The Journal of Molecular Diagnostics, Vol. 20, No. 4, July 2018

ELSEVIER

the Journal of Molecular Diagnostics
jmd.amjpathol.org

System for Informatics in the Molecular Pathology Laboratory

An Open-Source End-to-End Solution for Next-Generation Sequencing Clinical Data Management

Wenjun Kang,* Sabah Kadri,† Rutika Puranik,† Michelle N. Wurst,† Sushant A. Patil,† Ibro Mujacic,† Sonia Benhamed,† Nifang Niu,† Chao Jie Zhen,† Bekim Ameti,† Bradley C. Long,† Filipo Galbo,† David Montes,† Crystal Iracheta,† Venessa L. Gamboa,† Daisy Lopez,† Michael Yourshaw,‡ Carolyn A. Lawrence,‡ Dara L. Aisner,‡ Carrie Fitzpatrick,† Megan E. McNerney,†§¶ Y. Lynn Wang,† Jorge Andrade,*§ Samuel L. Volchenboum,*§ Larissa V. Furtado,† Lauren L. Ritterhouse,† and Jeremy P. Segal†‖

From the Center for Research Informatics, the Division of Genomic and Molecular Pathology,† Department of Pathology, and the Department of Pediatrics,§ Biological Sciences Division, The University of Chicago, Chicago, Illinois; the Department of Pathology and Colorado Center for Personalized Medicine,‡ University of Colorado Anschutz Medical Campus, Aurora, Colorado; the Comprehensive Cancer Center,¶ The University of Chicago Medicine, Chicago, Illinois; and the Informatics Subdivision Leadership,‖ Association for Molecular Pathology, Bethesda, Maryland*

Accepted for publication March 29, 2018.

Address correspondence to Jeremy P. Segal, M.D., Ph.D., Division of Genomic and Molecular Pathology, Department of Pathology, University of Chicago, 5841 S. Maryland Ave., MC 1089, Room N-339, Chicago, IL 60637. E-mail: jsegal5@bsd.uchicago.edu.

Next-generation sequencing (NGS) diagnostic assays increasingly are becoming the standard of care in oncology practice. As the scale of an NGS laboratory grows, management of these assays requires organizing large amounts of information, including patient data, laboratory processes, genomic data, as well as variant interpretation and reporting. Although several Laboratory Information Systems and/or Laboratory Information Management Systems are commercially available, they may not meet all of the needs of a given laboratory, in addition to being frequently cost-prohibitive. Herein, we present the System for Informatics in the Molecular Pathology Laboratory (SIMPL), a free and open-source Laboratory Information System/Laboratory Information Management System for academic and nonprofit molecular pathology NGS laboratories, developed at the Genomic and Molecular Pathology Division at the University of Chicago Medicine. SIMPL was designed as a modular end-to-end information system to handle all stages of the NGS laboratory workload from test order to reporting. We describe the features of SIMPL, its clinical validation at University of Chicago Medicine, and its installation and testing within a different academic center laboratory (University of Colorado), and we propose a platform for future community co-development and interlaboratory data sharing. *(J Mol Diagn 2018, 20: 522—532; https://doi.org/10.1016/j.jmoldx.2018.03.008)*

Over the past few years, laboratories increasingly have adopted next-generation sequencing (NGS) technologies for molecular diagnostics in clinical oncology because of the expanding diversity of diagnostic, prognostic, and therapeutic genomic markers that require assessment in the context of various malignancies.[1] Onboarding NGS technologies into the laboratory and keeping up with the intense pace of change in oncology diagnostics via continuous test evolution can be immensely challenging. The most commonly addressed NGS-associated obstacles relate to the complexity of the underlying molecular biology applications and the scale and processing of the primary sequencing data to uncover meaningful tumor-related anomalies.[2–4]

Supported by the Biological Sciences Division at the University of Chicago, the Institute for Translational Medicine, and NIH Clinical and Translational Science Awards grant UL1 TR000430 (awarded to the Division of Genomic and Molecular Pathology, The University of Chicago).

W.K. and S.K. contributed equally to this work.

Disclosures: None declared.

https://doi.org/10.1016/j.jmoldx.2018.03.008

However, a less-appreciated problem is the general organization of the laboratory and the management of laboratory data and information flows, which can become urgent and compelling as the laboratory scale grows with few or no straightforward solutions.

Proper management of NGS diagnostic assays requires the organization of large amounts of information about patients, specimens, laboratory processes, and process status, as well as storage and management of genetic variants, interpretations, and reports. Often, laboratories use spreadsheets and e-mails to organize these data, but these methods can be insecure and are inefficient as volumes inevitably increase. Laboratory Information Systems (LISs) and/or Laboratory Information Management Systems (LIMSs) are not new to the molecular pathology laboratory, but the need for specialized systems is greatly heightened by the complexity of oncology NGS sample management, library preparation, sequencing, and data interpretation, compared with more traditional molecular pathology assays and workflows.[5–7] Some of the most challenging areas are as follows. i) Specimens: molecular oncology specimen workflow processes are difficult in general, requiring review of perhaps multiple specimens, block selections, management of recuts, and assessment of adequacy and tumor purity. NGS analysis may compound these difficulties because of potentially more stringent specimen requirements compared with single-gene tests. ii) Workflow tracking: compared with PCR-based molecular pathology assays, NGS laboratory workflows may be highly variable (eg, amplicon versus hybrid capture) and may require a large number of steps over multiple days, potentially with more than one technologist participating in the preparation. NGS also has the unique feature of library pooling before sequencing, based on planned complementarity of sample-specific barcode sequences. Thus, sequencing batches typically include multiple sample libraries, which may be a problematic piece of logic to manage for many traditional molecular laboratory information systems. iii) Bioinformatics processing: every laboratory that performs clinical NGS uses either commercially available or custom data processing pipelines, which may vary significantly, raising the issue of whether and to what degree this aspect of the laboratory should or could be integrated into an information management system. As pipelines are updated, it also is critical to track the pipeline version that was used to process each specimen. iv) Interpretation and reporting: for an NGS clinical laboratory to function properly, it is essential to have a support platform for reviewing and interpreting final NGS data and creating reports. Historical variants and interpretations need to be archived and should be searchable to allow for easy review of new cases, and there is a need to assemble all relevant case information into a final document for reporting. v) Overall case management: to prevent confusion and minimize turnaround time, the status of each of these steps needs to be continuously tracked such that laboratory staff can quickly determine which samples require which processing step. As laboratory volume increases, the difficulty of maintaining awareness of the status of every specimen and analyzed data set in the laboratory grows, and because of the complexity of NGS workflows, it ultimately can become unmanageable without an effective status tracking system. Team sign-out organization also can be problematic, and a mechanism for clear assignment of responsibility for case review and completion can be extremely beneficial.

As the available options were investigated, it was found that the available commercial LIMS/LIS options did not meet all of our requirements and also frequently were extremely cost-prohibitive. As a result, a modular end-to-end information system was generated to handle this workload to cover all stages from test order to reporting. During the development process, workflow tracking and data management issues common across molecular pathology laboratories were focused on avoiding implementation of logic unique to our laboratory whenever possible, in the interest of creating a system with the greatest potential to support the continued evolution in more than one laboratory. Here, we present the System for Informatics in the Molecular Pathology Laboratory (SIMPL), a free and open-source LIS/LIMS system for nonprofit molecular pathology NGS laboratories, developed at the Genomic and Molecular Pathology Division at the University of Chicago Medicine (UCM-GMP). We also describe its features, clinical validation of the system at UCM-GMP, its installation, testing within a different academic center laboratory, and propose options for possible future community co-development and interlaboratory data sharing. The authors should be contacted to obtain a copy of the SIMPL codebase.

Materials and Methods

System Design

SIMPL is a web-based LIS implemented largely in Django (Django, *https://www.djangoproject.com*, last accessed December 19, 2017), programmed in Python, because of its straightforward architecture and approachable database design. It was developed as a result of UCM-GMP's collaboration with the University of Chicago Center for Research Informatics, and takes advantage of powerful and secure infrastructure that was already available. In UCM-GMP's configuration, SIMPL runs on virtual machines within a large secure computing cluster maintained by the Center for Research Informatics, following the organization's Information Technology security policies, which are based on the NIST 800-53 Cybersecurity Framework (*https://www.nist.gov/cyberframework*, last accessed November 1, 2017). The main software runs on a web server connected to a database server running MySQL software version 5.6.36 (Oracle Corporation, Redwood City, CA) (Supplemental Figure S1). SIMPL incorporates Secure Sockets Layer encryption and allows for Lightweight Directory Access Protocol (LDAP) authentication for user login. This allows

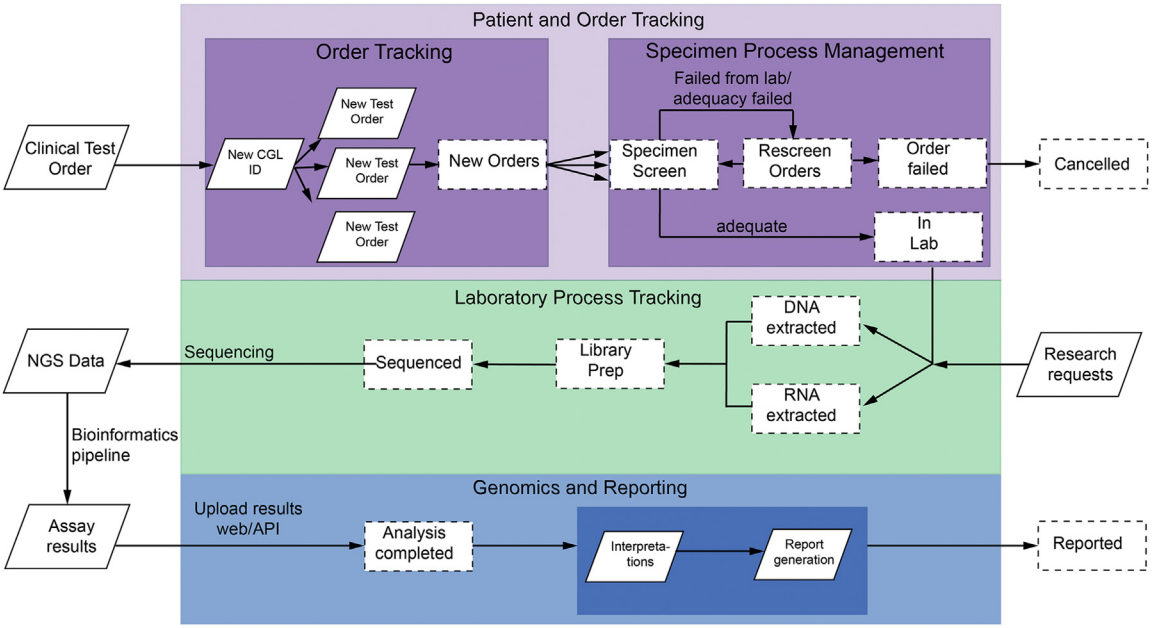

Figure 1 The three main modules of SIMPL managing are as follows: i) Patient and Order Tracking (purple), ii) Laboratory Process Tracking (green), and iii) Genomics and Reporting (blue). **Dashed rectangles** represent the status gates for each test order. API, application program interface; CGL, Clinical Genomics Laboratory; Lab, laboratory; NGS, next-generation sequencing; Prep, preparation.

straightforward connection to existing hospital user verification systems for security and password management. A part-time employee is responsible for maintaining security and functional updates to SIMPL. SIMPL is designed to help manage molecular pathology information management across the entirety of the laboratory testing process, including pre-analytic, analytic, and postanalytic phases. This includes recording patient information and associated NGS test orders, specimen tracking processes, DNA/RNA extraction, library preparation, and sequencing batches, as well as storage of variants (and other result types) from the assay performed. Interpretations can be added to each genomic result, and previous interpretations can be searched, copied, or modified to assist with ongoing analysis. The system has the ability to autogenerate editable reports for each patient test including patient and specimen details, results, interpretations, and general information about the test. In addition to the clinical module, SIMPL also is designed to support some functionality for research samples via a research module (see *Research Module*), because NGS clinical laboratories often participate simultaneously in clinical care and translational scientific projects. Figure 1 shows a schematic representation of the three main modules of the system, with the first module handling patient, order, and specimen information (patient and order tracking); the second module handling laboratory process batch information (laboratory process tracking); and the last module handling interpretation and reporting (genomics and reporting). Supplemental Figure S2 shows the dashboard of the SIMPL web interface, which is the screen seen by the user after logging in. This screen shows a snapshot of all of the samples currently in process by the laboratory and is dynamically updated. The plus sign separates the clinical and research samples.

SIMPL was beta-tested over a period of 2 years, during which each module of the system was incrementally developed, tested, and improved using mock data in a test environment on a development server. This system is now clinically live at the Molecular Pathology Laboratory at UCM-GMP.

Patient, Order, and Specimen Management

In SIMPL, limited protected health information is stored for each patient in the system including the full name, medical record number, date of birth, and sex. Each patient can be assigned to one or multiple categories, each with a three-letter prefix that decides the internal unique identifier for this patient in its category. At UCM-GMP, patients receive the designation CGL (for Clinical Genomics Laboratory) or other customized prefixes for particular research projects, determined within the research module described below. A patient thus may have multiple linked identifiers. Every time a new patient is added to SIMPL in a specific category, the system automatically increments and assigns the next available number in the category. Figure 1 shows the various status gates that each test order may proceed through in dashed boxes. At any given point, a user logged into the system can access an order and check the status of the order. The "Order Status" section in Figure 2 shows the progression of an example test order. Users also can ask for reports detailing which cases are awaiting particular steps in the process.

SIMPL allows each subject to receive multiple test orders, either on the same or separate specimens. Duplicate orders

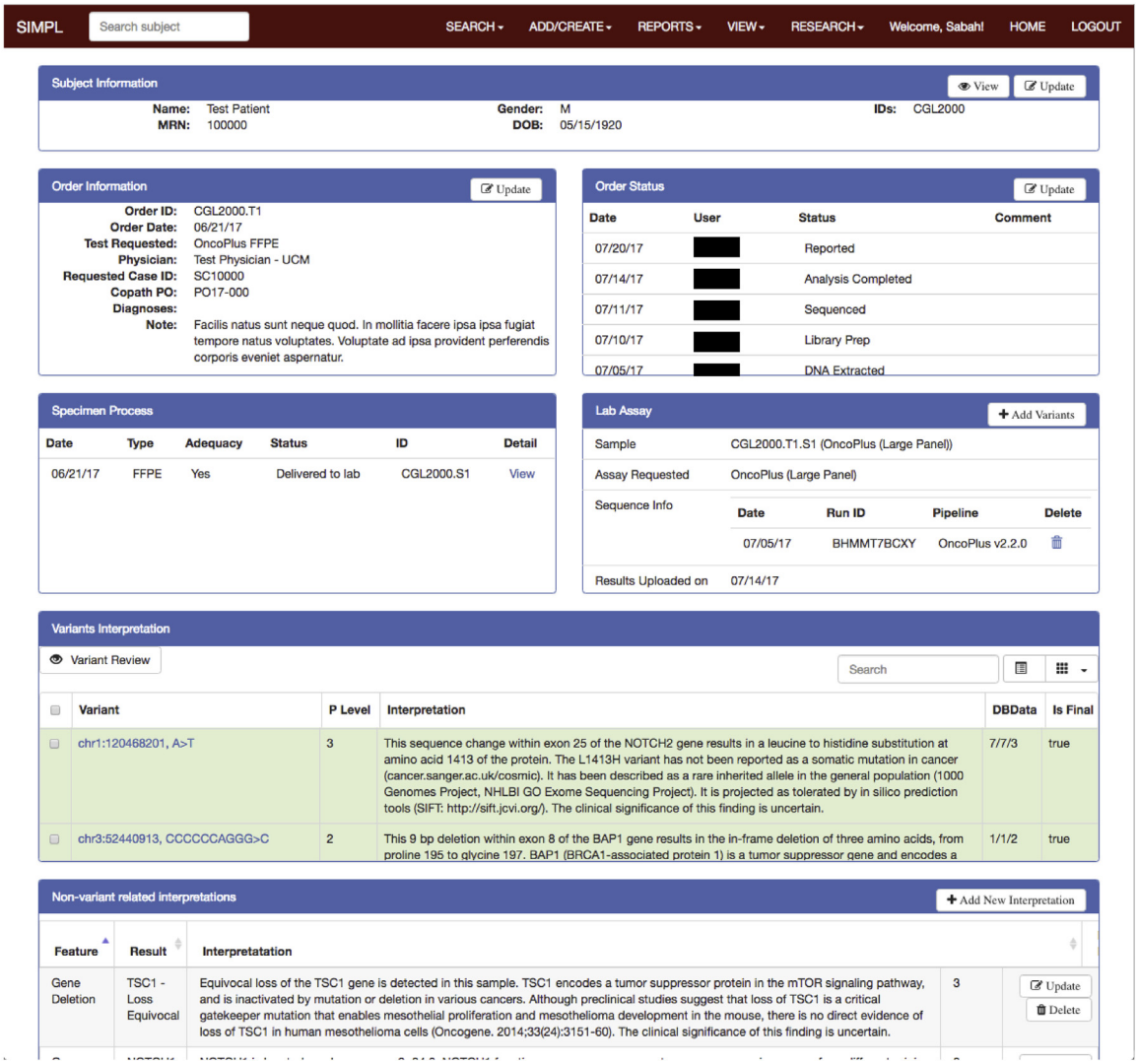

Figure 2 Test order overview page in SIMPL. This example shows mock data generated on the development server during testing of the system. The panels in order from **top** to **bottom** and **left** to **right** show patient information under "Subject Information," order summary under "Order Information," the status history of the order under "Order Status," list of specimens for the test order under "Specimen Process," run specific information under "Lab Assay," interpretations of variants and nonvariants selected for reporting under "Variants Interpretation" and "Non-variant related interpretations," respectively, and the top part of the text box for the order report under "Order Report," along with the status of the report and the date on which it was uploaded to CoPath.

can be detected based on the associated CoPath (Cerner, Kansas City, MO) IDs. Recorded order information includes the date, requesting physician, hospital, test, and diagnosis (International Classification of Diseases, 10th revision code from the order). The system is designed to accept orders from multiple hospitals, and thus stores basic information about the hospitals and the physicians. After a test order is generated for a specific patient, the test order number is automatically incremented in the system, starting with T1 for the first order. The status of the order is set to "New Orders" at this time (Figure 1). The order entry process is currently manual in SIMPL, performed by accessioning staff. In a future update, an interface between SIMPL and other hospital information systems may be included.

The test orders are linked to specific specimen processes belonging to each subject (patient) and each specimen can be tracked in SIMPL (on the "Specimen Screen"). Multiple specimens from the same patient can be linked back to the patient, and the specimen process screen in SIMPL can be used to see the previous specimens tested in a dropdown menu. Logic has been introduced to the web server such that depending on the type of specimen source (formalin-fixed, paraffin-embedded; peripheral blood; cytology smear; bone marrow), certain fields related to the specimen process are mandatory. As an example, the collection date is mandatory for the blood and bone marrow samples, whereas the tumor cell percentage is mandatory only for cytology smears and formalin-fixed, paraffin-embedded specimens (a separate free-text descriptive field is available for blood and bone marrow specimens to describe associated hematopathology findings). Recut request and receipt dates as well as overall reviews of specimen adequacy can be recorded. SIMPL also is equipped

to handle special cases in which instead of the specimen the laboratory directly receives DNA/RNA or the specimen type is unknown. OncoTree specimen and diagnosis classifications are built-in and can be recorded for each specimen to facilitate retrospective data mining (OncoTree, *http://oncotree.mskcc.org*, last accessed December 19, 2017). The specimen process in SIMPL allows for recording of all information associated with this process, before the decision is made to either deliver the specimen to the laboratory extraction ("In Lab") or to fail the specimen. Specimens also may be failed from the laboratory if, for example, the DNA yield is suboptimal. In such cases, either a new specimen process may be generated if another potential specimen exists ("Rescreen Orders"), or the entire test may be canceled ("Cancelled" status in SIMPL) (Figure 1).

Laboratory Process Management

SIMPL performs batch level management of specimens as they move through the laboratory for extraction, library preparation, and sequencing (Figure 1). Specimen processes that pass adequacy checks can be included in batches for nucleic acid extraction, with the requirement for DNA versus RNA based on the recorded features of the laboratory assay. After extraction, test status is updated to "DNA extracted" or "RNA extracted." The extracted samples then are available for batches of library preparation ("Library Prep"), which then are available for batches of sequencing runs ("Sequenced"). For each batch, the laboratory technician (or operator) creating the batch is logged into the system. If a test is ordered on a specimen that previously was extracted/tested, the system automatically alerts the user of this situation. At UCM-GMP, the technician then checks whether adequate material is available and can decide to either make a new extraction of the specimen or use the previously extracted material.

Genomics Results Interpretation and Reporting

Once the samples are sequenced and data are available, all bioinformatics pipelines are run on secure high-performance computing clusters hosted at the University of Chicago. Currently, the pipeline processing and data management in SIMPL are kept separate, but may be linked in the future. The sequencing data for each test order is processed according to the latest version of the clinically validated bioinformatics pipeline specific for the test on a high-performance computing cluster, and the pipeline versions are logged in SIMPL when the data are uploaded along with other run-specific metadata. The "Lab Assay" section in Figure 2 shows the run-specific information of an example test order in SIMPL. The "+ Add Variants" button shown in Figure 2 can be used to perform variant uploads through the web interface. The genomics module of SIMPL is used after the assay results are available ("Analysis completed"). The data can be uploaded individually for each sample using the web interface or using an application program interface for batch uploads. The system can handle both variant and nonvariant results, which include copy number, fusion, and other structural rearrangements reported by UCM NGS clinical assays.

Each variant is stored as a combination of chromosome, position, reference, and mutation, and sample-specific variants store the pipeline version generating that variant as well as depth information at that genomic position and the variant allele frequency. Each variant is also linked to its annotation, specific to the annotation software. Variant calls are annotated and converted to Human Genome Variation Society nomenclature using Alamut Batch software version 1.4.4 (Interactive Biosoftware, Rouen, France), which also pulls from publicly available databases such as COSMIC (*http://www.sanger.ac.uk/cosmic*, last accessed December 19, 2017),[8] NCBI dbSNP,[9] Scale Invariable Feature Transformation (SIFT) algorithm,[10] and so forth. Older assays at UCM-GMP were annotated using ANNOVAR,[11] and SIMPL can store these annotations as well, but the database can be easily modified to adapt to a center's annotation system.

Although research samples are stored in SIMPL, genomic results are not stored for research samples because these are not interpreted in the system and thus do not influence the database statistics used for result interpretation. However, users have the ability to upload results for research samples as well.

After the bioinformatics pipelines are completed, results are uploaded to SIMPL as described above, and the order status is changed to "Analysis Completed." At this time, the cases are available for assignment by the pathologists through a case assignment window (Supplemental Figure S3). The case reviewers and the pathologists review the primary data, make interpretations of the detected variants, and assemble and sign out a final report detailing the case findings along with comments or recommendations that may be appropriate within the context of each patient's disease process. The case can be assigned to one reviewer to draft a report and one molecular pathologist to finalize and sign out the report. Once a case is assigned (Supplemental Figure S3), the case will be available for the assigned users and the assignment status and related comments are tracked by SIMPL. When the primary reviewer completes the case and finalizes their report, an e-mail is generated automatically and sent to the case pathologist, and the case will appear in the work list of the case pathologist. When the case is officially signed out in SIMPL, its status will change from "Analysis Completed" to "Reported," and it will drop from the queue of the case pathologist.

The overall workflow for case review includes the following: i) variant review and creation/modification of interpretations, ii) creation/modification of any necessary nonvariant interpretations, iii) report generation and review, and iv) case completion/sign-out (Figure 3).

The variant review window is implemented as a scrollable window, with all column headings allowing sorting or filtering using arrow buttons, selectors, or blank fields (Figure 3A). For example, to remove common inherited variants one might filter out variants present at appreciable

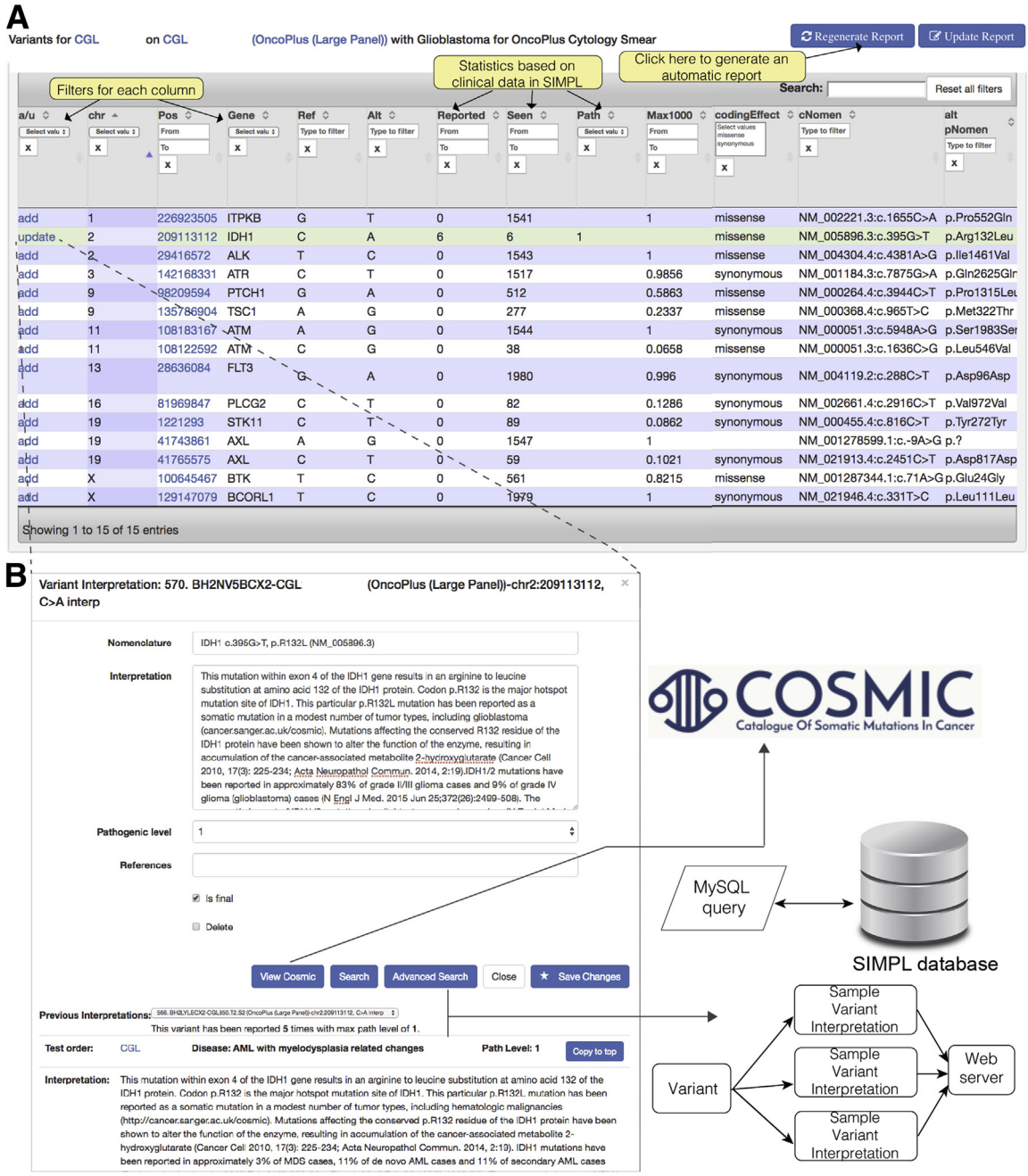

Figure 3 Variant Interpretation in SIMPL. **A:** The variant review page in SIMPL for an example case. **Yellow boxes** highlight some features of this page. The top right button can be used to regenerate the report when interpretations are modified. **B:** The variant interpretation window for an example *IDH1* mutation in **A** when selected. The "View Cosmic" embedded link opens up the variant in COSMIC while the search buttons can be used to query the SIMPL database for previous interpretations of the same variant.

frequency (eg, 1%) using the Max1000 (1000 Genomes Project Max allele frequency field) field by entering "0.01" in the "To" field. Variants then may continue to be reviewed as per assay-specific guidelines. Any variants deemed worthy of reporting may have interpretations assigned to them using the edit button. The interpretation window prepopulates with the Human Genome Variation Society nomenclature from the annotation, which can be edited by the pathologist, and provides a box to add interpretive text and a drop-down choice of pathogenic rating (Figure 3B). If the same variant has been seen in the laboratory before, the previous interpretation will autopopulate to the bottom of the window for review, and there is an advanced search button that allows for searching of previous interpretations by gene, diagnosis, pathogenic rating,

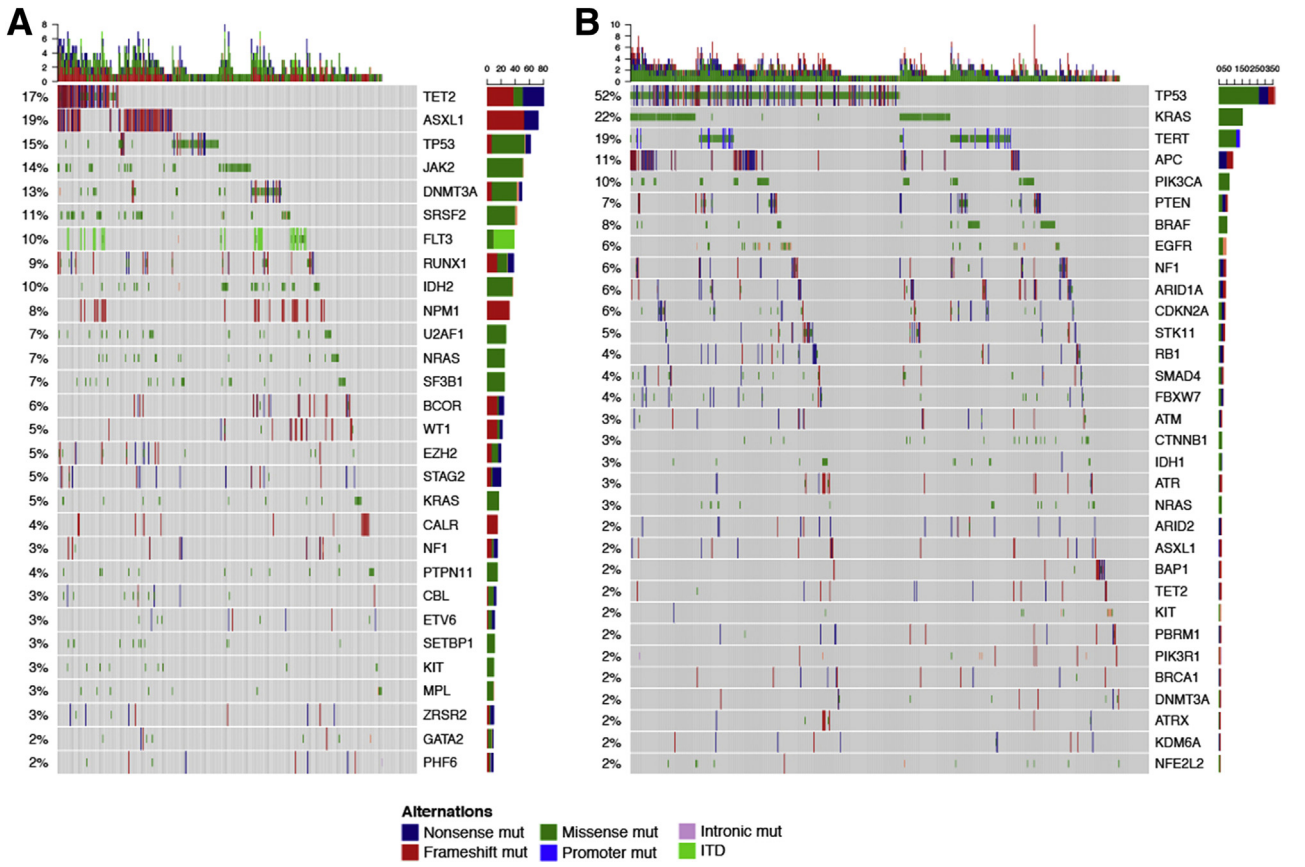

Figure 4 Matrix of pathogenic variants in the top 25% of genes in SIMPL for all hematologic (**A**; *N* = 377) and solid tumor (**B**; *N* = 708) samples that have been run on UCM-OncoPlus. ITD, internal tandem duplication; mut, mutation.

and so forth. Interpretations can be finalized using an "Is final" button.

The use of a database to store this information increases the scope for data mining projects, and value for data sharing in larger genomic data-sharing initiatives such as the GEnetics of Nephropathy—an International Effort (GENIE) consortium.[12] An example is shown in Figure 4, which shows the top 25% genes with pathogenic mutations in all cases run on UCM-OncoPlus,[2] split by hematologic (Figure 4A) and solid tumor (Figure 4B) cases. The knowledge base stores the more recently updated annotation for each variant along with every interpretation that has ever been reported for the variant. The knowledge base can be queried by authorized individuals using advanced search pages generated on the website, based on pathogenic level, OncoTree classifications, interpretive text, coding change, and so forth.

Nonvariant interpretations are any interpretation of an identified genomic anomaly as a result of the test that is not a variant. Potential nonvariant findings include copy number abnormalities, gene fusions, rearrangements, and so forth. These interpretations contain the type of anomaly and a free text box in which to enter the proper nomenclature for the finding, along with a clinical interpretation and a pathogenic rating.

Report Generation

Reports can be generated automatically in SIMPL by clicking "Generate Report," which becomes available after the results are uploaded and until a report is "Finalized." SIMPL uses a predefined report template specific to the clinical test and populates it with pertinent case information including diagnosis, specimen information, and all saved interpretations from the SIMPL database into a single text-based document, which is available for review and editing. Only laboratory directors have the privileges to edit the report templates for each test (see the *User Roles and Groups* section below). UCM-GMP uses only text-based reporting because of the nature of the hospital information systems, but modifying the SIMPL code to instead produce a formatted PDF report would be quite straightforward.

System Architecture

SIMPL is a web application that was built on top of the Django Web Framework (version 1.11.9), which was developed in Python. The front end (client side) was written using HTML 5, CSS 3.0, and JavaScript. The major user interface was developed using Bootstrap 3.3 (*https://getbootstrap.com*,

last accessed June 8, 2018) and jQuery 2.2 (*http://www.cs.ubc.ca/labs/spl/projects/jquery*, last accessed June 8, 2018). The back end (server side) was scripted using Django version 1.11 on Python 3.5. MySQL version 5.7 was selected as the relational database management system to store all user-generated data. Users could be authenticated either through one or more institutional LDAP servers or through local user accounts stored in SIMPL. Elasticsearch version 2.3 (Elasticsearch BV, Mountain View, CA) was used as the search platform for generalized full-text searches. All of the software used in SIMPL is open-source. Hardware requirements are modest and the system runs in Windows/IIS (Microsoft, Redmond, WA) and Linux/Apache (Apache Software Foundation, Forest Hill, MD) or Nginx (*https://www.nginx.com*, last accessed June 8, 2018) webserver environments. Supplemental Figure S1 shows the diagram of the system architecture and software environment of SIMPL. The SIMPL database and code are structured so that all user interactions with the system are logged, allowing for retrospective evaluation of all SIMPL user activity since the implementation of the system. This feature is extremely helpful for helping troubleshoot laboratory problems or errors. The SIMPL database is backed up nightly to redundant tape drive systems housed within the Center for Research Informatics.

User Roles and Groups

SIMPL follows a test order from creation to reporting, and thus the users of the system vary from those entering patient information, laboratory technicians entering laboratory process information, bioinformaticians to upload test results, and pathologists for interpretation and reporting. Thus, SIMPL was designed to manage user permissions and privileges by the user's role and group, respectively, which is similar to the Unix system. The various user groups defined in SIMPL include the "User admin" group, which are users who have privileges to add new user accounts and disable the access of existing users, the "OncoTree management group" to manage entries related to the OncoTree classification, and the "Job admin group" to assign cases for reviewers and molecular pathologists. SIMPL also defines a role for each user, such as "pathologist," "bioinformatician," "lab technician," and so forth, and there is an ability to restrict access to certain parts of the database for certain roles. Any laboratory-specific role restrictions can be implemented.

Research Module

Many clinical oncology NGS laboratories participate in a combination of direct patient care testing and translational research projects because NGS clinical assays often are useful for cohort tumor profiling and other types of projects. Adding these specimens to laboratory work lists and tracking their parallel processing can be challenging, thus we have implemented a research module in SIMPL that allows for batch uploading of research project lists and coordinate tracking, with custom-coded sample prefixes for each project. Research specimens involve different hard-coded logic and appear on different sections of recommended work lists, but can be merged with clinical work by trained medical technologists to produce joint extraction, library preparation, or sequencing batches.

Results

Clinical Validation

By using the College of American Pathologists' requirements for validation of any LIS or computational system as a framework, all of the features, tasks, and test cases required for the clinical validation of a system such as SIMPL were identified. To assist other laboratories with integration, a template for such a validation can be provided for future users who develop their own systems or adapt SIMPL for their Clinical Laboratory Improvement Amendment laboratories. To clinically validate a system that operates at this magnitude, it is essential to establish the security, disaster recovery, access restrictions, audit logging, and data integrity at each level.

Over a period of 2 years, each module, feature, and web-based logic was beta-tested using mock data in a test environment on a development server. Figure 2 shows an example of a mock case that was generated as part of approximately 2000 cases on the development server in this process. Before SIMPL, all patient, specimen, and order information, as well as dates and statuses, were logged in a cumbersome Microsoft Excel sheet. For 6 months before and 3 months after validation, patient information was logged simultaneously in SIMPL, as well as in the pre-SIMPL Excel (Microsoft) spreadsheet, and the data integrity of the two was confirmed manually. The only errors found were attributed to human errors in the form of typos or spelling mistakes.

In addition to these, a separate part of the clinical validation included prospective manual entry of mock case data into the development server to ensure data integrity and production of proper clinical reports. Ten test cases were generated to reflect the variety of molecular pathology caseload at UCM-GMP, including both hematologic and solid tumor cases. Orders were placed and specimen processes were entered per usual practice. Specimens were brought through extraction, library preparation, and sequencing batches successfully. Mock variants were uploaded for each case, and preset interpretations were produced for a set of variants. Finally, reports were generated for each case and inspected to ensure that all expected information and interpretive comments were included.

After the system was thoroughly tested and data integrity was ensured, all historical data, including the variants and the clinical reports, were uploaded to the SIMPL database to make it up-to-date. Currently, SIMPL stores all information at UCM-GMP regarding test orders, specimens, laboratory processing, variants, and interpretations/reports associated with more than 3200 test orders for more than 2500 individual patients.

Deployment

As a trial of the open-source model to assess feasibility of cross-site installation and modification, the SIMPL codebase was shared with the Molecular Correlates Laboratory (MCL) at the University of Colorado (Denver, CO). Currently, the system is in the process of being deployed and customized for use with their existing NGS test systems. Establishing SIMPL in a development environment, such as a Linux or Mac OS X (Apple, Cupertino, CA) desktop computer, is a relatively straightforward process for a programmer with Python experience. MCL was able to install the system in a desktop test environment connected to the University LDAP system within just a day or two to begin working on altering certain SIMPL features and logic to support a few of the unique features of their laboratory. The most challenging aspect of this deployment is not related to software but rather working through site-specific infrastructure and security limitations. Security and Health Insurance Portability and Accountability Act compliance may require some time and effort. One must plan for obtaining administrative approvals, implementing firewall rules for interconnection between the SIMPL web server, its database server, user's computers, and storage systems, and configuring file permissions. Back-up strategy and methodology also must be determined.

The codebase contains documentation for all required software and packages required to set up a working copy of SIMPL. The generalized steps include the following: i) cloning the SIMPL source code and establishing version control as desired; ii) installing operating system−specific prerequisites for Python modules (eg, on Linux systems, various openldap-related and database-related packages may need to be installed); iii) configuring a database for SIMPL to use, such as the open-sourced MySQL or MariaDB (Menlo Park, CA); iv) setting up a Python virtual environment for SIMPL, which enables version sandboxing, along with the latest version of Python and other required packages outlined in the documentation, including Django; v) updating the configuration, database, and LDAP files and settings with site-specific settings.

Once configuration is complete, SIMPL can be run using simple Python scripts, by populating the database, creating an administrative user, and starting the web application. The user then can log-in to the administrative site using the super user credentials from above to add/update user accounts with privileges to log into and use SIMPL. With the site available for use, the source code then can be modified to customize the system as desired.

The built-in Django web server is not recommended for production use and cannot handle secure https connections. Choosing and deploying web servers, taking into account server load, redundancy requirements, ease of configuration, and other factors will require considerable planning and configuration effort. One popular solution is to use the fast, lightweight web Nginx server as a front end, which serves static files and acts as a reverse proxy. Nginx then directs dynamic requests to a Gunicorn version 19.4.5 (*http://gunicorn.org*, last accessed June 8, 2018) web server gateway interface server that relays them to SIMPL.

Discussion

Despite many rapid advances in NGS diagnostics, much of the recent history of this field has involved redundant efforts across many sites to build test workflows and information systems. This has proven costly and has stifled growth and/or prevented implementation at many laboratories. This siloed approach also has favored reference laboratories and larger academic centers at the expense of smaller operations that may lack the required resources and expertise to build complex automated molecular biology workflows and information systems independently.

Among the myriad impediments to establish and grow an in-house NGS program, laboratory information management may be the most problematic owing to a significant lack of affordable comprehensive options. It is important for the full scope of test phases to be modeled in a system to allow for proper incorporation of specimen data in the report as well as adequate data mining and turnaround time tracking, among many other things. Genomic oncology diagnostic workflows, in particular, often include extremely complex specimen evaluations, involving multiple rounds of slide and recuts review, tumor cell percentage assessments, and adequacy evaluations; as such, the ability to record all of this history for each case is critical. Likewise, effective variant storage, review, interpretation, and reporting are major obstacles for many laboratories. Unfortunately, few or none of the commercially available options allow for adequate mirroring of both up-front specimen management and the resulting downstream variant databasing and reporting. Beyond the preconfigured commercial options, other companies offer varying degrees of customized solutions. These are as expensive, if not more expensive, than the preconfigured solutions. However, although they offer the possibility of a system that more closely reflects the particulars of the client laboratory, it is always possible that the end result will underdeliver on previous promises. With either preconfigured or customized solutions, laboratories can spend hundreds of thousands of dollars and still end up with systems that do not adequately support key laboratory processes.

Because of the high cost and/or general unsatisfactory nature of most commercial options, UCM-GMP chose to develop a laboratory information system from the ground up, as other groups have done. However, in the course of efforts to develop SIMPL, the focus was on modeling generic molecular pathology workflows and limiting the incorporation of logic or systems specific to the UCM-GMP laboratory. From the outset, it was the intention to share this system broadly, and to provide laboratories with a new and affordable option for laboratory information management.

There were multiple options by which this system could be propagated among the community beyond simply commercializing and competing with existing companies, including hosting or cloud services. In the end, an open-source approach was chosen because of the concern for having to provide for each individual laboratory's unique needs, and the desire to grow more of a community development network. Certainly, there is a modest degree of expertise and resources that are needed to clone a copy of SIMPL and tailor it to the unique needs of a new laboratory. However, these requirements are far less than what is needed to design a system from scratch, and the labor involved is also extremely affordable compared with the cost of onboarding a commercial system. As an example, the University of Colorado MCL is currently in the process of onboarding this software, and two programmers were able to install and connect the system to their institution's LDAP system within a day to be able to log-in to SIMPL, after which they were able to immediately begin work on desired modifications. UCM-GMP has shared subsequent software modifications with the MCL, who also will be sharing their modifications with UCM-GMP. It is hoped that the MCL will be in a position to launch within a few months, compared with the 2+ years that UCM-GMP invested initially (which is not an atypical experience). In our experience, most of the effort required to set up SIMPL involves regulatory compliance, and an institute with an existing secure infrastructure in place can deploy and maintain the system with a part-time software engineer/programmer.

It should be noted that in some ways, SIMPL is still a work in progress. Although we consider the preanalytical and postanalytical portions of the software to be relatively mature, the analytical laboratory portion covers only basic process steps (eg, status and batching for in-laboratory, extraction, library preparation, sequencing), but does not yet model plate maps or understand detailed variable molecular biology preparations. SIMPL does not currently support Health Level Seven interface, although we have plans to do so in the future. At the UCM-GMP laboratory, SIMPL's in-laboratory tracking is augmented with additional spreadsheets, and this has been a reasonable solution for a volume of a few thousand NGS cases per year. This phase also could be supplemented by an additional system. SIMPL is being updated and improved continuously in the laboratory, but the ideal would be that other laboratories could take on the challenge of implementing additional modules, and thus the software could continue to grow in capabilities as a result of a community effort. Furthermore, leveraging data warehouse capabilities available at the University of Chicago, we aim to integrate genomic findings from SIMPL with detailed clinical information in the future.

It remains to be determined to what degree other laboratories might adopt SIMPL. Bringing a laboratory information system on board is a big decision that requires careful consideration. It also may be daunting to consider instituting a homegrown or semi-homegrown system. Still, there are interesting opportunities that could arise from multilaboratory sharing. For example, having multiple laboratories on the same system would make the sharing of de-identified variants and variant interpretations fairly straightforward, allowing laboratories to benefit from each other's expertise and experience. This also could facilitate easy integration of groups of laboratories into larger genomic data sharing initiatives, such as GENIE.[12]

Recent payer trends suggest that we are entering an era of questionable reimbursement, and laboratories should look for ways to find efficiencies and minimize redundant efforts. Molecular pathology information systems have been historically very expensive to implement, either by purchasing an existing solution or creating one in-house. They also often are inadequate to cover all aspects of laboratory workflow and data management. As such, it is the hope of UCM-GMP that SIMPL may be regarded as a useful and affordable option that other laboratories in the community may adopt and that, together, SIMPL can be developed further with additional features and functionality over time.

Acknowledgments

We thank Thorbjorn Axelsson (Information Technology Operations and Infrastructure); Andy Brook and the entire University of Chicago Center Research Informatics systems team for their outstanding support with the servers, systems, and documentation; and Dr. Vinay Kumar for his vision and support.

Supplemental Data

Supplemental material for this article can be found at *https://doi.org/10.1016/j.jmoldx.2018.03.008*.

References

1. Kamps R, Brandão RD, van den Bosch BJ, Paulussen ADC, Xanthoulea S, Blok MJ, Romano A: Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. Int J Mol Sci 2017, 18. pii: E308
2. Kadri S, Long BC, Mujacic I, Zhen CJ, Wurst MN, Sharma S, McDonald N, Niu N, Benhamed S, Tuteja JH, Seiwert TY, White KP, McNerney ME, Fitzpatrick C, Wang YL, Furtado LV, Segal JP: Clinical validation of a next-generation sequencing genomic oncology panel via cross-platform benchmarking against established amplicon sequencing assays. J Mol Diagn 2017, 19:43−56
3. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF: Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015, 17:251−264

4. Pritchard CC, Salipante SJ, Koehler K, Smith C, Scroggins S, Wood B, Wu D, Lee MK, Dintzis S, Adey A: Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn 2014, 16:56−67
5. Roy S, Durso MB, Wald A, Nikiforov YE, Nikiforova MN: SeqReporter: automating next-generation sequencing result interpretation and reporting workflow in a clinical laboratory. J Mol Diagn 2014, 16: 11−22
6. Aronson SJ, Clark EH, Babb LJ, Baxter S, Farwell LM, Funke BH, Hernandez AL, Joshi VA, Lyon E, Parthum AR, Russell FJ, Varugheese M, Venman TC, Rehm HL: The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. Hum Mutat 2011, 32:532−536
7. Sharma MK, Phillips J, Agarwal S, Wiggins WS, Shrivastava S, Koul SB, Bhattacharjee M, Houchins CD, Kalakota RR, George B, Meyer RR, Spencer DH, Lockwood CM, Nguyen TT, Duncavage EJ, Al-Kateb H, Cottrell CE, Godala S, Lokineni R, Sawant SM, Chatti V, Surampudi S, Sunkishala RR, Darbha R, Macharla S, Milbrandt JD, Virgin HW, Mitra RD, Head RD, Kulkarni S, Bredemeyer A, Pfeifer JD, Seibert K, Nagarajan R: Clinical genomicist workstation. AMIA Jt Summits Transl Sci Proc 2013, 2013:156−157
8. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA: COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 2011, 39:D945−D950
9. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001, 29:308−311
10. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009, 4:1073−1081
11. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010, 38:e164
12. AACR Project GENIE Consortium: AACR Project GENIE: powering precision medicine through an international consortium. Cancer Discov 2017, 7:818−831