Article published online: 2020-12-31



# Indispensability of Clinical Bioinformatics for Effective Implementation of Genomic Medicine in Pathology Laboratories

Srikar Chamala[1] Siddardha Majety[2] Shesh Nath Mishra[2] Kimberly J. Newsom[1] Shaileshbhai Revabhai Gothi[2] Nephi A. Walton[3] Robert H. Dolin[4] Petr Starostik[1]

[1] Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida, United States
[2] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, United States
[3] Intermountain Precision Genomics, St. George, Utah, United States
[4] Elimu Informatics, Richmond, California, United States

Address for correspondence Srikar Chamala, PhD, Department of Pathology, Immunology and Laboratory Medicine, University of Florida, P.O. Box 100275, Gainesville, FL 32610, United States (e-mail: schamala@ufl.edu).

ACI Open 2020;4:e167–e172.

## Abstract

Patient care is rapidly evolving toward the inclusion of precision genomic medicine when genomic tests are used by clinicians to determine disease predisposition, prognosis, diagnosis, and improve therapeutic decision-making. However, unlike other clinical pathology laboratory tests, the development, deployment, and delivery of genomic tests and results are an intricate process. Genomic technologies are diverse, fast changing, and generate massive data. Implementation of these technologies in a Clinical Laboratory Improvement Amendments-certified and College of American Pathologists-accredited pathology laboratory often require custom clinical grade computational data analysis and management workflows. Additionally, accurate classification and reporting of clinically actionable genetic mutation requires well-curated disease/application-specific knowledgebases and expertise. Moreover, lack of "out of the box" technical features in electronic health record systems necessitates custom solutions for communicating genetic information to clinicians and patients. Genomic data generated as part of clinical care easily adds great value for translational research. In this article, we discuss current and future innovative clinical bioinformatics solutions and workflows developed at our institution for effective implementation of precision genomic medicine across molecular pathology, patient care, and translational genomic research.

**Keywords**

- bioinformatics
- genome
- medical informatics
- pathology
- diagnostic tests
- clinical laboratory information systems
- electronic health records
- molecular testing
- next-generation sequencing
- cancer

## Background and Significance

Traditional genetic testing involves single analyte analyses based on polymerase chain reaction or Sanger sequencing to detect mutations in the human genome. Technological improvements in deoxyribonucleic acid (DNA) sequencing like next-generation sequencing (NGS) and drop in sequencing costs over the past decade have made it feasible to routinely conduct high-throughput genomic testing to investigate large portions of the genome in patient care. NGS-based clinical genomic testing poses several unprecedented challenges to pathology laboratories that include handling high-throughput data, storage, data acquisition costs, analytical complexity and cost, data interoperability standards,

**received**
June 29, 2020
**accepted after revision**
November 4, 2020

**DOI** https://doi.org/10.1055/s-0040-1721480.
**ISSN** 2566-9346.



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

**Table 1** Challenges posed by next-generation sequencing-based clinical genomic testing in molecular pathology laboratories

| Feature | Clinical genomics – next-generation sequencing |
|---|---|
| High-throughput | • Illumina NovaSeq platform can generate 6 Tb of genomic sequence data in 44 hours |
| Storage | • 8–3000 Gb of genomic sequence data per day per NovaSeq sequencer<br>• What data to store?<br>• How long to store?<br>• Enduring continued expense of storage<br>• IT planning and support for storage space and security |
| Data acquisition cost | • Labor-intensive – specimen DNA library preparation process is long and complex<br>• Several days for sequencing run<br>• Up to several thousands of dollars for reagents |
| Analytical cost | • Custom development of custom analysis pipelines<br>• High-performance computing facilities<br>• One to several bioinformatician(s) – expensive |
| Data interoperability standards | • BED, VCF, BAM, CRAM, etc. (fairly well established)<br>• HL7 FHIR Genomics, GA4GH standards (evolving)<br>• LIS integration (evolving) |
| Clinical genomic data reporting and archiving | • Scanned images<br>• Discrete data reporting to EHR<br>• Analytic and clinical validity<br>• Genomic Archiving Communication System (GACS) server<br>• Enterprise genomic database and storage |

Abbreviations: DNA, deoxyribonucleic acid; EHR, electronic health record; HL7, Health Level Seven; IT, information technology; LIS, Lab Information System.

and clinical reporting (►Table 1).[1] Increased complexity of multianalyte testing makes it even harder for clinicians to interpret the analytic validity (e.g., ability of a test to predict the presence of a variant) or clinical validity (e.g., ability of a variant to predict the presence of a disease) of laboratory-generated results. Successful development and deployment of clinical genomic tests at clinical laboratories and delivery of genomic test results into electronic health record (EHR) systems require expertise in clinical bioinformatics, an interdisciplinary field that integrates knowledge of molecular medicine, laboratory medicine, bioinformatics, and health informatics. In this article, we will discuss clinical bioinformatics implementation of University of Florida (UF) Health cancer genetic test panel (GatorSeq) including developing clinical grade genome analysis software pipelines, automated communication of genomic test results into EHR/Lab Information System (LIS), representation of genomic data in EHR/LIS, and genomic data archiving.

## Clinical Grade Genome Analysis Software Development

There are numerous bioinformatics software packages for genome-wide analysis which are themselves changing at a high rate. There are commercial off-the-shelf genomic assays and software solutions. However, they are designed for a broad customer base and have less interest in addressing the needs of a particular laboratory or institution. This lack of flexibility requires most clinical genomic testing laboratories to build their own genomic data analysis pipelines and workflows which integrate appropriate bioinformatics software packages.

Building custom clinical bioinformatics workflows is complex and requires integration of multiple platforms/servers, several dozens of bioinformatics software with their dependencies, and several network storages. For effective operation and maintenance of clinical bioinformatics pipelines, it is critical to adopt software development frameworks that are easily portable, reliable, reproducible, and scalable. To address this, we used Nextflow workflow manager[2] in conjunction with containerization technology (Docker) and version control (Bitbucket) tool (►Fig. 1).[3,4] There have been several hundred workflow managers developed over the past decade[5] with different strengths and purposes.[6] Out of these we chose Nextflow based on ease of use, its compatibility with our computing environments, containerization technologies, version control software, and positive feedback from our colleagues and collaborators who have used it before.

Key components for building a clinical grade bioinformatics software workflow are portability, reliability, reproducibility, and scalability. The principle of portability is the ability to move (easily) software pipeline from one system to another and being agnostic to underlying computing infrastructure[7] (e.g., from local high-performance system to Amazon Web Services). Nextflow allows us to do this by simply changing a configuration file. Reliability is "high probability of failure-free software operation for a specified period of time in a specified environment"[8] and consistently being able to capture failed workflow steps. Nextflow enables reliability by allowing modularization and automation of workflow analysis steps and diagnosis of failed steps. Once diagnosed and corrected, Nextflow allows running from the failed step onward rather than rerunning the whole pipeline.

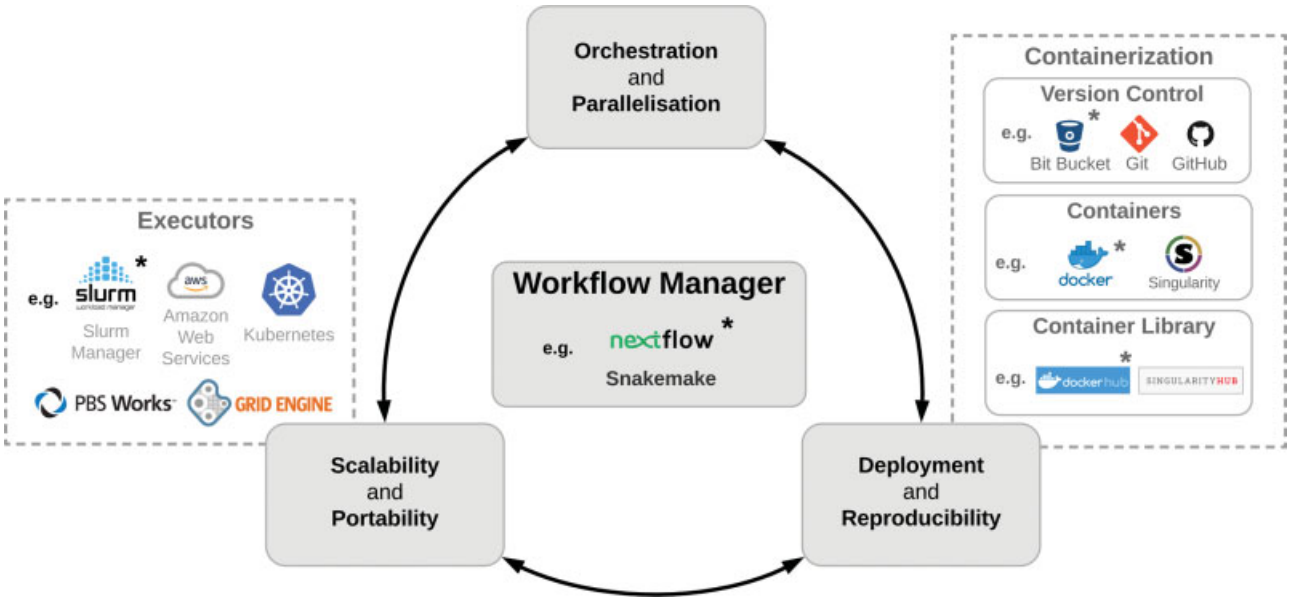


**Fig. 1** Abstraction of underlying execution system using workflow manager (Nextflow) in conjunction with containerization (Docker) and revision control (Bitbucket) tools. The tools marked with asterisk are the ones that we used in our clinical bioinformatics workflows.

Reproducibility is captured by the formula "same data + same analysis = same evidence."[9] All the bioinformatics software and operating system environments in Nextflow are embedded into a Docker[10] container and are hosted on the Docker Hub[11] (►**Supplementary Fig. S1**, available in the online version). Containerization allows replication of software packages, their dependencies, and operating system environments. One can specify a Docker container version/url in the Nextflow configuration file to run upon. Additionally, Nextflow code is version controlled using Bitbucket.[12] To reproduce the analysis one can simply provide Nextflow run command with Bitbucket's Nextflow repository bioinformatics code version number (►**Supplementary Fig. S2**, available in the online version). Nextflow also provides the ability to scale the clinical bioinformatics workflow by increasing or decreasing computing resources by modifying the configuration file without modifying the underlying workflow.[2]

## Clinical Genomics Data in EHR/Lab Information System

Optimal representation/interoperability of genomic data within/between EHR(s) for effective clinical care is still a difficult task. Major challenges in representing genomic variants in EHR are: (1) several hundreds of thousands of genetic variants per patient, (2) changing clinical interpretation guidelines and variant reclassifications, (3) lack of evolved standards and adaptability by EHR/Lab Information System (LIS) vendors, and (4) lack of adequate training for clinicians to manage genetic testing results

### Storing Genomic Variants as Laboratory Values

Currently, the most prevalent practice of reporting from genomic testing laboratories and storing in LIS/EHR is using scanned images or PDF files. For example, at UF Health genomic test results reported from external laboratories are stored as scanned PDF files in Epic's EHR (►**Supplementary Fig. S3**, available in the online version). Limitations of this approach are (1) lack of easy access to data which is typically buried in dozens of patient test orders, (2) lack of searchability for retrospective analysis or reinterpretation, and (3) lack of support for clinical decision making.

We addressed the above limitation of data presentation and accessibility by storing as discrete data fields in the form of laboratory values. This is similar to how complete blood count test panels contain multiple result components like red blood cell count, platelet count, hematocrit, etc. Likewise, we represent each gene as a result component and its corresponding value is the genomic variant. For example, in our cancer gene panel we have 177 genes that are tested. Our laboratory information system (Epic Beaker) has 177 result components each corresponding to one gene (►**Fig. 2A**). The genetic variants in these corresponding genes are reported as values (►**Fig. 2A**). We report only actionable variants and usually only one actionable variant per gene. In the rare circumstance that a gene contains more than one actionable variant, we report them as semicolon delimited values. When the genomic results are signed out in Epic Beaker, only genes with clinically actionable genomic variants will appear in the patient chart (►**Fig. 2B**). The advantage of this method is that (1) results are summarized and easily accessible to the clinician via Result Review, (2) also, a link to full genomic test report is conveniently available in the Result Review section (►**Fig. 2B**), (3) data are searchable for retrospective analysis, and (4) clinical decision support (CDS) is possible at the gene level. The limitations of this approach are that (1) it is cumbersome to handle exome sequencing or large gene panel results, (2) requires lots of scrolling due to multiple entry fields which makes it time consuming to search, (3) manual data entry of genetic variants is still required which makes it prone to error

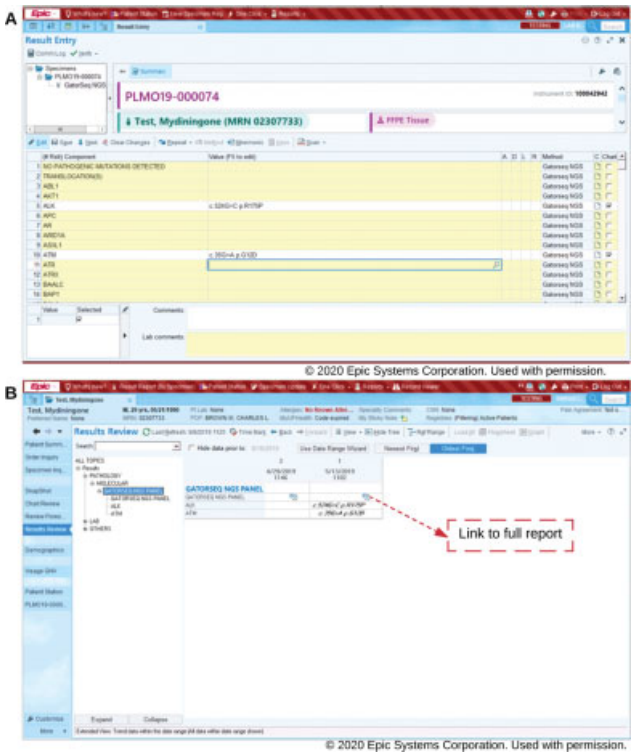


**Fig. 2** (**A**) Genomic result data entry view in laboratory information system (Epic Beaker). (**B**) Genomic results in the electronic health record (EHR) Patient Chart (Epic Care).

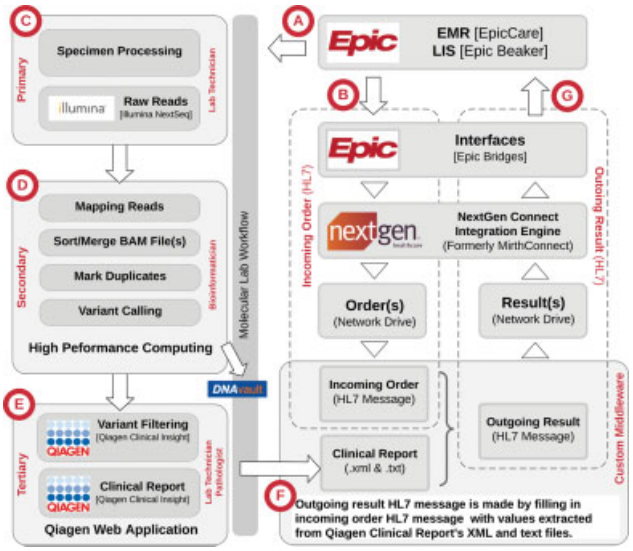


**Fig. 3** Technical overview of cancer gene panel testing workflow including custom middleware developed for automatically interfacing genomic data with Lab Information System (LIS)/electronic health record (EHR).

(this is addressed below via custom middleware solution), and (4) it cannot be used for whole genome sequencing assays as it may result genetic variants in nongenic regions. As our laboratory moves from "traditional" genetic testing to NGS, we are finding the inability to conveniently manage any more than a limited set of genomic findings within the EHR increasingly problematic. On top of this, we are seeing growing interest in genetics by nongeneticists (e.g., primary care providers), who often ask that we provide more concise actionable recommendations. These limitations are prompting us to explore new genomics-EHR integration capabilities, as described in the below sections.

Only clinically actionable genetic variants are being stored into the EHR/LIS systems. For in-house cancer genomic testing, we are retaining all the genetic variants including actionable variants in flat files and also storing them in a queryable in-house genomic database Web application called DNA Vault (see ►**Fig. 3**, Step D and ►**Supplementary Fig. S4**, available in the online version). This is currently being used for quality control purposes by molecular pathology team. We are in plans to build a robust genomic data storage and access infrastructure that could be accessible to both clinicians as well as research community at UF Health.

## Custom Middleware for Cancer Genomic Testing

Clinical laboratory testing instruments at UF Health are typically connected to the Data Innovations (DI) instrument manager (Data Innovations, South Burlington, Vermont, United States), which will autotransmit the results to Epic Beaker via NextGen Connect Integration Engine (formerly MirthConnect). However, not all instruments have DI compatible software drivers. In these cases, a laboratory technologist is required to enter results manually into Beaker LIS, a process which is labor-intensive and error-prone. One such platform without native DI connectivity support is QIAGEN Clinical Insight (QCI) Interpret (Web application hosted on Qiagen's cloud computing infrastructure)[8,9] software that we use for clinical genomic variant interpretation and genomic test reporting. We developed a custom middleware solution for automatically interfacing the genomic test results from the QCI Interpret into EPIC Beaker as laboratory test values represented as described above. Our overall approach is shown in ►**Fig. 3**.

In ►**Fig. 3**, Step A, as soon as genomic testing is ordered and a specimen is collected, it appears on the pathology laboratory worklist in Epic Beaker LIS. We configured our electronic interface software EPIC Bridges and NextGen Connect for simultaneously sending an order Health Level Seven (HL7) message (►**Fig. 3**, Step B) to a network folder. Once the specimen is received in the laboratory the DNA is processed (►**Fig. 3**, Step C). The resulting raw DNA sequencing data are run through our clinical grade custom bioinformatics pipeline on high-performance computing infrastructure (►**Fig. 3**, Step D). This will output genomic variants (Variant Call Format file format) which are automatically uploaded to QCI Interpret. Laboratory technologist(s) and molecular pathologist(s) will use QCI Interpret to shortlist the genomic variants based on their known evidence of clinical actionability. These clinically actionable genomic variants (in accordance with the cancer genomic reporting guidelines provided by the Association for Molecular Pathology, American Society of Clinical

Oncology, and College of American Pathologists[13]) and descriptions of their clinical impact (including associated clinically actionable drugs) are output as XML and text files (►Fig. 3, Step E), which are then processed by our middleware solution, which matches the output to the HL7 incoming order messages, and generates an outgoing HL7 result message (►Fig. 3, Step F). This outgoing HL7 message is placed on the network drive and automatically picked up by NextGen Connect and pushed into Beaker (►Fig. 3, Step G).

## Future Direction and Discussion

To address the limitations of PDF and variants as laboratory results described above, EHR vendors are enhancing their products in anticipation of structured genomic findings (e.g., Epic's genomics indicators), and HL7 Version 2 messaging and HL7 FHIR Genomics reporting standards are maturing. Epic's genomics indicators module has a data structure based on the HL7 2.51 clinical genomics report standard.[14–16] This module allows for the storage of discrete variants which are searchable, and upon which decision support and patient/provider facing information can be accessed. This module does not support storing entire sequences and all genetic phenotypes must be predefined and developed for every result that is returned. At the time of writing there is no support for Infobuttons or other standard API that would allow for access to external knowledge sources for genetic phenotypes. This requires that we develop and maintain these knowledge resources internally. Methodology for reporting and returning negative results is still under development, as the dynamic nature of gene sequencing and interpretation does not currently support a definitive negative result outside of specific variant confirmation.

Large research projects such as eMERGE[17] and CSER[18] are exploring the use of FHIR Genomics, and HL7 FHIR is gaining wide traction, as are apps based on the SMART-on-FHIR platform.[19–22] But FHIR in and of itself will not be a complete solution—NGS can identify thousands to millions of variants, whose clinical significance can change over time as our knowledge evolves. Today's EHRs are not equipped to manage such a large volume of (dynamic) results. One approach being explored to address this latter issue is to store genomic data outside the EHR,[23] in a genomic data server, also referred to as a Genomic Archiving and Communication System (GACS).[24–26] A GACS stores sequence data generated from a sequencing laboratory and is analogous in many ways to a Picture Archiving and Communication System, which stores image files that are not suitable to store directly in an EHR. This trend has led the Office of the National Coordinator's Sync for Genes project to emphasize the need for pilots that test GACS integration with EHRs.[27] HL7 has recently begun formalizing a set of FHIR-based operations, that can serve up normalized genomic data from a GACS, regardless of how that data was natively structured. We are in the planning stages of adopting a FHIR-based GACS solution, which will complement our current approach by enabling population queries, genomic reanalysis, novel genomics-EHR and CDS integration strategies, and several other scenarios.

### Authors' Contributions

All authors participated in the conceptualization and design of the clinical genomics and bioinformatics protocol and workflows described in this manuscript. S.C., S.M., K.N., S.R.G., and R.D. lead the informatics implementation efforts. S.C., N.W., and R.D. took the lead in writing the manuscript with all authors' input. All authors provided critical feedback and approved the final version.

### Funding

None.

### Conflict of Interest

None declared.

### Acknowledgments

We thank Grady Jacobs, Ashley Chandler, Dawn Blood, Greg D. Mullersman, and other members from the UF Health enterprise IT services for their contributions to the development of the EPIC-related informatics implementations detailed in this paper. We thank Tanmay Lele for feedback on this manuscript. The authors gratefully acknowledge Vektra Casler, for contribution to refine some of the figures used in this manuscript.

## References

1 Hart SN. Will digital pathology be as disruptive as genomics? J Pathol Inform 2018;9:27
2 Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35(04):316–319
3 Strozzi F. Scaling bioinformatics analysis using Nextflow and AWS. Accessed February 20, 2020 at: https://www.slideshare.net/FrancescoStrozzi/nextflow-and-aws-batch-gccbosc-2018-104095847
4 Di Tommaso P. Reproducible Computational Pipelines with Docker and Nextflow. Accessed February 20, 2020 at: https://www.slideshare.net/insideHPC/reproducible-computational-pipelines-with-docker-and-nextflow
5 Common Workflow Language Computational Data Analysis Workflow Systems. Accessed June 8, 2020 at: https://s.apache.org/existing-workflow-systems%0A
6 Larsonneur E, Mercier J, Wiart N, et al. Evaluating workflow management systems: a bioinformatics use case. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE 20182773–2775
7 Di Tommaso P. Nextflow meets Dockstore. 2018. Accessed May 26, 2020 at: https://www.nextflow.io/blog/2018/nextflow-meets-dockstore.html%0A
8 Pan J. Software Reliability. 1999. Accessed May 26, 2020 at: https://users.ece.cmu.edu/~koopman/des_s99/sw_reliability/
9 Ryan P, Schuemie M. Reliability, replication and reproducibility: examples and perspectives. 2017. Accessed May 26, 2020 at: https://www.ohdsi.org/wp-content/uploads/2015/04/OHDSI-replication-and-reproducibility-Ryan-Schuemie-31jan2017.pdf
10 Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014;2014:2
11 Docker Inc Docker Hub. 2014. Accessed May 26, 2020 at: https://hub.docker.com/

12 Atlassian Inc Bitbucket. 2008. Accessed November 30, 2020 at: https://www.atlassian.com/software/bitbucket
13 Li MM, Datto M, Duncavage EJ, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017;19(01):4–23
14 HL7 International HL7 Version 2.5.1 Implementation Guide Lab Results Interface(LRI), Release 1, STU Release 3–US Realm. 2017. Accessed June 8, 2020 at: https://www.hl7.org/documentcenter/public/wg/clingenomics/20180403-V2LRI-Ch.5CGandCodeSystemTables.pdf
15 Sutton JA, Wix KK, Amusan AA, et al. Experiences and lessons learned from a new functionality for pharmacogenomics CDS in the electronic health record. 2018. Accessed June 8, 2020 at: https://zerista.s3.amazonaws.com/item_files/1663/attachments/427873/original/amia_summit_2018_poster.pdf
16 Caraballo PJ, Sutton JA, Giri J, et al. Integrating pharmacogenomics into the electronic health record by implementing genomic indicators. J Am Med Inform Assoc 2020;27(01):154–158
17 eMERGE Consortium. Electronic address: agibbs@bcm.edu eMERGE Consortium. Harmonizing clinical sequencing and interpretation for the eMERGE III network. Am J Hum Genet 2019;105(03):588–605
18 Wynn J, Lewis K, Amendola LM, et al. Clinical providers' experiences with returning results from genomic sequencing: an interview study. BMC Med Genomics 2018;11(01):45
19 Alterovitz G, Warner J, Zhang P, et al. SMART on FHIR Genomics: facilitating standardized clinico-genomic apps. J Am Med Inform Assoc 2015;22(06):1173–1178
20 Warner JL, Rioth MJ, Mandl KD, et al. SMART precision cancer medicine: a FHIR-based app to provide genomic information at the point of care. J Am Med Inform Assoc 2016;23(04):701–710
21 Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23(05):899–908
22 Swaminathan R, Huang Y, Moosavinasab S, Buckley R, Bartlett CW, Lin SM. A review on genomics APIs. Comput Struct Biotechnol J 2015;14:8–15
23 Walton NA, Johnson DK, Person TN, et al. Genomic data in the electronic health record. Adv Mol Pathol 2019;2:21–33
24 Dolin RH, Boxwala A, Shalaby J. A pharmacogenomics clinical decision support service based on FHIR and CDS Hooks. Methods Inf Med 2018;57(S 02):e115–e123
25 Masys DR, Jarvik GP, Abernethy NF, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. J Biomed Inform 2012;45(03):419–422
26 Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA 2013;309(12):1237–1238
27 Alterovitz G, Brown J, Chan M, et al. Enabling clinical genomics for precision medicine via HL7 fast healthcare interoperability resources. 2017. Accessed June 10, 2020 at: https://www.healthit.gov/sites/default/files/sync_for_genes_report_november_2017.pdf