ICS 35.240.80
C 07

团　　体　　标　　准

T/CHIA 42.2-2023

长非编码 RNA 和蛋白相互作用注释标准
第 2 部分：RIP-seq 和 CLIP-seq 的数据分析方法

Specifications for interaction between long non-coding RNA and proteins

Part 1：Computational analysis pipeline of RIP-seq and CLIP-seq

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 I
引　言 ..... 错误！未定义书签。
1 范围 1
2 规范性引用文件 1
3 术语和缩略语 1
4 RIP-seq 数据处理流程 1
5 CLIP-seq 数据分析流程 2

前　言

本标准按照GB/T 1.1—2020给出的规则起草。

T/CHIA 42-2023《长非编码RNA和蛋白相互作用注释标准》分为以下3个部分：

——第1部分：RIP-seq和CLIP-seq的实验方法流程

——第2部分：RIP-seq和CLIP-seq的数据分析方法

——第3部分：长非编码RNA及其相互作用蛋白质的功能注释

本标准为T/CHIA 42-2023的第2部分。

本标准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。

本标准主要起草单位：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）、中国科学院计算技术研究所和北京睿博解码生物科技有限公司。

本标准主要起草人：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、赵屹、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

《长非编码RNA和蛋白相互作用注释标准 第2部分：RIP-seq和CLIP-seq的数据分析方法》旨在规范现有长非编码RNA和蛋白质相互作用实验技术RIP-seq和CLIP-seq的数据分析流程，为长非编码RNA和蛋白质相互作用的数据处理提供一套术语规范、定义明确、语义语境无歧义的标准。

长非编码 RNA 和蛋白相互作用注释标准
第 2 部分：RIP-seq 和 CLIP-seq 的数据分析方法

1 范围

本标准规定了RIP-seq和CLIP-seq的数据分析流程规范。

本标准适用于指导研究人员进行RIP-seq和CLIP-seq等捕获长非编码RNA与蛋白相互作用的高通量测序数据分析。

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标准必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本标准；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

GB/T 29859-2013 生物信息学术语

GB/T 30989-2013 高通量基因测序技术规程

GB/T 35890-2018 高通量测序数据序列格式规范

T/CHIA 21.1—2021 组学样本处理与数据分析标准 第 1 部分：全基因组测序数据分析

T/CHIA 21.5—2021 组学样本处理与数据分析标准 第 5 部分：转录组测序数据分析

3 术语和定义

GB/T 29859-2013、T/CHIA 21.5—2021中界定的以及下列术语和定义适用于本标准。

3.1

峰区域识别 peak calling

使用reads的分布情况对RNA与蛋白相互作用的区域进行预测的方法。获取到的reads密集的区域成为峰（peak）。

3.2

IP 组

指的是用相应的抗体进行免疫沉淀的产物，是与响应蛋白相互作用的RNA。

3.3

Input 组

指的是样本中的总RNA，属于阳性对照。

4 RIP-seq 数据处理流程

4.1 质量控制

利用测序数据质控软件对测序数据进行质量评估，使用去接头软件去除低质量reads和连续的低质量片段，去掉接头序列。包括对测序质量的评估、去污染、去低质量、N比例、

GC含量、重复情况、序列长度分布情况、碱基平衡情况等，可以通过质控汇总软件批量显示质量控制结果。

4.2 比对

通过质控的有效测序数据利用基因组比对软件比对到参考基因组，并对比对结果进行排序。

4.3 reads 统计与表达量的计算

使用表达量统计软件对比对到各个RNA注释的reads进行统计，计算每个RNA的reads数量。然后使用表达谱分析软件计算每个RNA的表达量。

4.4 富集度计算与目标蛋白相互作用 RNA 集合的选取

使用表达谱分析软件计算IP组和Input组的表达量比率（fold change）和FDR值。同时计算对照组和Input组的表达量比率（fold change）和FDR值，对两组样本富集的RNA进行过滤，选择fold change大于2且FDR小于0.05的RNA作为富集RNA。将对照组中富集的RNA从实验组富集的RNA集合中去除，获得与目标蛋白相互作用的RNA集合。

5 CLIP-seq 数据分析流程

5.1 质量控制

利用测序数据质控软件对测序数据进行质量评估，使用去接头软件去除低质量reads和连续的低质量片段，去掉接头序列。包括对测序质量的评估、去污染、去低质量、N比例、GC含量、重复情况、序列长度分布情况、碱基平衡情况等，可以通过质控汇总软件批量显示质量控制结果。

5.2 比对

通过质控的有效测序数据利用基因组比对软件比对到参考基因组，并对比对结果进行排序。

5.3 结合位点识别

使用结合位点识别软件对比对获得的bam文件进行peak calling，获得每个蛋白结合的位点位置。对注释文件和位点位置取区间交集，可以获得与目标蛋白相互作用的RNA集合。