ICS 35.240.80
C 07

团 体 标 准

T/CHIA 41.2-2023

非编码 RNA 注释标准
第 2 部分：基本信息注释

Specifications for annotation of non-coding RNA
Part 2：Basic information

2023-11-14 发布 2024-02-01 实施

中国卫生信息与健康医疗大数据学会 发布

目　次

前　言

T/CHIA 41-2023《非编码RNA注释标准》分为以下3个部分：

——第1部分：分类信息；

——第2部分：基本信息注释；

——第3部分：功能及疾病注释。

本部分为T/CHIA 41-2023的第2部分。

本部分按照 GB/T 1.1-2020给出的规则起草。

本部分由中国科学院生物物理研究所提出，由中国卫生信息与健康大数据学会归口。

本部分主要起草单位：中国科学院生物物理研究所、中国科学院北京基因组研究所（国家生物信息中心）、浙江大学、复旦大学、清华大学、中国人民解放军总医院、北京蛋白质组研究中心、中国科学院微生物研究所、北京大学人民医院、中国科学院上海营养与健康研究所、中南大学、空军军医大学（第四军医大学）和北京睿博解码生物科技有限公司。

本部分的主要起草人：陈润生、何顺民、宋廷瑞、张鹏、周红红、王晓娜、方向东、金力、何昆仑、李亦学、张学工、段会龙、周水庚、渠鸿竹、赵思琪、钱颖、王霞、吕旭东、朱云平、马俊才、杨忠、石乐明、吴松峰、吴林寰、王振、陈先来、贾志龙、张昭军、娄晓敏、阮修艳、单广乐、乔楠、刘登辉、丁子建。

引　言

《非编码 RNA 注释标准 第 2 部分：基本信息注释》旨在为非编码 RNA 的注释提供一套术语规范、条理清晰、意义明确、语义语境无歧义的基本信息的标准规范，保证注释信息的一致性，并防止注释者在注释过程中遗漏非编码 RNA 的重要信息，便于对非编码 RNA 已知知识的获取，保证信息的有效交换、有效挖掘分析和共享。

非编码 RNA 注释标准

第 2 部分：基本信息注释

1 范围

本标准规定了非编码RNA注释中的基本信息。

本标准适用于对非编码 RNA 注释。

2 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准的条款。凡是注日期的引用文件，其随后所有的修改单（不包括勘误的内容）或修订版均不适用于本标准。凡是不注日期的引用文件，其最新版本适用于本标准。

GB/T 29859-2013 生物信息学术语

3 术语和定义

GB/T 29859-2013 中界定的以及下列术语和定义适用于本文件。

3.1

fasta 格式 fasta format

一种基于文本的、用于表示核苷酸序列或氨基酸序列的格式，详情见附录 7.1。

3.2

非编码 RNA 基因 non-coding RNA gene

非编码 RNA 来源的基因组 DNA 区域，与蛋白编码基因类似，一个非编码 RNA 基因可能转录出多条相似但不同的非编码 RNA 转录本。

3.3

HUGO 基因命名委员会 HUGO Gene Nomenclature Committee

对人类基因组上包括蛋白编码基因、非编码基因等在内的所有基因提供一个唯一的、标准的、可以广泛传播的名称。

3.4

NCBI 物种分类数据库 Taxonomy Database

该数据库提供了 NCBI 公共序列数据库中所有物种的分类和命名法。

3.5

正链 forward strand

双链互补 DNA 分为正链和负链，参考基因组数据一般都只记录正链，即从数据库中

得到的基因组 fasta 格式文件都是正链的碱基序列。

3. 6

正义链 sense strand

DNA 上与 RNA 转录本相同方向的链，又称编码链、有义链。

3. 7

反义链 antisense strand

与正义链互补的一条核苷酸序列链，又称非编码链、无义链、模板链。

3.8

一基坐标 one-based coordinate

从 1 开始计数的位置表示方式，即核酸序列第一个碱基为 1，之后的碱基依次加 1，是最直观的表示方式。位置区间表示为起始位点与终止位点两个位置的组合，区间长度等于终止减起始的差值+1。

3.9

零基坐标 zero-based coordinate

从 0 开始计数的位置表示方式，即核酸序列第一个碱基为 0，之后的碱基依次加 1，是计算机领域常用的表示方式。位置区间的表示需要额外把终止位点数值+1，也就是终止位点的表示与一基坐标一致，起始位点的表示比一基坐标小 1，区间长度正好等于终止减起始的差值。零基坐标也可以理解为碱基之间空隙的表示。

4 缩略语

下列缩略语适用于本文件。

DNA：脱氧核糖核酸（Deoxyribonucleic Acid）。

RNA：核糖核酸（Ribonucleic Acid）。

HGNC：HUGO基因命名委员会（HUGO Gene Nomenclature Committee）。

ID：识别号（Identity）。

NCBI：美国国家生物技术信息中心（National Center for Biotechnology Information）。

nt：核苷酸（nucleotide）。

txid：物种分类号（taxonomy ID）。

5 物种分类

物种参照 NCBI 的物种分类法进行分类，见表 1。

表 1 物种大类别划分

分类	英文名称	物种数目
古生菌	Archaea	728

分类	英文名称	物种数目
细菌	Bacteria	20,719
真核生物	Eukaryota	461,904
真菌	Fungi	49,082
后生动物	Metazoa	240,317
植物	Viridiplantae	159,251
病毒	Viruses	4,631

注: 完整谱系可通过 NCBI Taxonomy 网址: (https://www.ncbi.nlm.nih.gov/Taxonomy/) 进行检索，详情见附录 7.2。

6 非编码 RNA 基本信息注释元数据

6.1 元数据描述

本部分采用摘要表示的方式定义和描述元数据。摘要内容包括以下 7 个属性：中文名称、定义、英文名称、数据类型、约束/条件、词表、备注。

6.1.1 中文名称

指元数据元素或元数据实体的中文名称。

6.1.2 定义

描述元数据元素或元数据实体的基本内容，给出信息资源某个特性的概念和说明。

6.1.3 英文名称

元数据元素或元数据实体的英文名称，一般用英文全称

6.1.4 数据类型

元数据元素的数据类型，对元数据元素的有效值域和允许对该值域内的值进行有效操作的规定。

6.1.5 约束条件

说明元数据实体或元数据元素是否必须选取的属性。包括必选、可选。

必选（M）：表明该元数据元素或元数据实体必须选择。

可选（O）：根据实际应用可以选择也可以不选的元数据元素或元数据实体。如果一个可选元数据实体未被使用，则该实体所包含的元素（包括必选元素）也不选用。可选元数据实体可以有必选元素，但只当可选实体被选用时才成为必选。

6.1.6 词表

对元数据元素或元数据实体的可填内容进行规范或建议（根据需要选用）

6.1.7 备注

对元数据元素或元数据实体的进一步说明（根据需要选用）

6.2 元数据

本节给出非编码 RNA 基本信息注释元数据的定义

6.2.1 非编码 RNA 基因

非编码 RNA 基因的注释信息包括非编码 RNA 的基因识别号，名称，别名，物种，类型，基因组位置，转录本列表，以及功能及疾病。

表 2. 非编码 RNA 基因元数据

中文名称	定义	英文名称	数据类型	约束条件	词表	备注
基因识别号	识别具体非编码 RNA 基因的唯一标识符	Gene identity	字符串	必选（M）	自定义或者用已有数据库的编号，如：NONHSAG000001.2	可作为本元数据关键字
名称	非编码 RNA 的标准名称	Gene symbol	字符串	可选（O）	可参考 HGNC 命名，如 XIST	
别名	非编码 RNA 的其它常用或曾用名称	Alias	字符串	可选（O）		
物种	非编码 RNA 来自何种生物	Organism	元数据	必选（M）	参照 NCBI 物种分类数据库和本标准的物种元数据定义	
类型	非编码 RNA 所属分类	Category	字符串	必选（M）	依据本标准第 1 部分分类信息的定义，如：lncRNA、snRNA 等	
基因组位置	非编码 RNA 在基因组上的位置	Genome position	元数据	必选（M）	依据本标准的基因组位置元数据定义	
转录本列表	非编码 RNA 基因包含的所有转录本	Transcripts	元数据列表	必选（M）	依据本标准的转录本元数据定义	每个基因至少包含一条转录本
功能及疾病	非编码 RNA 功能及与疾病关系的注释	Function and disease	元数据	可选（O）	依据本标准第三部分的功能及疾病注释元数据定义	

6.2.2 物种

物种元数据用来定义非编码 RNA 来源的物种信息，包括物种的分类号，物种名和常用名。

表 3. 物种元数据

中文名称	定义	英文名称	数据类型	约束条件	词表	备注
分类号	物种的唯一识别号	Taxonomy identity	字符串	必选（M）	参照 NCBI 物种分类数据库，如 txid9606	可作为本元数据关键字
物种名	物种的标准拉丁文名称	Species	字符串	必选（M）	参照 NCBI 物种分类数据库，如 Homo sapiens	
常用名	物种的常用俗名	Common name	字符串	可选（O）	如 Human，Mouse 等	

6.2.3 基因组位置

基因组位置信息包括基因组版本，染色体，染色体起始位点，染色体终止位点，链，外显子数目和外显子列表。

表 4. 基因组位置元数据

中文名称	定义	英文名称	数据类型	约束条件	词表	备注
基因组版本	基因组位置所依据的参考基因组组装版本	Assembly	字符串	必选（M）	参考 NCBI 物种分类数据库和 UCSC 基因组浏览器中的参考基因组命名，如 GRCh38，hg19，mm10 等	
染色体	RNA 所在的染色体名称	Chromosome	字符串	必选（M）	基因组序列 fasta 文件中的序列名，如 chrX，1，NC_000001.11 等	存在带 chr 和不带 chr 的形式，可用参考序列 ID
染色体起始位点	RNA 在染色体上的 5’端位置	Chromosome start site	整数	必选（M）	基因在染色体上的最小位置，一基坐标或零基坐标皆可	与转录本的链方向无关
染色体终止位点	染色体上的 3’端位置	Chromosome stop site	整数	必选（M）	基因在染色体上的最大位置	
链	RNA 序列是	Strand	字符	必选	+, -, .	不确定链方向的情

中文名称	定义	英文名称	数据类型	约束条件	词表	备注
	否来自基因组的正链		串	（M）		况下用“.”
外显子数目	长非编码RNA含有的外显子数目	Exon number	整数	可选（O）		转录本相关信息
外显子列表	RNA转录本的所有外显子的区间位置	Exons	字符串	可选（O）	如11000-12000,14000-16000,17000-18000	基因或只有单个外显子的转录本可不填,范围不超过起始到终止位点的区间

6.2.4 转录本

转录本信息包括转录本识别号，名称，所属基因识别号，基因组位置，长度和序列。

表 5. 转录本元数据

中文名称	定义	英文名称	数据类型	约束条件	词表	备注
转录本识别号	识别具体转录本的唯一标识符	Transcript identity	字符串	必选（M）	自定义或者用已有数据库的编号，如：NONHSAT000002.2	可作为本元数据关键字
名称	转录本名	Transcript name	字符串	可选（O）		
基因识别号	转录本所属的基因	Gene identity	字符串	必选（M）	本转录本所属的非编码RNA基因识别号	
基因组位置	转录本在基因组上的位置	Genome position	元数据	必选（M）	依据本标准的基因组位置元数据定义	如转录本存在剪接,需提供外显子列表
长度	转录本的长度	Length	整数	必选（M）	不包含 intron 的成熟转录本长度	
序列	转录本的核酸序列	Sequence	fasta	可选（O）	不包含 intron 的成熟转录本的正义链序列	DNA或RNA(T换U)序列皆可

7. 附录

7.1 fasta 文件格式

fasta 文件中每个序列单元都是从“>”开始到下一个“>”结束。由“>”开头的任意文字说明，用于序列标记，单个序列的标识必须具有唯一性以区分每条序列，一般为一行；第二行开始为序列本身，可以由多行组成，只允许使用核苷酸或氨基酸编码符号。不同物种的基因组文件多以 fasta 格式进行存储。如图 1 所示：

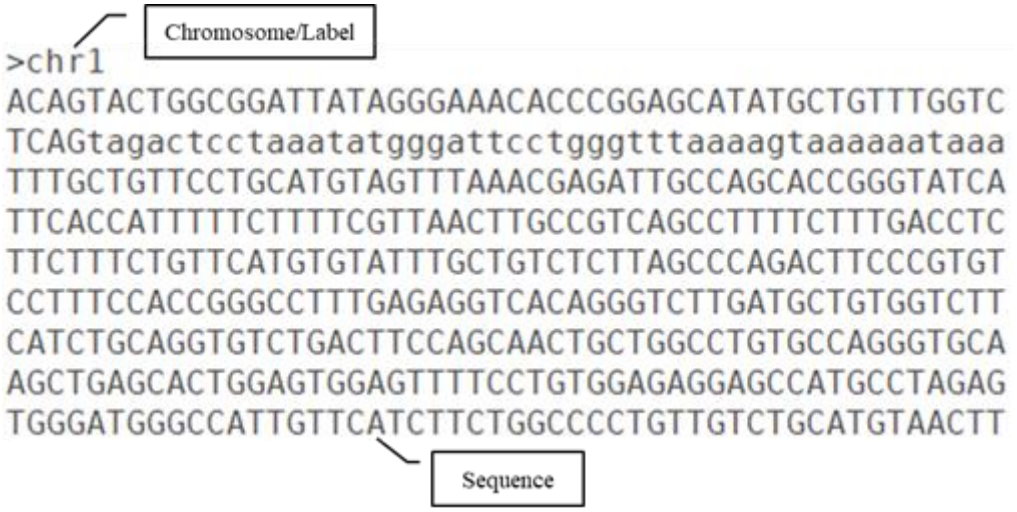

图 1 fasta 文件示例

7.2 NCBI 物种分类数据库使用方法

进入 NCBI Taxonomy 数据库网站：(https://www.ncbi.nlm.nih.gov/Taxonomy/) 进行检索。在搜索框中输入物种名后点击“Search”，例如：输入 human。在搜索结果页面点击“Homo sapiens”跳转到“Taxonomy Browser”页面，继续点击“Homo sapiens”可得到更详细的信息，包括 txid、完整的谱系、基因组信息等（图 2）。

图 2 搜索物种的详细信息