Genomic Sequence Variation Markup Language (GSVML) for Global Interoperability of Clinical Genomics Data

Jun Nakaya[a], Keisuke Ido[a], Kaei Hiroi[a], Woosung Yang[a], Michio Kimura[b]
[a] *Graduate School of Medicine, Kobe University, Kobe, Japan*
[b] *School of Medicine, Hamamatsu University, Hamamatsu, Japan*
junnaka@med.kobe-u.ac.jp

Abstract

Objective: Aiming to make good use of internationally accumulated genomic sequence variation data that is a result of the explosive amount of recent genomic researches, the development of the interoperable data exchanging format and its international standardization are the demands. The GSVML is focusing on genomic sequence variation data and human health application domain such as the gene based medicine or the pharmacogenomics.
Design and Method: We developed the GSVML through eight steps having the use case analysis and the domain investigations. By focusing on the design scope to human health application and genomic sequence variation, we tried to eliminate the ambiguity and to give the practicability. Basically we intended to follow and to satisfy the requirements derived from the use case analysis of human based clinical genomic applications. By the database investigations, we tried to minimize the redundancy of the data format, while we tried to maximize the data covering range. We tried to have communication ability and interface ability to the other markup languages with intention to exchange various omics data among different kinds of omics researchers or facilities. The interface ability with the developing clinical standards such as the Health Level Seven Genotype information model were analyzed in detail.
Results: The Genomic Sequence Variation Markup Language (GSVML) is developed. The GSVML is human health oriented and has three categories as variation data, direct annotation, and indirect annotation. The variation data category is the required category, while the direct annotation category and the indirect annotation category are optional. The annotation categories contain the omics and clinical information and have internal relations. Six use cases and three criteria in human health application were examined. Eleven data elements and three criteria are examined as the data format for genomic sequence variation data exchange. The data format of five international SNP databases and six markup languages were investigated. An interface ability to Health Level Seven Genotype Model was investigated in terms of 317 items.
Conclusion: The development of the data exchanging format for the genomic sequence variation data exchanging with focusing on human health application including researches is the demand. The GSVML can be the pinpointed answer for this demand. The international standardization of the GSVML is also the demand and is in progress. The GSVML has an ability to enhance the genomic sequence variation data utilization internationally by providing a standardized platform for both clinical and research applications.

1. Introduction

In current electronic world, we have been dealing with clinical data and image data as major data types for healthcare. Storming into the post genomic era, we are urged to handle the additional huge and internationally overswollen genomic data additionally besides clinical data and image data. The huge number of experimental data from the recent explosive amount of sequence variation researches especially SNP researches produced the overswollen data and many databases with various types of data formats world widely. Nowadays there are so many kinds of omics data having relation with the sequence variation data around the world and they are waiting for their effective utilization especially for human health. To utilize them truly, opening the way to exchange sequence variation data around the world independently from the types of data format is the entry point and is the first hurdle. Providing the globally interoperable data format is in urgent demand for managing, analyzing, and utilizing these genomic data.

This paper intends to show the development process and to provide the outline of the interoperable data exchanging format for genomic sequence variation data as a Markup Language in human health domain without forcing to change any existing database schema with an eye to the international standardization.

As the genomic data, we have genome sequence, genomic sequence variation, and other genome based data as expression data, proteomics data, molecular network, etc. In this paper, we are focusing on the genomic sequence variation. Among the genomic sequence variation, we started with SNP and gave priority to it because of the following three reasons.

1. SNP is the most researched sequence variation for human health.
2. In the current context, SNP data is overswollen around the world with various types of data formats.
3. SNP data already has a great impact for human applications as the Gene Based Medicine and the pharmacogenomics.

Considering that genomic sequence variation especially SNP has its significant meaning in the gene based medicine and the pharmacogenomics for human health domain, the sequence variation data exchanging format is the key to enhance the gene based clinical research and the gene based medicine. In the present circumstances, SNP is expected to be a key to understand human response to the external stimuli such as alien invasions, therapies, and the environmental interactions [1]. In case of alien invasion, the bacterial infection is an example and a response to an infection is different among individuals. In case of therapy, the degree of a side effect of a drug is different among patients. These responses are also different at various environments. Considering that SNP is the major and simple polymorphism in human health genomic research domain, setting a starting point to it and expanding the specification from it to the other sequence variation data will be reasonable. Under such circumstances, the informational approaches to the variation data are getting more important to support genomic research and genomic medicine [2]. Starting with SNP, we widened the covering range of GSVML from SNP to the other sequence variation data. GSVML is focusing on DNA sequence variation data in human health domain. To utilize the accumulated variation data among many facilities around the world, the point of the data exchanging is in the messaging and its interoperable data format. In the current context, Markup Language is the reasonable choice to make an answer for this need.

The Markup Language is a set of symbols and rules for their uses when doing a markup of a document [3]. As for Markup Language development, the first standardized markup language was SGML (standard generalized markup language) [4] which has strong similarities with troff and nroff text layout languages supplied with Unix systems. HTML (Hypertext Markup Language) was based on SGML [5]. XML (Extensible Markup Language) is a pared-down version of SGML, designed especially for Web documents [6]. XML acts as the basis for XHTML (Extensible HTML) [7] and WML (Wireless Markup Language) [8] and for standardized definitions of system interaction such as SOAP (Simple Object Access Protocol) [9]. By contrast, text layout or semantics is often defined in a purely machine-interpretable form, as in most word processor file formats [10]. Markup Language for the biomedical field based on XML is in developing for these decades to enhance the exchange data among researchers. BSML (Bioinformatic Sequence Markup Language) [11], SBML (Systems Biology Markup Language) [12], Cell ML (Cell Markup Language) [13], Neuro-ML (Neuro Markup Language) [14] are the examples. Polymorphism Mining and Annotation Programs (PolyMAPr) [15] is centric on SNP and tries to achieve mining, annotation, and functional analysis of public database as dbSNP [16], CGAP [17], and JSNP [18] through programming.

As a result of recent explosive amount of SNP researches, the huge number of experimental SNP data have been accumulated in many databases with various types of data formats and are waiting for utilization in drug discovery, clinical diagnosis, and clinical researches. We investigated the diversity of databases of dbSNP, JSNP, CGAP, and ALFRED [19]. The international standardization organizations like Health Level Seven (HL7), International Standardization Organization (ISO), DICOM and JPEG have been keeping efforts to standardize the medical data. GSVML project is in the process of the standardization at ISO TC215 [20]. GSVML project also has a relation with HL7 efforts as for genomic data message handling. Health Level Seven Clinical Genomics Special Interest Group (HL7 CG SIG) [21] is summarizing the clinical use cases for general genomic data. We investigated the interface ability of GSVML with HL7 genotype model [22] and the SNP use case of the Japanese millennium project [23] at the HL7 CG SIG

Based on these contexts and investigations, this paper elucidates the requirements, the issues, and the development process of the GSVML with an eye to the international standardization. Then this paper shows the design and the specification of the GSVML.

2. Design and Method

2.1. Eight Steps Development

We developed the GSVML through following eight steps:

Step 1: We set the required elements and the needs according to the use case analyses. Prior to the development, we elucidated the specification needs for genomic sequence variation data exchanging among the human health facilities. We set the elements and the issues according to the classified use cases. To clear up the elements and the issues that we should solve, we limited the use case of SNP data in the human health related fields. Here the human health related fields indicate the human applications including clinical practice, clinical trial, and translational research. We prepared six use cases for three typical criteria. Four use cases are concerning about the clinical practice, and one use case for each clinical trial and translational research.

Step 2: We designed the initial basic structure and DTD

Step 3: We investigated the applicability of existing Markup Languages to the sequence variation data derived requirements. Investigated markup languages were MAGE-ML, BSML, SBML, Cell ML, and RNAML [24] to the use case needs. We also investigated the PolyMAPr program that is not a markup language but a SNP centric analyzer. Comparing with existing Markup Languages, we set the position of GSVML in the domain of the Markup Language.

Step 4: We refined the initial structure and DTD.. And we designed the XML Schema as more validated information model.

Step 5: We investigated the existing SNP databases. By investigating the existing SNP databases data formats, we confirmed the specification needs for data exchanging format of SNP data. We extracted the sharable format based on the data format diversity investigation. By extracting the common format and their characteristics, we try to trim the fat of data format off. Investigated databases were JSNP, dbSNP, HGVBase, ALFRED, and Human SNP Database.

Step 6: We checked the interface ability to the Health Level Seven Genotype Model. As Health Level Seven is one of the good choices of the clinical standardizations, we checked an interface ability of GSVML to HL7 genotype informational model. To use GSVML actually both in clinical application and in clinical research application, having an interface ability to HL7 will be the good user-friendly point.

Step 7: We redefined the needs to GSVML and its demanded elements. Based on the above investigations, we set the position of GSVML in the field of the Markup Language and also we cleared up its data specification needs.

Step 8: We refined the initial structure, DTD, and XML Schema.

We have done the design work in collaboration with HL7 CG SIG. Aspiring to become the ISO standard, "to and fro" processes between design work, measurement work, and the standardization process were needed.

2.2. Design Principles

By focusing on the design scope to human health application and genomic sequence variation, we tried to eliminate the ambiguity and to give the practicability. Basically we intended to follow and to satisfy the requirements derived from the use case analysis for clinical genomic applications. Based on this standing point, we designed the initial version of GSVML that was called as PML (Polymorphism Markup Language).

To satisfy the data exchange needs among the variation databases, we tried to minimize the redundancy of the data format while we tried to maximize the data covering range for the databases. This compelled to optimize the size and the specification of the sharable data format of GSVML.

To have communication with other markup languages in biomedical field with intention to exchange various omics data among different kinds of omics researchers or facilities, we tried to have communication ability and interface ability to the other markup languages. These efforts modified the GSVML data structure and improved the interface ability.

Considering that the use cases of the GSVML are mainly in clinical fields, the interface ability with the existing clinical standards such as the Health Level Seven is important. We tried to give GSVML an interface ability to the Health Level Seven Clinical Genomics SIG Genotype information model.

3. Measurement

3.1. Use Case Analysis

We analyzed the use cases and the required elements, and the indicated elements derived from these use cases and requirements for the GSVML. Here some use cases are supposed. In case of SNP application, the SNP associated genes are in the SNP annotation. The clinical information and observations are included in the clinical annotation of indirect annotation. All kinds of omics data including proteomics data are included in the omics annotation of indirect annotation. The demands to these elements are different among the use cases. As an example, the Omics annotation as indirect annotation is necessary to the great extent possible for the Gene therapy among MD and other paramedicals.

In translational research, the SNP data is exchanged among hospitals, research institutes, MDs, researchers, and the pharmaceutical company. In this case, individual SNP data is sent/received with individual clinical data and the other additional data that are needed to specify the experiment and the patients. The number of the demanded data is almost several dozen, while the parameters for each individual are many.

In clinical trial, the number of the demanded data depends on the clinical phase. Early phases do not need many individuals but need many parameters, while late phases are in opposite.

In clinical practice, individual SNP data is sent/received with individual clinical data. For more advanced diagnosis, individual genomic data including omics data are demanded. For the prescription derived from pharmacogenomics, the SNP data will not be exchanged in most cases. The exchanging data will be the prescription, reasons, and its annotations. For gene therapy, individual SNP data is sent/received with individual clinical data and individual genomic data. For disease prevention based on the individual polymorphism, individual SNP data is sent/received with individual clinical data.

In the general Use Case of GSVML in clinical scene, every actor such as MD, Hospital can exchange data smoothly without forcing to change their existing database schema through GSVML. In the same way, the researches can exchange their genomic sequence variation data without any pain. As an example, in case of genetic diagnosis, the individual SNP data is exchanged among the facilities as hospitals, Medical Laboratories. This data is also exchanged among the persons such as MDs, Laboratory Analysts, Counselors, and in some case patient him/herself. Here individual SNP data is encapsulated with the individual clinical data and his/her omics data in some cases for further examination. To analyze this individual SNP data, the individual SNP data need to be compared with the database derived SNP data that have various types of data formats.

Figure 1 is an example work flow of the SNP analysis [25]. This is the case of the "Japanese National Millennium Project". This project tries to find the dominant SNPs or genes for the 5 Life Style Diseases. This is the case for the Diabetes Mellitus. In this case SNP data and additional information are exchanged among the facilities. Here we have not only the SNP data but also the clinical data and SNP annotations as the clinical data, the Omics annotations, and the environmental data.

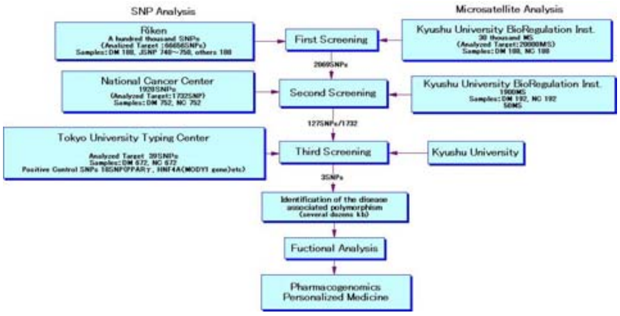
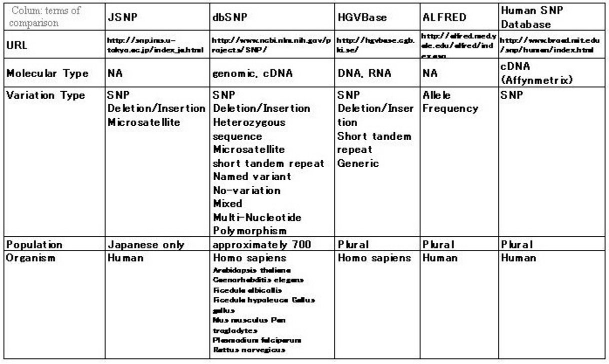


Figure 1. An example work flow of SNP analysis.

3.2. Diversity of SNP Databases

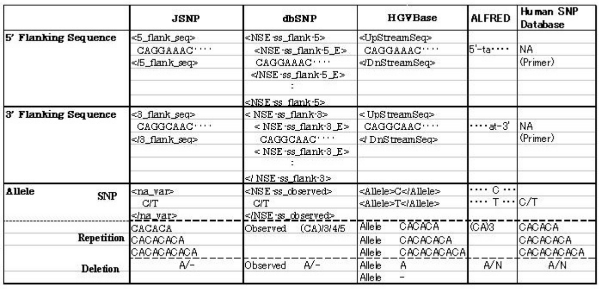
Table 1 shows the results from the diversity analysis among the international existing databases. As an example of the molecular type, each databases uses the word "cDNA" or "RNA". They have almost the same meaning in the way of the sequence, while the experimental preparation is different. As an other example of the Organism, the Homo sapiens and the human have almost the same meaning, while the representations is different.

Table 1. Diversity analysis of international databases. (quoted from [25])



Colum: terms of comparison	JSNP	dbSNP	HGVBase	ALFRED	Human SNP Database
URL	http://snp.ims.u-tokyo.ac.jp/index_ja.html	http://www.ncbi.nlm.nih.gov/projects/SNP/	http://hgvbase.cgb.ki.se/	http://alfred.med.yale.edu/alfred/index.asp	http://www.broad.mit.edu/snp/human/index.html
Molecular Type	NA	genomic, cDNA	DNA, RNA	NA	cDNA (Affynmetrix)
Variation Type	SNP Deletion/Insertion Microsatellite	SNP Deletion/Insertion Heterozygous sequence Microsatellite short tandem repeat Named variant No-variation Mixed Multi-Nucleotide Polymorphism	SNP Deletion/Insertion Short tandem repeat Generic	Allele Frequency	SNP
Population	Japanese only	approximately 700	Plural	Plural	Plural
Organism	Human	Homo sapiens Arabidopsis thaliana Caenorhabditis elegans Ficedula albicollis Ficedula hypoleuca Gallus gallus Mus musculus Pan troglodytes Plasmodium falciparum Rattus norvegicus	Homo sapiens	Human	Human

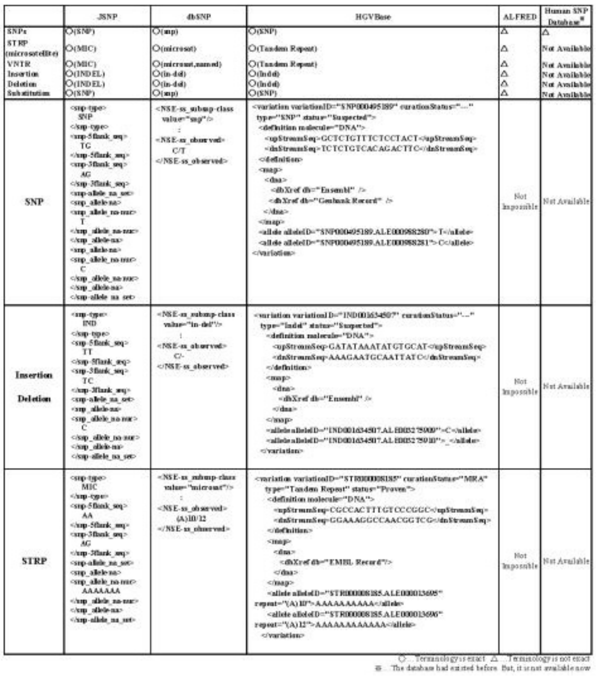
Table 2. Diversity of data representation in SNP databases. (quoted from [25])



	JSNP	dbSNP	HGVBase	ALFRED	Human SNP Database
5' Flanking Sequence	<5_flank_seq> CAGGAAAC···· </5_flank_seq>	<NSE-ss_flank-5> <NSE-ss_flank-5_E> CAGGAAAC···· </NSE-ss_flank-5_E> : <NSE-ss_flank-5>	<UpStreamSeq> CAGGAAAC···· </DnStreamSeq>	5'-ta····	NA (Primer)
3' Flanking Sequence	<3_flank_seq> CAGGCAAC···· </3_flank_seq>	< NSE-ss_flank-3> < NSE-ss_flank-3_E> CAGGCAAC···· < NSE-ss_flank-3_E> : </ NSE-ss_flank-3>	< UpStreamSeq> CAGGCAAC···· </ DnStreamSeq>	····at-3'	NA (Primer)
Allele SNP	<na_var> C/T </na_var>	<NSE-ss_observed> C/T </NSE-ss_observed>	<Allele>C</Allele> <Allele>T</Allele>	···· C ··· ···· T ···	C/T
Repetition	CACACA CACACACA CACACACACA	Observed (CA)/3/4/5	Allele CACACA Allele CACACACA Allele CACACACACA	(CA)3	CACACA CACACACA CACACACACA
Deletion	A/-	Observed A/-	Allele A Allele -	A/N	A/N

Table 2 shows the diversity of data representation in the SNP databases. The Row represents the international SNP databases. The Column represents the comparison terms. There are also many diversities in representation for the SNP data. As an example, the Representations for the 5' and 3' flanking sequences are completely different among the SNP databases. As an other example, the representations for the Allele about the SNP representation, repetition representation, and the deletion representation are different among the SNP databases. To exchange the data efficiently among these databases internationally, we need to standardize these data exchanging format from these data representation level.

Table 3. Diversity of sequence variation data representation in variation databases. (quoted from [25])



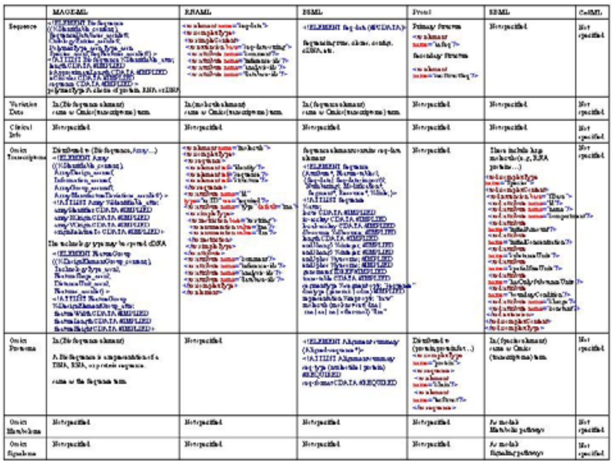
	JSNP	dbSNP	HGVBase	ALFRED	Human SNP Database※
SNPs	○(SNP)	○(snp)	○(SNP)	△	△
STRP (microsatellite)	○(MIC)	○(microsat)	○(Tandem Repeat)	△	Not Available
VNTR	○(MIC)	○(microsat,named)	○(Tandem Repeat)	△	Not Available
Insertion	○(INDEL)	○(in-del)	○(indel)	△	Not Available
Deletion	○(INDEL)	○(in-del)	○(indel)	△	Not Available
Substitution	○(SNP)	○(snp)	○(SNP)	△	Not Available
SNP	<snp-type> SNP </snp-type> <snp-5flank_seq> TC </snp-5flank_seq> <snp-3flank_seq> AG </snp-3flank_seq> <snp-allele_na_set> <snp_allele-na> <snp_allele_na-nuc> T </snp_allele_na-nuc> </snp_allele-na> <snp_allele-na> <snp_allele_na-nuc> C </snp_allele_na-nuc> </snp_allele-na> </snp-allele_na_set>	<NSE-ss_subsnp-class value="snp"/> : <NSE-ss_observed> C/T </NSE-ss_observed>	<variation variationID="SNP000495189" curationStatus="---" type="SNP" status="Suspected"> <definition molecule="DNA"> <upStreamSeq>GCTCTGTTTCTCCTACT</upStreamSeq> <dnStreamSeq>TCTCTGTCACAGACTTC</dnStreamSeq> </definition> <map> <dna> <dbXref db="Ensembl" /> <dbXref db="Genbank Record" /> </dna> </map> <allele alleleID="SNP000495189.ALE000988280">T</allele> <allele alleleID="SNP000495189.ALE000988281">C</allele> </variation>	Not Impossible	Not Available
Insertion Deletion	<snp-type> IND </snp-type> <snp-5flank_seq> TT </snp-5flank_seq> <snp-3flank_seq> TC </snp-3flank_seq> <snp-allele_na_set> <snp_allele-na> <snp_allele_na-nuc> C </snp_allele_na-nuc> </snp_allele-na> </snp-allele_na_set>	<NSE-ss_subsnp-class value="in-del"/> : <NSE-ss_observed> C/- </NSE-ss_observed>	<variation variationID="IND001634507" curationStatus="---" type="Indel" status="Suspected"> <definition molecule="DNA"> <upStreamSeq>GATATAAATATGTGCAT</upStreamSeq> <dnStreamSeq>AAAGAATGCAATTATC</dnStreamSeq> </definition> <map> <dna> <dbXref db="Ensembl" /> </dna> </map> <allele alleleID="IND001634507.ALE003275909">C</allele> <allele alleleID="IND001634507.ALE003275910">_</allele> </variation>	Not Impossible	Not Available
STRP	<snp-type> MIC </snp-type> <snp-5flank_seq> AA </snp-5flank_seq> <snp-3flank_seq> AG </snp-3flank_seq> <snp-allele_na_set> <snp_allele-na> <snp_allele_na-nuc> AAAAAAA </snp_allele_na-nuc> </snp_allele-na> </snp-allele_na_set>	<NSE-ss_subsnp-class value="microsat"/> : <NSE-ss_observed> (A)10/12 </NSE-ss_observed>	<variation variationID="STR000008185" curationStatus="MRA" type="Tandem Repeat" status="Proven"> <definition molecule="DNA"> <upStreamSeq>CGCCACTTTGTCCCGGC</upStreamSeq> <dnStreamSeq>GGAAAGGCCAACGGTCG</dnStreamSeq> </definition> <map> <dna> <dbXref db="EMBL Record"/> </dna> </map> <allele alleleID="STR000008185.ALE000013695" repeat="(A)10">AAAAAAAAAA</allele> <allele alleleID="STR000008185.ALE000013696" repeat="(A)12">AAAAAAAAAAAA</allele> </variation>	Not Impossible	Not Available

○…Terminology is exact △…Terminology is not exact
※…The database had existed before. But, it is not available now

Table 3 shows the diversity of sequence variation data representation. Almost all terms have diversities in representation about the variation data. The column lists the international SNP databases. The row shows the comparison terms of variation data.

3.3. Markup Language representation comparisons

Table 4. Comparison among Markup Languages. (quoted from [25])



	MAGE-ML	RNAML	BSML	ProML	SBML	CellML
Sequence	<!ELEMENT BioSequence [illegible]	<xs:element name="seq-data"> [illegible]	<!ELEMENT Seq-data (#PCDATA)> Sequencing runs, clones, contigs, cDNA, etc.	Primary Structure <xs:element name="AaSeq"/> Secondary Structure <xs:element name="SecStrucSeq"/>	Not specified	Not specified
Variation Data	In (BioSequence element) same as Omics(transcriptome) term.	In (molecule element) same as Omics(transcriptome) term.	In (Sequence element) same as Omics(transcriptome) term.	Not specified	Not specified	Not specified
Clinical Info	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified
Omics Transcriptome	Distributed to (BioSequence, Array...) <!ELEMENT Array [illegible] The technology type may be spotted cDNA <!ELEMENT FeatureGroup [illegible]	<xs:element name="molecule"> [illegible]	Sequence element contains seq-data element <!ELEMENT Sequence [illegible]	Not specified	These include large molecules (e.g., RNA proteins...) <xsd:complexType name="Species"> [illegible]	Not specified
Omics Proteome	In (BioSequence element) A BioSequence is a representation of a DNA, RNA, or protein sequence. same as the Sequence term.	Not specified	<!ELEMENT Alignment-summary (Aligned-sequence*)> <!ATTLIST Alignment-summary seq-type (nucleotide | protein) #REQUIRED seq-format CDATA #REQUIRED	Distributed to (protein, proteinInfo...) <xs:complexType name="protein"> [illegible]	In (Species element) same as Omics (transcriptome) term	Not specified
Omics Metabolome	Not specified	Not specified	Not specified	Not specified	As model Metabolic pathways	Not specified
Omics Signalome	Not specified	Not specified	Not specified	Not specified	As model Signaling pathways	Not specified

Table 4 shows the results of comparison among Markup Languages. The column lists the Markup Languages and the row shows the comparison terms. This time we investigated the markup languages as MAGE-ML, RNAML, BSML, ProML, SBML, and CellML. The compared terms are Sequence, Variation Data, Clinical Info, Transcriptome, Proteome, metabolome, signalome, and other Omics data. The results shown in Table 4 can be summarized as follows:

- All Markup Languages can describe the DNA sequence data, but the representations are different.

- Some can describe the variation data, but the details of it were difficult at every markup languages. Some does not have the descriptive ability of it.

- The proteomics information can be described by almost all Markup Languages, but the terms are different.

- The definitions of the basic vocabulary are different among Markup Languages. As an example, the word "Species" means chemical classification at SBML, while it means biological classification at the other Markup Languages.

- No Markup Languages have descript ability or expandability of clinical annotative data at this time.

- No Markup Languages have ability of interface to HL7 Genotype Model at this time.

3.4. GSVML requirement

The pit fall of genomic data handling is in the lack of the genomic sequence variation centric data exchanging format. GSVML is centric on genomic sequence variation, human, and clinical use. All of its needs and specifications are derived from these directions. Fundamentally GSVML should have the sharable representations for genomic sequence variation data such as allele, type, position, length, and region. These representations also should have expandability to the possible other sequence variation data. The annotations of variations such as variation associated genes, individual sequence, experimental assay are essential to understand the basis and the situation of the genomic sequence variation. To understand the clinical significance or to use in clinical situations, the peripheral annotations of variations such as clinical observation, phenotype are necessary to determinate the meanings of variations.

3.5. Interface Analysis to Health Level Seven

The interface of GSVML to HL7 was examined through comparing the terms of the HL7v3 Ballot 10 Genotype model and the GSVML.

The difference is typically reflected in the difference of the entry point, structure and content.

The entry point to GSVML is the variation locus. In contrast, the entry point to HL7v3 Genotype Model is the genotype (genetic locus).

In GSVML, variation data and its annotations are hierarchically categorized in three criteria as variation data, direct annotation, and indirect annotation. The variation data is associated with genotype, alleles, and sequences in the variation data criterion. The annotations as omics and clinical concerns are described hierarchically in the direct annotation criterion or the indirect annotation criterion. In HL7 Genotype model, the main elements are genotype, allele, variation, expression, sequence, and phenotype. Genotype is associated with a pair of alleles.

GSVML has the variation centric information model, while HL7v3 has the genotype centric information model. Both HL7v3 and GSVML have the genetic information and their supportive annotations, while the scope is different. GSVML describes the annotative information with hierarchical classification in one information model, while HL7v3 describes it with the multiple information models that are described with relatively planar classifications.

4. Result

Genomic Sequence Variation Markup Language (GSVML) version 1 was created.

The main role of GSVML in markup language context is to offer the sharable data format for exchanging genomic sequence variation data among the facilities that have the various types of data formats.

The envisioned application of GSVML is in human health domain, this indicates that GSVML standardization need harmonization with the clinical information models and other omics information models.

4.1. GSVML structure

The outlined structure of GSVML is shown in Figure 2. GSVML is consist of three data criteria as variation data, direct annotation, and indirect annotation.

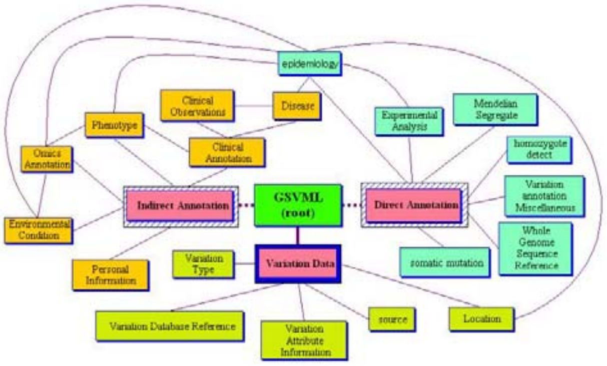


Figure 2. Outlined structure of GSVML. (quoted from [25])

The variation data criterion describes the straightforward variation data as allele, type, position, length, region, etc (Figure 3).

Some elements in the variation data criterion as variation type, location, variation attribute, and source are required, while the other elements are optional. Here database reference structure is defined as a dbref type that is a user defined complex type. The associated gene element is coupled with location element and the epidemiology element.

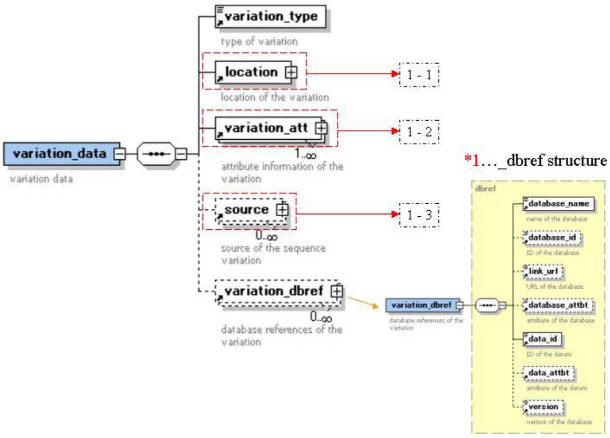


Figure 3. Variation data criterion of GSVML (quoted from [25])

The direct annotation criterion describes the attached data of variation data as whole genome sequence, mendelian segregate, homozygote detect, somatic mutation, experiment analysis, epidemiology, and miscellaneous (Figure 4). All of the direct annotation elements are optional. The experiment analysis element is consisted of the two categorized elements as variation-identify and variation-characterize. The variation-identify element has child elements to describe the experimental background to identify the variation data. The variation-identify element also has elements for publication and submitter, and recursive reference is allowed between the publication elements and submitter elements. The variation-characterize element has child elements to describe clinical statistics or genetic statistics. All kinds of statistical methodologies are allowed, while the typical elements as p-value, linkage disequilibrium index, descendent index, and maximum lod score are defined as the isolated elements. The epidemiology category gathered the statistical elements from all of GSVML elements, and it describes the statistical data. This category includes the associated gene from variation data, the disease epidemiology from indirect annotation, population and frequency from direct annotation. Each element of the disease epidemiology element is defined in the indirect annotation criterion.

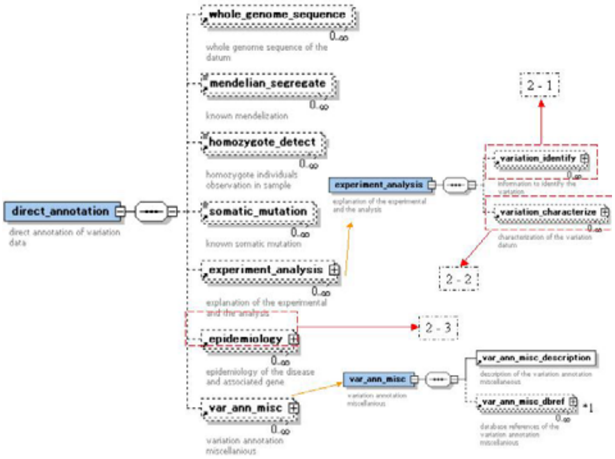

whole_genome_sequence
0..∞
whole genome sequence of the datum
mendelian_segregate
0..∞
known mendelization
2 - 1
homozygote_detect
0..∞
homozygote individuals observation in sample
experiment_analysis
explanation of the experimental and the analysis
variation_identify
information to identify the variation
direct_annotation
direct annotation of variation data
somatic_mutation
0..∞
known somatic mutation
variation_characterize
0..∞
characterization of the variation datum
experiment_analysis
0..∞
explanation of the experimental and the analysis
2 - 2
2 - 3
epidemiology
0..∞
epidemiology of the disease and associated gene
var_ann_misc_description
description of the variation annotation miscellaneous
var_ann_misc
variation annotation miscellaneous
var_ann_misc_dbref
*1
0..∞
database references of the variation annotation miscellanious
var_ann_misc
0..∞
variation annotation miscellanious

Figure 4. Direct annotation criterion of GSVML (quoted from [25])

The indirect annotation criterion describes the explanatory/higher-level information of variation data as the omics data, the clinical information, and the environmental data (Figure 5). The personal information is defined with personal description element and database reference. GSVML supposes all situations of describing the personal information, while at the most cases the personal information is encrypted or numbered. The phenotype element and the omics element allow broad data type to be able to describe it in any types of data formats. The clinical annotation category has disease element, clinical observation element, and database reference. The disease element is consisted of minimum set of disease descriptions, disease epidemiology, and database reference. Some elements in the disease description category are coupled with its expression probability. This probability elements are referred in the epidemiology data section. The database reference allows to describe with the other type of onotological description like SNOMED-CT. The clinical observation element is followed to SOAP description, while the almost all elements are diverted from the disease elements. The family history element is defined under the clinical observation element. The family history element for each family member is coupled with personal information, phenotype, and clinical annotation. To describe the character of each member, the recursive description is allowed.

These data criteria have relations to each other internally.

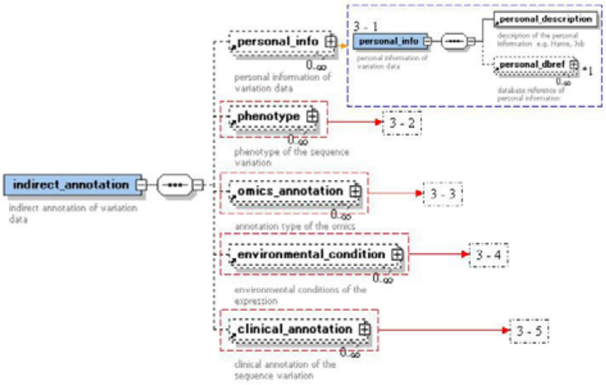

indirect_annotation
indirect annotation of variation data
personal_info
0..∞
personal information of variation data
3 - 1
personal_info
personal information of variation data
personal_description
description of the personal information e.g. Name, Job
personal_dbref
*1
0..∞
database reference of personal information
phenotype
0..∞
phenotype of the sequence variation
3 - 2
omics_annotation
0..∞
annotation type of the omics
3 - 3
environmental_condition
0..∞
environmental conditions of the expression
3 - 4
clinical_annotation
0..∞
clinical annotation of the sequence variation
3 - 5

Figure 5. Indirect annotation criterion of GSVML (quoted from [25])

5. Discussion

In the current context, annotative information around the genomic sequence variation data is increasing and is getting to embed the information holes. The variation data themselves are also increasing and resulted in various databases. This trend is typical in SNP data. Here the pit fall of genomic sequence variation data handling is in the lack of the sequence variation centric data exchanging formats. Historically many markup languages and programs are developed to handle the genomic information. However, there have been no SNP centric or sequence variation centric markup languages so far. GSVML can be addressed as the first genomic sequence variation centric markup language.

From application side, GSVML is human health centric. Considering that SNP is the highly researched polymorphism and has the great impact especially in human health domain, we can say that GSVML has the greatest potential to be the pinpointed ML for human healthcare. On the other hand, setting the applications to practical human health means to handle the direct or indirect annotative information. Here direct annotation indicates general information such as SNP associated genes, and indirect annotation indicates all of omics information and clinical information. To understand the situation of each patient, we need these kinds of additional information. For this reason, the development of GSVML need harmonization with the clinical standardization organizations such as Health Level Seven, International Organization for Standardization (ISO). The development work of GSVML collaborated with the Health Level Seven Clinical Genomics SIG work. The standardization effort of GSVML in ISO is in process, the improvement raised from the standardization process of ISO activities will be fed back and the design will be changed if need at the next version of GSVML. The "to and fro" process between the design work and the standardization process will continue to reflect the demands in future.

GSVML intends to apply for exchanging messages related to human health. In development and standardization of GSVML in this application domain, we kept an eye on the patient safety, the clinical efficiency, and the medical costs. For the patient safety, the conservation and the secrecy of patient information are important. The sharable data format and the standardization activity of GSVML can contribute to the data conservation of the domain field internationally, and the public key infrastructure will also need these kinds of the sharable data format. For the clinical efficiency, the simplicity and the easy understandability are important. The structure of the GSVML is hierarchically classified from end-user's view such as clinicians or researchers. The information model of the GSVML adopted the element-based-definition to simplify the usage of the GSVML. For the medical costs, the installation ability is important. Providing the GSVML with DTD and XML schema will be a good offer for installation at current context. The GSVML designed with intention to adopt the end-user understandable classification and the simplified information technology.

GSVML can be used for the clinical variation data exchanging among various facilities having various types of data formats. In the greater framework of clinical data standardization, GSVML will play a part of describing the variation data and its necessary information. At the version 1, we validated the annotative information such as clinical information or omics information with intentional roughness to accept the various representations of the end-user's descriptions. The many efforts to standardize the data format of these annotative information are going on. If these efforts reach a stage where one can rest, the more detailed validation of the annotative information will be our future work.

6. Conclusion

GSVML is in the demand for genomic sequence variation data exchanging. The GSVML is the sharable data exchanging format to exchange the genomic sequence variation data and the annotative information among the facilities having various types of data formats. The envisioned applications of GSVML are in human health domain, and the GSVML are demanded to equip a harmonization with clinical information and omics information as annotations of variation data. The GSVML can enhance the genomic sequence variation data utilization internationally by providing a sharable platform for data exchanging.

7. Acknowledgements

We thank the Health Level Seven Clinical Genomics Special Interest Group and the International Organization for Standardization TC 215 Working Group who gave us the valuable advises.

8. References

[1]. Holden AL. 2002. The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. Biotechniques Suppl: 22-24, 26.

[2]. Elias Zerhouni, "Medicine. The NIH Roadmap." Science. 2003 Oct 3;302(5642):63-72.

[3]. Cognitive Science Princeton University, “Overview for Markup Language,” internet article of http://www.cogsci.princeton.edu/cgi-bin/webwn2.0?stage=1&word=Markup+Language, 1998

[4]. International Organization for Standardization, ISO 8879: Information processing --Text and office systems -- Standard Generalized Markup Language (SGML), ([Geneva]: ISO, 1986)

[5]. T. Berners-Lee and Dan Connolly, "HyperText Markup Language Specification -- 2.0", RFC 1866. Proposed Standard , Nov. 1995.

[6]. W3C recommendation, "Extensible Markup Language (XML) 1.0 (Second Edition)", internet article of http://www.w3c.org/TR/2000/REC-xml-20001006, 1998

[7]. W3C recommendation, "XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition) A Reformulation of HTML 4 in XML 1.0", internet article of http://www.w3.org/TR/xhtml1/, Jan. 2000

[8]. W3C recommendation, "WAP Forum - W3C Cooperation White Paper ", internet article of http://www.w3.org/TR/1998/NOTE-WAP-19981030, 1998

[9]. W3C recommendation, "Simple Object Access Protocol (SOAP) 1.1", internet article of http://www.w3.org/TR/2000/NOTE-SOAP-20000508/ , 2000

[10]. Flying Boat Mobile Communications, “Glossary of Terms relevant to Mobile Communications,” internet article of http://homepages.nildram.co.uk/~jidlaw/pages/glossary.html, 2004.11.

[11]. Laurent SS. Biggar RJ. 1999 "Inside SMLDTDs: Scientific and Technical." Berkeley, CA: McGraw-Hill.

[12]. Hucka M, Finney A, Sauro HM, Bolouri H, et al. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models." Bioinformatics. 2003 Mar 1;19(4):524-31.

[13]. Hedley, W.J., Nelson, M.R., Bllivant, D.P. and Nielson, P.F. "A short introduction to CellML." Phil. Trans. Roy. Soc. London A, 359, 1073-1089, 2001.

[14]. Goddard NH, Hucka M, Howell F, Cornelis H, Shankar K, Beeman D. "Towards NeuroML: model description methods for collaborative modelling in neuroscience." Philos Trans R Soc Lond B Biol Sci. 2001 Aug 29;356(1412):1209-28.

[15]. Freimuth RR, Stormo GD, McLeod HL. "PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis." Hum Mutat. 2005 Feb;25(2):110-7.

[16]. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

[17]. Buetow KH, Edmonson MN, Cassidy AB. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet 21: 323-325.

[18]. Ohnishi Y, Tanaka T, Yamada R, Suematsu K, Minami M, Fujii K, Hoki N, Kodama K, Nagata S, Hayashi T, Kinoshita N, Sato H, Kuzuya T, Takeda H, Hori M, Nakamura Y. 2000. Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population. Hum Genet 106: 288-292.

[19]. Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ. "ALFRED: a Web-accessible allele frequency database".Pac Symp Biocomput 2000.:639-50.

[20]. International Organization for Standardization Technical Committee 215, http://www.iso.org/iso/en/stdsdevelopment/tc/

[21]. Health Level Seven Clinical Genomics Special Interest Group, internet article of http://www.hl7.org/Special/committees/, Since Sep. 2002

[22]. HL7 information model of genotype (HL7 POCG_DM000023), http://www.hl7.org/special/Committees/clingenomics/docs.cfm

[23]. Yoshida T. "[SNP project in the Millennium Genome Project, Japan]" Gan To Kagaku Ryoho. 2002 Jun;29(6):963-7.

[24]. Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, et al. "RNAML: a standard syntax for exchanging RNA information." RNA. 2002 Jun;8(6):707-17.

[25] International Organization for Standardization, TC215. Genomic sequence variation markup language. Available from: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=43182&scopelist=PROGRAMME. <Accessed July 11, 2006>