ICS 35.240.80
CCS C 07

中华人民共和国国家标准

GB/T 40419—2021

健康信息学
基因组序列变异置标语言(GSVML)

Health informatics—Genomic Sequence Variation Markup Language (GSVML)

(ISO 25720:2009,MOD)

2021-10-11 发布　　2022-05-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言 …… Ⅲ
引言 …… Ⅳ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　GSVML 定位 …… 3
5　GSVML 结构 …… 3
6　GSVML 的 DTD 和 XML 模式 …… 19
附录 A（资料性）　GSVML 开发情况说明 …… 20
A.1　GSVML 开发需求分析 …… 20
A.2　GSVML 开发过程 …… 21
A.3　基本参考资料 …… 22
附录 B（规范性）　GSVML 的 DTD …… 35
附录 C（规范性）　GSVML 的 XML 模式 …… 53
参考文献 …… 95

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件使用重新起草法修改采用ISO 25720：2009《健康信息学　基因组序列变异置标语言（GSVML）》。

本文件与ISO 25720：2009相比做了下述结构调整：

——按照GB/T 1.1—2020的规定对原文件中的引言、第2章、第5章、第6章和附录C的内容进行梳理、整合和删除重复内容，调整为本文件的引言、第4章、第5章、第6章和附录A，原文件的第3章和第4章调整为本文件的第2章和第3章，原文件的附录A和附录B调整为本文件的附录B和附录C。

本文件与ISO 25720：2009的技术性差异及其原因如下：

——按照GB/T 1.1—2020的规定对本文件范围进行了修改；

——将原文件中第3章的“EN 13606(all parts)”调整为本文件中第2章的ISO 13606(所有部分)；

——删除了原文件中的4.5、4.13、4.14、4.17、4.24和4.27，增加了术语“3.11 基因组序列变异置标语言”。

本文件做了下列编辑性改动：

——删除了原文件的参考文献的第20项和22项。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国标准化研究院提出并归口。

本文件起草单位：中国标准化研究院、深圳华大生命科学研究院、常州南京大学高新技术研究院、北京航空航天大学、珠海鼎基标准技术有限公司、国家卫生健康委科学技术研究所、中国医药企业管理协会、浙江树人大学、四川大学华西医院、南京吉芮康生物科技研究院有限公司、中国国际工程咨询有限公司、煦标医药科技(上海)有限公司、深圳统标科技有限公司、湖南君科再生医学科技有限公司、汕头市信德嘉生物科技有限公司、潮州和德生物技术有限公司。

本文件主要起草人：任冠华、张冀聪、王福清、王然、华子春、马旭、王媛媛、李雪、乔宝良、陆胤、许恒、舒洋、胡敏进、易静薇、魏笑、李锦轩、范志伟、许莉、唐秀丹。

引　言

基因组序列变异研究的爆炸式增长产生了海量的实验数据，并以各种类型的数据格式储存在世界各地的众多数据库中。为了能有效地管理、分析和利用这些数据，当务之急就是要对数据进行标准化以实现全球范围内的交换与共享。国际标准制定组织针对这些数据已经或正在制定相关的标准，HL7（Health Level Seven，健康信息交换与传输标准）是针对临床数据制定的标准，DICOM（Digital Imaging and Communcation in Medicine，医学数字成像和通信）和 JPEG（联合图像专家组）是针对影像数据制定的标准，而基因组序列变异置标语言（Genomic Sequence Variation Markup Language，GSVML）是鉴于基因组序列变异——特别是 SNP（Single Nucleotide Polymorphism，单核苷酸多态性）和 STRP（Short Tandem Repeat Polymorphism，短串联重复多态性）对于改善人类健康的基因医学和药物基因组学具有重大的作用，并且它们是针对基因组数据、尤其是人类相关的 DNA 变异数据所制定的标准。GSVML 开发情况说明见附录 A。

本文件为人类健康的基因组序列变异数据提供了一种数据交换格式，主要是针对 SNP 和 STRP 的案例给出了 GSVML 的规定。SNP 和 STRP 是人类健康相关研究中主要的和简单的多态性，可以其为中心将本文件的应用扩展到其他序列变异数据中。

健康信息学
基因组序列变异置标语言(GSVML)

1 范围

本文件确定了GSVML的定位,规定了GSVML结构,给出了GSVML的DTD和XML模式。

本文件适用于人类健康领域(包含临床实践、预防医学、转化研究和临床研究)中在不强制改变数据库模式情况下的基因组序列变异数据交换。

本文件适用对象为与人类健康相关的物种,如人、细胞系和临床前试验动物。本文件不适用于其他生物学物种。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

ISO 13606(所有部分) 健康信息学 电子健康记录通信(Health informatics—Electronic healthcare record communication)

3 术语和定义

下列术语和定义适用于本文件。

3.1

执行者 actor

对系统提供刺激的事物或人。

注:执行者包括人和其他半自动的物品(如机器、计算机任务和系统)。

3.2

生物信息序列置标语言 bioinformatic sequence markup language;BSML

用于生物信息数据的可扩展语言规范和容器。

3.3

细胞置标语言 cell markup language;cell ML

为基于计算机的生物模型提供一种标准方法进行表示和交换的可扩展置标语言。

3.4

癌症基因剖析工程 cancer gene anatomy project;CGAP

包含用于人类和小白鼠各种肿瘤组织的基因组表达数据,并提供用于获取基因组数据的方法和试剂信息的数据库。

3.5

医学数字成像和通信 digital imaging and communication in medicine;DICOM

医学信息学领域中用于医学成像设备(如放射学成像)与其他系统之间进行数字信息交换,并确保其互操作性的标准。