论 文

RNA结合蛋白的CLIP-seq数据的生物信息学 分析流程建立及初步应用

谢正尧, 臧敏, 李鑫, 徐宇君*

南京医科大学, 生殖医学国家重点实验室, 南京 211166

* 联系人, E-mail: xuyujun@njmu.edu.cn

收稿日期: 2020-12-11; 接受日期: 2021-02-01; 网络版发表日期: 2021-03-08

国家自然科学基金(批准号: 3177652, 31970792)资助

摘要 近年来, 紫外交联免疫沉淀结合高通量测序(UV cross-linking immunoprecipitation followed by high-throughput sequencing, CLIP-seq)成为鉴定RNA结合蛋白(RNA-binding proteins, RBP)的靶标序列和结合位点的新技术, 为研究RNA结合蛋白功能、解析其分子机制提供了强有力的工具。尽管国际上已有应用CLIP技术的报道, 但是实验方法周期长、技术性强, 尤其是后续的数据处理和生物信息学分析令没有计算机基础和专职生物信息分析人员的实验室望而却步。基于2016年Van Nostrand等人优化的CLIP方法, 本团队建立并完成了多个RBP在组织和细胞中的CLIP测序及数据的生物信息学分析流程, 并验证了其可行性。本研究整合了多家实验室已公开发表的分析软件和脚本, 对人前列腺癌DU145细胞系中的PUMILIO1(PUM1)蛋白CLIP-seq数据进行分析, 总结出对于蛋白功能分析更有效的命令和参数, 并建立了完整的生物信息学分析流程。运用该流程寻找PUM1蛋白在人前列腺癌发展中的潜在机制, 本团队识别出PUM1在DU145细胞系中的1189个靶标, 并发现PUM1主要结合其靶标mRNA的3'非翻译区(untranslated region, UTR)和转录终止位点(transcription termination sites, TTS)区域, UGUA-HAUA(H代表A/U/C)基序(motif)为PUM1蛋白识别的特征性序列。PUM1可能通过结合靶标mRNA的3' UTR对靶标进行转录后调控, 参与癌细胞的生长、分化等过程以及相关的癌症通路中。本团队还发现, PUM1在人前列腺癌细胞系中能够与其他物种或组织中类似的靶标结合, 并且结合方式高度一致, 这与该家族蛋白在RNA结合功能域和功能上的高度保守性相符合。因此, 本文报道的技术方法和生物信息学分析流程可以为有兴趣开展RBP的靶标鉴定的研究人员提供一个一站式的方法学服务平台, 也有助于推动转录后调控在疾病和发育中的机制研究和RBP的RNA调控等新兴领域的发展。

关键词 RNA结合蛋白, 紫外交联免疫沉淀, 高通量测序, 生物信息学, PUM1, 前列腺癌, 3' UTR

RNA作为一类非常重要的生物分子, 以messenger RNA(mRNA), transfer RNA(tRNA), small nucleolar RNA(snoRNA), long non-coding RNA(lncRNA), Piwi-

interacting RNA(piRNA), microRNA(miRNA)和circular RNA(circRNA)等多种形态参与到生命活动的各个方面^[1~3]。RNA结合蛋白(RNA-binding proteins, RBPs)是

引用格式: 谢正尧, 臧敏, 李鑫, 等. RNA结合蛋白的CLIP-seq数据的生物信息学分析流程建立及初步应用. 中国科学: 生命科学, 2021, 51: 450-461
Xie Z Y, Zang M, Li X, et al. A bioinformatics pipeline for processing CLIP-seq data of RNA-binding proteins and its application (in Chinese). Sci Sin Vitae, 2021, 51: 450-461, doi: 10.1360/SSV-2020-0403

一类参与细胞RNA成熟、稳定、运输和降解等过程的蛋白质,可以直接与RNA结合,也可以作为大分子蛋白复合物的组成部分与RNA结合.在细胞核中,RBPs可以改变mRNA前体的加工,产生不同编码和非编码的RNA序列;在细胞质中,RBPs通常情况下通过与靶标RNA的3'非翻译区(untranslated region, UTR)相互作用影响mRNA的定位、翻译和稳定性^[4-6].

RBPs通过特定的、具有保守性的RNA结合功能域(RNA binding domain, RBD)结合多种靶标RNA,在转录和转录后基因表达的调控中发挥重要作用.RBD是RBPs中负责与RNA结合的功能单位.多个或多种RBD经常出现在单个RBP中,这些模块化排列可以协调并增强与RNA的结合(表1)^[7].最常见的RNA结合功能域包括RRM, KH domain, PUF repeat, Helicase和PAZ domain等.尽管很多RBPs在多个组织表达,但是仍有部分RBPs显示组织特异性表达,如生殖器官睾丸和卵巢中的DAZ家族蛋白^[8-10]、PIWI蛋白^[11]、NANOS蛋白^[2,12]、VASA蛋白等.近来的RBPs研究表明,除了带有经典的RNA结合功能域的RBPs,很多RBPs并不具备已知的RNA结合功能域,但是它们仍然能够与RNA结合.对于这些非经典RBPs的鉴定,必须依赖有效的RNA结合免疫共沉淀等新技术才能确定.通过免疫共沉淀结合高通量测序方法鉴定核酸蛋白复合物(ribonucleoprotein particle, RNP)已经成为一项研究RBPs的生理功能和分子机制的关键技术.

研究RBPs对RNA的调控功能对于阐明重要生命活动过程以及疾病发生的复杂分子机制具有重要意义,明确RBPs的结合靶标及其结合位点的实验方法是理解该研究的关键^[13].研究表明,多数RBPs并非仅结合一个或几个靶标RNAs,而是结合数十数百个靶标RNAs,每个RBP通过结合这些靶标协调调控下游靶标RNA的稳定性和/或翻译^[14].紫外交联免疫沉淀(UV cross-linking immunoprecipitation, CLIP)技术因为能够全面识别RBP在生理条件下的RNA底物而被广泛使用^[15],尤其是对于全面认识RBPs参与的下游靶标调控群很有帮助.通过与高通量测序(high throughput sequencing)结合,利用生物信息学方法可以找到RBP结合的靶标基因和对应motif的相关信息,在单核苷酸分辨率上明确RNA结合位点.CLIP实验技术方法近年来被不断改进,目前国内外使用较为普遍并且得到广泛认可的新型CLIP技术是eCLIP.CLIP技术改进的目的是

更加准确、方便地得到RBP直接结合的RNA片段以用于高通量测序(表2),然而,测序结果的生物信息学分析对于国内绝大多数可以完成CLIP实验的实验室来说,技术门槛较高且暂时缺乏详细的分析策略.

本研究旨在通过构建基于Linux系统的生物信息学分析流程,提供参考软件和脚本以及关键步骤的阐述和终端命令,使其能够适配大部分CLIP技术类型的双端测序,并提出后续分析的思路 and 方向,以期为更多对RBP感兴趣、有能力做CLIP实验却不善于生物信息学分析的传统实验室(wet lab)研究人员提供高通量测序后续的计算机分析策略,并根据分析得到RBP对靶标的结合特征,为后续揭示RBP的生物学功能提供线索和依据.

1 分析流程

分析平台计算机的配置CPU为E5-2630v4,内存128 G, GPU为GTX 1080Ti.本生物信息学分析流程(图1)基于Linux系统(Ubuntu 16.04),使用的分析工具均来自自己公开发表的文献或软件脚本(表3),需要提前自行安装配置.本流程整体上主要用于处理双端测序数据,但是其中许多分析工具也可以处理单端测序数据.因为样本的来源为人前列腺癌细胞系,因此流程中所有数据库均使用人参考基因组hg38(GRCh38)的对应版本数据库.统计学方面的数据分析和绘图使用Excel和Graphpad软件.

本文提供的参考命令和参数(网络版补充数据1)均是以双端测序eCLIP数据为分析对象,虽然经过了来自多种CLIP技术的大量样本的多次调整与验证,但是根据不同的数据特点以及实际需求仍然需要对参数进行个性化调整.不同类型CLIP技术的双端高通量测序数据的建库方式不同,因此在数据的预处理上需要根据建库情况以及返回的原始数据形式进行相应的调整,流程中的其他步骤基本通用.相关的专有名词解释在网络版补充数据2进行了阐述,作为对接下来正文内容的补充说明.

1.1 原始数据质量控制

得到原始数据(本研究的实验数据文件为PUM1_DU145_R1.fastq, PUM1_DU145_R2.fastq, PUM1Input_DU145_R1.fastq, PUM1Input_DU145_R2.fastq)后,

表 1 RNA结合功能域概述(基于Corley等人^[7])
Table 1 Summary of RNA binding domains (Based on Corley et al.^[7])

RNA结合功能域	所属蛋白家族	平均长度(aa)	靶标RNA结合特征
RNA recognition motif (RRM)	非常普遍, 常见于核糖核蛋白(heterogeneous RNP, hnRNP)、剪接因子	90	通过RNP motif与ssRNA(single-stranded RNA)的2~8 nt序列进行互作, RBP中的多个RRM可以独立或协同增加结合特异性
K homology (KH)	非常普遍, 常见于核糖核蛋白、翻译调控蛋白	70	通过GXXG motif结合ssRNA的4 nt序列, RBP中的多个KH结构域(I 型)可以独立或协同增加结合特异性
Double-stranded RNA binding domain (dsRBD)	RNA酶(RNases)、RNA依赖性腺苷脱氨酶(ADARs)、Dicer酶	65	特异识别结合dsRNA(double-stranded RNA)的A型RNA螺旋结构或者环结构附近的特定碱基序列
Zinc finger (ZnF)	非常普遍, 常见于转录因子、METTL酶	30	与RNA相互作用的ZnF亚型包括CCHC(zinc knuckle), CCCH, CCCC(RanBP2)和CCHH亚型, C和H分别为配位锌原子的半胱氨酸残基和组氨酸残基; zinc knuckle识别结合RNA的茎环结构(结合环以及茎中磷酸盐骨架中的碱基), CCCH和CCCC亚型识别结合3 nt的重复序列, CCHH亚型功能丰富, 结合单链或双链RNA
Pumillo-like repeat (PUM)	PUF蛋白	334	高度保守的36个氨基酸形成8个α螺旋重复, 形成凹形RNA结合面, 通过识别8 nt的motif结合ssRNA
Pentatricopeptide repeat (PPR)	RNA编辑蛋白	35×n, 1<n<30	2~30个重复形成螺旋管状支架, 每个重复与ssRNA的2个碱基特异结合
Pseudouridine synthase and archaeosine transglycosylase (PUA)	RNA修饰酶、代谢酶	66~98	通常通过富含甘氨酸的序列结合dsRNA及其邻近的环
thiouridine synthases, RNA methylases and pseudouridine synthases (THUMP)	tRNA修饰酶	100~110	通常位于RNA修饰结构域和N-端铁氧化还原蛋白样结构域(NFLD)附近, 特异性识别tRNA的3'-CCA末端和邻近的茎结构
YT521-B homology (YTH)	YTH家族m ⁶ A readers	100~150	特异结合未修饰的腺苷上的m ⁶ A
Helicase	DExH/D-box, Ski2-like, RIGI-like, NS3, UPF1-like RNA结合解旋酶	350~400	解旋酶家族有2个保守的recombinase A (recA)-like子结构域(subdomain), 通过recA-like domains或者多聚解旋酶形式围绕被结合的RNA
Cold shock domain (CSD)	冷休克蛋白(cold shock protein)、Y-box结合蛋白(Y-box/Y-box binding protein)	70	通过保守的RNP1和RNP2 motif结合ssRNA的3或4 nt序列, 含有CSD的RBP在它们识别的序列类型上差别很大
Ribosomal S1-like (S1)	核糖体蛋白, 翻译起始因子, RNA酶 II、多核苷酸磷酸化酶(PNPase)	70	结合单链或双链RNA, 有时会与其他PDB共同组合出现在RBP中
Sm and like-Sm (Sm/Lsm)	U1剪接体蛋白、Hfq蛋白	80	含有该motif的蛋白常发生多聚化, 组合成剪接体中小核糖核蛋白(snRNPs)的蛋白核心来结合RNA
La motif (LAM)	La蛋白、La相关蛋白	90	La蛋白通过LAM-RRM联合区域紧密结合聚合酶III转录的小RNA的3'末端UUU-OH元件
Piwi-Argonaute-Zwille (PAZ)	AGO(Argonaute)蛋白、Dicer酶	170	主要结合靶标RNA中5 nt左右的序列
P-element Induced Wimpy Testis (PIWI)	AGO蛋白	290	主要结合靶标RNA中3~5 nt的序列
Intrinsically disordered region (IDR)	多数RBP	不定	虽然缺乏结构但在RBP中占比很高, 具有保守性, 可以作为RBP中唯一的RBD结合靶标RNA, 并且可以促进蛋白质-RNA的相互作用, 对RNA序列的依赖性很小

表 2 不同CLIP技术的比较

Table 2 Comparisons of different CLIP methods

主要CLIP技术种类	发表时间	优点/特点	局限性	首次发表时的Peak Caller
HITS-CLIP ^[16]	2008	首次将高通量测序技术与CLIP技术相结合, 256 nm紫外交联和RNA酶消化步骤使该技术得以捕捉体内瞬时的相互作用	一般只运用于系统组织水平, 紫外交联在穿透性、效率等方面有一定局限性, 分辨率低, 固有噪声背景影响大	Pyicoclip
PAR-CLIP ^[17]	2010	使用4SU等光学活性核糖核酸衍生物标记并用365 nm紫外交联, 提高了交联效率和分辨率, 交联介导的突变位点提供了准确的结合位点信息, 减少了背景序列的干扰	PCR扩增方法的偏好性使分辨率受限, 长时间使用会导致细胞中毒, 可能会影响实验结果	—
iCLIP ^[18]	2010	利用cDNA环化捕获截断的cDNA, 提升了分辨率, 在单核苷酸水平上精确确定RBP-RNA交联位点	反应步骤较多, 环化效率不高, 熟练操作较困难, 且需消耗大量的原材料	—
eCLIP ^[19]	2016	分两步添加接头, 降低了RNA损失, 极大提高了建库成功率, 减少了PCR重复, 靶标识别率显著提高; 改进了定量比较Input去除背景的方法	操作有一定难度, 成本较高	CLIPper
seCLIP ^[20]	2017	是eCLIP的简化改良版, 支持单端测序, 降低了成本	操作有一定难度, 成本较高	—
spyCLIP ^[21]	2019	在RBP融合SpyTag和SpyCatcher之间形成共价键, 极大提高RBP-RNA复合物纯度, 易于自动化; 避免了现有CLIP方法对抗体的依赖, 提高了数据的重复性和特异性, 大幅降低了实验操作的难度和时长	依赖预先构建的RBP-Spy融合蛋白的存在, 较难应用于人和动物组织实验	iCount

图 1 CLIP-seq分析流程图. 方框代表关键分析步骤, 椭圆代表关键输出数据

Figure 1 A flow chart for CLIP-seq analysis. Boxes represent key analysis processes, and ellipses represent key output data

表 3 CLIP-seq分析工具

Table 3 Tools for CLIP-seq analysis

分析过程	工具	版本
原始数据质量控制	FastQC	v0.11.8
数据预处理	Cutadapt	v.1.18
	gscripsts	1.1
去除PCR重复	gscripsts	1.1
生物学背景Input标准化		
分离重复序列	STAR	v2.6.1d
比对参考基因组		
获得富集峰	CTK	v1.1.3
统计学背景模型标准化		
CIMS		
CITS		
注释	HOMER	v4.10.4
motif分析		
其他	Samtools	0.1.19
	Picard	2.21.1
	Bedtools	v2.25.0

首先根据测序订单核对原始数据的信息。原始数据通常为fastq文件, 该文件中每4行代表一个Read的所有结果, 因此原始数据的总行数是4的倍数, 可以此进行初步检验。如果是双端测序, 同一个测序样本的Read1和Read2文件行数是相同的。随后抽取原始数据的部分reads进行随机抽查(通常用命令查阅前40行和最后40行的数据), 初步验证reads两端的接头是否符合相关CLIP技术的建库规则和特点。以eCLIP为例, 根据建库规则(图2A和B), 测序得到的原始数据应符合图2C中结构, 如果接头序列和结构有大量错误则说明测序前的CLIP实验中可能存在问题。可使用FastQC^[22]软件对原始数据的测序质量做全面的统计和评价。

1.2 数据预处理

本文中委托的测序公司根据样本特异性接头对索引序列(index)进行了识别并对样本进行了区分。有些公司完成测序后会提供一个或一组混合有多个样本的文件作为测序结果, 这种情况需要自行根据index进行样本的区分。对于原始数据, 首先使用gscripsts^[19]中的demux_paired_end.py根据eCLIP建库时紧靠着靶标RNA序列上游的条形码序列(barcode)进行文库分拆

(demultiplexing), 所有不含barcode的reads直接在该步骤被过滤掉。同时, 该脚本可以将建库时的随机序列(random-mers)等相关信息提取到read名称(read name)中作为reads的唯一识别码, 作为后续去除PCR重复的依据。

随后去接头的过程对于后续的两次比对至关重要。根据不同CLIP技术的建库规则和特点, 使用Cutadapt^[23]进行个性化的参数设定可以满足绝大多数接头形式的reads修剪。有时运行一轮不足以去掉两端的接头, 需要更换参数再运行一次, 直至完全去除接头。此时, 可以使用FastQC软件再次对接头进行检查, 也可以手动抽查一些reads对去接头前后的序列进行比较, 以此来判断接头是否去除干净。

在eCLIP中, 数据经过预处理, 最终只留下干净的靶标RNA序列供后续的比对与分析(图2C)。

1.3 分离重复序列

通过STAR^[24]软件比对相应物种的Repbase数据库(本文使用的是Human RepeatMasker, version hg38)分离出比对上的重复序列(repetitive elements)(rRNA等大量的重复序列reads可能会造成虚假的现象)。这些结合到重复序列上的reads多数都是因为重复序列的高表达产生的背景信号, 但不排除其中有真实的结合到重复序列上的信号, 这部分数据可以单独分析, 需要更加量化的分析处理。本文不再细述。有兴趣的读者, 可以参考文献[25]。

1.4 比对参考基因组

使用上一步未比对上重复序列的reads, 通过STAR软件比对参考基因组(本文使用的是UCSC, 人参考基因组版本hg38)。通过调整比对参数来控制比对的严格程度使其适中, 过低或者过高的比对率都会影响后续的分析结果。过低的比对率可能会导致遗漏大量的结合位点和靶标, 使得motif和靶标的分析结果与实际情况产生较大误差, 并且造成数据量的浪费; 过高的比对率可能会带来许多假阳性的结合位点, 影响结果准确度。

1.5 去除PCR重复

在构建文库的PCR扩增步骤, 会人为产生靶标RNA的重复克隆, 这些PCR重复序列(PCR duplicates)

注：RC D70X/D50X Primer=D70X/D50X Primer上的一段反向互补序列，属于linker/adaptor相关序列的一部分

图 2 eCLIP建库与数据预处理总览。A: IP样本的建库(以D501和D701 PCR引物为例); B: SMInput样本的建库(以D501和D701 PCR引物为例); C: 数据预处理各步骤的效果。N代表随机碱基，包括5位和10位的随机序列(NNNNN和NNNNNNNNNN)。有@符号的单元格表示在read名称中新增加的信息

Figure 2 Summary of eCLIP library construction and data preprocessing. A: Library construction of IP samples (D501 and D701 PCR primers as examples); B: library construction of SMInput samples (D501 and D701 PCR primers as examples); C: preprocessing results of each step. N stands for random base, including 5-mers and 10-mers of random-mers (NNNNN and NNNNNNNNNN). Cells with @ sign indicate newly added information in the read names

都具有以下特征：含有相同的5-mer和10-mer随机序列。使用每个reads独特的随机序列作为依据，去除因PCR扩增造成的重复reads，可以使用gscripts中的barcode_collapse_pe.py作用于bam文件，也可以使用CTK (CLIP tool kit)^[26]中的fastq2collapse.pl作用于fastq文件。

如果对双端测序Read1和Read2的fastq文件分别进行去除PCR重复后，再使用STAR进行Rebase和参考基因组比对会出现Read1和Read2不一致的报错，因此可以在比对之前使用Sickle(<https://github.com/najoshi/sickle>)的PE模式对低质量、不一致的reads进行过滤。

1.6 reads最终处理

通过Picard工具对上一步去除PCR重复的bam文件进行排序后，再使用Samtools^[27]提取Read2，后续所有的分析都是针对Read2的。最后通过CTK中的parseAlignment.pl获得比对的特异reads以及突变信息。

1.7 获得富集峰、富集峰的标准化的筛选

为了获得reads显著堆叠富集的基因组坐标，可以使用CTK中tag2peak.pl的算法对所有比对的特异reads进行peak calling，同时根据物种特异的统计学背景模型进行第一轮的标准化的筛选，筛选一些低于统计学模型上预期高度的富集峰(peaks)。第一轮标准化适用于

没有生物学背景对照的样本数据(这种数据无法进行第二轮标准化), 有生物学背景对照的数据可以选做。

如果不需要或者数据质量不支持进行第一轮标准化, 可以在tag2peak.pl的命令中删除与标准化相关的参数直接运行获得所有真实的peaks, 然后通过gscripts中的Peak_input_normalization_wrapper.pl同生物学背景SMInput进行第二轮标准化, 赋予每一个peak一个相对于背景信号的 P 值(P -value)和倍数(fold change, fd, -fold), 作为后续筛选peaks的依据。对于符合条件的数据第二轮标准化是必做的。推荐筛选标准为 P -value<0.05, fd至少要2倍以上, 两个筛选指标至少要选择一种进行筛选以获得有意义的peaks。经过统计学背景模型和生物学背景Input的两次筛选, 以获得最具有统计学和生物学意义的peaks, 可以尽可能降低peaks的假阳性。

1.8 注释、靶标分析和motif分析

根据每个peak对应的相对于Input背景的 P -value和fd信息, 筛选出符合要求的所有peaks, 使用HOMER^[28]软件中的annotatePeaks.pl对标准化后符合要求的peaks进行对应参考基因组的注释(annotation)。通过这些peaks对应的基因坐标, 可以得到它们所属的靶标基因信息和其所属的区域等信息, 以及RBP的靶标列表以及结合区域分布情况, 以此为基础可以进行后续相关的GO, KEGG等富集分析(本文使用的Metascape在线网站), 对RBP的靶标参与的生物学过程以及通路进行分析, 寻找可能的分子调控机制。当然顺序上是可以调整的, 即可以先对所有peaks进行注释再进行筛选和后续分析。关于RBP的motif分析, 可以使用HOMER软件中的findMotifsGenome.pl, 该脚本可以利用Peaks的基因坐标使用统计学算法推测RBP识别并结合的motif。

1.9 紫外交联突变位点分析

交联位点是RBP与靶标RNA直接紧密结合的位点, 紫外交联时RBP与RNA会在此处形成共价键, 在逆转录靶标RNA片段建库之前需要先消化掉RNA上的RBP, 交联位点会留下氨基酸残基, 在逆转录RNA时会发生突变, 其中缺失、插入、替换这三类点突变位点被称为紫外交联介导的突变位点(cross-linking induced mutation sites, CIMS), 截断位点被称为紫

外交联介导的截断位点(crosslink-induced truncation sites, CITS)。从原理上来说, 通过peaks找到的RBP结合位点与CIMS/CITS位点在基因坐标上应该是高度相关的。因此, 可以通过CIMS/CITS分析从单个碱基分辨率上去验证通过peaks找到的motif的可信性。还要考虑依赖peaks找到的motif有可能只是结合位点附近的富集序列, 并非RBP真正识别并结合的motif。在CTK的网站上有具体的获取CIMS/CITS位点的工具和参考命令(https://zhanglab.c2b2.columbia.edu/index.php/ECLIP_data_analysis_using_CTK)。利用CTK找到CIMS/CITS位点后, 可以获得以这些位点为中点的一定长度的序列(本研究默认使用-50/+50 nt), 在Anaconda Spyder中输入Python脚本代码(网络版补充数据3), 填充相关信息后可以分析motif周围CIMS/CITS位点的富集情况。如果motif上或者紧贴着motif的个别位点有显著富集, 则说明RBP直接结合该motif的可能性很大。

1.10 重复性分析与后续分析

CLIP-seq通常会有多个生物学重复的样本, 通常需要对3~4个生物学重复样本的靶标以及peaks的分布和高度等参数进行重复性分析, 以验证CLIP实验的质量。同时, 通过IGV软件对数据进行可视化, 观察靶标上的reads分布在各个重复样本之间的重复性。重复性不好的数据可能是实验导致, 也可能是测序导致, 对重复性相对于其他重复样本明显欠佳的样本, 需要考虑是否将其舍去。1.8节中提到的motif分析和靶标分析都可以对每个生物学重复独立进行然后比较分析结果的重复性, 当然, 更推荐选择多个重复性较好的重复样本中的peaks或靶标的交集进行motif分析和富集分析, 以得到更加可靠的结果。

此外, 高通量转录组测序, 如RNA-seq, 通常会对各个转录本上的reads进行标准化定量, 比如使用FPKM/RPKM(Fragments/reads per kilobase of exon model per million mapped fragments/reads)、TPM(transcripts per kilobase of exon model per million mapped reads)等作为表达量, 通过比较在两组之间的不同基因上的表达量差异找出差异基因, 多组CLIP-seq数据之间也可以这样分析, 但是标准化定量后的值反映的不是表达量, 而是结合量。经过标准化定量处理后可以聚类分析, 从而找到同一种RBP在不同组织或者条件下对靶标结合能力的异同点, 或者不同RBP在同

一组织或条件下的对靶标结合能力的异同点, 从而揭示相关RBP的结合和功能特征.

2 结果

Pumilio(Pum)/Puf家族蛋白在真核生物中广泛存在, 该家族蛋白普遍是通过一种序列特异性的方式结合单链靶标RNA的3' UTR区域来发挥功能^[29]. 使用本研究中的CLIP-seq分析流程, 本团队对人前列腺癌DU145细胞系中PUM1的CLIP-seq数据进行分析. 数据量的实时监测结果显示, 关键步骤的数据量变化趋势是逐步递减的, 说明该流程从数据处理的原理上来说是合理的(表4).

为了探究人前列腺癌DU145细胞系中PUM1作为保守的Pum家族蛋白成员在靶标mRNA中主要结合的区域, 本团队通过分析DU145细胞系RNA-seq数据(GEO: GSM4816848)得到了mRNA各区域的peaks占比, 用以近似代表该细胞系中表达的mRNA中各区域的占比(作为对照). 相较于5' UTR、启动子-转录起始位点(promoter-transcription start sites, promoter-TSS)和外显子(exon)区域, 靶标mRNA的3' UTR/TTS区域的peaks在CLIP中不但净占比高, 而且相较于对照是明显上升的, 而exon和5' UTR/promoter-TSS区域都是下降的(图3A, 网络版补充数据4A). 尽管PUM1结合的exon中的peaks净占比也很高, 但是该区域peaks的高度水平(peaks中的reads数)通过 t 检验计算是显著低于3' UTR和TTS的(图3B), 这提示PUM1在前列腺癌细胞中主要结合靶标mRNA的3' UTR/TTS区域, 该结论与Pum家族蛋白在其他物种或组织中的靶标结合特征是一致的.

通过对peaks位点和CITS位点的分析, 本团队找到了统计学上PUM1结合最显著的motif, UGUAAUA (H代表A/U/C, 图3C). 另外, 在该motif周围也存在大量CITS位点的富集, 而一个随机的8-mer motif(AUC-

GAUCG)周围是没有CITS位点富集的, 提示该motif是PUM1直接结合的序列(图3D). 通过对靶标peaks的注释, 可以得到PUM1在前列腺癌细胞中潜在的候选靶标集合, 利用IGV^[30]软件可以对靶标上的reads进行可视化, 从而更直观地展示PUM1在靶标基因上结合特征(图3E). 本团队发现, 识别的候选lncRNA靶标*MA-LAT1*靠近5'端和mRNA靶标*SLC38A2*的3' UTR有显著高于Input水平的reads富集, 提示PUM1与这两个候选靶标RNA的特定区域的结合. 此外, 使用这些靶标基因的集合, 本团队对这些基因进行GO, KEGG通路富集分析(图3F), 找到了PUM1在前列腺癌细胞中可能参与的生物学过程和代谢通路, 以及对应的靶标基因, 这些都是揭开PUM1在前列腺癌细胞中作用机制的关键.

在HEK 293细胞中, 在生长因子的刺激下, PUM1能更好地与*CDKN1B* (p27)的3' UTR结合, 诱导其与miR-221和miR-222的关联, 从而抑制p27的表达, 这种抑制是进入细胞周期所必须的^[31]. 在小鼠中, 也有报道Pum1结合*Cdkn1b*基因的3' UTR并抑制它, 促进G1-S转化, 增加细胞增殖^[32]. 在前列腺癌细胞中, 通过本次CLIP-seq分析依旧在PUM1的靶标中找到了*CDKN1B*, 并且通过RIP(RNA-binding protein immunoprecipitation)-qPCR(quantitative real-time PCR)对PUM1与该靶标的互作结合进行了实验验证(网络版补充数据4B). 通过生物信息分析与实验相结合, 对PUM1保守靶标*CDKN1B*的捕获与验证, 一方面符合PUM1的保守性, 证明了流程的可靠性; 另一方面, PUM1通过结合*CDKN1B*以及其他相关的靶标不仅参与细胞增殖生长、分化发育等生物学过程, 而且根据本次的结果, PUM1有可能结合调控*CDKN1B*以及其他相关的靶标参与到蛋白多糖、miRNA的癌症相关通路.

以上结果仅作为演示本研究的生物信息学流程, 展示该方法可行性的个例, 完整的靶标鉴定和motif分析研究需要完成至少3次生物学重复的eCLIP实验, 但

表 4 PUM1 CLIP-seq输出数据情况总结

Table 4 Summary of PUM1 CLIP-seq output data

reads数				peaks数		靶标基因数(≥ 4 fd)	
原始数据	分离重复序列后	比对参考基因组后	去除PCR重复后	总数	生物学背景Input标准化筛选后(≥ 4 fd)	全基因组	3' UTR和TTS
63482295	44028979	15050713	2377525	309887	257144	16858	5871

图 3 PUM1 CLIP-seq分析结果. A: 信号富集高于Input背景4倍和16倍的peaks在基因组上的区域分布情况; B: 靶标mRNA上各基因组区域的peaks高度水平; C: 分别使用peaks(≥ 4 倍)和CITS($P\text{-value}\leq 0.01$)位点前后200 bp范围内序列进行PUM1 motif分析; D: UGUHAUA motif周围CITS位点的富集分析(蓝线), 一个随机8位motif作为阴参(红线); E: PUM1候选靶标上reads的可视化展示, reads显著富集形成了peaks, 左上角标注了轨道中最高peaks高度的范围, 箭头表示转录本方向; F: 选取富集倍数大于16倍的位于3' UTR/TTS区域的peaks, 对其所属的1189个基因进行GO和KEGG分析

Figure 3 Results from PUM1 CLIP-seq bioinformatics analysis. A: Genomic distribution of peaks with more than a 4-fold and 16-fold enrichment against the size-matched input; B: peak height in different genomic regions; C: PUM1-binding motif analysis within 200 bases upstream and downstream of the peaks (≥ 4 fd) and CITS ($P\text{-value}\leq 0.01$) sites, respectively; D: enrichment analysis of CITS sites around the UGUHAUA motif (blue line), with a 8-mer random motif as a negative control (red line); E: visualization of reads on 2 identified targets of PUM1. Reads that were enriched significantly form peaks. Height range of the peaks in the track is displayed in the upper left corner, and the arrow represents the transcriptional direction. F: GO and KEGG analyses of 1189 targets which locate in 3' UTR/TTSs with a fold change of more than 16

是生物信息学流程分析都是可以按照以上方法, 在获得多次生物学重复的peaks交集和靶标交集后进行后续的靶标和motif分析。

为了更好地验证该流程对于不同来源、不同RBP的CLIP-seq数据分析的普适性与可靠性, 本团队通过下载ENCODE数据库(<https://www.encodeproject.org>)中K562细胞系RBFOX2蛋白的eCLIP数据(ENCSR-756CKJ)及其对应的SMInput数据(ENCSR0511XX), 找到了与已有报道一致的UGCAUG motif(网络版补充数据4C)。因此, 本流程可以分析不同来源的RBP的CLIP-seq数据并找到其特异的已知的结合motif, 具有普适性与可靠性。

3 讨论

本研究通过融合多种高通量转录组测序生物信息学分析软件和脚本并进行优化与补充, 构建了一套新的、从本土化改良后的eCLIP实验流程(网络版补充数据5和6)到原始数据处理到结果输出的简洁又全面的CLIP-seq生物信息学分析流程, 提供了经过多次分析实验得到的关键步骤的原创终端命令参数方案(网络版补充数据1)以及原创的CIMS/CITS可视化分析脚本(网络版补充数据3), 尽可能地让缺乏生物信息学基础的研究人员也可以对照本文流畅地运行整个分析流

程, 拿到最有价值的CLIP-seq分析结果, 这有助于CLIP技术在国内缺少生信专业人员的传统生物学实验室的进一步推广, 并且通过CLIP-seq的分析能够揭示更多在全基因组层面RBP与靶标结合和作用机制的特征。

利用该流程, 本研究初步分析了人前列腺癌DU145细胞系中PUM1对靶标RNA的结合特征。根据一个生物学重复的CLIP-seq分析的结果, 本团队发现, PUM1在前列腺癌细胞中更倾向于结合3' UTR/TTS上保守的UGUAHAUA motif, 这个结果与已知的在多物种中Pum家族蛋白结合的motif一致或者相似, 都包括最核心的UGUA序列^[33]。虽然最富集的CITS位点在UGUAHAUA上代表着PUM1与该motif的直接结合, 但是本团队注意到motif周围的CITS位点的出现比正常生理环境偏多, 这可能与癌细胞环境中靶标突变的概率增加有关。

目前在多个物种中有报道的Pum家族蛋白参与的生物学过程主要有胚胎发育^[34,35]、神经系统的功能和发育^[36-38]、干细胞与生殖细胞的维持^[38-41]、rRNA进程以及核糖体发生^[42,43]以及趋化性的细胞运动^[44]等, 本团队的GO富集分析结果能够找到类似的生物学过程。这样的结果既符合Pum家族蛋白的序列与功能的高度保守性, 也验证了本研究分析流程的可靠性。在后续的研究中, 本团队会对更多重复进行分析以得到更准确的靶标和功能线索并进行验证。

致谢 感谢哥伦比亚大学医学中心的张朝林教授和冯会娟博士对CLIP-seq分析流程构建的指导。感谢加州大学Gene W. Yeo实验室, 尤其是Eric L. Van Nostrand博士在数据方法方面的分享和指导。感谢山东农业大学信息科学与工程学院农业大数据研究中心的孙凯对于分析流程建立和分析脚本方面的帮助。感谢南京信息工程大学的徐军教授团队和马凯特大学计算机科学部的耿宝宝对实验室生信平台构建的帮助。感谢南京医科大学生殖医学国家重点实验室的赵婷婷老师对于文章的修改建议和帮助。

参考文献

- 1 Sharp P A. The centrality of RNA. *Cell*, 2009, 136: 577-580
- 2 Zhang C, Gao L, Xu E Y. LncRNA, a new component of expanding RNA-protein regulatory network important for animal sperm development. *Semin Cell Dev Biol*, 2016, 59: 110-117
- 3 Chen L L, Yang L. Regulation of circRNA biogenesis. *RNA Biol*, 2015, 12: 381-388
- 4 Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*, 2014, 15: 829-845
- 5 Scotti M M, Swanson M S. RNA mis-splicing in disease. *Nat Rev Genet*, 2016, 17: 19-32
- 6 Van Nostrand E L, Freese P, Pratt G A, et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 2020, 583: 711-719

- 7 Corley M, Burns M C, Yeo G W. How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol Cell*, 2020, 78: 9–29
- 8 Xu E Y, Moore F L, Pera R A. A gene family required for human germ cell development evolved from an ancient meiotic gene conserved in metazoans. *Proc Natl Acad Sci USA*, 2001, 98: 7414–7419
- 9 Vangompel M J W, Xu E Y. The roles of the DAZ family in spermatogenesis: more than just translation? *Spermatogenesis*, 2011, 1: 36–46
- 10 Shah C, Vangompel M J W, Naeem V, et al. Widespread presence of human BOULE homologs among animals and conservation of their ancient reproductive function. *PLoS Genet*, 2010, 6: e1001022
- 11 Watanabe T, Lin H. Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs. *Mol Cell*, 2014, 56: 18–27
- 12 Kato Y, Katsuki T, Kokubo H, et al. Dazl is a target RNA suppressed by mammalian NANOS2 in sexually differentiating male germ cells. *Nat Commun*, 2016, 7: 11272
- 13 Xiong B Y, Ye Y F, Cui L Y, et al. Advances on protein-RNA interaction identification technology (in Chinese). *Curr Biotechnol*, 2020, 10: 217–225 [熊冰钰, 叶均瑶, 崔凌云, 等. 蛋白质-RNA相互作用鉴定技术研究进展. *生物技术进展*, 2020, 10: 217–225]
- 14 Keene J D. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 2007, 8: 533–543
- 15 Ule J, Jensen K B, Ruggiu M, et al. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 2003, 302: 1212–1215
- 16 Licatalosi D D, Mele A, Fak J J, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 2008, 456: 464–469
- 17 Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 2010, 141: 129–141
- 18 König J, Zarnack K, Rot G, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 2010, 17: 909–915
- 19 Van Nostrand E L, Pratt G A, Shishkin A A, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, 2016, 13: 508–514
- 20 Van Nostrand E L, Nguyen T B, Gelboin-Burkhart C, et al. Robust, cost-effective profiling of RNA binding protein targets with single-end enhanced crosslinking and immunoprecipitation (seCLIP). *Methods Mol Biol*, 2017, 1648: 177–200
- 21 Zhao Y, Zhang Y, Teng Y, et al. SpyCLIP: an easy-to-use and high-throughput compatible CLIP platform for the characterization of protein-RNA interactions with high accuracy. *Nucleic Acids Res*, 2019, 47: e33
- 22 Andrews S. Babraham Bioinformatics -FastQC A Quality Control tool for High Throughput Sequence Data. 2013
- 23 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 2011, 17: 10–12
- 24 Dobin A, Davis C A, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013, 29: 15–21
- 25 Van Nostrand E L, Pratt G A, Yee B A, et al. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol*, 2020, 21: 90
- 26 Shah A, Qian Y, Weyn-Vanhenryck S M, et al. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, 2017, 33: 566–567
- 27 Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, 25: 2078–2079
- 28 Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 2010, 38: 576–589
- 29 Spassov D S, Jurecic R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life*, 2003, 55: 359–366
- 30 Robinson J T, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*, 2011, 29: 24–26
- 31 Kedde M, van Kouwenhove M, Zwart W, et al. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*, 2010, 12: 1014–1020
- 32 Lin K, Qiang W, Zhu M, et al. Mammalian Pum1 and Pum2 control body size via translational regulation of the cell cycle inhibitor *Cdkn1b*. *Cell Rep*, 2019, 26: 2434–2450.e6
- 33 Nishanth M J, Simon B. Functions, mechanisms and regulation of Pumilio/Puf family RNA binding proteins: a comprehensive review. *Mol Biol Rep*, 2020, 47: 785–807
- 34 Murata Y, Wharton R P. Binding of pumilio to maternal *hunchback* mRNA is required for posterior patterning in *Drosophila* embryos. *Cell*, 1995, 80: 747–756

- 35 Lin K, Zhang S, Shi Q, et al. Essential requirement of mammalian *Pumilio* family in embryonic development. *Mol Biol Cell*, 2018, 29: 2922–2932
- 36 Chen G, Li W, Zhang Q S, et al. Identification of synaptic targets of *Drosophila pumilio*. *PLoS Comput Biol*, 2008, 4: e1000026
- 37 Muraro N I, Weston A J, Gerber A P, et al. Pumilio binds *para* mRNA and requires Nanos and Brat to regulate sodium current in *Drosophila* motoneurons. *J Neurosci*, 2008, 28: 2099–2109
- 38 Zhang M, Chen D, Xia J, et al. Post-transcriptional regulation of mouse neurogenesis by Pumilio proteins. *Genes Dev*, 2017, 31: 1354–1369
- 39 Naudin C, Hattabi A, Michelet F, et al. PUMILIO/FOXP1 signaling drives expansion of hematopoietic stem/progenitor and leukemia cells. *Blood*, 2017, 129: 2493–2506
- 40 Forbes A, Lehmann R. Nanos and Pumilio have critical roles in the development and function of *Drosophila* germline stem cells. *Development*, 1998, 125: 679–690
- 41 Crittenden S L, Bernstein D S, Bachorik J L, et al. A conserved RNA-binding protein controls germline stem cells in *Caenorhabditis elegans*. *Nature*, 2002, 417: 660–663
- 42 Yu J, Cho H C, Huang Y, et al. Human Puf-A, a novel component of 90S pre-ribosome, links ribosome biogenesis to cancer progression. *bioRxiv*, 2018
- 43 Wang Z, Sun X, Wee J, et al. Novel insights into global translational regulation through Pumilio family RNA-binding protein Puf3p revealed by ribosomal profiling. *Curr Genet*, 2019, 65: 201–212
- 44 Hotz M, Nelson W J. Pumilio-dependent localization of mRNAs at the cell front coordinates multiple pathways required for chemotaxis. *Nat Commun*, 2017, 8: 1366

A bioinformatics pipeline for processing CLIP-seq data of RNA-binding proteins and its application

XIE ZhengYao, ZANG Min, LI Xin & XU YuJun

State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing 211166, China

UV cross-linking immunoprecipitation followed by high-throughput sequencing (CLIP-seq) has emerged to be a powerful tool in dissecting mechanisms of RNA-binding proteins (RBP). Based on a recently developed eCLIP method by Yeo lab at UCSD, we performed target identifications and motif analyses on several RNA-binding proteins and established a bioinformatic pipeline for CLIP-seq data of RNA-binding proteins. By integrating the analysis software and scripts from different sources and analyzing RNA-binding protein PUMILIO1(PUM1) CLIP-seq data from a human cancer cell line, we identified 1189 RNA targets and found that PUM1 preferentially bound to the 3' untranslated region (3' UTR) and transcription termination sites (TTS) of its target mRNAs. The motif UGUAAUA (H stands for A/U/C) is consistent with the canonical motif sequence of PUM1 for target recognition. Hence PUM1 may be involved in the proliferation and differentiation of cancer cells by regulating target gene expression at the post-transcriptional level via binding to the 3' UTR.

RNA-binding protein, UV cross-linking immunoprecipitation, high throughput sequencing, bioinformatics, PUM1, prostate cancer, 3' UTR

doi: 10.1360/SSV-2020-0403