Implementing semantic interoperability of gene expression profiles using the HL7 FHIR standard

Florian AUER[a,1], Zhibek ABDYKALYKOVA[a], Dominik MÜLLER[a] and Frank KRAMER[a]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Germany*

Abstract. The characterization of diseases using high-throughput omics technologies has increasing relevance for individualized treatment decisions. Especially gene expression profiles can capture significant molecular differences that can foster patient stratification and pave the way towards precision medicine. Electronic health records already evolved to capture genomic data within clinical systems and standards like FHIR enable sharing within, and even between institutions. However, FHIR only provides profiles tailored to variations in the molecular sequence. Although expression patterns are neglected in FHIR, they are equally important for decision making. Here we provide an exemplary implementation of gene expression profiles of a microarray analysis of patients with acute myeloid leukemia using an adaption of the FHIR genomics extension. Our results demonstrate how FHIR resources can be facilitated in bioinformatics-based decision support systems or used for the aggregation of molecular genetics data in multi-center clinical trials.

Keywords. FHIR, interoperability, omics, gene expression

1. Introduction

Measuring the gene expression in patient samples provides more detailed insights into the molecular conditions of the underlying disease. Over the years, high-throughput technologies have evolved to be used in routine clinical diagnostics and foster individualized treatment. At the same time digitalization in healthcare systems advanced to electronic health records (EHR) already capturing genomic data. Interoperability and data sharing between systems and institutions gain more importance with commonly accepted standards like Fast Healthcare Interoperability Resources (FHIR) [1] as a foundation.

FHIR divides the information into modular and extensible components, as well as adapts widely established web standards and the RESTful architecture principle for the sharing of EHRs. Building upon this standard the FHIR Genomics extension enables the inclusion of genomics data, however, the included profiles are tailored to cover variations in the molecular sequence while expression patterns are neglected. Moreover, recommendations for the realization of gene expression results in FHIR are

[1] Corresponding Author: Florian Auer, IT Infrastructure for Translational Medical Research, Alter Postweg 101, 86159 Augsburg, Germany; E-mail: florian.auer@informatik.uni-augsburg.de

lacking. Nevertheless, these insights are important for decision support and translational research.

Here we provide a feasible FHIR implementation for gene expression profiles from a microarray analysis and demonstrate the interoperability of the resulting FHIR resources with an interactive web application.

2. Methods

2.1. Gene expression data

The data set examines a dose-limiting side effect in patients diagnosed with acute myeloid leukemia (AML) that are treated with chemotherapy [2]. In particular mucositis, DNA damage within the oral mucosa caused by the chemotherapy is investigated based on the derived gene expression profiles. The samples are collected from punch buccal biopsies from five AML patients pre- and post-chemotherapy, and three healthy controls for comparison.

Microarray analysis was performed using Human Genome U133 Plus 2.0 Array (Affymetrix, Santa Clara, CA) with GRCh38.p13 (Genome Reference Consortium Human Build 38, Ensembl release 99) as a reference, followed by a Robust Multichip Average (RMA) normalization of the raw data. Subsequently, Linear Models for Microarray data (LIMMA) [3] and Significance Analysis of Microarrays (SAM) [4] was applied to identified genes potentially affected by the presence of AML, or predictive of oral mucositis after chemotherapy. The authors made the data available at the EBI Expression Atlas [5] portal by the ID E-GEOD-10746.

We chose this gene expression data set because the conducted analysis represents a typical bioinformatics workflow resulting in several gene expression profiles from the same and different individuals that enable disease classification and patient stratification into risk groups [6].

2.2. Adaption in FHIR resources

The central element within FHIR to capture real-world concepts in healthcare systems is the *Patient* resource. The study itself evolves around patient treatment therefore all subsequent patient-specific results, and resources implementing those refer to this base element. Detailed information about the sample donors was not included in the original data set to preserve the anonymity of the participants, instead, we used artificially generated data using Synthea™ [7] to create *Patient* resources as reference.

The medical condition of the AML patients was captured by the *Condition* resource to distinguish them from the healthy donors of the control samples. The single samples are captured by the *Specimen* resource and serve as a link to distinguish between samples collected from the same patient, namely pre- and post-chemotherapy.

The gene expression values are generated based on the GRCh38.p13 reference genome and were measured for each sample. Since all gene expression profiles use the same reference database, the single genes contained in the reference genome were included as *MolecularSequence* resources. The actual expression values are treated as single measurements that are stored as *Observation* resources with the *Observation-geneticsGene* extension referring to the corresponding gene symbol. The mapping

details of the data to properties of *MolecularSequence* and *Observation* resources are presented in table 1.

Table 1. Mapping of FHIR resources and corresponding elements from source files.

FHIR resource property	Property type	Table column/source
MolecularSequence		***gff3***
identifier	Identifier	gene_id and Name
type	code (aa | dna | rna)	"dna"
referenceSeq.chromosome	CodableConcept (chromosome-human)	seqid
referenceSeq.orientation	code (sense | antisense)	strand
referenceSeq.referenceSeqId	CodableConcept (Ensembl)	gene_id and ID
referenceSeq.windowStart	integer	start
referenceSeq.windowStart	integer	end
Repository.url	uri	*Ensembl url* + gene_id
Observation		***expression csv***
identifier	Identifier	Gene ID
status	code (ObservationStatus)	"final"
Gene	CodableConcept (HGNC)	Gene Name
extension: observation-geneticsGene		
code	CodableConcept (LOINC codes)	"48018-6"
subject	Reference (Patient)	*Patient ID*
specimen	Reference (Specimen)	*Specimen ID*
derivedFrom	Reference (MolecularSequence)	*Molecular Sequence ID*

Although the *Patient* resource is referenced directly within the *Observation* resource, the *Specimen* resource is still required to differentiate between the different samples of the same patient. Using the *MolecularSequence* resource as a reference avoids redundancy in common genomic information and simplifies the retrieval of the gene expression values for one particular gene across the different samples. An overview of the references between the resources is shown in figure 1.

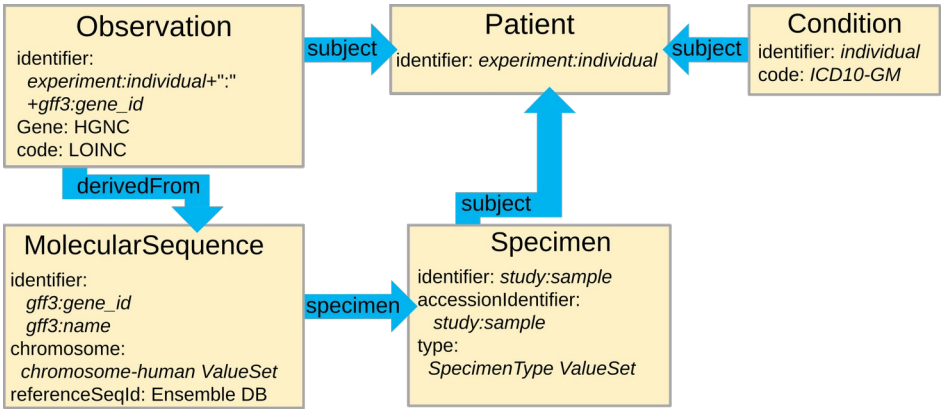

Figure 1. *Overview of the extended FHIR resources and their references between each other as well as to external databases.*

An in-house installation of the dockerized HAPI FHIR server [8] was used for storing the created resources. We developed a web application that uses the FHIR REST API to retrieve and display the FHIR resources to demonstrate a minimalistic decision support system.

2.3. Data and material availability

All necessary software to reproduce the results is publicly available on GitHub at https://frankkramer-lab/gene-expression-on-fhir. This includes scripts to download the used data sets from the official platforms, the setup of the HAPI FHIR server as a docker container and data import and the hosting of the web application. Additionally, the python code for data processing and uploading to the FHIR server, as well as the source code of the web application is provided within the repository. Our code is licensed under the open-source GNU General Public License Version 3 (GPL-3.0 License), which allows free usage and modification for anyone.

Secure hosting of a public FHIR back end was not possible, therefore a demonstration of the web application with hard-coded excerpts of the FHIR resource data is hosted as a static service using the GitHub pages functionality which can be accessed at https://frankkramer-lab.github.io/gene-expresssion-on-fhir.

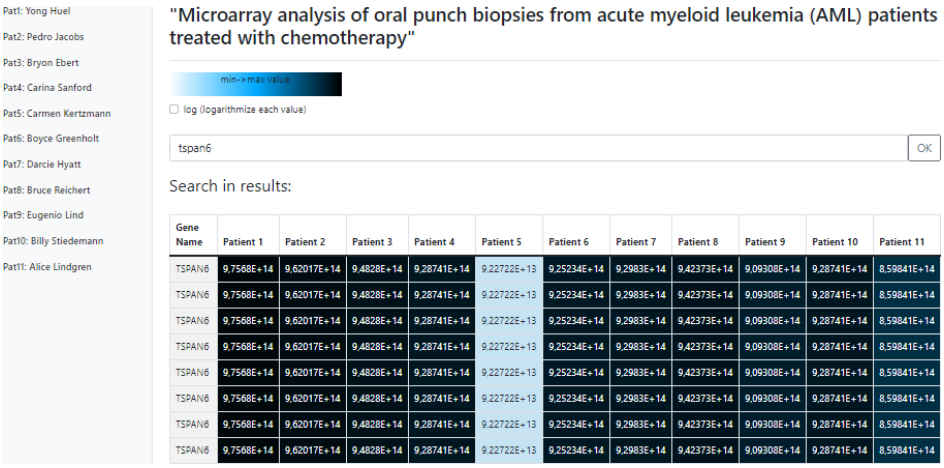

Figure 2. *Web application using the created FHIR resources to show the gene expression between the samples of different patients.*

3. Results

The original data translated to 252,684 resources stored on our FHIR server. For performance improvements, not all gene ids in the reference genome (60,617 ensemble entries) were encoded in FHIR but only those present in the gene expression data. A detailed overview of the created resources and the time requirements is shown in table 2. The web application demonstrates the usage of the created resources obtained directly from a FHIR server.

Table 2. Summary of FHIR resources and time required to upload to the FHIR server

FHIR resource	Number of objects	Time for creation
Patient	8	~1sec
Condition	5	~1sec
Specimen	11	~2sec
MolecularSequence	21,055	~5min
Observation	231,605	~20min

The web application demonstrates the usage of the created resources: Those are obtained directly from the FHIR server, then linked and assembled into a visual representation of the gene expression across the patient samples (figure 1).

4. Discussion

Through our addition to the FHIR Genomics extension, we were able to include genomic profiling) data. Since only excerpts of the molecular data are necessary for detailed investigation, FHIR encoded gene expression profiles are suitable for usage in web-based applications. Furthermore, we were able to demonstrate the integration capabilities of FHIR encoded genomic profiles in decision support systems. Further improvements could consist of consolidation of the outcome of the analyses, e.g. significantly differentially expressed genes between samples, as DioanosticReport resources.

5. Conclusion

Our results demonstrate how FHIR resources can be facilitated for the clinical exchange of expression profiles. These resources also can be used in decision support systems or for patient assessment. The further extension of genomic features in FHIR allows the opportunity to establish the currently missing standard for aggregation of molecular genetics data in multi-center clinical trials. Within this context, patient stratification across health institutions due to omic profiling presents a complex scenario to carry out multi-center clinical trials.

References

[1] D. Bender and K. Sartipi, “HL7 FHIR: An agile and RESTful approach to healthcare information exchange,” in Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems, Jun. 2013, pp. 326–331. doi: 10.1109/CBMS.2013.6627810.

[2] J.-L. C. Mougeot, F. K. Bahrani-Mougeot, P. B. Lockhart, and M. T. Brennan, “Microarray analyses of oral punch biopsies from acute myeloid leukemia (AML) patients treated with chemotherapy,” Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, vol. 112, no. 4, pp. 446–452, Oct. 2011, doi: 10.1016/j.tripleo.2011.05.009.

[3] G. K. Smyth, “limma: Linear Models for Microarray Data,” in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds. New York, NY: Springer, 2005, pp. 397–420. doi: 10.1007/0-387-29362-0_23.

[4] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” Proc Natl Acad Sci U S A, vol. 98, no. 9, pp. 5116–5121, Apr. 2001, doi: 10.1073/pnas.091062498.

[5] M. Kapushesky et al., “Gene Expression Atlas at the European Bioinformatics Institute,” Nucleic Acids Research, vol. 38, no. Database, pp. D690–D698, Nov. 2009, doi: 10.1093/nar/gkp936.

[6] H. Döhner et al., “Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet,” Blood, vol. 115, no. 3, pp. 453–474, Jan. 2010, doi: 10.1182/blood-2009-07-235358.

[7] J. Walonoski et al., “Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record,” Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 230–238, Mar. 2018, doi: 10.1093/jamia/ocx079.

[8] Y. K. R. Sánchez, S. Demurjian, and M. S. Baihan, “Achieving RBAC on RESTful APIs for Mobile Apps Using FHIR,” 2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), 2017, doi: 10.1109/MobileCloud.2017.22.