLUSION Leveraging both computational and manual curation processes in combination with reusable templates empowers automation and scalability for both the MCG knowledgebase and MCG website.

JCO Clin Cancer Inform 5:995-1004. © 2021 by American Society of Clinical Oncology

INTRODUCTION

Identification of genomic abnormalities is critical for diagnosis and treatment of many cancers.[1,2] As new biomarkers are identified and molecular testing panels expand, evaluating the clinical significance of variants becomes progressively more difficult. Current guidelines recommend patients be managed on clinical trials,[3] many of which include molecularly targeted interventions and biomarker-based inclusion criteria.[4-6] Recognizing the rapid expansion of precision oncology, in collaboration with clinicians and experts, we developed My Cancer Genome (MCG),[7] the first resource to provide publicly available, detailed information on cancer-related genetic variants geared toward clinicians.[8-11] MCG has distinguished itself through its inclusion of clinical trials and organization of content along cancer-relevant cell-signaling pathways.[9,12-14] However, the site is also used by researchers, pathologists, and diagnostic labs developing interpretive reports, as well as patients and caregivers.

When MCG was launched in 2011, the precision oncology field was still narrow enough to manually curate actionable molecular variants in cancer and their related drugs. As the number and types of clinically actionable variants expanded, timely curation of these variants required a more scalable approach. Although the website was originally hard-coded and required developer time to implement all changes, we quickly transitioned first to a Microsoft SharePoint–enabled content management system and then to a custom-developed replacement for SharePoint. However, as these content management systems used a wiki style that enabled contributions by subject matter experts around the world, this approach still did not offer the scalability required to match the field. To transform the website into an appropriately scalable resource, we developed and adopted an approach to

ASSOCIATED CONTENT

Appendix

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on August 26, 2021 and published at ascopubs.org/journal/cci on September 23, 2021: DOI https://doi.org/10.1200/CCI.21.00084

CONTEXT

Key Objective

To develop a scalable precision oncology knowledgebase that expands with the field of precision oncology.

Knowledge Generated

Assertion models for structured, manually curated diagnostic, therapeutic, and prognostic assertions and reconfigurable templates for generating natural language statements and visualizations with these and other structured assertions are present within the My Cancer Genome (MCG) knowledgebase. Combining these assertions and templates allowed rapid scaling of the website to more than 30,000 pages in an automated manner.

Relevance

Previous versions of the MCG website used a wiki-style model in which site size was limited by curator time. Updating the website to an assertion-based data model enables automated scalability, which reduces the curator time required to generate MCG content and allows clinicians and other site users to access content related to a much broader range of cancer-related variants and disease types.

biomarker, disease, drug, and trial curation on the basis of assertions, defined here as cause-and-effect statements supported by a source (such as professional guidelines or the primary literature) that can be computationally queried.

In a previous paper,[15] we described an assertion model for clinical trials that organize clinical trial biomarker eligibility criteria, arms, and interventions as a set of structured assertions within a custom-developed knowledge management system (KMS). Here, we define our assertion models for therapeutic, prognostic, and diagnostic biomarkers and describe how we leveraged these assertions in combination with publicly available data sets to generate content that is structured to promote reusability, reconfiguration, and computation. Finally, we describe how our dynamic template model transforms these assertions into natural language statements (ie, human-readable statements) and display them as text and visualizations on My Cancer Genome website,[7] in a modular, scalable manner. The assertion data model described herein provides a crucial public resource that is both human- and machine-readable. Our model also enables integration of MCG assertions with applications beyond the website, such as landscape analyses, interpretive reports for pathologists, structured integration with electronic health records, therapeutic landscape analyses, clinical trial planning, and molecular tumor board reports.

METHODS

Developing an Assertion-Based Approach to Page Generation

The first iterations of the MCG website required manual generation of each page on the website, which facilitated individualization of each page but was time-intensive for contributors, hindering scalability with the expanding precision oncology field. To promote scalability, we developed a data model on the basis of assertions that can be computationally queried. To build the information architecture to support assertion generation (Appendix Fig A1), several structured and semistructured knowledge resources were imported into a web-based KMS and processed to support automated and manual assertions. For example, functional assertions were supported by processing data from dbNSFP, UniProt, and the Universal Transcript Archive.[16-19] The prevalence of genomic alterations for specific cancer types is calculated from data imports from AACR Project Genomics Evidence Neoplasia Information Exchange (GENIE).[20] The KMS imports clinical trials documents from ClinicalTrials.gov daily to facilitate manual clinical trial assertion generation.[15] Alteration, disease, and drug ontologies were manually synthesized within the KMS and updated using data from publicly available ontologies, hierarchies, and nomenclature systems (Table 1).

Classes of Assertions in the MCG Knowledgebase

Assertions in the MCG knowledgebase can be grouped into two classes: (1) manually curated and (2) calculated. Manually curated assertions are extracted from documents with unstructured data by trained curators using the concepts in the KMS ontologies (Table 1). In the context of the MCG website, manually curated structured assertions may be displayed either individually or as part of a set of calculated assertions. Calculated assertions are computationally generated from both manually curated structured assertions and structured data from external sources to create functional and prevalence data for biomarkers, diseases, and clinical trial eligibility criteria and interventions (Table 2). This assertion-based approach empowers computational querying and data use in downstream applications, such as the MCG website.

Therapeutic assertions. Therapeutic assertions in the MCG knowledgebase connect biomarker status to predicted drug response in the context of specific diseases and therapies (Fig 1A). Most therapeutic assertions in the MCG knowledgebase are manually curated from US Food and Drug Administration labels and National Comprehensive Cancer Network guidelines; however, our model also allows curation from other sources, including guidelines from

TABLE 1. Ontologies Used to Generate Assertions for the My Cancer Genome Knowledgebase

Ontology	Description	Input(s)
Alterations	Organized by gene and alteration type Alteration types include mutations, copy number and protein expression changes, and cytogenetic abnormalities Associated with their chromosomal locations, genes, exons, and coding regions for protein domains	Manually curated from source documents using HGVS and ISCN nomenclature and validated against the UTA when possible
Alteration groups	Sets of alterations collected because of either the commonality of use or to facilitate nesting logic Relationship between alterations within an alteration group indicated using ALL/ANY/NONE logic operators May contain both alterations and other alteration groups	Manually curated from professional guidelines and published literature
Diseases	Malignant and benign neoplastic diagnoses organized in parent-child relationships that generally align with the NCIt Annotated with disease description, synonyms, and codes within common nomenclature systems	DiseaseOntology.org, NCIt, OncoTree, WHO Classification of Tumors, SNOMED CT, and UMLS
Genes	Originally imported from RefSeq and now maintained against UTA	RefSeq and UTA
Drugs	Cancer-related drugs relevant to assertions May be annotated with drug description, gene targets, drug class, parent-child relationships, synonyms, and drug codes within nomenclature systems	NCIt and clinical trial documents
Drug groups	Sets of drugs connected by the ANY operator	Manually curated from clinical trial documents

NOTE. Ontologies were manually maintained in the knowledge management system and updated using data from publicly available ontologies, hierarchies, and nomenclature systems described in the Input(s) column.

Abbreviations: HGVS, Human Genome Variation Society; ISCN, International System for Human Cytogenetic Nomenclature; NCIt, NCI Thesaurus; RefSeq, NCBI Reference Sequence Database; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms; UMLS, Universal Medical Language System; UTA, Universal Transcript Archive.

ASCO, National Institute for Health Care and Excellence, the Scottish Medicines Consortium, and the British National Formulary, as well as peer-reviewed clinical trial literature. These assertions are published both alone and in combination with other assertions as text on the MCG website. Additionally, the computationally queryable nature of these structured therapeutic assertions allows easy visualization and analysis for potential applications beyond the website, such as inclusion in next-generation sequencing interpretive reports, molecular tumor board reports, the electronic health record, therapeutic landscape analyses, and clinical trial planning.

Prognostic and diagnostic assertions. Our data model allows manual curation within the KMS of structured prognostic and diagnostic assertions from professional guidelines and the primary literature. Prognostic assertions indicate the effect of the biomarker on the prognosis of a specific disease (Fig 1B). Diagnostic assertions identify biomarkers that are considered characteristic or diagnostic of a particular cancer type (Fig 1C). Rule definition for our diagnostic assertion model is ongoing; our current diagnostic assertion data model focuses on hematologic malignancies because of the complexity of data modeling requirements for solid tumor malignancies. Prognostic and diagnostic assertions currently in the KMS are derived almost exclusively from the National Comprehensive Cancer Network guidelines and WHO classifications; we are currently integrating these assertions with the website.

Trial eligibility groups, treatment contexts, and cohorts. These assertion types have been extensively discussed in our previous publication.[15] Briefly, unstructured eligibility criteria and treatment arm information for cancer-related trials on ClinicalTrials.gov and other clinical trial sources (Appendix Fig A1) are imported into the KMS. Structured eligibility groups, treatment contexts, and cohort linkages are manually constructed from these documents. Eligibility groups consist of biomarker, disease, and disease context (eg, primary, locally advanced, metastatic, etc) information, whereas treatment contexts contain therapy and therapeutic context (eg, neoadjuvant, adjuvant, first-line, etc) information. Eligibility groups are connected to relevant treatment contexts via cohort linkages, which indicate the patient population eligible for each trial arm. When updated clinical trial documents are imported for previously curated trials with a recruiting or not yet recruiting status on ClinicalTrials.Gov, the new documents are manually reviewed and assertions are updated as necessary. Only clinical trials with these statuses are actively curated; however, we do not eliminate trials from our data set when enrollment closes, so curations of closed trials are also available on the website. To reduce the number of false positives, website users can filter the clinical trial search results by recruiting status.

Clinical trial biomarker, disease, and treatment trends. Biomarker, disease, and treatment prevalence in clinical trials are calculated from manually curated eligibility criteria, treatment arm, and cohort assertions in the KMS.

TABLE 2. Classes of Assertions in the My Cancer Genome Knowledgebase

Assertion Class	Description	Assertion Source	Data Source(s)	Example	Count
Therapeutic assertions	Statement of the therapeutic efficacy (or lack of efficacy) of a treatment in a disease with biomarker features	Manually curated	FDA, NCCN, NICE, and the primary literature	Non–small cell lung cancers with an EGFR T790M mutation may be sensitive to osimertinib (FDA and NCCN)	364
Prognostic assertions	Statement of prognostic effect of a specific biomarker when observed in a specific disease	Manually curated	NCCN guidelines and the primary literature	FLT3 ITD mutations confer an unfavorable prognostic effect in acute myeloid leukemia	164
Diagnostic assertions	Statement of diseases associated with molecular alterations	Manually curated	WHO Classification of Tumours	Observation of the PML-RARA fusion in patients with myeloid neoplasms may indicate a diagnosis of acute promyelocytic leukemia	150
Trial eligibility groups	Statement of disease and biomarker eligibility rules for an interventional cancer trial	Manually curated	Clinical trial documents	Patients with HER2-positive breast cancer may be eligible to enroll on clinical trial (NCT03821233)	11,488
Trial treatment contexts	Statement of a treatment regimen used in an interventional cancer trial	Manually curated	Clinical trial documents	Patients may be treated with ZW49 on clinical trial (NCT03821233)	13,031
Trial cohorts	Statement of disease and biomarker eligibility groups that are eligible for a specific treatment regimen in an interventional cancer trial	Manually curated	Clinical trial documents	Patients with HER2-positive breast cancer may be treated with ZW49 on clinical trial (NCT03821233)	31,596
Functional assertions	Statement of alteration effect on protein function	Imported from external data sources	UTA, UniProt, and dbNSFP	EGFR L858R is predicted to be deleterious (SIFT) or pathogenic (ClinVar)	1,239
Frequency assertions	Statement of the frequency with which a specific biomarker is observed in a specific disease	Calculated from external database	AACR Project GENIE	EGFR is altered in 19.77% of non–small-cell lung carcinoma patients with EGFR L858R present in 5.43% of all patients with non–small cell lung carcinoma (AACR Project GENIE)	525,656

NOTE. Assertions in the MCG knowledgebase are manually curated and computationally calculated from the unstructured and structured sources described in the Data Source(s) column. Although prognostic and diagnostic assertions are currently curated as part of the MCG knowledgebase, they are not yet publicly available on the MCG website.

Abbreviations: dbNSFP, database for nonsynonymous SNPs' functional predictions; EGFR, epidermal growth factor receptor; FDA: US Food and Drug Administration; HER2, human epidermal growth factor receptor 2; ITD, internal tandem duplication; MCG, My Cancer Genome; NCCN: National Comprehensive Cancer Network; NICE: The National Institute for Health and Care Excellence; SIFT: Sorting Intolerant From Tolerant; UTA, Universal Transcript Archive.

MCG biomarker, disease, and drug pages provide graphical displays (Fig 2 and Data Supplement), enabling rapid assessment of distribution in clinical trials.

Functional assertions. Functional assertions in the KMS predict a genetic variant's effect on protein sequence, structure, and function. Content for functional assertions is directly imported from dbNSFP, UniProt, and the Universal Transcript Archive.[16-19] These assertions are displayed in the Overview section of MCG website pages for variants for which functional assertions exist in the KMS.

Frequency assertions. Frequency assertions describe the prevalence of a particular genetic variant in specific cancer diagnoses. Frequency assertions are calculated from prevalence data in the AACR Project GENIE public data set,[20] which contains clinically annotated next-generation sequencing data for more than 112,000 cancer samples. The frequency distribution of top diagnoses and gene alterations (divided into genes and specific alterations) is displayed on Biomarker pages in the MCG site. The frequency distribution of genes most frequently altered and genetic variants most frequently observed in sequencing reports for that diagnosis is displayed on MCG Disease pages.

Reconfigurable Templates for Information Transformation and Visualization

Utilization of these structured assertions for MCG and other downstream applications presented a critical challenge: how do we present this information in a manner that is both automated and readable? The MCG knowledgebase transforms machine-readable, automated assertions into human-readable natural language statements and visualizations on My Cancer Genome website.[7] This process relies on a series of reusable, reconfigurable templates that are populated by assertions in the knowledgebase and automatically updated as assertions are added or updated in the knowledgebase. Each template is composed of either a natural language statement or graph that queries the structured assertion set (Fig 2). Natural language statements transform these assertions into sentences that can be connected into paragraphs of text, whereas graphs visualize the distribution of the assertion type of interest.

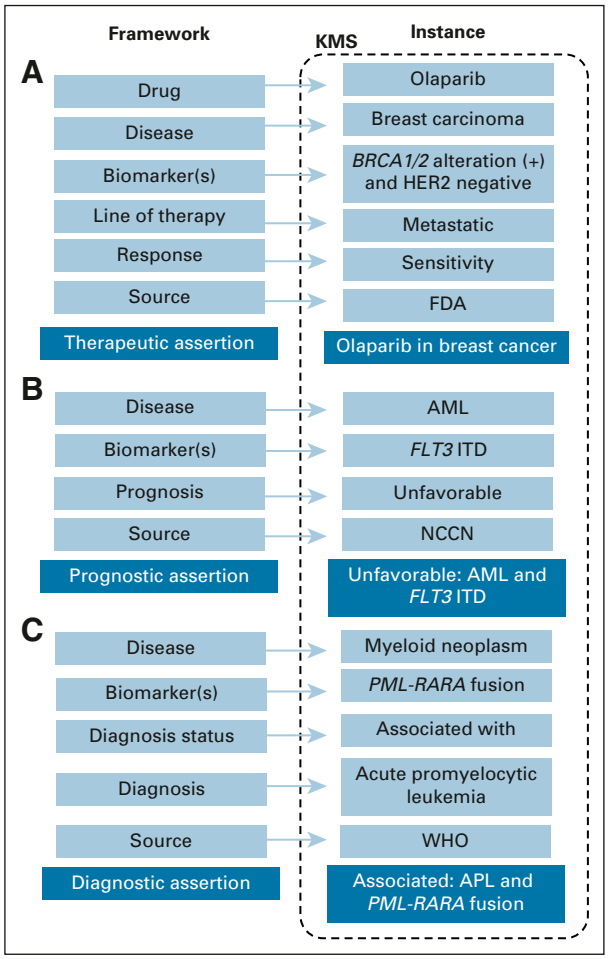

FIG 1. Therapeutic, prognostic, and diagnostic assertion curation workflows. A curator generates structured assertions from assertion sources. These assertions are constructed in the KMS using the (A) therapeutic, (B) prognostic, or (C) diagnostic assertion data models and are reported using natural language statements and visualizations in the My Cancer Genome website.[7] AML, acute myeloid leukemia; APL, acute promyelocytic leukemia; FDA, US Food and Drug Administration; KMS, knowledge management system; NCCN, National Comprehensive Cancer Network.

These modular templates can then be arranged in subject-specific configurations to generate web pages in an automated manner for biomarkers, diseases, and drugs with associated assertions.

RESULTS

In the original website, the MCG team manually curated and maintained content for 1,338 genes and alterations, 25 diseases, 697 drugs, and 19 pathways, in addition to a clinical trial search feature. Leveraging computational and manual curation processes in combination with dynamic templates allowed us to scale the website to more than 18,100 gene and alteration pages, 900 disease pages, 2,700 drug pages, and 9,100 clinical trial pages (approximately 7,400 open trials and 1,700 closed trials; Fig 3). Web pages on My Cancer Genome website[7] are updated with new data from many classes of assertions in the KMS every two weeks, as depicted in Table 3 and described in detail below.

Biomarker Pages

Individual biomarker pages are generated for all genetic variant and expression biomarkers (as well as selected immuno-oncologic and viral markers) that are associated with a curated clinical trial or therapeutic assertion and all alterations observed in five or more GENIE cases (Data Supplement; ALK fusion archive).[21] As prognostic and diagnostic assertions are added, biomarker and disease pages (see below) are added for all biomarkers and diseases associated with those assertions. For each biomarker with relevant functional and frequency assertions, an Overview section displays these assertions using natural language statements and visualizations. A Biomarker-Directed Therapies section highlights all therapeutic sensitivity and resistance assertions affiliated with that biomarker. Users can navigate directly from the therapeutic assertions to a list of related drug pages (see below) that provide further drug information. A Clinical Trials section displays graphs of the landscape of clinical trials with the biomarker as an inclusion criterion, and users can navigate directly to the list of clinical trials in these graphs. A Significance section at the bottom of the page details the biomarker, therapeutic assertion, and clinical trial prevalence for the biomarker on a disease-by-disease basis.

Disease Pages

Disease pages are generated for all diagnoses curated in clinical trials or therapeutic assertions (Data Supplement; Solid Tumor-archive).[22] Each disease page contains an Overview section that includes a definition from the NCI Thesaurus (NCIt) and, when applicable, graphs of frequency assertions for alterations observed in the disease in the GENIE data set. Therapeutic assertions curated with the disease are also displayed, as well as graphs of the alteration-based inclusion criteria and therapies curated on clinical trials associated with the disease. As with the biomarker pages, users can navigate directly from the disease page to a list of associated drug and clinical trial pages (see below). A Significance section at the bottom of the page details the gene alteration, therapeutic assertion, and clinical trial prevalence for that diagnosis on a gene-by-gene basis.

Drug Pages

Drug pages are generated for all drugs curated on clinical trials or therapeutic assertions (Data Supplement; TinyURL: Pembrolizumab-archive).[23] An Overview section displays key synonyms and the NCIt definition when available, and a Biomarker-Directed Therapies section displays related biomarker-based therapeutic assertions. A Clinical Trials section displays biomarker and disease inclusion criteria for trials in which patients receive the drug. Users can navigate directly from that section to a list of associated clinical trial pages (see below). A Drug Details

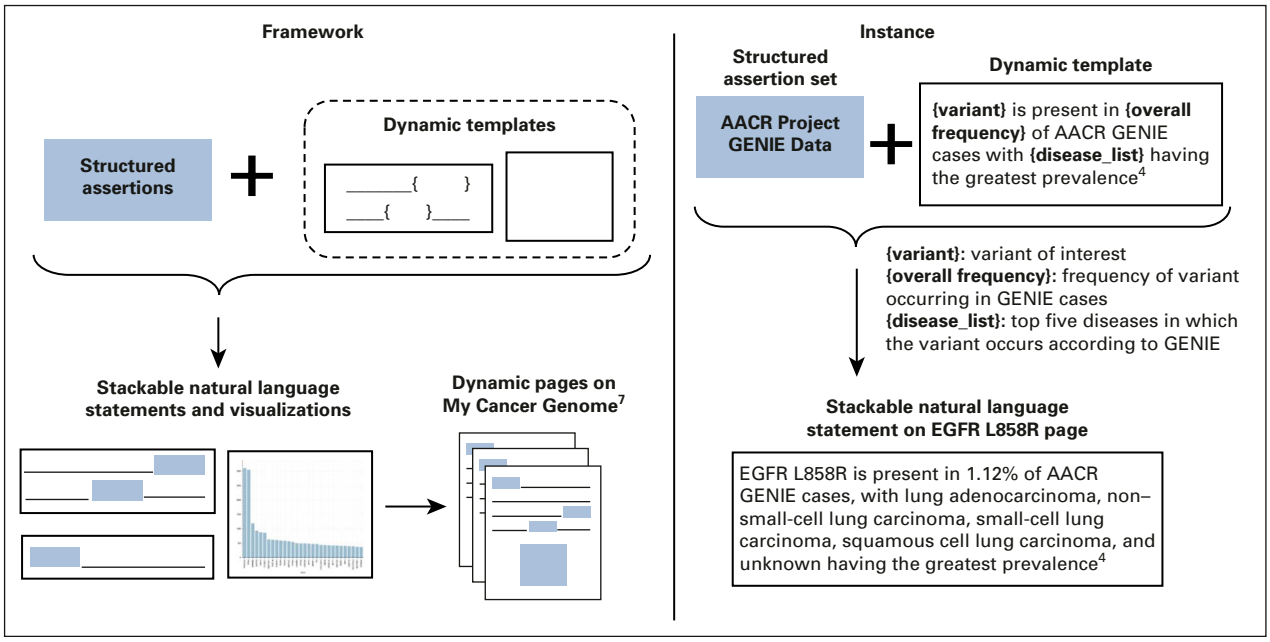

FIG 2. Automatically updated templates power dynamic page generation. Templates are content building blocks that can be stacked to generate pages that automatically update as assertions are curated and edited. Empty brackets and boxes denote regions in template statements into which assertions can be imported to create natural language statements describing a diversity of biomarker, disease, and therapeutic assertions. Blue boxes and bars indicate assertions that have been combined with templates to create natural language statements and visualizations. GENIE, Genomics Evidence Neoplasia Information Exchange.

section at the bottom of the page lists all synonyms, drug categories, gene targets, and NCIt, Unified Medical Language System, or Systematized Nomenclature of Medicine (SNOMED) codes curated for the drug.

Clinical Trial Pages

Clinical trial pages are generated for all clinical trials with curated disease and treatment information in the KMS (Data Supplement; TinyURL: NCT-archive).[24] For each trial, the eligibility criteria, treatment context, and cohort assertions are displayed at the top of the page, followed by selected portions of the clinical trial document from ClinicalTrials.Gov. This allows users to rapidly identify key biomarker- and diagnosis-related eligibility criteria and therapies available through the trial. New clinical trial documents and assertions are imported from the KMS every 2 weeks.

Pathway Pages

Pathway pages on the original version of the MCG website were developed for key cancer-related pathways as previously described.[9] Pathway pages on the updated MCG website retain the original MCG website structure. Development of dynamic templates and visualization from structured assertions in the KMS is ongoing for these pages.

DISCUSSION

Since its inception in 2011, the mission of the MCG knowledgebase has been to provide publicly available, clinician-accessible information regarding the clinical actionability of variants observed in cancer.[9-11] To parallel the accelerating pace of precision oncology discovery, we developed an assertion-based data model in combination with natural language and graph templates to enable dynamic, up-to-date generation of biomarker, disease, drug, and trial pages for the MCG website. This data model uses calculated and manually curated assertions derived from structured and unstructured data sources. Automated templates transform assertions into stackable natural language statements and visualizations. These templates can be combined and iterated to generate thousands of pages of dynamic content about cancer-related biomarkers, diseases, and drugs on the MCG website (Fig 3). Leveraging both computational and manual curation processes in combination with reusable templates frees human resources to focus on adding new content to the KMS, which can then be automatically updated and reused across the website. This approach allows automated direction of an enormous quantity of information into more than 30,000 pages that target the diverse needs of the clinicians, researchers, and other entities interpreting the clinical significance of genetic variants in a much more efficient manner than the original model. These 30,000 pages of content were generated in only 18 months; by contrast, using the original model, we generated and maintained content for only 1,338 genes and alterations and 25 diseases over a period of eight years. Per Google Analytics, the improved search functionalities on the current MCG website are much more heavily used than the

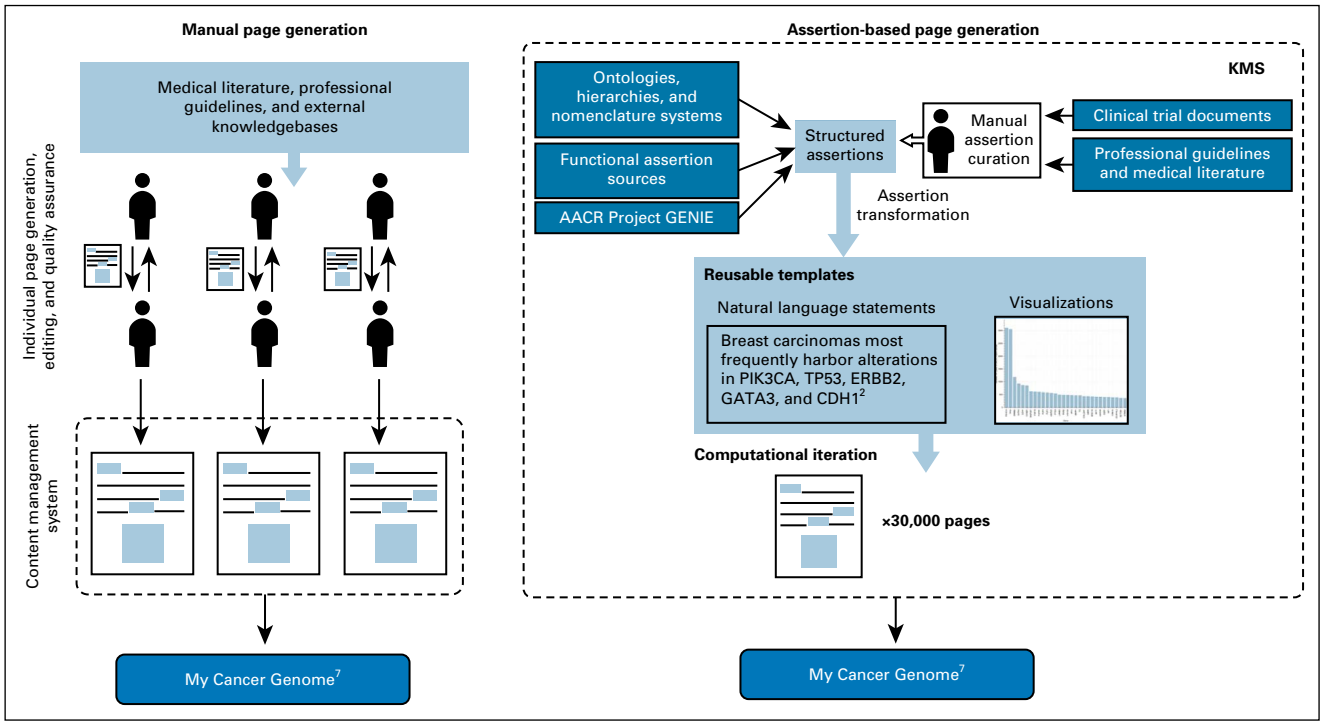

FIG 3. Manually generated pages versus dynamic templates. The model for the original My Cancer Genome website (left) required each page on the website to be manually written and updated, which was an extremely time-intensive process. Changing to an assertion-based model (right) allowed pre-existing structured data sets, ontologies, hierarchies, and nomenclature systems to be leveraged in combination with manually curated assertions from unstructured data sources to generate reusable, reconfigurable natural language statements and visualizations. These templates are combined and iterated to computationally generate autoupdating website pages in a scalable manner.

rudimentary search in the original MCG. Overall visits to site landing pages have remained stable across the transition from the original MCG to the current version (Data Supplement).

Although the amount of curator hours spent writing, editing, and formatting individual pages for My Cancer Genome website[7] has dramatically decreased, considerable curator time is still required to generate clinical trial, therapeutic, prognostic, and diagnostic assertions because all these assertions require information abstraction from unstructured data sources. For example, generation of structured eligibility criteria, treatment contexts, and cohorts from clinical trial documents can take anywhere from 10 minutes to over an hour of curator time per clinical trial depending on eligibility requirement complexity. Using an assertion-based model, manually generated structured assertions can be reused anywhere on the site, but curator time investment is still required to generate and maintain an assertion repository. Although using this assertion-based model allowed us to rapidly scale biomarker, disease, drug, and clinical content on the My Cancer Genome website,[7] not all page types adhere well to this model. Specifically, pages focused on the synthesis of information (eg, pathways pages) were challenging to propagate in an automated manner using solely assertion-based approaches and manually generated versions of these pages were imported from the original MCG website for the current website release.

Additional challenges associated with adopting an assertion-based model for the MCG website include

TABLE 3. Assertion Types Included in Each Dynamic Page Type on My Cancer Genome Website[7]

	Page Types			
Assertion Class	Biomarker	Disease	Drug	Clinical Trial
Therapeutic assertions	✓	✓	✓	
Prognostic assertions	*	*		
Diagnostic assertions	*	*		
Trial eligibility groups	✓	✓	✓	✓
Trial treatment contexts	✓	✓	✓	✓
Trial cohorts				✓
Functional assertions	✓			
Frequency assertions	✓	✓		

NOTE. Check marks indicate assertions currently on My Cancer Genome website.[7] Asterisks indicate assertions that are present in the knowledge management system and will be integrated in future versions of the website.

establishing appropriate quality control (QC) processes. Loading new versions of large external data sets or altering templates can result in significant website errors, and the QC manager needs a high degree of familiarity with the website, assertion model, and underlying data sets to perform cross-checks on the testing server before content deployment. We use at least one member of the team to manually review content during QC. Furthermore, to ensure accuracy of underlying assertions, we maintain a rigorous expert review process for manually curated assertions before publication.

In the future, we plan to convert pages that require manual information synthesis into hybrid pages rooted in subject matter expert–synthesized content supplemented with dynamic, templated, and assertion-based text and visualizations. For example, future versions of the pathway pages may include both manually generated figures and descriptions of the pathway's significance in cancer and dynamically updating descriptions of therapies and clinical trials targeting genes in the pathway that are derived from the assertion model. We also hope to supplement all page types with more assertion types, such as prognostic and diagnostic assertions, that will enhance the clinical actionability of website content. In future work, we plan to implement a level of evidence system so users can evaluate the trustworthiness of curated assertions published on the website.

Using a data model with assertion structures that are application-agnostic empowers use of MCG assertions in additional applications, such as landscape analyses, interpretive reports for pathologists, and molecular tumor board reports. In the future, these assertions could be incorporated with electronic health records to provide in situ clinical decision support for patients who test positive for key biomarkers. Additionally, assertions in the MCG knowledgebase could inform literature reviews, clinical trial planning, and policy decisions.

AFFILIATIONS

[1]Vanderbilt-Ingram Cancer Center, Nashville, TN
[2]Rush University Medical Center, Chicago, IL

CORRESPONDING AUTHOR

Mia Levy, PhD, Rush University Medical Center, 1725 W. Harrison St, Suite 809, Chicago, IL 60612; e-mail: mia_levy@rush.edu.

PRIOR PRESENTATION

Presented at the Cancer Genomics Consortium 2019 (oral presentation) August 11-14, 2019, Nashville, TN, and the AACR Annual Meeting 2020, virtual, April 24-29, 2020.

SUPPORT

Supported by foundation support from the Edward P. Evans Foundation, the Joyce Family Foundation, the Robert J. Kleberg, Jr and Helen C. Kleberg Foundation, Anonymous Foundation, the T. J. Martell Foundation, and Vanderbilt University Technology Maturation Funds; grant support from the Susan G. Komen foundation (SAC160070), the Institute of Museum and Library Services (IMLS LG-06-13-0180-13), the NHGRI–IGNITE I3P grant (NHGRI U01 HG007253), a GE Healthymagination award, an NCI Cancer Center Support Grant (P30CA068485), and a Pfizer Independent Grant for Learning and Change; and corporate charitable gifts from Bristol Myers Squibb and GenomOncology.

AUTHOR CONTRIBUTIONS

Conception and design: Marilyn E. Holt, Kathleen F. Mittendorf, Michele LeNoue-Newton, Neha M. Jain, Christine M. Lovly, Travis Osterman, Christine Micheel, Mia Levy
Financial support: Christine Micheel, Mia Levy
Administrative support: Ingrid Anderson, Christine Micheel, Mia Levy
Collection and assembly of data: Marilyn E. Holt, Kathleen F. Mittendorf, Michele LeNoue-Newton, Ingrid Anderson, Christine M. Lovly, Christine Micheel, Mia Levy
Data analysis and interpretation: Marilyn E. Holt, Kathleen F. Mittendorf, Neha M. Jain, Christine M. Lovly, Travis Osterman, Christine Micheel, Mia Levy
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Marilyn E. Holt
Research Funding: GenomOncology (Inst)
Travel, Accommodations, Expenses: GenomOncology

Kathleen F. Mittendorf
Research Funding: GE Healthcare (Inst)

Michele LeNoue-Newton
Employment: DaVita (I)
Stock and Other Ownership Interests: DaVita (I)
Research Funding: GE Healthcare (Inst)

Neha M. Jain
Research Funding: GE Healthcare (Inst)

Christine M. Lovly
Honoraria: Takeda, Pfizer, Cepheid, Foundation Medicine, AstraZeneca, Blueprint Medicines, Lilly, Genentech, Syros Pharmaceuticals, AstraZeneca, Amgen, Roche, EMD Serono, Daiichi Sankyo/Astra Zeneca, Puma Biotechnology
Consulting or Advisory Role: Novartis, Foundation Medicine, Takeda, Pfizer, Cepheid, Foundation Medicine, AstraZeneca, Blueprint Medicines, Syros Pharmaceuticals, Genentech, Lilly, Amgen, EMD Serono, Daiichi Sankyo/Astra Zeneca, Puma Biotechnology
Research Funding: Novartis, Xcovery, AstraZeneca
Travel, Accommodations, Expenses: Pfizer, Takeda, Foundation Medicine, Novartis, Cepheid, Genentech

Travis Osterman

Stock and Other Ownership Interests: Infostratix
Consulting or Advisory Role: eHealth, AstraZeneca, Outcomes Insights, Biodesix, MDoutlook, GenomOncology, Cota Healthcare, Flagship Biosciences
Research Funding: GE Healthcare, IBM Watson Health, Microsoft
Travel, Accommodations, Expenses: GE Healthcare

Christine Micheel

Consulting or Advisory Role: Roche, Genentech
Research Funding: GenomOncology (Inst), GE Healthcare (Inst)

Mia Levy

Employment: SeqTech Diagnostics (I)
Leadership: Personalis
Stock and Other Ownership Interests: Personalis, GenomOncology
Honoraria: Roche
Consulting or Advisory Role: Personalis, GenomOncology, Roche
Research Funding: GenomOncology
Patents, Royalties, Other Intellectual Property: Royalties from GenomOncology for licensing of My Cancer Genome content
Travel, Accommodations, Expenses: Roche

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors gratefully acknowledge the subject matter experts, editors, and My Cancer Genome team members, past and present, who have contributed to the My Cancer Genome knowledgebase and website, and Dr William Pao, who was a cofounder of My Cancer Genome. Additional thanks to GenomOncology, who built the technological framework required for the current version of the My Cancer Genome knowledgebase.

REFERENCES

1. Kumar-Sinha C, Chinnaiyan AM: Precision oncology in the age of integrative genomics. Nat Biotechnol 36:46-60, 2018
2. Malone ER, Oliva M, Sabatini PJB, et al: Molecular profiling for precision cancer therapies. Genome Med 12:8, 2020
3. Treatment by Cancer Type. https://www.nccn.org/guidelines/category_1
4. Biankin AV, Piantadosi S, Hollingsworth SJ: Patient-centric trials for therapeutic development in precision oncology. Nature 526:361-370, 2015
5. Jain NM, Culley A, Micheel CM, et al: Learnings from precision clinical trial matching for oncology patients who received NGS testing. JCO Clin Cancer Inform 5:231-238, 2021
6. Roper N, Stensland KD, Hendricks R, et al: The landscape of precision cancer medicine clinical trials in the United States. Cancer Treat Rev 41:385-390, 2015
7. My Cancer Genome: https://www.MyCancerGenome.org
8. Giuse NB, Kusnoor SV, Koonce TY, et al: Guiding oncology patients through the maze of precision medicine. J Health Commun 21:5-17, 2016 (suppl 1)
9. Taylor AD, Micheel CM, Anderson IA, et al: The Path(way) less traveled: A pathway-oriented approach to providing information about precision cancer medicine on My Cancer Genome. Transl Oncol 9:163-165, 2016
10. Kusnoor SV, Koonce TY, Levy MA, et al: My Cancer Genome: Evaluating an educational model to introduce patients and caregivers to precision medicine information. AMIA Jt Summits Transl Sci Proc 2016:112-121, 2016
11. Micheel CM, Anderson IA, Lee P, et al: Internet-based assessment of oncology health care professional learning style and optimization of materials for web-based learning: Controlled trial with concealed allocation. J Med Internet Res 19:e265, 2017
12. Li X, Warner JL: A review of precision oncology knowledgebases for determining the clinical actionability of genetic variants. Front Cell Dev Biol 8:48, 2020
13. Wagner AH, Walsh B, Mayfield G, et al: A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet 52:448-457, 2020
14. Cambrosio A, Campbell J, Vignola-Gagné E, et al: "Overcoming the bottleneck": Knowledge architectures for genomic data interpretation in oncology, in Data Journeys in the Sciences. Cham, Switzerland, 2020. pp 305-327
15. Jain N, Mittendorf KF, Holt M, et al: The My Cancer Genome clinical trial data model and trial curation workflow. J Am Med Inform Assoc 27:1057-1066, 2020
16. Liu X, Wu C, Li C, et al: dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat 37:235-241, 2016
17. Liu X, Jian X, Boerwinkle E: dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32:894-899, 2011
18. Hart R and Prlic A: Universal Transcript Archive Repository. Version uta_20180821. San Francisco, CA, Github, 2015. https://github.com/biocommons/uta
19. UniProt Consortium: UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 47:D506-D515, 2019
20. AACR Project GENIE Consortium: Powering precision medicine through an international consortium. Cancer Discov 7:818-831, 2017
21. "ALK Fusion". My Cancer Genome. 22 January 2021. https://www.mycancergenome.org/content/alteration/alk-fusion/
22. "Malignant Solid Tumor". My Cancer Genome. 26 January 2021. https://www.mycancergenome.org/content/disease/malignant-solid-tumor/
23. "Pembrolizumab". My Cancer Genome. 30 April 2021. https://www.mycancergenome.org/content/drugs/pembrolizumab/
24. "Clinical Trial: NCT04676477". My Cancer Genome. 30 April 2021. https://www.mycancergenome.org/content/clinical_trials/NCT04676477/

APPENDIX

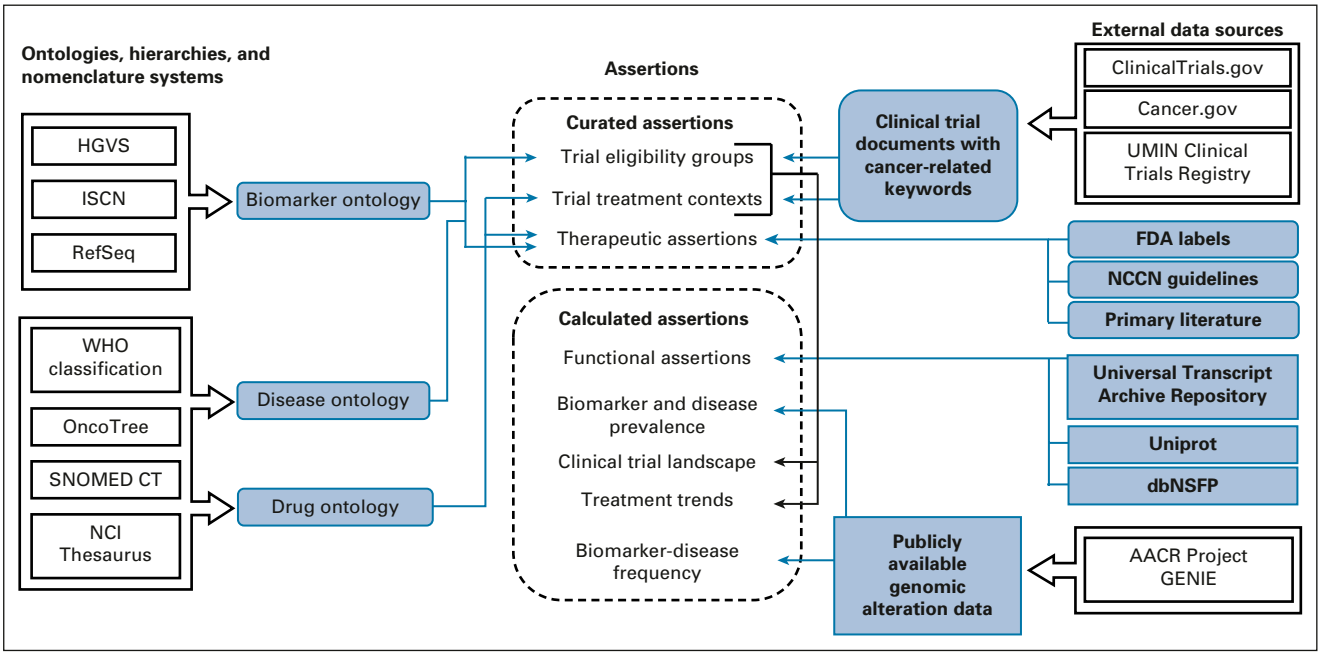

FIG A1. Information architecture of My Cancer Genome. Colored fields and arrows show the information flow from ontologies and data sets that directly contribute to manually curated and calculated assertions. Calculated assertions are defined as assertions that are computationally derived from manually curated structured assertions and structured data from external sources. Fields and arrows that are not colored are imported into an ontology or data set before assertion generation. Black, skinny (nonblock) arrows show information flow from curated assertions to calculated assertions (ie, how curated assertions are used to generate calculated assertions). Ontologies or data sets with rounded shapes are manually supplemented or annotated by MCG curators. Ontologies, nomenclatures, or data sets with square, rectangular shapes are not manually supplemented or annotated. Professional guidelines include guidelines from the NCCN, ASCO, NICE, BNF, and SMC. BNF, British National Formulary; dbNSFP, database for nonsynonymous SNPs' functional predictions; FDA, US Food and Drug Administration; GENIE, Genomics Evidence Neoplasia Information Exchange; HGVS, Human Genome Variation Society (nomenclature); ISCN, International System for Human Cytogenetic Nomenclature; MCG, My Cancer Genome; NCCN, National Comprehensive Cancer Network; NCIt, NCI Thesaurus; NICE, National Institute for Health Care and Excellence; RefSeq, NCBI Reference Sequence Database; SMC, Scottish Medicines Consortium; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms; UMIN, University Hospital Medical Information Network.

