**Genome Sequencing Report Exchange System**

Hwijun Kwon1, Shinwoong Lee1, Mingyu Kim2, Ilkon Kim\*

School of computer science, Kyungpook National University, Software Technology Research Center (SWRC)

41566, 80, Daehak-ro, Buk-gu

Daegu, Republic of Korea

fjrzlgnlwns@naver.com, tonizlee@naver.com, mingyu3317@gmail.com , ikkimgg@gmail.com

**Abstract**

**Sensational advances in genome analysis techniques such as the Next Generation Sequencing (NGS) have resulted in a large number of personalized medical services that utilize genetic information. However, genome sequencing reports are exchanged in nonstandard format such as PDF, Word, etc. This problem makes it difficult to use genomic information in medical systems. Therefore, to use genomic data in such system, it is necessary to make the genome sequencing report that can be linked to hospital system. we expressed genome sequencing data as Health Level 7 (HL7) Fast Health Interoperability Resource (FHIR) and ISO/TS 20428 To solve this problem. And then we implemented client/server system in the method of FHIR to test genome sequencing reports exchange. The report developed in this study is applicable to real hospital systems and the server/client systems are easier to implement and comply with standards.**

Keywords-component; HL7, FHIR, Genomics, NGS, Genome analysis report, ISO/TS 20428

**I. Introduction**

Thanks to the rapid development of the Next Generation Sequencing (NGS), the time and money spent on genetic testing have dramatically decreased [1]. Advances in these genomic tests have become the basis of the active use of genetic information in personalized medical treatments such as Precision Medicine (PM) [2].

PM aims at personalized medical care. To achieve PM's goal, it is needed to construct a genomic data cohort. A cohort is a sample collected according to a specific research purpose. For this reason, many countries that want to introduce PM system are collecting genomic data to make cohort. In this situation, the use of genome data is likely to increase. Figure 1 shows estimated usage and capacity of genomic data. It can be seen that the genomic data availability is extremely limited comparing with the data that can be produced.

The reasons for this problem are privacy, complexity of data, sponsor’s interest, difficulty of analysis and so on [3,4]. One of the most important reasons is interoperability on genome sequencing report.

|  |
| --- |
|  |

Fig. 1 Estimated size of Human genome sequencing capacity and avail-able sequencing data in dbGaP-Genomic data Sharing repository [3].

The process of Genome Sequencing report can be divided by 6 steps ( (1)Data production–(2)Processing and event detection –(3)Filtering, review and verification–(4)Annotation and functional prediction-(5)Interpretation and report generation- (6)Clinical application) [4]. Data generated in each process needs to be stored in a standardized manner to share and analyze. The overall data from each process is well standardized in many studies (e.g. Variant notation, Interpretation, Data integration, etc.) [6, 7, 8, 9].

Although there are various standards for expressing the results of sequencing, there is no technical standard for expressing structured form that can link patient clinical data and genome analysis results data in Electronic Medical Record (EMR) system. Most standards were exchanged through files such as PDF [5]. Even in the case of GSVML, it is difficult to develop system and get feedbacks.

This problem can cause interoperability problems. It is hard to link this report to the hospital information system. Even if it is linked, it cannot be used in system such as clinical support system [5]. Such a document forms, therefore, degrade the interoperability of hospital system.

So, in this study, we suggest standard genomic data format that can be linked in EMR system by using FHIR and ISO/TS 20428 and implement a system that can exchange this document with interoperability.

**II. Data Standardization**

The genomic analysis data elements defined in ISO/TS 20428 (2016) is largely divided into two parts, *summary* and *detailed contents*. Since many portion of summary contents overlap detailed contents, we just standardized *detailed contents*. All FHIR Resources were made in DSTU Version 2.

*A.* *Data Analysis*

The *detailed contents* are divided into *required fields* and *optional fields*. *Required fields* include key information related to genomic analysis data, such as sequencing information. *Optional field* contains subsidiary information that are not related to sequencing data such as medical examination history and racial information. Table 1 shows the data element presented by the two documents.

TABLE I

Data elements of required fields and Optional fields [10]

|  |  |
| --- | --- |
| **Required Fields** | **Optional Fields** |
| Clinical sequencing orders | Family History/Pedigree information |
| Information on subject of care | Reference sequence |
| Information of legally authorized person ordering clinical sequencing | Conditions of specimen that may limit adequacy of testing |
| Performing laboratory | Racial genomic information |
| Associated disease and phenotype | Detailed sequencing information |
| Biomaterial information | References |
| Genetic variations |  |
| Classification of variants |  |
| Recommended treatment |  |

*1) Required fields data*

The data elements of the *required fields* shown table 2 are defined in Table 2 of ISO / TS 20428 and each data element is matched to standardized metadata.

TABLE Ⅱ

Data elements and their metadata for required fields [10]

|  |  |
| --- | --- |
| **Data elements** | **Metadata** |
| **Clinical sequencing orders** | Clinical sequencing order code | Order code | LOINC |
| Information on sequencing order | TEXT |
| Data and time  | Order date | ISO 8601 |
| Specimen collection |
| Order received data |
| Report date |
| Addendum creation date |
| Specimen information | ISO/TS 22220:2011 |
| **Information on subject of care** | Identifiers | ISO/TS 22220:2011 |
| Name |
| Birth date | ISO 8601 |
| Sex | ISO/TS 22220:2011 |
| Ethnicity | HL7 v3 Code System Race |
| **Information of legally authorized person** **ordering clinical sequencing** | ISO/TS 27527:2010 |
| **Performing laboratory** | Basic information | TEXT |
| Information of report generator | TEXT |
| Information of legally confirmed person on sequencing report | ISO/TS 27527:2010 |
| **Associated disease and phenotype** | ICD |
| **Biomaterial information** | Type of sample | SPREC |
| Genomic source class in biomaterial | LOINC |
| Conditions of specimen | TEXT |
| **Genetic variations** | Gene symbols and names | HGNC |
| Sequence variation information | Notation | HGVS |
| Effects of variants | TEXT |
| Sequence variant ID | Database unique ID |
| **Classification of variants** | Pathogeny | ENUM(“Pathogenic”, “likely pathogenic”, “Unknown significance”, “likely benign”, “Benign”)) |
| Clinical relevant | ENUM(“Identified”, “Likely identified”, “Uncertain”, “Not identified”) |
| **Recommended treatment** | Medication | ISO 11615 |
| Clinical trial information | Clinical trial ID |
| Known protocols related to a variant | TEXT |
| Other recommendation | TEXT |

*2) Optional fields data*

ISO / TS 20428 defines the data elements of the *optional fields* as shown in Table 3.

Table Ⅲ

Data elements and their metadata for optional fields [10]

|  |  |
| --- | --- |
| Data elements | Metadata |
| Medical history | ICD |
| **Family history/Pedigree information** | HL7 v3 IG: Family History/Pedigree Interoperability |
| **Reference genome version** | Genome Regerence Consortium Human Genome release ID |
| Racial genome information | TEXT |
| **Genetic variation** | Gene symbols and names | HGNC |
| Sequence variation information | Notation | HGVS |
| Effects of variation | TEXT |
| Sequence variant ID | Database unique ID |
| HGVS version | HGVS version number |
| **Detailed sequencing information** | Clinical sequencing date | ISO 8601 |
| Quality control metrics | NUMERIC |
| Base calling information | Read depth | NUMERIC |
| Reference allelic depth |
| Alternative allelic depth |
| Allele frequency |
| Genotype |
| Sequencing platform information | Type of sequencers | TEXT |
| Library capture methods |
| Target capture methods |
| Read type | ENUM(“single-end”, “paired-end”) |
| Read length | TEXT |
| Analysis platform information | Alignment tools | TEXT |
| Variant calling tools |
| Other tools |
| Chromosome coordination system | ENUM(“zero-based”, “one-based”, “half-open”) |
| Annotation tools and databases | TEXT |
| **Reference** | TEXT |

*\*ENUM represents the contents should be chosen among the given category*

*B. Standardization Of Sequencing Report*

 Based on the data identified in data analysis section, we made FHIR resources with the *required fields* and *optional fields* data.

Selecting base Resource is the first step. The criteria for choosing a base resource is how many base resource can represent the data element and how similar is the purpose of use of the FHIR resource to that of the document.

 After selecting base resource, other resources are selected to represent remaining data element. Base resource is linked with other resources by using the Uniform Resource Identifier (URI). If there is any data elements that has not been expressed after such process, it must be expressed in FHIR extension element that is user-definable. This process is called profiling.

*1) Required Fields Profiling*

Fig. 1 shows profiling result of *required fields*. Order resource is selected as base resource to express clinical sequencing order information. And then, to express *information on subject of care*, *legally authorized person ordering clinical sequencing*, performing *laboratory* and other data elements, we used patient, *practitioner* and *organization*, user-defined extension resource respectively and connected them to the order resource through URI.

*2) Optional Fields Profiling*

 Fig. 2 shows the structured data elements of the *optional fields* as FHIR resource through profiling. *Optional fields* represent *family history/pedigree information* element by designating *family member history* resource as base resource. Other data elements such as *reference genome version* and *racial genome information* are expressed in family member history resource through extension.

Fig. 1. Required fields profiling

Fig. 3. A simple scenario for exchanging sequencing report



Fig. 2. Optional fields profiling

**III. Genome Sequencing Report Exchange System Implementation**

In this study, we implemented a system to exchange sequencing reports that were standardized previous section.

Prior to implementing the whole system, we designed a server / client system to analyze the scenario of exchange the sequencing report.

*A. Genome Sequencing Exchange Scenario Analysis*

In this process, there are three actor (EMR system, hospital system, and a sequencing facility). Fig. 3 shows a simple scenario for exchanging sequencing report. In the whole process of sequencing report exchange, a clinician asks for a genetic test to hospital in the EMR system first. The hospital receives the request and delivers it to the sequencing facility. And sequencing facility performs the sequencing test and sends the sequencing reports to FHIR server that may be in the hospital system. FHIR server store the report and notify hospital system. Finally a EMR client system receive the report from FHIR server.

As a result of this sequencing report transaction scenario, we design two client systems for EMR system and sequencing facility system and one FHIR server for FHIR resource exchange. It is very difficult to replace the existing hospital system with a new server in medical institutions such as hospital. For this reason, the FHIR server is designed not to replace the existing hospital server but to function as a module.



*B. Client Implementation*

 In this study, we develop EMR system and Sequencing facility system client system by using WEB programming language (JavaScript). All transmission processes are performed through Representational State Transfer (REST) API. Table 4 shows REST API used for exchanging sequencing report.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4 REST API used for exchanging sequencing report

|  |  |
| --- | --- |
| **Interaction** | **Message** |
| **request** | POST https://{endpoint url}/request/{requestType} |
| **create** | POST https://{endpoint url}/{resourceType} |
| **read** | GET https://{endpoint url}/{resourceType}/{id} |
| **update** | PUT https://{endpoint url}/{resourceType}/{id} |
| **delete** | DELETE https://{endpoint url}/{resourceType}/{id} |

 |

*1)* *EMR client*

Fig. 4 shows two functions of the EMR client system. The first one is the function of sending the request message to EMR system by using the REST method. The second one is the function of receiving the sequencing report from the FHIR Server.



Fig. 4 Function of EMR client system

*2) Sequencing Facility Client*

 Fig. 5 shows the function of sequencing facility client system. When a request message is sent to from an existing EMR server, sequencing facility client system creates FHIR resources based on the sequencing data extracted from the device. The sending report functions bind together these resources generated and deliver to the FHIR sever via POST message.



*C. FHIR Server Implementation*

Fig. 5. Function of sequencing facility client system

The FHIR server provides REST APIs. The client system communicates with the server through the REST APIs that the server defines. Through REST APIs provided by FHIR servers, the client can delegate necessary business services (Analysis) to the server.

 Through these business service, server performs directly interaction such as create, update, delete, read and so on as to FHIR resource in the DB. Fig. 6 shows a simple structure of the FHIR Server.

|  |
| --- |
| Fig. 6 Structure of FHIR Server |

**IV. Result**

 In this study, we analyzed the data elements of the ISO / TS 20428 standard and assign them to the FHIR resource. And we designed a system that can exchange these resources considering the actual usage environment. Based on this, the whole system was developed. We did not proceed with the analysis using the sequencing data, but confirmed that all the data is stored in the server formally and can be linked with the EMR system. In addition, since the developed FHIR server minimizes the change of the existing EMR system, it is easy to develop, easy to maintain and compatible with each other.

**V. Conclusion**

With advances in sequencing technology, genomic data has become essential data in clinical practice. Under these circumstances, it is important to develop a digitalized sequencing report that can be linked with EMR system. For this reason, this study developed FHIR resource with ISO/TS 20428 and designed FHIR server / client system for exchanging and using this FHIR resource and it worked well.

The FHIR server was developed as a modular system rather than modifying the entire existing medical system. When standardizing hospital information systems, such a functioning system can reduce developmental difficulties and economic burdens.

**Acknowledgment**

This study was supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005) and the MIST(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion)(2015-0-00912) and the Technology Innovation Program (10053584, Standardization of Structured Dielectric Analysis Test Report for Electronic Medical Records) funded By the Ministry of Trade, industry & Energy(MI, Korea).

**References**

[1] Joseph Henson, German Tischler, and Zemin Ning, “Next-generation sequencing and large genome assemblies”, 13(8): 901–915, 2014 March 20.

1. FDA, “The Precision Medicine Initiative”, <https://www.fda.gov/ScienceResearch/SpecialTopics/PrecisionMedicine/default.htm> .
2. Kovalevskaya NV, Whicher C, Richardson TD, Smith C, Grajciarova J, Cardama X, et al, "DNAdigest and Repositive: Connecting the World of Genomic Data", PLoS Biol 14(3): e1002418.doi:10.1371/journal.pbio.1002418, 2016 March 24.
3. Benjamin M Good, Benjamin J Ainscough, Josh F McMichael, Andrew I Su and Obi L Griffith, “Organizing knowledge to enable personalization of medicine in cancer”, Genome Biology, 2014 Aug 15.
4. Suyong Shin, “ISO/TS 20428:Structured Clinical Genome Analysis Report”, KOSMI(Korean Society of Medical Informatics) presentation,2017 Jun 23.
5. Nakaya, J. Genomic Sequence Variation Markup Language (GSVML). International Journal of Medical Informatics 79th volume(pp. 130-142), 2010 February.
6. ISO/NP 21393, Omics Markup Language (OML), 2016 Apr
7. ISO/NP 25720, Whole genome sequence markup language (WGML), 2016 Apr
8. Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology", 17(5): 405–424., 2015 November 01.
9. Shin Soo-Young. ISO/TS – 20428:2017 Health informatics -- Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records. International Organization for Standardization(ISO) Technical Committees(TC) 215; Health informatics, 2017 Mar.
10. Carlos Marcos, Arturo Gonzales, Mor Peleg, Carlos Cavero, “Solving the interoperability challenge of a distributed complex patient guidance system: a data integrator based on HL7's Virtual Medical Record standard”, 22:587–599, 2015 April 6
11. Mehdi Kchouk, Jean-François Gibrat, Mourad Elloumi, “An Error Correction Algorithm for NGS Data”, 10.1109/DEXA.2017.33, 2017 Aug.