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

Srikar Chamala1 Siddardha Majety2 Shesh Nath Mishra2 Kimberly J. Newsom1 Shaileshbhai Revabhai Gothi2 Nephi A. Walton3 Robert H. Dolin4 Petr Starostik1

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

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,
and clinical reporting (Table 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 integrate 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 in conjunction with containerization technology (Docker) and version control (Bitbucket) tool (Fig. 1). There have been several hundred workflows developed over the past decade with different strengths and purposes. 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 (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” 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.

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

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

Abbreviations: DNA, deoxyribonucleic acid; EHR, electronic health record; HL7, Health Level Seven; IT, information technology; LIS, Lab Information System.
Reproducibility is captured by the formula “same data + same analysis = same evidence.” All the bioinformatics software and operating system environments in Nextflow are embedded into a Docker container and are hosted on the Docker Hub (~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. 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.

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.
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) 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\textsuperscript{13}) and descriptions of their clinical impact (including associated clinically actionable drugs) are output as XML and text files (\textsuperscript{-}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 (\textsuperscript{-}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 (\textsuperscript{-}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 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.\textsuperscript{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\textsuperscript{17} and CSER\textsuperscript{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.\textsuperscript{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,\textsuperscript{23} in a genomic data server, also referred to as a Genomic Archiving and Communication System (GACS).\textsuperscript{24–26} A GACS stores sequence data generated from a sequencing laboratory and is analogous in many ways to a Picture Archiving and Communications 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.\textsuperscript{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**

5. Common Workflow Language Computational Data Analysis Workflow Systems. Accessed June 8, 2020 at: https://s.apache.org/existing-workflow-systems%0A


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


Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA 2013;309(12):1237–1238