Journal of the American Medical Informatics Association, 0(0), 2017, 1–7
doi: 10.1093/jamia/ocw148
Research and Applications

Research and Applications

The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations

Linda Huang,[1,2] Helen Fernandes,[1,3] Hamid Zia,[1,3] Peyman Tavassoli,[1,3] Hanna Rennert,[3] David Pisapia,[1,3] Marcin Imielinski,[1,3] Andrea Sboner,[1,2,3] Mark A Rubin,[1,3,*] Michael Kluk,[1,3,*] Olivier Elemento[1,2,*]

[1]Institute for Precision Medicine, [2]Institute for Computational Biomedicine, [3]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

*Corresponding authors: E-mails: rubinma@med.cornell.edu; mik9095@med.cornell.edu; ole2001@med.cornell.edu

Received 18 June 2016; Accepted 26 September 2016

Abstract

Objective: This paper describes the Precision Medicine Knowledge Base (PMKB; https://pmkb.weill.cornell.edu), an interactive online application for collaborative editing, maintenance, and sharing of structured clinical-grade cancer mutation interpretations.

Materials and Methods: PMKB was built using the Ruby on Rails Web application framework. Leveraging existing standards such as the Human Genome Variation Society variant description format, we implemented a data model that links variants to tumor-specific and tissue-specific interpretations. Key features of PMKB include support for all major variant types, standardized authentication, distinct user roles including high-level approvers, and detailed activity history. A REpresentational State Transfer (REST) application-programming interface (API) was implemented to query the PMKB programmatically.

Results: At the time of writing, PMKB contains 457 variant descriptions with 281 clinical-grade interpretations. The EGFR, BRAF, KRAS, and KIT genes are associated with the largest numbers of interpretable variants. PMKB's interpretations have been used in over 1500 AmpliSeq tests and 750 whole-exome sequencing tests. The interpretations are accessed either directly via the Web interface or programmatically via the existing API.

Discussion: An accurate and up-to-date knowledge base of genomic alterations of clinical significance is critical to the success of precision medicine programs. The open-access, programmatically accessible PMKB represents an important attempt at creating such a resource in the field of oncology.

Conclusion: The PMKB was designed to help collect and maintain clinical-grade mutation interpretations and facilitate reporting for clinical cancer genomic testing. The PMKB was also designed to enable the creation of clinical cancer genomics automated reporting pipelines via an API.

Key words: precision medicine, database, cancer genomics, clinical reporting, pathology, application-programming interface

BACKGROUND AND SIGNIFICANCE

A growing number of medical institutions have started genomic testing–driven precision medicine programs.[1–3] A critical component of clinical genomic testing is the generation of accurate and informative reports containing clinical-grade interpretations of genomic alterations. Such reports must not only list which variants and mutations were found in a given clinical sample, but also provide interpretations of these variants in the context of available and relevant clinical information. In cancer, the clinical significance and interpretation of somatic mutations and germline variants often depends on the tumor context, that is, tumor type and site. At Weill-Cornell Medicine's (WCM)

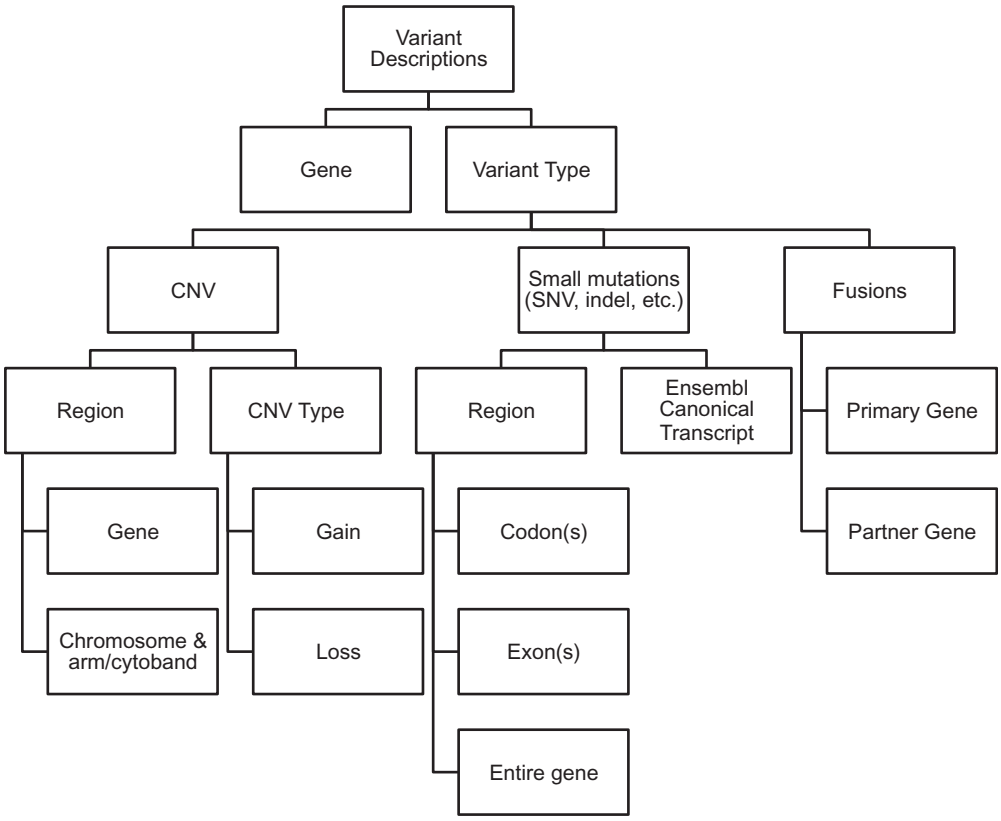

Fig. 1. Diagram of the variant description data types.

Institute for Precision Medicine, together with New York Presbyterian Hospital, we routinely conduct genomic sequencing using both targeted testing of hotspot mutations (AmpliSeq™ Cancer Hotspot Panel v2, Life Technologies) and whole-exome sequencing.[3] These tests are approved by the New York State Department of Health. After sequencing and analyzing the data using either commercial (Torrent Suite, Life Technologies) or custom (whole-exome sequencing) pipelines, the position and effect of mutations on coding sequences are determined using publicly available tools including the Ensembl Variant Effect Predictor (VEP),[4] snpEff,[5] and Annovar.[6] Molecular pathologists who sign out genomic testing reports need to interpret the clinical relevance of each annotated mutation, summarizing their findings in a molecular report. This is usually a tedious task that requires extensive literature curation. Some resources, such as ClinVar, have started cataloguing the clinical significance for variants relative to disease phenotypes.[7] To facilitate the task of interpreting cancer mutations, several online resources have attempted to curate and catalogue clinically relevant mutations. These include Washington University's CiViC DB (https://civic.genome.wustl.edu), MD Anderson's Personalized Cancer Therapy (https://pct.mdanderson.org), Vanderbilt-Ingram Cancer Center's MyCancerGenome (https://www.mycancergenome.org), and several others. While they are helpful resources, in our experience few of these databases contain clinical-grade interpretations actually applicable to clinical reporting. In some instances, mutation interpretations do not meet required levels of brevity and specificity. In some databases, mutations are not interpreted in the context of specific tumor types. In others, only point mutations and indels are catalogued, while common clinically relevant mutations such as gene fusions and copy number alterations/variations are not included. Several clinically critical features may be missing from these databases for integration into routine workflows, such as whether a variant is a pertinent negative in a given tumor type (i.e., a variant for which information regarding the accuracy of negative calls must be reported). Some databases may have limited ability to maintain up-to-date information content or may lack versioning. Finally, while some databases have application-programming interfaces (APIs) for integration into automated workflows, most do not. To overcome these collective limitations, we created the Precision Medicine Knowledge Base (PMKB). The PMKB is currently restricted to variant interpretations for oncology. It was designed in close collaboration with pathologists to ensure accurate and standardized terminology and workflows compatible with clinical use. Importantly, all interpretations are either written or approved by board-certified molecular pathologists. The PMKB's interpretations have been used in over 1500 AmpliSeq tests and 750 whole-exome sequencing tests and are accessed either directly via a Web interface or programmatically via an API.

MATERIALS AND METHODS

Design

PMKB began development in 2015, in order to aid pathologist signout of AmpliSeq 50-gene panel results, and was later expanded to support a broad array of features for signing out whole-exome sequencing reports, such as copy number variations, germline variants, pertinent negatives, and many more. PMKB was built using the Ruby on Rails Web application framework, chosen for its popularity as a platform for complex moderate-load Web applications, wide variety of open-source library extensions, and ease of use. Ruby on Rails provides an interface with a relational database by default, giving users the

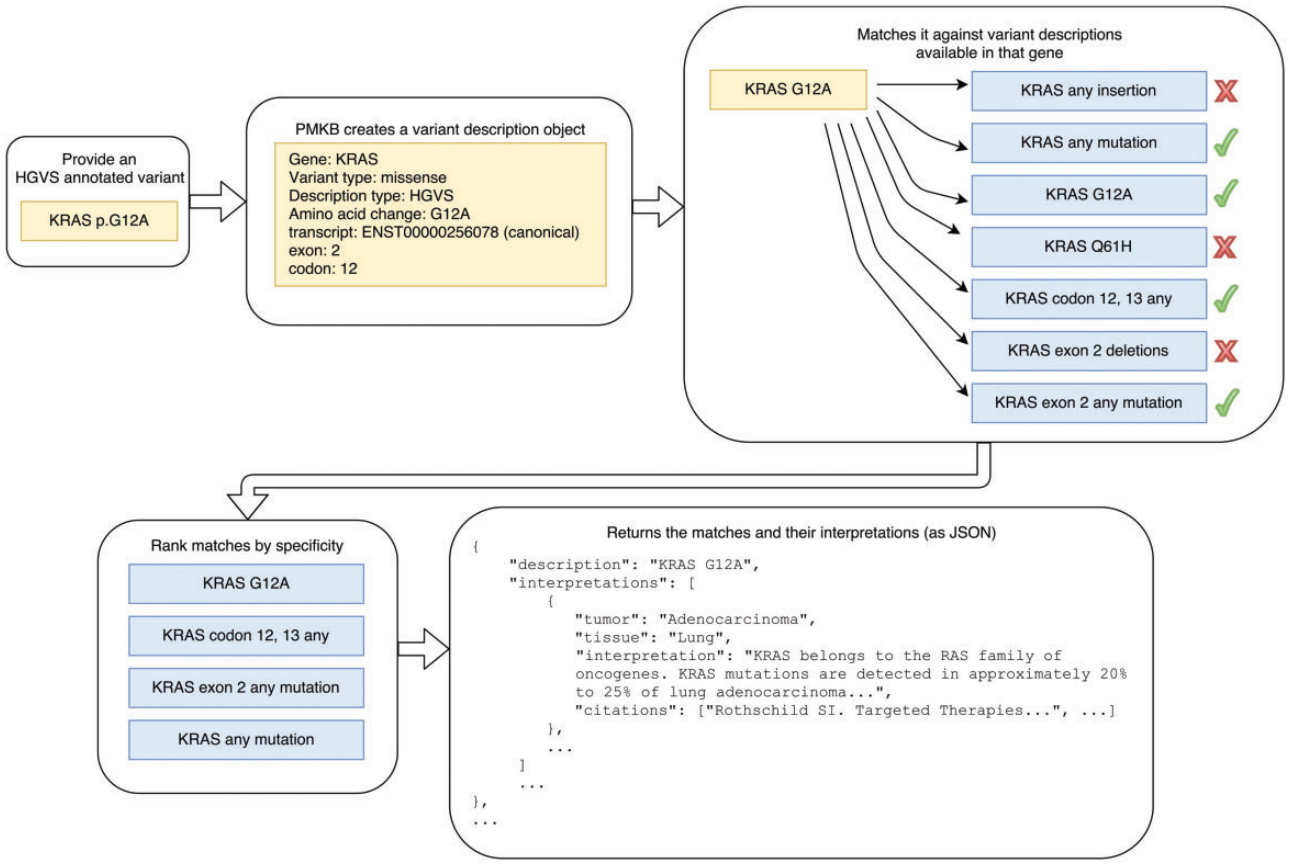

Fig. 2. Illustration of information returned by the PMKB API when querying a variant.

advantage of a database system that promotes data uniqueness and atomicity. PMKB uses a straightforward database schema that structures aspects of an interpretation in a modular way (Supplementary Figure S1). The design considerations of PMKB's design and data fields are 2-fold: (1) to provide data granularity in a way that makes automatic retrieval of interpretations possible, and (2) to provide a convenient experience for pathologists, who will be classifying variants and writing interpretations as part of clinical genomics sign-out.

Data granularity – interpretations

An interpretation in PMKB requires associations with 3 different kinds of elements as part of its identity: (1) gene-variant descriptions, (2) cancer and tumor-type descriptions, and (3) tissue-type descriptions. The interpretation object itself contains the textual interpretation, supported by relevant literature, and a numeric tier. The tier is a category indicating how clinically actionable an interpretation is. This type of association captures the level of specificity expected in a clinical-grade report's interpretations. When adding an interpretation to the PMKB, one can associate as many of each element (i.e., cancer or tumor and tissue types) as are applicable to the interpretation. For example, an interpretation for a mutation in BRAF may be relevant to several different tissue types. This multiple-association structure avoids repetition of interpretation text, since the mutation information is primarily linked to the cancer and tumor types. Associations can be as broad or as specific as necessary. For example, an interpretation could be specified for a variant in "any tumor type" or "any tissue" for more general comments. Separating the tumor type and tumor site also allows for more flexibility when generating new interpretations.

Data granularity – variant descriptions

We developed a system for variant description that incorporates existing standards such as those of the Human Genome Variation Society (HGVS)[8] (Figure 1). At the highest level, a variant is described by the gene it is associated with and a variant category. Variant categories include small, localized mutations (single nucleotide variants, indels), copy number alterations, and gene fusions. Descriptions of small, localized mutations such as single nucleotide variants and indels can be broken down into 2 groups: a specific mutation described using HGVS protein-change and DNA-change notation, and a gene region–based description. Gene regions can be further divided into specific codons, specific exons, and the entire gene (Figure 1). The variant type, eg, deletion, insertion, missense, nonsense, etc., must be specified. Fusions are described as a pair consisting of the primary gene and the partner gene. Copy number alterations are described as either a gain or loss of copy number, pertaining to either a gene or a chromosome region, using arm and cytoband notation, eg, 17p13.1. When variants are entered, PMKB automatically retrieves specific gene region information from Ensembl, based on Ensembl's canonical transcript for a gene and its GRCh37-based API.[9]

Separating variant descriptions into discrete fields facilitates the process of matching them against existing annotations. PMKB's REpresentational State Transfer (REST) API is set up to take a variant's HGVS protein notation as input and match that variant against multiple levels of variant descriptions, then return all relevant interpretations. For example, in the case of a mutation, PMKB's REST API would take KRAS p.G12A as input (Figure 2) and match it against all KRAS mutations in its database. This API query could return

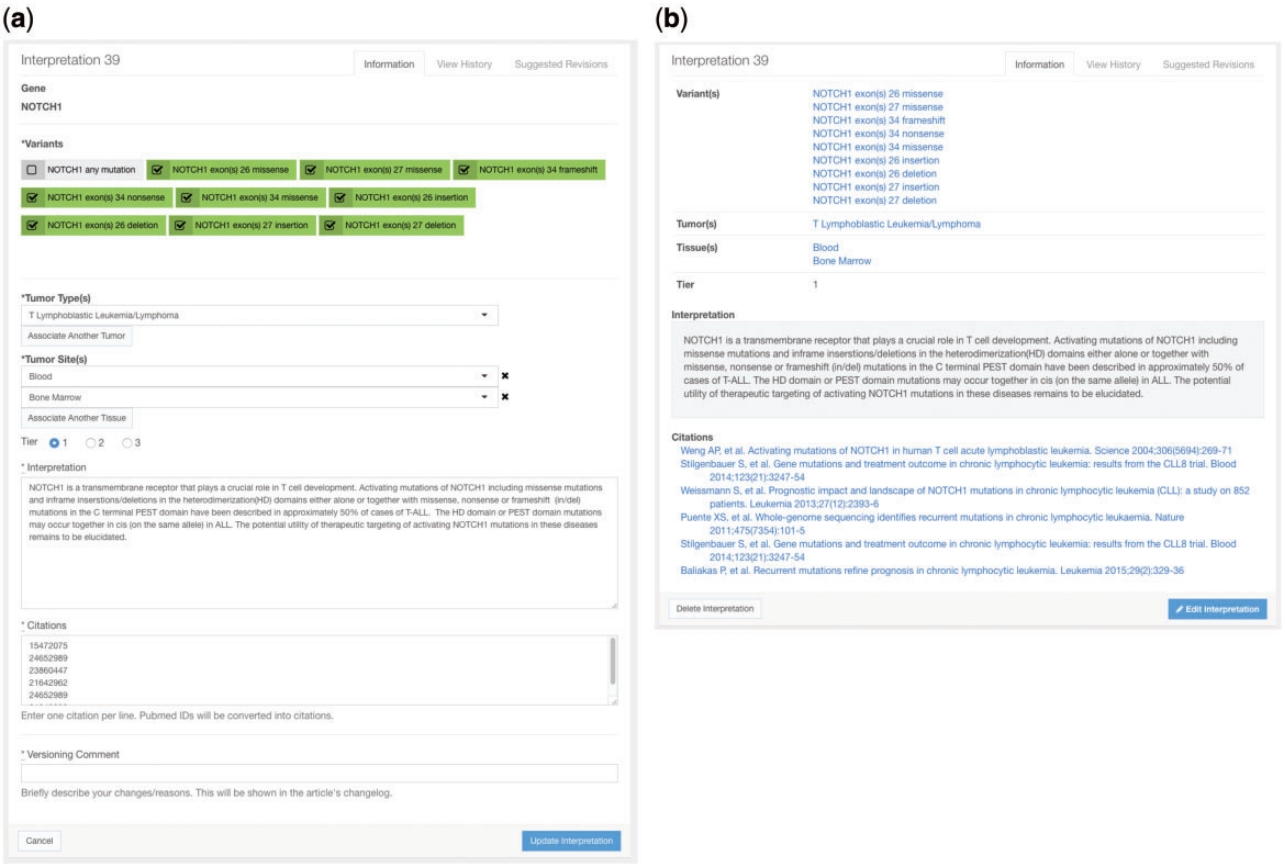

Fig. 3. Screenshots of (a) the interface for entering an interpretation in PMKB, and (b) summary view of an interpretation after entry/edit. This also demonstrates how PubMed IDs in the citations entry were resolved to complete citations in the display.

interpretations for KRAS p.G12A, KRAS codon 12 missense, KRAS exon 2 missense, KRAS exon 2 any mutation, and KRAS any mutation (Figure 2). These matches are ranked in order of specificity, based on the width of the sequence they fall in and whether the variant type is a match or "any." Extra fields in the search query can make PMKB return only those interpretations that are linked to a specific tumor type and tissue type, if desired.

Tumor type and tissue type objects

Tumor types correspond to possible diagnoses, such as melanoma or adenocarcinoma. Tissue types correspond to the primary site of a diagnosis. We have adopted a standard terminology for both tumor types and tissue types that was assembled by a team of highly experienced molecular pathologists. As part of the clinical genomics reporting process, patients' diagnoses need to be assigned according to this standardized terminology (such terminology may coexist with a free-text diagnosis). These lists of tumor types (https://pmkb.weill.cornell.edu/tumors) and tissues (https://pmkb.weill.cornell.edu/tissues) are expanded and edited as needed to accurately categorize new and evolving diagnoses.

RESULTS

User interface

The PMKB currently consists of a multiuser interface for entering, editing, browsing, and querying variants. Entering variant descriptions into PMKB is done via a hypertext markup language (HTML) form. For convenience, the user may enter a Catalogue of Somatic Mutations in Cancer (COSMIC) ID and autofill the form with an HGVS description from a locally stored version of the COSMIC database.[10] Otherwise, the user may choose a gene from a dropdown list or type text in a search box to search for any gene symbol accepted by the Human Genome Organization.[11] Once a gene is chosen, the user will choose a variant type, which determines what other fields are required. For small and localized mutations, the user can choose a description type — HGVS notation, codon, exon, or "anywhere in gene" — bringing up a field with the HGVS notation, codon range, or exon range, respectively. Additionally, the user can flag a variant description as germline using a separate checkbox. If "CNV" is selected as the variant type, the user will have to specify either gain or loss, and can select a checkbox if he or she wishes to use a chromosome-based location instead of a gene-based location. Chromosome-based locations use extra fields for the chromosome and cytoband. Choosing "rearrangement/fusion" for the variant type will bring up another field for partner gene that also allows choosing from any Human Genome Organization symbol. Finally, the user must provide a versioning comment that will be preserved in the change log for that entry. Once a variant is submitted, PMKB automatically adds region information using Ensembl's API, which is useful for putting variants in the proper standardized context.

The user interface for entering interpretations also uses an HTML form (Figure 3a). This form allows the user to first select from any gene in PMKB that has at least 1 variant description. The user will then see a list of variant descriptions for that gene and can check off as many as are relevant. This is a very powerful feature

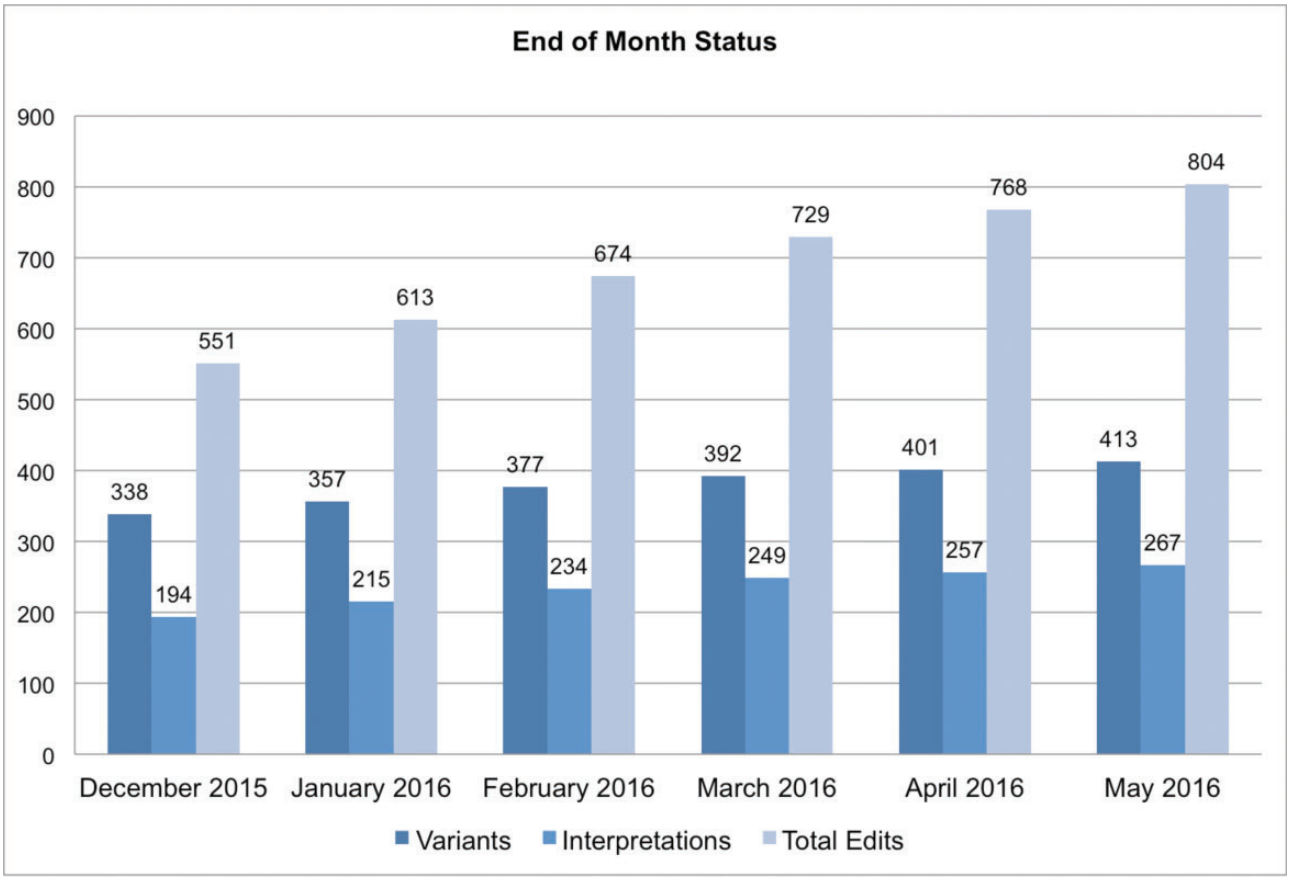

Fig. 4. Growth of the knowledge base over time. Entries created prior to December 2015 represent work that IPM pathologists had saved in Excel spreadsheets. Currently, a small team at the IPM makes contributions to PMKB, and expansion in PMKB's user base could result in a much higher rate of growth.

that greatly facilitates applying a single interpretation to many variants, including new variants that are encountered on a rolling basis that warrant the same interpretive comment. In addition, this feature also allows the user to easily edit/modify an interpretive comment already present in the knowledge base and apply that newly edited comment to 1 or more specific variants among the many variants present for that gene. Tumor types and sites are also available in dropdown lists, and users may add as many entries as they see fit. Radio buttons are used to select a tier, and the actual text of the interpretation goes into a specific text field. Citations are entered using PubMed IDs, 1 per line in a text box. After the interpretation is submitted, the PubMed API will be used to turn these IDs into citation strings, allowing for greater consistency and convenience in citation formatting (Figure 3b). All interpretations must be supported by at least 1 literature citation. Finally, the user should enter a versioning comment that will be preserved in the entry's change log. Interpretation pages provide links to associated variants, tumor types, tissue types, and PubMed entries, including links to external online resources such as Ensembl, COSMIC, PubMed, etc.

A search engine with auto-complete function helps query the PMKB for specific genes and variants. PMKB's repository of information has been growing steadily, thanks to continuous curation efforts by our molecular pathology team (Figure 4). At the time of writing, PMKB contains 457 variant descriptions with 281 interpretations (Figure 4). Genes including EGFR, BRAF, KRAS, and KIT are associated with the largest numbers of interpretable variants (Figures 5a and b). Adenocarcinomas are by far the largest tumor type, followed by acute myeloid leukemia and myelodysplastic syndromes (Figure 5c). The usage pattern and content are a result of the tumor types and tissue types that are currently being tested by the different platforms at our institution, and are expected to expand over time.

User roles

Within the PMKB application, users have 3 different levels of privilege: a high-level "approver," who can review and approve others' entries; standard users, who can submit edits; and guests, who cannot make changes. The first role is reserved for the PMKB's molecular pathologists, who can enter any changes to variant descriptions or interpretations directly for publication on the site. The second role is intended for general users, such as clinical fellows, medical students, members of the computational team, etc., who can create and edit variant descriptions and interpretations, but their new entries and modifications must be approved by an approver pathologist before being published on the site. A pathologist can also choose to edit another user's entry before approval. Since interpretations are used for clinical reporting, signing-out molecular pathologists can be guaranteed that all changes have been approved by a board-certified pathologist. All changes are also tracked in an audit log for each entry. Administrators and users in the first-tier role can restore any variant description or interpretation to a previous state based on the audit log. Auditing is an important part of clinical workflows, and PMKB can provide this capability in a robust capacity.

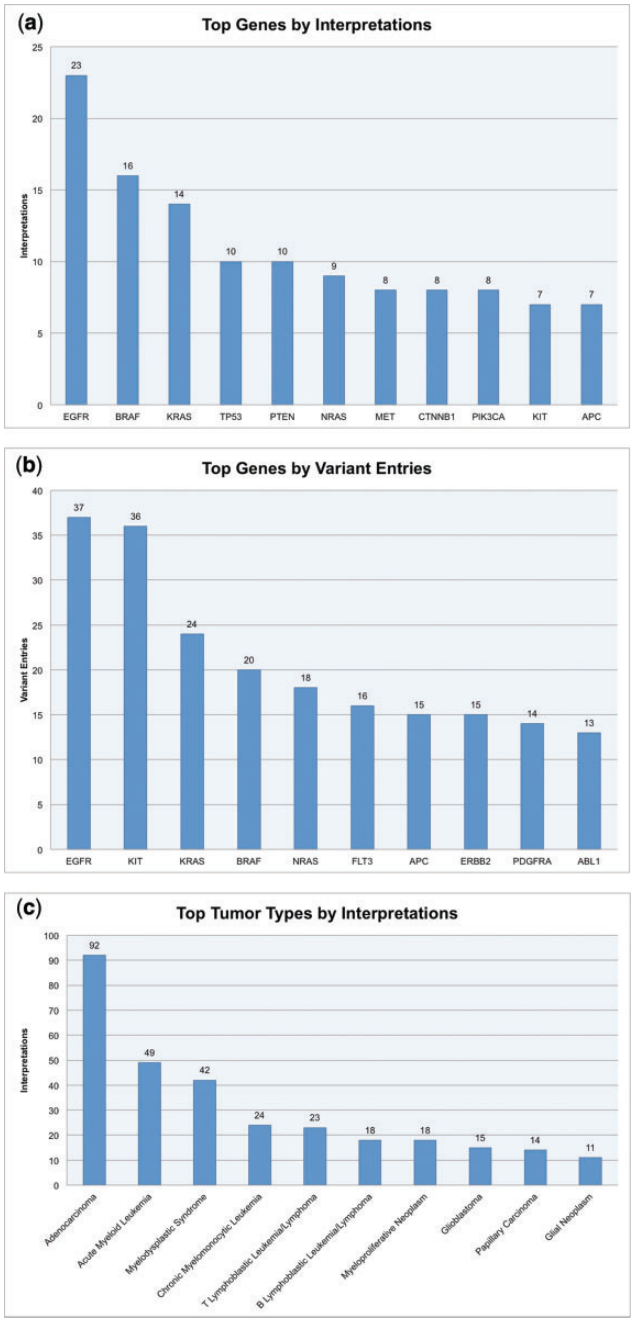

Fig. 5. (a) The top 10 genes in PMKB when ranked by number of variant descriptions in the database. (b) The top 10 genes in PMKB when ranked by number of interpretations in the database. The prevalence of certain genes in PMKB naturally corresponds to genes that have well-studied variants with possible diagnostic use. (c) The top 10 solid tumor types in PMKB when ranked by number of interpretations.

For potential collaborations outside of WCM, security assertion markup language (SAML)-based authentication allows a wider variety of users to access and make edits to PMKB. SAML is a format that allows for standardized exchange of authentication data, providing the capability to use institution-based logins within external Web applications. WCM is part of the SAML-based InCommon federation, and the registration of PMKB as an InCommon website opens it up to collaborations with pathologists and researchers at other institutions. As curating interpretations is a common activity related to genomic testing reports, this could potentially reduce repetition of that work between institutions.

An example of API integration

To illustrate integration with PMKB's API and the use of PMKB to retrieve interpretations, we describe a tool called the AmpliSeq Results Converter that was developed in the Institute for Precision Medicine (IPM) specifically to aid processing of variants in variant call format (VCF) for the AmpliSeq assay (Ion AmpliSeq™ Cancer Hotspot Panel v2, Life Technologies). The AmpliSeq Results Converter converts annotated VCFs into more human-friendly spreadsheets and text reports to facilitate reporting in electronic health records. The AmpliSeq Results Converter is a Python script that queries the PMKB based on a sample's tumor and primary site and variants' HGVS annotations within the VCF. Queries occur via the PMKB's REST API through hypertext transfer protocol (HTTP) requests. PMKB returns the most relevant interpretation for each request, and the AmpliSeq Results Converter generates a report suitable for upload into the laboratory information system Cerner Millennium Helix. The AmpliSeq Results Converter has greatly facilitated the laboratory's workflow by eliminating the need to manually copy interpretations from an Excel spreadsheet into the diagnostic report. Use of this pipeline allows the appropriate interpretations to be pulled automatically from the PMKB into Millennium Helix, our clinical reporting system. All final reports are reviewed by molecular pathologists prior to case signout to ensure that the proper associations are made between tumor type, tissue type, gene variant, and interpretive comment. The lab has used PMKB as its primary tool for entering and retrieving interpretation information starting in late December 2015. Use of PMKB's API has significantly facilitated the process of retrieving interpretations for variants, especially when the interpretations were generalized for gene regions (eg, EGFR exon 19) rather than specific point mutations (eg, EGFR p.L858R). Currently, PMKB's API for interpretation retrieval has been used for several hundred cases, and the PMKB is used routinely as part of the clinical workflow.

CONCLUSION

We have herein described the Precision Medicine Knowledge Base, an interactive online application for collaborative editing, maintenance, and sharing of structured clinical-grade cancer mutation interpretations. All interpretations are available free of charge to the community under the Creative Commons Attribution 4.0 Internation license and can be accessed either via the Web interface (https://pmkb.weill.cornell.edu) or programmatically via the existing API. Within our institution, the PMKB has already proven to be an enormously useful tool for storing and retrieving interpretation information amenable to use by pathologists and reporting pipelines. It has led to significant improvements in the laboratory workflow for the AmpliSeq 50-gene panel assay, saving considerable time and effort, and is currently being used to report WES results.[3]

Since many institutions face similar problems with reporting, it is our hope that the success of PMKB at Weill-Cornell Medicine/New York Presbyterian Hospital can be replicated and expanded with potential collaborators at other institutions. To support both larger assays in-house and potential traffic from collaborators, PMKB's infrastructure, SAML authentication capabilities, and API can readily be leveraged, while continuing to serve results reliably at

acceptable speeds. The speed of computation becomes highly relevant with larger panels of genes that can return calls for hundreds of different variants, an issue that can be addressed with further parallel processing and database query optimization within PMKB's API.

DISCUSSION

In addition to improving PMKB's technical performance, there is the challenge of increasing the output of the interpretations themselves. Quality and quantity of interpretations are what will make PMKB attractive to potential collaborators and will set a standard for future contributions to the knowledge base. Since interpretations must be written by qualified individuals, the PMKB software tries to ease the burden of storing and organizing interpretations, allowing pathologists to focus on writing interpretive comments and signing out cases. Along with writing new interpretations comes the labor of keeping interpretations and tier information up to date with current information on published literature, clinical trials, and US Food and Drug Administration drug approvals. Bringing in clinicians and physicians (currently only on the receiving end of the reporting process), collaborators, and outside contributors could significantly increase the output, quality, and maintainability of interpretations, but this must be managed carefully within PMKB to ensure that all edits are up to par. It is important to make sure that PMKB has tools for managing users and user-produced content to prepare for a growing user base.

PMKB will need to continue to adapt based on feedback from users, in order to make sure a variety of needs around reporting are met. Some examples are tumor type–specific pertinent negatives (already supported in PMKB but not yet systematically annotated) and ongoing improvements to interpretation content. Another future feature includes adding a "talk page"–style interface for users to communicate with one another about potential edits and entries, which will become important to track discussions in an expanded user base, especially when an interpretation needs to be brought up to date with new literature. Extensions of the API will provide more functions for retrieving interpretations and address some of the ongoing issues described above. Ideally, this knowledge base will serve as an open tool amenable to crowdsourcing of content over time by experts in specific subspecialties.

FUNDING

This work was supported by a CAREER grant from the National Science Foundation (DB1054964), National Institutes of Health grant R01CA194547, and the Hirschl Trust.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

LH designed and implemented PMKB and was responsible for all programming work. HF conceived the PMKB's early design. MK provided heavy design guidance throughout the project. MK, HF, HR, HZ, PT, DP, and MI contributed variant interpretations directly to PMKB. LH and OE drafted the manuscript. MK, HR, DP, and PT provided manuscript revisions. OE, MK, and MAR supervised work on PMKB. All authors approved publication of the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the Elemento lab and members of the Institute for Precision Medicine for their feedback and discussions.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Trans Med*. 2011;3(111):111ra21.
2. Van Allen EM, Wagle N, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med*. 2014;20(6):682–8.
3. Beltran H, Eng K, Mosquera JM, et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Onco.l* 2015;1(4):466–74.
4. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
5. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
7. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–85.
8. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37(6):564–69.
9. Yates A, Beal K, Keenan S, et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*. 2015;31(1):143–45.
10. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(Database issue):D805–11.
11. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(Database issue):D1079–85.