



紫外交联和蛋白免疫共沉淀技术的演进与创新*

赵佳敏¹⁾ 卢城江²⁾ 杨明²⁾ 那顺布和^{1)**} 王纲^{2)**}⁽¹⁾ 内蒙古大学生命科学学院, 呼和浩特 010021; ⁽²⁾ 复旦大学生命科学学院, 上海 200438

摘要 RNA结合蛋白(RBPs)在细胞中广泛存在,在多种生物学过程中发挥着重要作用。紫外交联和蛋白质免疫共沉淀(UV cross-linking and immunoprecipitation, CLIP)技术是一种用于鉴定蛋白质直接结合RNA的技术。随着高通量测序技术的发展,CLIP技术已成为鉴定RNA结合蛋白所结合的具体RNA种类和序列的关键方法。为了满足不断增长的科研需求,CLIP技术经历了多次改进,衍生出多种版本,如HITS-CLIP、PAR-CLIP和iCLIP等。日益完善的CLIP技术在揭示新型蛋白质-RNA相互作用、解析RNA修饰和调控机制等方面具有重要价值,在RNA生物学研究中不可或缺。本文详细回顾了这些技术的演进与创新,比较了它们的优势与局限性,探讨了近年来CLIP技术在非经典蛋白质-RNA相互作用的研究中的应用,并展望了CLIP技术在未来RNA生物学中的创新应用与发展方向。

关键词 紫外交联和蛋白质免疫共沉淀, RNA结合蛋白, RNA, 测序

中图分类号 Q-33, Q5-33

DOI: 10.16476/j.pibb.2024.0400

CSTR: 32369.14.pibb.20240400

蛋白质与RNA的相互作用是mRNA转录、剪接、运输、翻译和降解等所有RNA活动的基础,调控了细胞的生存、复制、分化和逆境抵抗等生命活动的方方面面^[1]。细胞内存在大量的RNA结合蛋白(RNA-binding protein, RBP)。这些RBP一部分可以通过经典的RNA结合结构域,如RNA识别基序(RNA recognition motif, RRM)、K同源结构域(K homology domain, KH)、冷休克结构域(cold shock domain, CSD)、类Sm结构域(like-Sm domain, LSM)等^[2]直接结合RNA,另一部分可以通过非经典的RNA结合结构域^[3]实现与RNA的结合。紫外线交联和蛋白质免疫共沉淀(UV cross-linking and immunoprecipitation, CLIP)技术是鉴定RBP所结合的RNA的种类和序列的一种常用技术。CLIP技术的核心原理是通过紫外线照射细胞,使RBP与其直接结合的RNA之间形成共价交联,随后对靶蛋白进行免疫沉淀或亲和纯化,进而将交联到靶蛋白上的RNA进行高通量测序。得到的测序结果不仅可以分析出目标蛋白质所具体结合的RNA种类和序列,还可以得到蛋白质在RNA上结合的位置以及基序(motif)等信息。CLIP的具体操作步骤大致分为细胞紫外交联、细

胞裂解、靶蛋白的免疫沉淀、RNA纯化与测序等多个步骤(图1)。CLIP-Seq技术起源于1990年代后期^[4],随着二代测序技术(next generation sequencing, NGS)的出现,CLIP技术得到了极大的发展,从早期的高通量测序结合RNA免疫共沉淀(high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation, HITS-CLIP)^[5]、光激活核苷增强交联及免疫共沉淀(photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation, PAR-CLIP)^[6]、单核苷酸分辨率交联和免疫共沉淀(individual-nucleotide resolution CLIP, iCLIP)^[7],再到近些年的增强型交联和免疫共沉淀(enhanced CLIP, eCLIP)^[8]、简化型交联和免疫共沉淀(simplified CLIP, sCLIP)^[9]、基于spyTag的交联和免疫共沉淀(spyTag-based CLIP, SpyCLIP)^[10]、变性交联和免

* 国家重点研发计划(2022YFA0806200)和国家自然科学基金(32030028)资助项目。

** 通讯联系人。

那顺布和 Tel: 0471-4996885, E-mail: bnashun@imu.edu.cn

王纲 Tel: 13916396401, E-mail: gwang_fd@fudan.edu.cn

收稿日期: 2024-09-11, 接受日期: 2024-11-26

疫共沉淀 (denaturing CLIP, dCLIP)^[11], 以及交联、连接和杂交测序 (cross-linking, ligation, and sequencing of hybrid, CLASHs)^[12] 等, CLIP-seq 技术在不断进行演进和优化, 整体朝着操作更简单

高效、测序结果覆盖度更高、实验误差更低的方向发展。本综述对CLIP实验中的多个重要步骤进行了梳理和分析, 并总结了技术的变化及其未来的发展方向。

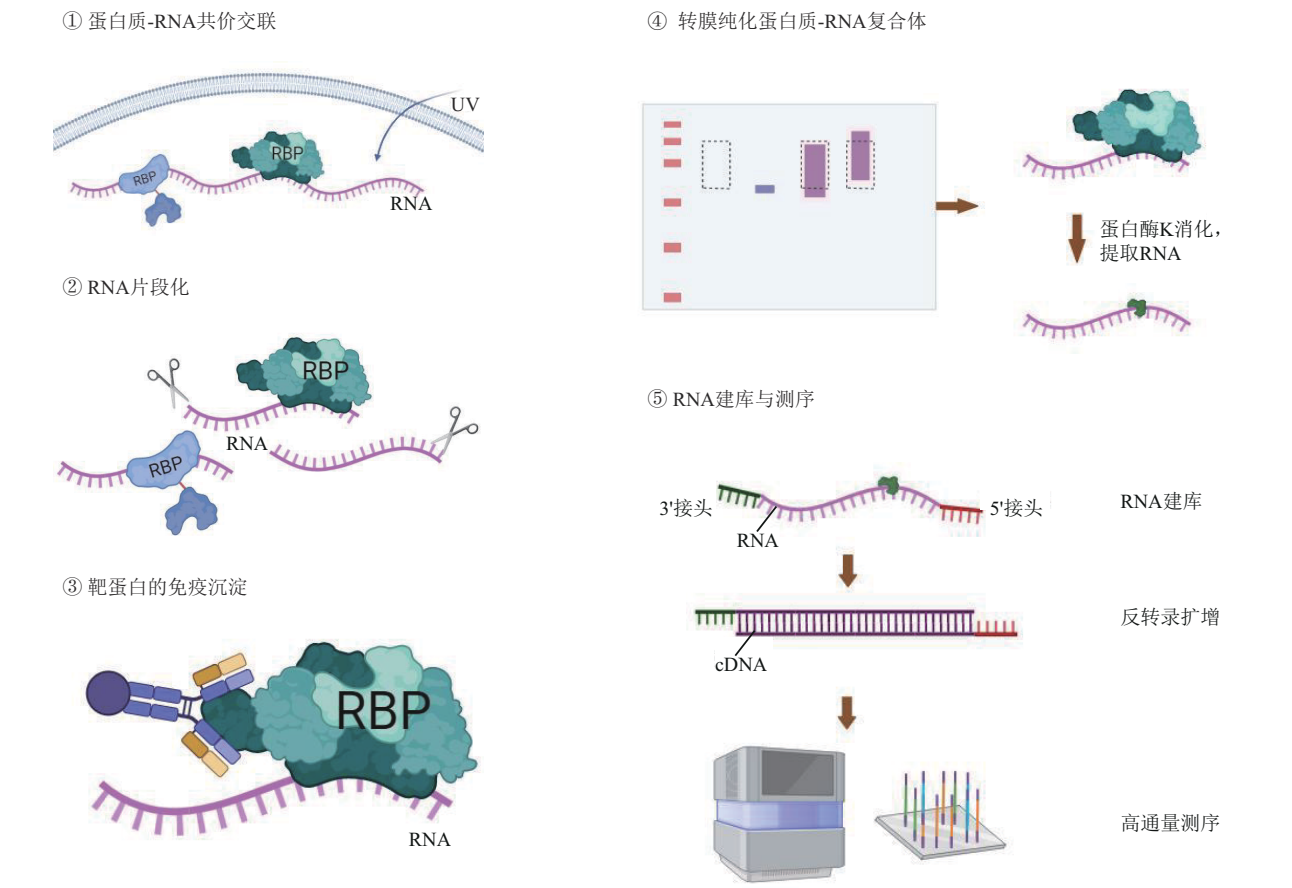


Fig. 1 Core steps of UV cross-linking and immunoprecipitation (CLIP)
图1 交联免疫沉淀 (CLIP) 的核心步骤 (<https://app.biorender.com>)
RBP: RNA结合蛋白 (RNA binding protein)。

1 靶蛋白-RNA复合体的富集与纯化

1.1 紫外交联

紫外交联步骤是CLIP实验的第一步。高能紫外线(如254 nm波长的UVC紫外线)激发了部分核苷酸(主要是尿嘧啶核苷酸^[3, 13-16])和部分氨基酸(主要是苯丙氨酸(phenylalanine, F)、赖氨酸(lysine, K)和甘氨酸(glycine, G)等^[3, 13-16])的原子基团, 使它们进入高能活跃状态, 在空间距离足够近的条件下, 它们发生共价结合, 释放能量回到稳定状态^[17]。紫外交联使得RBP和其直接结合的RNA之间形成共价键, 共价键的形成使得蛋白质与RNA之间的结合可以承受

严苛的清洗条件, 以便尽可能地清除背景RNA和非特异性结合的RNA^[4]。

影响蛋白质和RNA之间共价交联效率的因素有两个——紫外线的波长和交联总能量。使用的常见紫外波长有254 nm和365 nm, 如HITS-CLIP^[5]、iCLIP^[7]和eCLIP^[8]使用254 nm的UVC紫外线, 但UVC交联的效率较低^[6]。PAR-CLIP^[6]使用365 nm的UVA紫外线或者特殊情况下使用310 nm紫外线波长^[18](图2)。365 nm的UVA波长较长, 但需要配合4-硫尿苷(4-thiouridine, 4-sU)、6-硫基鸟嘌呤(6-thioguanine, 6-sG)等核苷酸类似物使用^[6]。4-sU和6-SG可以加入到细胞培养基中并被细胞吸收随后被掺入到新合成的RNA中。这类

核苷类似物在 365 nm 的紫外线下具有极高的交联效率,其交联蛋白的效率是 254 nm 下的普通核苷酸的交联效率的数百倍^[6]。尽管 PAR-CLIP 有更高的紫外交联效率,但是因为需要额外加入 4-sU,增加了操作的复杂度,因此,基于 UVC 紫外线的交联仍广泛应用在大部分 CLIP 技术中。

交联能量也是决定蛋白质-RNA 交联效率的重要因素,不同的 CLIP 方法使用的能量不一样,例如 PAR-CLIP 使用 UVA 辐射下的 360 mJ/cm² 的能量强度^[6],iCLIP 使用 UVC 辐射下 150 mJ/cm² 的能量强度^[7],eCLIP 和 dCLIP 使用 UVC 辐射下 400 mJ/cm² 的能量强度^[8, 11]。总体上来讲,目前所有的 CLIP 方法使用的交联能量强度分布在 100~500 mJ/cm² 之间。能量的选择需要根据具体的适用操作环境来优化和选择。由于紫外线穿透能力不强,即便是在同样的能量强度数值下,样品的厚度、样品溶液的深度、样品距离紫外灯管的距离都会影响实际的交联效果。在实际操作中,这些因素都应该予以考虑,确保实验的可重复性。例如,紫外线在水溶液中的穿透能力很差,因此在对细胞进行紫外交联时,是否保留覆盖在细胞上的培养基或者磷酸盐缓冲液(phosphate-buffered saline, PBS)是一个需要着重考虑的因素。在 PAR-CLIP 实验中,细胞在完全去除培养基的条件下进行紫外交联^[6];在 iCLIP、eCLIP 等其他 CLIP 方法中,细胞在 PBS 中进行紫外交联^[7],而加入的 PBS 的量在不同方法之间并不一致。我们建议在完全去掉培养基或者仅加入恰好覆盖细胞表面的 PBS 的情况下进行紫外交联。因为紫外线的波长较短,溶液越深,抵达细胞的紫外线越少,交联效果越差。第一次实验之前,建议用电泳或者蛋白质印迹法(Western blot)检测目标蛋白质的条带位移(shift)程度的方式筛选出最佳的交联能量。比如,将不同的 UV 交联条件处理的细胞进行 Western blot,在转膜时保留上样孔和浓缩胶(上层胶),对整张膜进行典型的 RBP(如 HNRNPC)在上样孔和浓缩胶处的迁移条带的检测。交联效果越好,则迁移条带的颜色越深。

1.2 靶蛋白富集

CLIP 实验本质上是通过对靶蛋白的富集和纯化来获得其直接结合的 RNA。目前,大部分 CLIP 实验采用 Protein G 或 Protein A 磁珠(beads)偶联靶蛋白抗体的方法来对靶蛋白进行免疫沉淀(immunoprecipitation, IP)。磁珠自身结合的 RNA 以及靶蛋白的相互作用蛋白所结合的 RNA 会造成

严重的背景 RNA 污染,因此,后续需要严苛地清洗以降低背景 RNA。大部分 CLIP 方法采用含有高浓度的盐离子、去垢剂甚至一定浓度蛋白质变性剂的缓冲液进行磁珠的清洗,因而对抗体结合抗原强度的要求较高。当无法获得可用于免疫沉淀的高质量抗体时,可以将靶蛋白质接上标签序列,随后利用标签抗体或者相应的亲和基质富集靶蛋白质。

可供融合表达的标签根据其后续的纯化方式分为两种。第一种为“免疫结合标签”。Flag、HA 等小肽标签,通过与对应的抗体结合来实现靶蛋白的免疫沉淀。其本质依然是基于抗原-抗体的结合,这种结合能够承受的清洗条件的严苛程度有限,与直接使用靶蛋白的抗体在清洗条件的选择上没有本质区别。第二种是“亲和结合标签”。生物素(Biotin)-亲和素(Avidin)之间的结合强度是目前已知的自然界最强大的配体-受体之间的结合能力,而 HaloTag 蛋白与底物小分子之间可以形成共价键,进而可以承受最严苛的蛋白质变性剂的清洗。这一类标签的适用于不基于抗原-抗体结合的纯化方式,可以承受极其严苛的清洗条件,因此可以带来更低的背景。下面着重对标签的类别与特点进行分析和总结。

1.2.1 免疫结合标签

常见的 Flag、HA、V5 等免疫结合标签利用了抗体与抗原相互结合的特性^[19]。它们是来自于病毒或者低等多细胞生物的外源肽段,一般与所研究的细胞内的蛋白质不具有同源性,因此特异性很强。另一方面,这些多肽标签经过多年的商业化改造和筛选,已经获得了很高亲和力的单克隆抗体。商业化的单克隆抗体能够很高效地富集靶蛋白质,而且能抵抗轻度到中度的变性溶液环境,在以往的研究中使用较多。但由于抗体与抗原的结合高度依赖于二者的三维结构,而这种结构无法承受高浓度的蛋白质变性剂处理。因此,对于含有强结合蛋白的靶蛋白来说,这种方法不是最佳选择,因为这些结合蛋白也可能与 RNA 结合,从而导致假阳性结果。

1.2.2 亲和结合标签

亲和结合标签(affinity tag)一般是基于配体-受体结合的标签,可以承受更高强度的蛋白质变性剂的清洗,包括:BioTag、Strep-tag、6×His-tag 及 HaloTag 等。BioTag 是常见的亲和结合标签,其本质是基于 Biotin 与 Avidin 高强度结合的系统^[20]。BioTag 又叫 AviTag,AviTag 是一段长度为 15 个氨

基酸的短肽, 可以被来自于大肠杆菌的 BirA 蛋白特异性识别^[21]。BirA 是生物素化酶, 可以在 AviTag 短肽上添加 Biotin^[22], 后续可以使用链霉亲和素 (Streptavidin) 磁珠进行特异性富集。Biotin-Streptavidin 之间的结合是目前已知的生物界最强的非共价结合体系, 其解离常数 (K_d) 可达 1.015 mol/L , 比“抗体-抗原”之间的亲和常数高大约 100 万倍^[23]。Biotin 与 Streptavidin 结合后, 可以承受高浓度的十二烷基硫酸钠 (sodium dodecyl sulfate, SDS)、尿素和盐离子^[24]。dCLIP 采用 BioTag 系统, 使用 8 mol/L 尿素、2% SDS 和 1 mol/L NaCl 依次对 IP 后的磁珠进行清洗, 获得了极高的信噪比^[11, 25], 即优良的信号叠加较小的背景信号。

Strep-tag 标签属于一种结构类似 Biotin 结构的小肽, 可以与改造自结合 Biotin 的 Streptavidin 蛋白的链霉亲和素类似物 (Streptactin) 蛋白特异性结合^[11]。Strep-tag 与 Streptactin 的结合属于配体-受体结合的类型^[26]。Strep-tag 序列来自于细菌并且解离常数非常小, 因此 Strep-tag 与 Streptactin 的结合具有高特异性和高亲和性的特点, 可以承受尿素等蛋白质变性剂的清洗, 适用于多种场景的蛋白质富集^[27]。应用 Strep-tag 的 CLIP 方法也能较好排除非特异性结合 RNA 和背景 RNA 的干扰^[28]。

$6\times\text{His}$ 标签内的组氨酸能与镍离子等金属离子发生螯合作用。金属螯合作用能够抵抗高强度的变性剂作用, 如 4% 的 SDS、 6 mol/L 的盐酸胍和 8 mol/L 的尿素^[29]。应用极端的变性条件, His-tag 纯化能够高效地去除靶蛋白质的相互结合蛋白质、背景蛋白质、间接结合 RNA 和背景 RNA, 达到很高的特异性。不过, $6\times\text{His}$ 标签也有一定的缺点, 例如组氨酸本身就是细胞内蛋白质的重要组分, 一些蛋白质也含有多个组氨酸的相似序列, 这会导致非特异蛋白质的结合, 但无法通过改变溶液环境去除。采用 $6\times\text{His}$ 标签的交联和 cDNA 分析 (cross-linking and analysis of cDNA, CRAC) 技术, 是最早的采用彻底蛋白质变性策略的 CLIP 技术^[29]。CRAC 使用 6 mol/L 盐酸胍, 一种极强的蛋白质变性剂来进行 His-beads 的清洗, 取得了很高的信噪比^[29]。

HaloTag 蛋白质改造自去卤代烷酶 (haloalkane dehalogenase), 改造后的去卤代烷酶可以与卤代烷小分子实现共价结合^[30], 使用提前偶联有卤代烷小分子的磁珠即可实现对靶蛋白的变性纯化。共价结合的蛋白质能够对抗所有的变性剂,

高效去除背景信号。但高特异性的代价是引进了很大的靶蛋白, 带来了不可忽视的空间位阻和分子量负担。例如 HaloTag 蛋白具有 33 ku 的分子质量, 比部分 RNA 结合蛋白的分子质量都大。过大的分子质量带来了潜在的靶蛋白质功能基团被封闭以及细胞定位异常等风险。省略凝胶的连接依赖型交联和免疫共沉淀 (gel-omitted ligation-dependent CLIP, Gold-CLIP) 利用 HaloTag 蛋白质能共价结合卤烷烃的特点, 将靶蛋白共价富集到含有卤化烷烃的基质上, 后续通过包括 8 mol/L 盐酸胍、 8 mol/L 尿素、10% SDS 在内的严苛清洗条件, 取得了极佳的信噪比和覆盖度^[31]。CRAC 也采用 HaloTag 并取得了类似的效果^[32]。

SpyCLIP 则避免了 HaloTag 的缺点。SpyCLIP 利用 SpyTag 和 SpyCatcher 之间的共价结合进行靶蛋白富集, 其中 SpyTag 是短肽, 可以连接在靶蛋白上^[10]。SpyTag 和 SpyCatcher 可以互相结合并自发形成分子内共价键^[33]。将 SpyCatcher 共价交联到磁珠上后, 其共价键结合体系可以承受饱和浓度的蛋白质变性剂的处理, 类似 HaloTag。值得注意的是, SpyTag 只有 13 个氨基酸, 属于短肽, 类似 Flag、HA 等标签, 与 HaloTag 相比, 对靶蛋白的空间结构影响较小^[33]。因此, 基于 SpyTag 体系的 CLIP 是较为理想的鉴定直接蛋白质-RNA 相互作用的技术方法^[10]。不过 SpyCLIP 需要纯化具有高活性的 SpyCatcher 蛋白, 并将 SpyCatcher 蛋白共价交联到磁珠上, 这增加了实验操作的工作量和复杂性。

1.2.3 纯化方式的选择

尽管使用共价结合的标签可以获得更低背景的测序结果, 但是外源引入的标签可能会对蛋白质的正常功能造成干扰。具体的选择需要根据具体的靶蛋白来确定。在能够获得高质量的可用于 IP 的抗体的情况下, 可以优先使用抗体进行 IP。否则, 建议使用 StrepTag、SpyTag 等小肽标签进行靶蛋白富集纯化。在条件允许的情况下, 为了获得更精确的结果, 可以使用包括抗体、标签配体等多种方式同时进行 CLIP 实验。

1.3 磁珠的清洗

CLIP 实验操作中会引入大量的背景 RNA^[34], 这些 RNA 主要是与靶蛋白相互作用的蛋白质所结合的 RNA, 此外, 还会有非特异性吸附到富集基质 (如磁珠) 上的 RNA, 这也构成了背景 RNA 的一部分。在 CLIP 实验中, 通常会使用洗涤缓冲液

(wash buffer) 清洗免疫沉淀步骤中的磁珠, 以尽可能地去除背景 RNA。由于清洗步骤是 CLIP 实验中降低假阳性的关键步骤, 因此, 洗涤缓冲液的选择至关重要。

不同的 CLIP 方法中使用的洗涤缓冲液强度不同。大致可以分为低强度、中等强度和高强度 3 种。

PAR-CLIP 使用低强度的洗涤缓冲液, 对应的洗涤缓冲液中的“强度”组分是 300 mmol/L KCl、0.5% (v/v) NP40^[6]。300 mmol/L KCl 属于较高浓度的盐溶液, 0.5% NP40 属于非变性去垢剂。上述组合有利于去除部分弱相互作用蛋白以及部分背景 RNA, 一定程度提高信噪比, 但是对靶蛋白紧密结合一些相互作用蛋白则没有作用, 对与基质非特异结合的 RNA 也没有很好的去除效果^[34]。

iCLIP 使用中等强度的洗涤缓冲液, 其洗涤缓冲液中含有 1 mol/L NaCl、1% Igepal CA-630、0.1% SDS、0.5% 脱氧胆酸钠 (sodium deoxycholate)^[7] 等几种成分。后 3 个去垢剂的组成强度与染色质免疫沉淀 (chromatin immunoprecipitation, ChIP) 实验中的细胞裂解液的强度相近, 与 1 mol/L 的 NaCl 组合使用, 其条件严苛程度基本接近大部分抗体的承受极限。eCLIP 等后续类似 CLIP 方法使用与 iCLIP 相似的洗涤缓冲液成分。尽管这类洗涤缓冲液已经可以去除大部分靶蛋白的相互作用蛋白以及背景 RNA, 但是, 如果想去除结合力更强的相互作用蛋白以及尽可能地去除所有背景 RNA, 则需要使用更高浓度的离子型去垢剂 (如 SDS) 或者饱和浓度的盐酸胍和尿素等蛋白质强变性剂, 以及更高浓度的盐溶液, 这种极为严格的清洗条件已经超过了抗体-抗原体系的承受能力。因此, 需要更改靶蛋白富集体系。

dCLIP 使用高强度的洗涤缓冲液, 或者称为“变性级”洗涤缓冲液。这得益于其采用的“生物素-链霉亲和素”富集系统^[11]。其使用的洗涤缓冲液条件是 8 mol/L 尿素、2% SDS 和 1 mol/L NaCl, 属于近乎完全变性^[11]。此条件能完全去除间接结合的蛋白质, 从而去除间接结合的 RNA 和背景 RNA, 取得更好的信噪比。但是细胞内有一定内源水平的 Biotin 及其修饰蛋白质, 它们会带来一定的背景。共价连接和亲和纯化 (covalent linkage and affinity purification, CLAP) 在靶蛋白上融合了 HaloTag 标签, 同样使用了“高强度”的洗涤缓冲液。在 90°C 3min、2% 十二烷基肌氨酸钠

(sodium lauroyl sarcosinate) 和 8 mol/L 尿素这种极为严苛的清洗条件下, CLAP 获得了极佳的信噪比, 排斥了多种被非变性 CLIP (如 iCLIP、eCLIP、HITS-CLIP) 所误识别的蛋白质-RNA 相互作用^[32]。

总体而言, 洗涤缓冲液的选择需要根据所使用的免疫沉淀体系而定, 尽可能在可选的范围内选择更强的洗涤缓冲液。

1.4 RNA 片段化

为了获得靶蛋白所结合 RNA 的序列和位置信息, 一般需要将目标 RNA 进行片段化处理, 即切除不结合靶蛋白的 RNA 区段, 保留结合区段。一般使用 RNA 酶或者广谱性核酸酶将 RNA 片段化。核酸内切酶种类包括 MNase、RNase T1、RNase A 和 RNase I^[35] 等。MNase 是全能核酸酶, 能够同时切割 RNA 和 DNA, 没有切割序列特异性。RNase T1 特异性在鸟嘌呤 (guanine, G) 之后进行切割, 具有 RNA 切割偏好性。RNase A 在胞嘧啶 (cytosine, C) 和尿嘧啶 (uracil, U) 之后发生切割, 也具备 RNA 切割偏好性。而 RNase I 则没有切割偏好性, 可以切割 RNA 的任何位置^[36]。

RNase A 因为难以终止其消化过程, 在后续的实验操作流程中都陆续被替换成其他种类的 RNA 酶或广谱性核酸酶。CLIP-seq 使用 15 U (酶活性单位) 和 0.2 U 的 MNase 浓度对 RNA 进行切割测试, 最终选择 0.2 U 的 MNase 浓度进行实验, 切割的 RNA 大小范围在 50~100 nt 之间^[37]。值得注意的是, MNase 切割 RNA 是离子依赖性的, 很容易被 EDTA 等离子螯合剂所失活^[37]。PAR-CLIP 使用终浓度为 10⁸ U/L 的 RNase T1 进行 RNA 片段化, 反应条件是 22°C, 15 min^[6]。iCLIP 使用母液浓度为 10⁸ U/L 的 RNase I, 使用浓度为 1/50~1/250 (稀释度), 反应条件为 37°C, 5 min^[7]。eCLIP 对 RNase I 的使用终浓度进行了梯度测试和优化, 最终选定 4 000 U/L 的终浓度进行 RNA 片段化, 取得了良好的效果^[8]。sCLIP 先使用 RNase T1 (1 : 1 000), 10 min, 37°C 的条件对细胞裂解液中的 RNA 进行轻微的片段化, 然后在磁珠的重悬液中使用 1 : 1 000 RNase I 在 37°C 处理 5 min 对 RNA 进一步片段化^[9]。dCLIP 也使用母液浓度为 10⁸ U/L 的 RNase I 对 RNA 进行片段化^[11]。

每种不同 CLIP 实验使用的酶种类、酶浓度和酶反应时间都不一样, 因此建议在实际操作中进行浓度梯度测试。RNA 切割效果可以通过示踪性核

昔酸标记的方式进行。PAR-CLIP 使用 γ -32P-ATP 对 RNA 末端进行标记, 然后通过放射自显影观察 RNA 大小分布^[6]; sCLIP 用生物素化的腺苷二磷酸 (ADP) 对 RNA 末端进行标记, 然后通过 Western blot 的方式检测 RNA 的长度信号^[9]。

1.5 膜转印和切膜回收

在大多数的 CLIP 方法的操作步骤中, 在对靶蛋白结合的 RNA 片段化之后, 下一个步骤是将蛋白质进行洗脱并通过十二烷基硫酸钠聚丙烯酰胺凝胶电泳 (SDS-PAGE) 分离, 随后靶蛋白被转印到硝酸纤维素膜 (NC 膜) 上。随后, 靶蛋白及其上方的一段膜被切割下来, 并进行膜上的蛋白质消化和 RNA 的提取。这一步的目的—一方面是评估 RNA 的片段化程度, 另一方面是进一步去除背景 RNA。因为吸附在磁珠上的背景 RNA 无法跟随 RBP 转到 NC 膜上, 所以以此种膜转印的方式去除。这主要基于 NC 膜的特性——NC 膜对核酸的结合远低于常用于核酸研究的尼龙膜, 但是 NC 膜拥有对蛋白质的最大结合能力。此外, 由于黏附在磁珠上的接头序列较短, 因此会在靶蛋白条带的下方出现。

基于以上原理, 随之而来的关键问题是靶蛋白或者靶蛋白结合的 RNA 的可视化。在早期的 CLIP^[4] 以及 HITS-CLIP^[5] 中, 使用放射性的 ³²P 标记的核苷酸来标记靶蛋白上结合的 RNA 片段, 转膜后使用放射自显影技术来可视化 RNA 条带或区域。在红外交联和免疫共沉淀 (infrared-CLIP, irCLIP) 中^[38], 近红外染料结合的接头序列被连接到靶蛋白结合的 RNA 上, 取代了带有危险性的放射性核苷酸。但是, 近红外染料需要专门的成像设备, 这限制了其应用。sCLIP^[9] 则选取一部分 IP 样品用 Biotin-ADP 进行 RNA 标记, 这样可以使用常用的 ECL 发光体系进行示踪, 简化了操作难度。在 eCLIP^[8] 中, 可视化方法被进一步简化。eCLIP 采用一部分 IP 样品进行 Western blot 的方式, 并根据经验选择靶蛋白条带上方的一段距离进行切膜, 这进一步简化了操作。

然而, 由于基于抗体的 IP 无法彻底去除相互作用蛋白, 另一个隐忧是, 如果靶蛋白的结合蛋白也是一个 RBP, 就会造成额外的假阳性。大部分 CLIP 方法认为这些相互作用蛋白结合的 RNA 可以通过 SDS-PAGE 加转膜然后选择性回收特定分子质量范围内 RNA 的方法来去除。但是, 如果一个相互作用蛋白拥有和靶蛋白相近甚至相同的分子质量, 这些污染 RNA 依然无法去除。因此, 采用含

有一定蛋白质变性剂的缓冲液在 IP 层面上进行清洗, 才是彻底去除背景 RNA 和非特异性 RNA 的最佳方法。在这种方法有效的前提下, 冗长的 SDS-PAGE 和转膜并选择性回收 RNA 的步骤将不再必需, 大大简化了实验操作需要的步骤和时间。

目前, 已经有部分 CLIP 方法省略了繁琐的膜转印与切膜回收步骤。这些方法主要是采用了高强度结合标签或共价结合标签的方法。因为膜转印的前提是蛋白质能够从磁珠上被成功洗脱下来。如果蛋白质无法从磁珠上洗脱下来, 则无法进行 SDS-PAGE。Gold-CLIP 在 RNA 片段化并添加 3'端接头后, 直接进行蛋白酶 K 的酶切, 随后回收 RNA 并进行后续建库^[31]。TLC-CLIP 同样省略了 PAGE 和转膜步骤, 并发现此步骤的省略并没有显著影响数据的质量, 反而极大地节省了时间^[39]。Fbio-CLIP 采用 Biotin-Flag 串联纯化模式, 也省略了 PAGE 和转膜步骤^[40]。SpyCLIP 由于采用了共价结合的 SpyTag, 得以承受变性缓冲液的清洗, 同样省略了 PAGE 和转膜步骤^[10]。

2 RNA建库

2.1 接头连接

建库是二代测序的专用术语, 指的是对核酸进行片段化处理、末端修复以及接头的连接, 从而为上机测序做好准备。RNA 建库包括 RNA 的反转录和测序接头的连接。CLIP 实验富集得到的 RNA 长度通常在 20~100 nt 之间, 属于小片段 RNA, 其建库方式可分为两大类: 线性建库和环化建库。最早的 CLIP 采用的是线性化连接接头的方式。总体流程是, RNA 在磁珠上进行片段化之后, 通过 SDS-PAGE 和转膜回收, 然后在 RNA 的两侧通过 T4-RNA 连接酶连接上 RNA 接头^[4]。随后在改进版本中, RNA 片段在磁珠上进行 3'接头的连接, 5'接头在 SDS-PAGE 和转膜回收 RNA 之后进行, 这样规避了接头二聚体和 T4 RNA 连接酶中携带的细菌 RNA 的残留^[17]。早期的接头用于 PCR 扩增以及基因芯片的杂交。后来随着二代测序的发展, 适用于二代测序仪的接头被应用到上述流程中, 由此发展出第一代高通量的 CLIP-seq 方法学——HITS-CLIP^[5] 和 CLIP-seq^[37]。

不过, HITS-CLIP 的建库方式存在严重效率上的不足。这种不足首先来源于接头之间的互相连接以及 RNA 的首尾自连 (图 2)。PAR-CLIP 通过采用类似 microRNA 建库的方式解决了这个问题^[41]。由

于商业化的 microRNA 建库试剂盒中采用了含有特殊突变的 T4 RNA 连接酶以及特殊处理的接头序列, 避免了接头之间的自连和目的 RNA 的首尾相连^[42]。PAR-CLIP 采用的接头连接方式极大地降低了副反应的发生^[42]。此外, PAR-CLIP 没有采用在磁珠上进行 3'端接头连接的方式, 而是在提取好最终的 RNA 片段后, 3'端接头和 5'端接头在溶液中依次连接的方式^[6]。CRAC 技术也采用类似的线性化建库方法^[29]。

与 PAR-CLIP 同年发表的 iCLIP 则采用了环状建库的方式^[7]。与 HITS-CLIP 和 PAR-CLIP 不同的是, iCLIP 的 3'端接头为串联组合型接头 (图2), 其中一段为用于 3'端 PCR 的接头序列, 一段为用于 5'端 PCR 的接头序列, 这两段序列由一段限制性内切酶 *Bam*HI 识别序列连接^[7], 串联型接头最后连接上一段随机标签 (barcode) 序列以标记最初的 cDNA。通过上述串联型接头进行反转录后, 使用

CircLigase 环化酶将上述 3'端接头为串联组合型接头的 3'端与 cDNA 的 5'端连接起来, 实现环化。随后, 加入一段与 *Bam*HI 识别序列互补的单链 DNA, 随后使用 *Bam*HI 限制性内切酶进行线性化。这样, 线性化后的 cDNA 两端都实现了接头连接^[7]。这种建库方式一方面实现了 cDNA 水平上的两侧接头连接, 另一方面可以通过接头连接的序列位置识别在反转录过程因交联位点残留的寡聚氨基酸的空间位阻效应造成的反转录中断的位置, 这一位置被认为是 RNA 直接结合蛋白质的具体位置, 进而实现精准的特定 RBP 在 RNA 上的结合序列位置定位和基序确认^[7]。iCLAP^[28]、Gold-CLIP^[31]、RNA 杂交和单核苷酸分辨率紫外交联免疫沉淀 (RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation, HiCLIP)^[43] 以及 irCLIP^[38] 也采用了类似的环化连接方法。

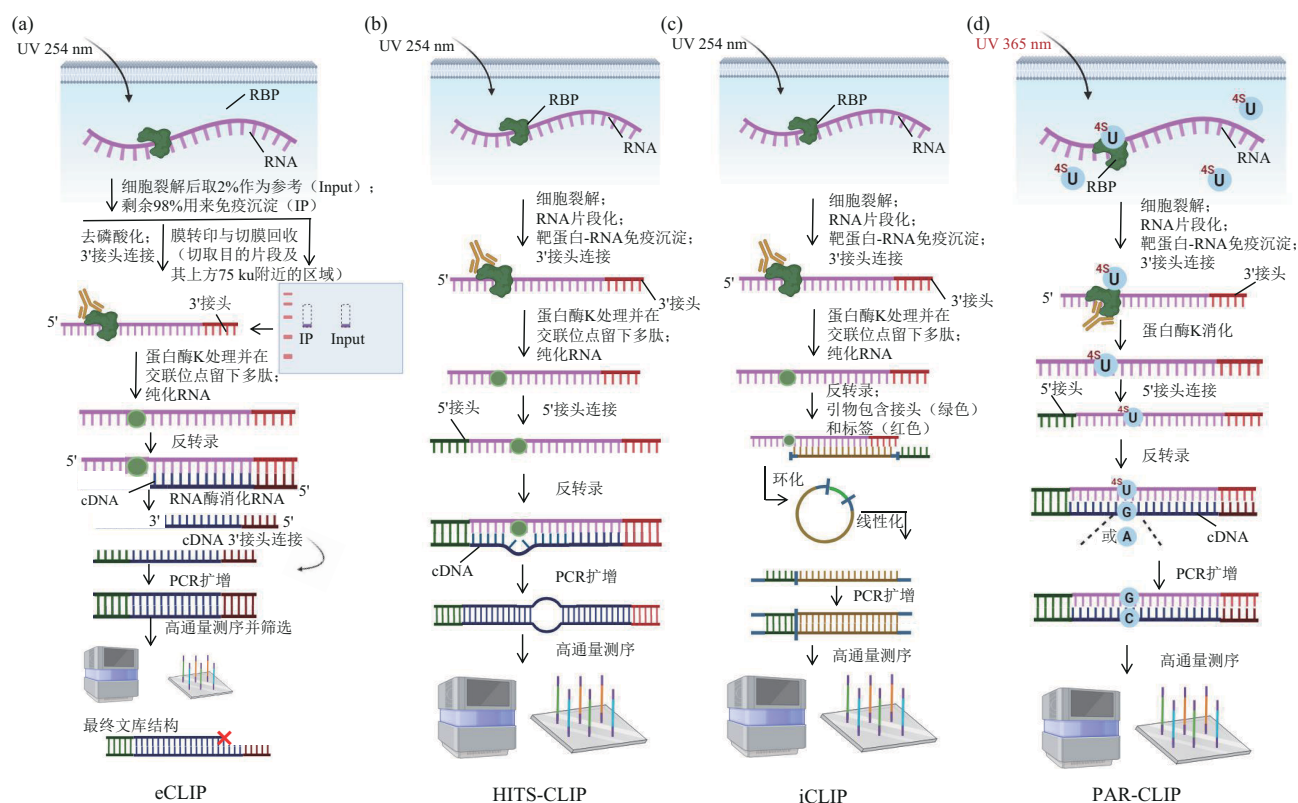


Fig. 2 Experimental workflows of eCLIP, HITS-CLIP, PAR-CLIP and iCLIP (<https://app.biorender.com>)

图2 eCLIP、HITS-CLIP、PAR-CLIP和iCLIP的实验流程图 (<https://app.biorender.com>)

(a) eCLIP将单个接头连接在交联RNA片段的3'端, 然后在反转录后将第二个接头连接到cDNA的3'端。PCR扩增截断的和可通读的片段, 随后高通量测序以及分析数据。(b) HITS-CLIP则是在交联RNA片段的3'端和5'端加接头后反转录。(c) iCLIP同eCLIP一样在反转录后将第二个接头连接到cDNA的3'端, 并加入一个标签 (barcode), 先成环后线性化, 利用在交联位点引起的cDNA终止来识别RBP-RNA相互作用的位置。(d) PAR-CLIP使用4-sU (4-thiouridine) 和365 nm的紫外线形成RNA与RNA结合蛋白 (RBP) 复合物, 最后利用胸苷 (T) 到胞苷 (C) 在cDNA中的转换来确定相互作用位置。

eCLIP 作为 ENCODE 计划的官方 CLIP 技术, 对线性化建库方法进行了优化。通过提升 T4 RNA 连接酶的浓度 (1 ul/10 ul 体积)、加入高浓度 PEG8000 (终浓度 18%) 和 DMSO (终浓度 3.2%) 的方式, 分别将 3'接头和 5'接头的连接效率提升至大于 90% 和大于 70% [8]。加入高浓度 PEG8000 能进一步提升 T4 RNA 连接酶的局部浓度, DMSO 能破坏 RNA 二级结构, 有利于 RNA 线性化进行连接。dCLIP 采用商业化的经过优化的小 RNA (small RNA) 建库试剂盒进行 DNA 建库 [11]。FBioCLIP-seq 为了避免 5'接头连接的低效性, 选择直接反转和扩增文库, 在后续扩增中加入 5'文库接头 [40]。CLASH 依赖于纯化与选定蛋白质结合的 RNA-RNA 双链, 通过连接这两条 RNA 双链形成嵌合 RNA, 然后对这些嵌合 RNA 进行逆转录, 它在实验方法上与 CRAP 类似, 但优化了 RNA-RNA 双链的回收 [12]。因此, 采用线性化的建库方式时, 需要优化 RNA 接头的连接效率, 以减少连接不足带来的 RNA 文库复杂度降低和引物二聚体扩增。

此外, 还有的 CLIP 技术对整个建库流程进行优化, 以提高文库完整度和特异性。例如, SpyCLIP 优化了整个线性化建库流程, 在 3'接头上加入 Biotin, 以便反转录能够在磁珠上发生, 简化了操作流程和减少电泳转膜纯化的损失 [10]。iCLIP、eCLIP 和 spyCLIP 等大部分 CLIP 方法均在接头中引入分子条形码 (unique molecular

identifier, UMI) 标签 [7-8, 10]。UMIs 是独一无二的身份码, 能够消除 PCR 扩增带来的过度复制 (over-duplication), 有利于去除 PCR-bias 和评价 PCR 质量, 实现 CLIP 在 RNA 水平的定量分析。UMIs 的引进是建库方式的一大提升。sCLIP 在 linker 中引起 T7 启动子, 直接对 RNA 文库进行扩增, 减少 cDNA 连接效率不足而引起的文库复杂度下降的问题 [9]。EasyCLIP 和 TLC-CLIP 在 3'-linker 中引入 polyA 的结构, 有利于后续 RNA 纯化, 简化了纯化操作, 减少了文库损失 [39, 44]。可见, RNA-linker 被赋予了越来越多的功能, 以不断简化建库流程, 提高文库复杂度。

在两种建库方法中, 线性化建库正成为主流。环化建库的方法操作繁琐、效率较低, 且容易造成较大的样品损失 [8]。在 2020 年发表的 iCLIP2 改进版中, 最终抛弃了环化建库的方法, 选择了类似 eCLIP 的 on-beads 线性化建库方法 [45]。

2.2 建库方式的演进

RNA 的建库是 CLIP 实验中的关键步骤之一。总体上来看, 每一种 CLIP 方法都有专属的建库方式。尽管建库方法是多样化的, 但是都包括“接头的连接”、“反转录”、“RNA 的可视化”、“背景 RNA 的去除”等步骤。建库方法的改进主要是在这些步骤的操作方式、顺序甚至去留上进行了多种优化改变。表 1 总结了目前不同的 CLIP 方法中采用的建库方式的核心特点和差异。

Table 1 Differences in library preparation for CLIP-seq methods
表1 CLIP-seq的建库方式差异

名称	建库方式	方法学特点
HITS-CLIP [46]	线性建库: RNase A消化, 去磷酸, on-beads加3' RNA接头 (adapter), 5'同位素标记分离目的RNA, 蛋白酶K消化, 纯化RNA加5' RNA接头, 反转录, 文库扩增PCR	第一个CLIP-seq方法学
CLIP-seq [37]	线性建库: Mnase消化, 去磷酸, on-beads加3' RNA接头, 5'同位素标记分离目的RNA, 蛋白酶K消化, 纯化RNA加5' RNA接头, 反转录, 文库扩增PCR	与HITS-seq相似, RNA部分酶切程度更加可控, Mnase能被金属离子螯合剂迅速失活
PAR-CLIP [6]	线性建库: RNase T1消化, 去磷酸, 5'同位素标记分离目的RNA, 蛋白酶K消化, 纯化RNA, 加3' RNA接头, 加5' RNA接头, 反转录, 文库扩增PCR	利用4-sU等核苷类似物, 发生T-C的转化, 精准定位到发生交联的核苷酸位点
iCLIP [7]	环化建库: RNase I消化, 去磷酸, on-beads加3' RNA接头以及5'同位素标记, SDS-PAGE分离目的RNA, 蛋白酶K部分消化, 带有双接头、酶切位点和barcode的引物进行反转录, 尿素-PAGE进行cDNA分离, cDNA环化, 酶切位点退火配对成双链, 限制性内切酶酶切, PCR扩增	RNA和DNA文库都经过大小筛选, 步骤较多
CLAP [28]	环化建库, 流程同iCLIP	使用Strep/His-tag双tag亲和纯化系统, 先Strep-tag纯化, 后8 mol/L尿素条件下His-tag纯化, 去除间接相互作用RNA和背景RNA能力强

续表1

名称	建库方式	方法学特点
irCLIP ^[38]	环化建库: RNase I和RNase A共同消化, 去磷酸, 3'连接带有Biotin和荧光染料的preA-DNA-adapter, SDS-PAGE分离RNA-蛋白质复合物, 切出目标RNA, 回收RNA, 消化被交联的蛋白质, 反转录, 环化cDNA, 第一轮PCR, 分选目的大小片段, 第二轮扩增PCR	SDS彻底裂解细胞, 无放射性RNA标记, 荧光标记RNA, 方便快速定量和质检富集的RNA。无内切酶切割, 环化cDNA通过PCR进行线性化
eCLIP ^[47]	线性建库: RNase I消化, 去磷酸, on-beads加3' RNA接头, SDS-PAGE分离目的RNA (切取目的片段及其上方75 ku附近的区域), 蛋白酶K彻底消化, RNA纯化, 反转录, cDNA 3'接头连接, PCR扩增和文库大小筛选	无RNA标记, 根据RNA大小计算预测分布, 提高了3' RNA接头的连接效率和cDNA 3' 接头连接效率, 流程比iCLIP简单, 文库损耗小
HiCLIP ^[43]	环化建库: RNase I消化, 去磷酸, on-beads加被末端封闭的A和B两种3' RNA接头, 去除B接头的末端封闭, 环化连接B接头和目的A RNA, SDS-PAGE分离RNA, 纯化RNA, 反转录, 酶切位点退火配对成双链, 限制性内切酶酶切, PCR扩增	鉴定结合到靶蛋白的RNA二级结构或结合到同一个蛋白质的两种RNA
seCLIP ^[48]	线性建库, 同eCLIP, 测试不同的反转录酶对于交联位点的跨越能力	不同的反转录酶对交联位点的跨越能力不同, 反应不一样的交联位点信息
FLASH ^[49]	环化建库: RNase I消化, T4PNK去磷酸5'磷酸, 3'连接带有两种barcode和随机序列的FLASH接头, 蛋白酶K处理, 所有的RNA样本混合纯化, 反转录, 去除RNA, 环化cDNA, PCR扩增文库	根据不同的tags区分不同来源的样本, 混合所有样本的方式减少了样品之间的误差, 简化了后续操作, 是最快捷的建库方法学
sCLIP ^[9]	线性建库: RNase T1消化, 生物素标记RNA, 蛋白酶K消化RNA, 纯化RNA, 多聚腺苷酸化RNA, 带有oligo (dT)、随机标签 (barcode)、接头和T7启动子的引物进行反转录, cDNA进行体外转录, 体外转录RNA 3'进行接头连接, cDNA合成和PCR文库扩增	在连接接头前对RNA进行扩增, 保证文库复杂度, 省略部分纯化步骤
dCLIP ^[11]	线性建库: RNase I消化, 去磷酸, 5'同位素标记分离目的RNA, 蛋白酶K消化, 纯化RNA加3'和5' RNA接头, 反转录, 文库扩增PCR	完全变性的条件去除背景, 建库流程比较简单
SpyCLIP ^[10]	线性建库: RNase I消化, 去磷酸, 连接生物素标记的3'接头, on-beads反转录, 得到cDNA, 带有UMIs的接头连接到cDNA 3'-end, PCR扩增	允许完全变性的清洗条件, on-beads操作, 省去凝胶分离纯化步骤; 建库引进UMIs, 有利于去除PCR产物中的重复序列
EasyCLIP ^[44]	线性建库: 去磷酸, 连接带有特定荧光标记、polyA的3'接头, 连接带有另一种荧光标记的5'接头, SDS-PAGE分离RNA, oligo (dT) 纯化RNA, 反转录, PCR	不同荧光标记的RNA可判断两步接头连接的效率, 操作步骤较为简化, 文库复杂度保存较好

3 对照实验设计

对照实验是CLIP实验中用于有效校验并滤除假阳性结果的重要策略。有恰当而优良的对照才能排除背景信号, 减少实验误差。对比不同的CLIP-seq的对照方式, 可以归纳为以下几种类型:

- a.使用靶蛋白质敲除的细胞或组织作为阴性对照组^[50], 使用无免疫源性的IgG^[51-53]或者细胞内不存在的蛋白质的靶标抗体 (如Flag等标签抗体) 作为阴性对照组^[5, 37-38, 54-55]。
- b.在使用IgG作为对照但无法得到足够的RNA文库时, 可以选择一些随机序列生成一个cluster (一组具有相似特征的生物序列或者数据点的集合) 和motifs (在DNA、RNA或蛋白质序列中反复出现的特定短序列或结构模式), 与实验组的motif进行比较^[44, 46]。

c.在相同实验操作下, 选择某种已知RNA结合蛋白的CLIP结果作为对照, 以说明抗体富集的RNA的特异性^[56]。不过这种方法使用频率并不高。

d.为了证明UV交联蛋白质-RNA的特异性, 使用没有UV照射的条件作为对照^[57]。

e.借鉴ChIP-seq的方式, 使用未富集的RNA, 即Input作为对照^[10, 58-62]。

对于外源过表达靶蛋白质, 因为选用了载体进行过表达, 并且携带标签序列, 使用的对照方式比较局限, 一般是以表达标签序列自身或者标签序列融合绿色荧光蛋白 (green fluorescent protein, GFP) 的方式作为对照^[31]。此外, 也可以用不紫外交联的同一种细胞作为对照组^[6]。表达不带有标签的靶蛋白作为对照组也是一种对照方法^[29, 63]。

综上, 对照组可以根据实际的实验设计进行选

择。选择的原则是控制变量, 最大化减少对照组与实验组的差异, 最理想情况下只有一个变量。如此才能避免实验产生假阳性。另外, 阳性对照也非常重要, 在初次进行 CLIP-seq 时, 一个很好的阳性对照是评价实验是否成功的重要标准。阳性对照通常选择已知的、高丰度的并且容易富集的靶蛋白质^[47, 64-65]。

4 CLIP测序数据的峰值分析

与转录组测序和全基因组测序不同的是, CLIP 测序产生的序列读段 (reads) 往往聚集分布在 RNA 的某个或某些位置, 这是因为 RBP 往往具有特定的 RNA 结合位点或序列模式。这一点与 ChIP-seq 较为类似。在 CLIP 测序数据的前期比对中, 一般使用常用的 FastQC^[66] 工具进行质控, 使用 cutadapt^[67] 进行去接头处理, 使用 gscripts^[8] 进行去重复以及标准化, 使用 STAR^[8] 或者 Bowtie2^[68] 进行序列比对, 使用 MACS2^[69] 等工具进行峰值识别 (peak calling), 使用 Homer^[70] 进行 motif 分析, 使用 CLIP 工具包 (CLIP tool kit, CTK)^[71] 等进行 RNA 结合位点分析。其中, 在峰值识别和 RNA 结合分析方面, CLIP 测序数据需要使用专门为此类数据开发的软件进行分析。尽管使用基于模型的 ChIP-seq 分析 2 (model-based analysis of ChIP-seq 2, MACS₂) 等常用于 ChIP-seq 的工具也可以进行峰值识别, 但是与专门性软件相比, 其准确性和特异性依然具有较大差距。

目前, 已经开发出多种软件用于 CLIP 测序数据的峰值识别。一般而言, 峰值识别分为两个步骤, 第一个是确定富含序列读段的区域, 第二个是计算所选区域的统计显著性。Piranha^[72] 将基因组序列划分为多个小区段, 使用零膨胀负二项分布来计算每个区段的显著性, 以区分背景信号, 从而提高峰值识别的准确性。然而, 这种方法可能会漏掉一些表达量较低基因的 RNA。ASPeak^[73] 使用 RIP-input 或者 RNA-seq 的数据来首先评估 RNA 表达的分布, 对于不表达的 DNA 区域, 则不再计算显著性。CLIPper^[74] 使用泊松分布来计算显著性, 并考虑到峰值 (peak) 的高度宽度等因素。Pyicoclip^[75] 和 iCount^[7] 则通过在用户定义的区域内存机分配序列读段来计算显著性。

此外, 一些特殊的 CLIP 方法拥有独特的峰值识别方法。PAR-CLIP 技术采用了基于 4-sU 的交联策略。在细胞培养基中加入 4-sU, 4-sU 会掺入到

新合成的 RNA 中。在 365 nm 的紫外线的照射下, 4-sU 可以与结合的蛋白质高效交联。4-sU 还有一个特点: 在反转录过程中, 与蛋白质交联的 4-sU 会被反转录成 G 而不是 A, 从而在 RNA 链的测序数据中出现 T-C 转化的突变。含有这种突变的序列读段会被认为是真实的峰值。这种独特的模式使得峰值的寻找不需要复杂的峰值识别算法。PARalyzer^[7] 便是专门用于分析 PAR-CLIP 测序数据的工具。后续发现, 即便是不发生交联, 部分 4-sU 依然有概率在反转录时出现 T-C 转换。为了消除这部分假阳性, BMix^[76] 和 wavCluster^[77] 两个软件被开发出来用于过滤非实验因素造成的 T-C 转换。

此外, 还有不少工具专门被开发出来以实现 CLIP 测序数据的峰值识别。MiClip^[78] 在峰值识别过程中采用了隐马尔可夫模型。pyCRAC^[79] 针对 HITS-CLIP、PAR-CLIP 和 CRAC 优化了全套分析代码。PureCLIP^[80] 可以用来分析基于反转录截短来定位交联位点的 CLIP 方法的数据, 如 eCLIP 等。HITS-CLIP 和 PAR-CLIP 方法会在测序得到的序列读段中引入突变, 甚至导致某些碱基的删除。而 iCLIP 和 eCLIP 等技术则会在交联位点处形成截短的 cDNA。PIPE-CLIP^[81] 和 CTK 两种工具, 同时在软件中提供了基于上述两种模式的峰值识别和交联位点筛选的功能。

对于同一个数据, 不同的软件分析后得出的结果可能有所不同。因此, 在条件允许的情况下, 建议综合使用多种工具进行分析。

5 RNA序列基序分析

分析靶蛋白结合在 RNA 上的具体位置是 CLIP 实验的重要目的之一。根据对结合位置的分析, 可以得到靶蛋白的 RNA 序列 motif, 这是关于靶蛋白结合 RNA 的偏好性的重要信息。

在早期的 CLIP 以及 HITS-CLIP 中, 由于无法鉴定蛋白质结合在 RNA 上的具体核苷酸位置, 因此尽管可以基于 RNA 片段进行 motif 分析, 但精确度较低。数据分析发现, 反转录酶在通过交联位点时, 容易发生突变。因此, 后期发展出来的基于 HITS-CLIP 的算法可以根据突变频率进行更精确的 motif 分析^[82]。

PAR-CLIP 由于在反转录层面实现了 T-C 转化, 因此可以方便地进行 motif 分析。在 iCLIP 和 eCLIP 以及类似的方法中, 由于蛋白酶 K 对靶蛋白进行消

化后, 共价结合在RNA的最后一个或几个氨基酸无法被消化去除, 因此反转录酶遇到未消化的氨基酸残基会大概率停顿脱落而生成截短的cDNA, 截短位点反应了交联位点的信息^[7-8, 83]。这样便可以进行较为精准的motif分析。

MEME^[84]是常用的进行CLIP数据motif分析的软件。需要注意的是, motif所处的序列可能并不是正好处在以交联位点为中心的短序列内, 也有可能在交联位点的两侧或附近。因此, 为了更准确地寻找motif, 需要以交联位点为中心向两侧进行扩展。这可以通过BEDTools^[85]的getfasta功能来实现。此外, 有些motif的筛选还需要考虑RNA的二级结构。MEMERIS^[86]、RNAcontext^[87]、Zagros^[88]、CapR^[89]、Graphprot^[90]、ssHMM^[91]等软件均考虑了RNA的二级结构。

6 总结与展望

CLIP技术的本质是鉴定目的蛋白质所结合的RNA以及蛋白质在RNA上的结合位置。为了实现这一点, 需要减少操作过程中RNA的损失以及增加被成功建库的RNA的种类和数量, 与此同时, 还要尽可能减少背景和非特异性结合的RNA。所有的CLIP方法中的改进和优化都围绕着这两个核心目标展开。

在DNA层次上, 与CLIP技术类似的是ChIP技术。ChIP-seq技术的本质是对靶蛋白结合的DNA序列进行富集和测序。有意思的是, 在基因组和转录调控领域的研究中常用的ChIP技术(包括ChIP-qPCR和ChIP-seq), 并没有演进出太多的方法学改进型或变种, 这与形形色色、种类繁多的CLIP技术方法学形成了鲜明对比。造成这种差异核心原因在于, ChIP实验采用的甲醛交联可以在高温下被逆转, 也就是所谓的“逆转交联”; 但是UV造成的交联无法被逆转, 在RNA需要被逆转录才能被测序的前提下, 交联位点对逆转录的阻碍造成了RNA建库的复杂性。因此, 对建库方法的改进一直是CLIP技术发展中最重要热点。除了“无法逆转交联”这一特点外, 由于CLIP操作中产生的RNA片段较短(20~100 bp), 因此不太适合RNA-seq中采用随机引物逆转录方法。大部分CLIP方法学采用由实验者来合成接头序列以及在RNA和cDNA的层次上分别进行两端接头的连接, 这些操作无法导致CLIP的RNA建库利用成熟的商业化建库试剂盒, 繁琐的建库步骤也是CLIP区别

于ChIP-seq的另一大特点。

尽管CLIP技术操作较为繁琐, 但经过多年的发展已经逐渐趋于成熟。多样化的CLIP技术方法学可以给不同实验条件提供最佳的方法学选择。与DNA结合蛋白不同, 在RBP领域, 存在大量的不具有经典RNA结合结构域的RBP^[92]。Baltz等^[93]通过oligo(dT)纯化所有与mRNA相互作用的蛋白质, 并且通过严格的洗涤方式破坏非直接的蛋白质-RNA相互作用, 最后鉴定出约800个结合mRNA的蛋白质。让人惊讶的是, 其中有约1/3的蛋白质从未被标注为RBP, 约15%的蛋白质在预测算法中从未被判定具备潜在的RNA结合能力。其中包括一些转录因子、蛋白激酶和一些没有被注释功能的蛋白质。它们没有具备一定特征的结构域, 因此以非常复杂多样化的方式结合RNA。Castello等^[2]用同样的oligo(dT)纯化方式, 得到与mRNA直接结合的蛋白质, 并通过RNA水解和胰酶酶切, 得到含有RNA残留的肽段, 最后通过质谱的方式鉴定出上千个直接结合RNA的蛋白质区域。同样的, 这些结合RNA的蛋白质也有一半没有被注释为RBP。以上这些研究和其他研究进一步证实了诸多没有被发现的RBP的存在^[13]。为了鉴定细胞内所有的RBP, Trendel等^[3]建立了纯化全细胞中所有直接结合RNA的蛋白质的方法。因为鉴定分辨率限制和蛋白质丰度偏差, 他们最后在MCF7中鉴定出了1207个高可信度RBP, 在HEK293中鉴定出了1357个高可信度蛋白质, 并且其中也有一部分是未被注释的有RNA结合能力的蛋白质。有趣的是, 他们发现有一群染色质重塑因子被鉴定为RBP, 其中TP53已经被证实可以结合RNA。进一步说明有许多潜在的RBP未被发现和研究。

由于不存在经典的RNA结合结构域, 对于新发现的这些RBP, 其结合RNA能力的真实性、强弱以及特异性存在争议。CLIP技术的多样化发展为严格验证新型RBP的真实性提供了较好的选择。PRC2复合体是一个典型的例子。PRC2是一个催化组蛋白H3K27me3修饰的多亚基蛋白复合体, 曾被多篇研究认定为可以结合RNA^[94]。但是由于前期研究是基于体外实验来证明其RNA结合特性的, 因此PRC2结合RNA的能力一直被质疑^[95]。RIP-seq技术曾被用来鉴定PRC2的结合蛋白^[96]。RIP-seq的本质是通过传统的蛋白质免疫共沉淀(co-immunoprecipitation, Co-IP)的方式拉下靶蛋

白, 并对拉下来的RNA进行高通量测序^[97]。由于Co-IP的条件温和, 导致拉下的靶蛋白结合有大量的相互作用蛋白, 这些相互作用蛋白也有可能结合RNA。因此, RIP-seq的结果存在严重的假阳性风险, 与可以承受较强蛋白质变性条件的CLIP相比, RIP-seq的结果不足以采信。dCLIP由于使用了Biotin-Streptavidin结合系统, 可以承受8 mol/L尿素、2% SDS等严苛的清洗条件, 因此可以最大程度上避免假阳性的产生^[11]。Rosenberg等^[25]使用dCLIP技术测序了PRC2亚基EZH2结合的RNA, 发现EZH2对含有G4结构的RNA具有较强的结合倾向性, 这一发现也被后续的结构生物学研究证实^[98]。然而, Guo等^[32]使用CLAP技术同样对EZH2可能结合的RNA进行了测序, 却并没有发现EZH2可以结合含有G4结构的RNA, 比如经典的XIST-RNA。CLAP同样可以承受8 mol/L尿素、2% SDS以及1 mol/L NaCl的清洗。这反映出, 针对同一个蛋白质, 不同的CLIP方法由于其操作细节的差异, 可能得出完全不同的结果。因此, 对于RBP尤其是非经典RBP的研究, 选择多种CLIP方法进行实验, 有助于得到更统一、更真实的结果。

在传统的CLIP技术中, 由于操作步骤本身带来的背景信号问题一直是最严重的问题。Guo等^[32]使用常用的eCLIP技术对外源表达的EZH2和GFP蛋白分别进行了CLIP测序实验。发现GFP蛋白结合的RNA的峰值呈现出特异性分布。但是, GFP作为外源蛋白, 没有任何报道证明其可以结合RNA。此外, 作者使用标签抗体在不表达标签融合PRC2蛋白亚基的样品中进行了eCLIP, 得到的峰值分布与在表达有标签蛋白的样本中得到的结果一致。这意味着, 常用的CLIP技术中, 对照样品组(比如IgG对照组)存在严重的背景信号问题, 这种背景信号似乎不能被常用的eCLIP洗涤液以及电泳转膜切膜等步骤所去除。这种现象在针对PRC2复合体^[94]、CTCF^[99]、YY1^[100]等非经典RBP的研究中尤其突出。

除了上述背景信号问题, CLIP实验的另一大问题是抗体的特异性。抗体特异性不佳会导致严重的假阳性结果。这一问题的改进依赖于开发高度特异性的抗体。然而, 开发高度特异性的抗体成本高、周期长, 不具有广泛的可行性。此外, 抗体与靶蛋白的结合还受制于靶蛋白的相互作用蛋白构成的空间位阻。某种特异性抗体可能在体外与纯化的单体蛋白质正常结合, 但在活细胞内, 靶蛋白通常

会与多种相互作用蛋白及RNA结合在一起。互作分子形成的空间位阻可能阻碍特异性抗体与抗原位点的结合。这是基于抗体的生物学技术开发中常见的一种局限性。这种局限性不仅存在于CLIP技术中, 也广泛存在于所有基于非变性条件下的抗体依赖型生物学实验中。因此, 在细胞系以及模式动物中采用CRISPR-Cas9等技术敲入(knock-in)可以承受蛋白质变性剂的标签(如spyTag、HaloTag等)是一个更具有广泛可行性的选择。

在CLIP技术未来的改进中, 我们认为有以下几大方向值得期待: 第一, 为更加彻底地去除背景信号, 应该引入包括SDS、盐酸胍、尿素等蛋白质变性剂在内的洗涤液。这意味着要么开发具有高结合力的抗体, 要么使用能够承受高强度变性剂的亲和标签。第二, 考虑到RNA的副本数量(copy number)多样性和动态性, 不能简单地使用类似ChIP-seq的标准化(normalization)方法。因此, CLIP测序结果中定量的准确性问题一直无法很好地解决。我们建议在建库的过程中, 使用等体积RNA样品建库并使用带有UMI标签的接头, 或者在初始的等体积样品中, 掺入一定比例的外源RNA(spike-in)。UMI或spike-in的引入将显著提高CLIP测序中定量结果处理的便利性、准确性。第三, 减少CLIP的操作步骤。我们期待未来的改进版本能够省略电泳和转膜步骤, 并实现在一个Ep管中完成反转录和基于cDNA的接头连接, 从而进一步增加文库的复杂性, 减少目标RNA的遗漏比例。

蛋白质-RNA相互作用是生命活动的基本形式之一, 越来越多的研究提示, 牢固持久或瞬时短暂的蛋白质-RNA相互作用几乎无处不在, 影响着蛋白质和RNA的诸多功能。CLIP技术经过多年的发展, 不断得到完善, 将在RNA结合蛋白研究领域发挥更重要的作用。

参 考 文 献

- [1] Müller-McNicol M, Neugebauer K M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet*, 2013, **14**(4): 275-287
- [2] Castello A, Fischer B, Frese C K, *et al.* Comprehensive identification of RNA-binding domains in human cells. *Mol Cell*, 2016, **63**(4): 696-710
- [3] Trendel J, Schwarzl T, Horos R, *et al.* The human RNA-binding proteome and its dynamics during translational arrest. *Cell*, 2019, **176**(1/2): 391-403.e19

- [4] Ule J, Jensen K B, Ruggiu M, *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 2003, **302**(5648): 1212-1215
- [5] Licatalosi D D, Mele A, Fak J J, *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 2008, **456**(7221): 464-469
- [6] 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**(1): 129-141
- [7] 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**(7): 909-915
- [8] 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**(6): 508-514
- [9] Kargapolova Y, Levin M, Lackner K, *et al.* sCLIP-an integrated platform to study RNA-protein interactomes in biomedical research: identification of CSTF2tau in alternative processing of small nuclear RNAs. *Nucleic Acids Res*, 2017, **45**(10): 6074-6086
- [10] 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**(6): e33
- [11] Rosenberg M, Blum R, Kesner B, *et al.* Denaturing clip, dCLIP, pipeline identifies discrete RNA footprints on Chromatin-associated proteins and reveals that CBX7 targets 3' UTRs to regulate mRNA expression. *Cell Syst*, 2017, **5**(4): 368-385.e15
- [12] Kudla G, Granneman S, Hahn D, *et al.* Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc Natl Acad Sci USA*, 2011, **108**(24): 10010-10015
- [13] Kramer K, Sachsenberg T, Beckmann B M, *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods*, 2014, **11**(10): 1064-1070
- [14] Brimacombe R, Stiege W, Kyriatsoulis A, *et al.* Intra-RNA and RNA-protein cross-linking techniques in *Escherichia coli* ribosomes. *Methods Enzymol*, 1988, **164**: 287-309
- [15] Hockensmith J W, Kubasek W L, Vorachek W R, *et al.* Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *J Biol Chem*, 1986, **261**(8): 3512-3518
- [16] Shetlar M D, Christensen J, Hom K. Photochemical addition of amino acids and peptides to DNA. *Photochem Photobiol*, 1984, **39**(2): 125-133
- [17] Ule J, Jensen K, Mele A, *et al.* CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*, 2005, **37**(4): 376-386
- [18] Williams G D, Townsend D, Wylie K M, *et al.* Nucleotide resolution mapping of influenza A virus nucleoprotein-RNA interactions reveals RNA features required for replication. *Nat Commun*, 2018, **9**(1): 465
- [19] Vandemoortele G, Eyckerman S, Gevaert K. Pick a tag and explore the functions of your pet protein. *Trends Biotechnol*, 2019, **37**(10): 1078-1090
- [20] Wilchek M, Bayer E A. Introduction to avidin-biotin technology. *Methods Enzymol*, 1990, **184**: 5-13
- [21] Fairhead M, Howarth M. Site-specific biotinylation of purified proteins using BirA. *Methods Mol Biol*, 2015, **1266**: 171-184
- [22] Ioannou M, Papageorgiou D N, Ogryzko V, *et al.* Mammalian expression vectors for metabolic biotinylation tandem affinity tagging by co-expression in *cis* of a mammalian codon-optimized BirA biotin ligase. *BMC Res Notes*, 2018, **11**(1): 390
- [23] Wilchek M, Bayer E A. The avidin-biotin complex in bioanalytical applications. *Anal Biochem*, 1988, **171**(1): 1-32
- [24] Dundas C M, Demonte D, Park S. Streptavidin-biotin technology: improvements and innovations in chemical and biological applications. *Appl Microbiol Biotechnol*, 2013, **97**(21): 9343-9353
- [25] Rosenberg M, Blum R, Kesner B, *et al.* Motif-driven interactions between RNA and PRC2 are rheostats that regulate transcription elongation. *Nat Struct Mol Biol*, 2021, **28**(1): 103-117
- [26] Schmidt T G M, Skerra A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat Protoc*, 2007, **2**(6): 1528-1535
- [27] Mishra V. Affinity tags for protein purification. *Curr Protein Pept Sci*, 2020, **21**(8): 821-830
- [28] Wang Z, Kayikci M, Briesse M, *et al.* iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol*, 2010, **8**(10): e1000530
- [29] Granneman S, Kudla G, Petfalski E, *et al.* Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci USA*, 2009, **106**(24): 9613-9618
- [30] Los G V, Encell L P, McDougall M G, *et al.* HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS Chem Biol*, 2008, **3**(6): 373-382
- [31] Gu J, Wang M, Yang Y, *et al.* GoldCLIP: gel-omitted ligation-dependent CLIP. *Genomics Proteomics Bioinformatics*, 2018, **16**(2): 136-143
- [32] Guo J K, Blanco M R, Walkup W G 4th, *et al.* Denaturing purifications demonstrate that PRC2 and other widely reported chromatin proteins do not appear to bind directly to RNA *in vivo*. *Mol Cell*, 2024, **84**(7): 1271-1289.e12
- [33] Reddington S C, Howarth M. Secrets of a covalent interaction for biomaterials and biotechnology: SpyTag and SpyCatcher. *Curr Opin Chem Biol*, 2015, **29**: 94-99
- [34] Reyes-Herrera P H, Speck-Hernandez C A, Sierra C A, *et al.* BackCLIP: a tool to identify common background presence in PAR-CLIP datasets. *Bioinformatics*, 2015, **31**(22): 3703-3705
- [35] Kishore S, Jaskiewicz L, Burger L, *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, 2011, **8**(7): 559-564
- [36] Wang Z, Tollervey J, Briesse M, *et al.* CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked

- to proteins *in vivo*. *Methods*, 2009, **48**(3): 287-293
- [37] Yeo G W, Coufal N G, Liang T Y, *et al*. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 2009, **16**(2): 130-137
- [38] Zarnegar B J, Flynn R A, Shen Y, *et al*. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Methods*, 2016, **13**: 489-492
- [39] Ernst C, Duc J, Trono D. Efficient and sensitive profiling of RNA-protein interactions using TLC-CLIP. *Nucleic Acids Res*, 2023, **51**(13): e70
- [40] Bi X, Shen X. Transcriptome-wide mapping of protein-RNA interactions. *Methods Mol Biol*, 2020, **2161**: 161-173
- [41] Hafner M, Landgraf P, Ludwig J, *et al*. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 2008, **44**(1): 3-12
- [42] Raabe C A, Tang T H, Brosius J, *et al*. Biases in small RNA deep sequencing data. *Nucleic Acids Res*, 2014, **42**(3): 1414-1426
- [43] Sugimoto Y, Vigilante A, Darbo E, *et al*. hiCLIP reveals the *in vivo* atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 2015, **519**(7544): 491-494
- [44] Porter D F, Miao W, Yang X, *et al*. easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat Commun*, 2021, **12**(1): 1569
- [45] Buchbender A, Mutter H, Reymond Sutandy F X R, *et al*. Improved library preparation with the new iCLIP2 protocol. *Methods*, 2020, **178**: 33-48
- [46] Chi S W, Zang J B, Mele A, *et al*. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 2009, **460**(7254): 479-486
- [47] Conway A E, Van Nostrand E L, Pratt G A, *et al*. Enhanced CLIP uncovers IMP protein-RNA targets in human pluripotent stem cells important for cell adhesion and survival. *Cell Rep*, 2016, **15**(3): 666-679
- [48] van Nostrand E L, Shishkin A A, Pratt G A, *et al*. Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions. *Methods*, 2017, **126**: 29-37
- [49] Aktaş T, Avcı İ, Maticzka D, *et al*. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature*, 2017, **544**(7648): 115-119
- [50] Saito Y, Yuan Y, Zucker-Scharff I, *et al*. Differential NOVA2-mediated splicing in excitatory and inhibitory neurons regulates cortical development and cerebellar function. *Neuron*, 2019, **101**(4): 707-720.e5
- [51] 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**(7818): 711-719
- [52] Luna J M, Barajas J M, Teng K Y, *et al*. Argonaute CLIP defines a deregulated miR-122-bound transcriptome that correlates with patient survival in human liver cancer. *Mol Cell*, 2017, **67**(3): 400-410.e7
- [53] Nowakowski T J, Rani N, Golkaram M, *et al*. Regulation of cell-type-specific transcriptomes by microRNA networks during human brain development. *Nat Neurosci*, 2018, **21**(12): 1784-1792
- [54] Moy R H, Cole B S, Yasunaga A, *et al*. Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense. *Cell*, 2014, **158**(4): 764-777
- [55] Lebedeva S, Jens M, Theil K, *et al*. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell*, 2011, **43**(3): 340-352
- [56] Darnell J C, Van Driesche S J, Zhang C, *et al*. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 2011, **146**(2): 247-261
- [57] Lee A S Y, Kranzusch P J, Cate J H D. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature*, 2015, **522**(7554): 111-114
- [58] Liu N, Dai Q, Zheng G, *et al*. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*, 2015, **518**(7540): 560-564
- [59] Körtel N, Rücklé C, Zhou Y, *et al*. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6ABoost machine learning. *Nucleic Acids Res*, 2021, **49**(16): e92
- [60] Sundararaman B, Zhan L, Blue S M, *et al*. Resources for the comprehensive discovery of functional RNA elements. *Mol Cell*, 2016, **61**(6): 903-913
- [61] Nussbacher J K, Yeo G W. Systematic discovery of RNA binding proteins that regulate microRNA levels. *Mol Cell*, 2018, **69**(6): 1005-1016.e7
- [62] Chu Y, Kilikevicius A, Liu J, *et al*. Argonaute binding within 3'-untranslated regions poorly predicts gene repression. *Nucleic Acids Res*, 2020, **48**(13): 7439-7453
- [63] Potts A H, Vakulskas C A, Pannuri A, *et al*. Global role of the bacterial post-transcriptional regulator CsrA revealed by integrated transcriptomics. *Nat Commun*, 2017, **8**(1): 1596
- [64] Kim Y, Park J, Kim S, *et al*. PKR senses nuclear and mitochondrial signals by interacting with endogenous double-stranded RNAs. *Mol Cell*, 2018, **71**(6): 1051-1063.e6
- [65] 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**(1): 90
- [66] Brown J, Pirrung M, McCue L A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, 2017, **33**(19): 3137-3139
- [67] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*, 2011, **17**(1): 10
- [68] Langmead B, Salzberg S L. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 2012, **9**(4): 357-359
- [69] Gaspar J M. Improved peak-calling with MACS₂. *BioRxiv*, 2018. DOI:10.1101/496521
- [70] 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**(4): 576-589
- [71] Shah A, Qian Y, Weyn-Vanhentenryck S M, *et al*. CLIP tool kit

- (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, 2017, **33**(4): 566-567
- [72] Uren P J, Bahrami-Samani E, Burns S C, *et al.* Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 2012, **28**(23): 3013-3020
- [73] Kucukural A, Özadam H, Singh G, *et al.* ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics*, 2013, **29**(19): 2485-2486
- [74] Lovci M T, Ghanem D, Marr H, *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol*, 2013, **20**(12): 1434-1442
- [75] Althammer S, González-Vallinas J, Ballaré C, *et al.* Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics*, 2011, **27**(24): 3333-3340
- [76] Golumbeanu M, Mohammadi P, Beerenwinkel N. BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. *Bioinformatics*, 2016, **32**(7): 976-983
- [77] Comoglio F, Sievers C, Paro R. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, 2015, **16**: 32
- [78] Wang T, Chen B, Kim M, *et al.* A model-based approach to identify binding sites in CLIP-Seq data. *PLoS One*, 2014, **9**(4): e93248
- [79] Webb S, Hector R D, Kudla G, *et al.* PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biol*, 2014, **15**(1): R8
- [80] Krakau S, Richard H, Marsico A. PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol*, 2017, **18**(1): 240
- [81] Chen B, Yun J, Kim M S, *et al.* PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol*, 2014, **15**(1): R18
- [82] Zhang C, Darnell R B. Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol*, 2011, **29**(7): 607-614
- [83] Huppertz I, Attig J, D'Ambrogio A, *et al.* iCLIP: protein-RNA interactions at nucleotide resolution. *Methods*, 2014, **65**(3): 274-287
- [84] Bailey T L, Boden M, Buske F A, *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 2009, **37**(Web Server issue): W202-W208
- [85] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010, **26**(6): 841-842
- [86] Hiller M, Pudimat R, Busch A, *et al.* Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*, 2006, **34**(17): e117
- [87] Kazan H, Ray D, Chan E T, *et al.* RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, 2010, **6**(7): e1000832
- [88] Bahrami-Samani E, Penalva L O F, Smith A D, *et al.* Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res*, 2015, **43**(1): 95-103
- [89] Fukunaga T, Ozaki H, Terai G, *et al.* CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biol*, 2014, **15**(1): R16
- [90] Maticzka D, Lange S J, Costa F, *et al.* GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 2014, **15**(1): R17
- [91] Heller D, Krestel R, Ohler U, *et al.* ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. *Nucleic Acids Res*, 2017, **45**(19): 11004-11018
- [92] Bose D A, Donahue G, Reinberg D, *et al.* RNA binding to CBP stimulates histone acetylation and transcription. *Cell*, 2017, **168**(1/2): 135-149.e22
- [93] Baltz A G, Munschauer M, Schwanhäusser B, *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, 2012, **46**(5): 674-690
- [94] Davidovich C, Wang X, Cifuentes-Rojas C, *et al.* Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol Cell*, 2015, **57**(3): 552-558
- [95] Betancur J G, Tomari Y. Cryptic RNA-binding by PRC2 components EZH2 and SUZ12. *RNA Biol*, 2015, **12**(9): 959-965
- [96] Zhao J, Ohsumi T K, Kung J T, *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell*, 2010, **40**(6): 939-953
- [97] Zambelli F, Pavesi G. RIP-Seq data analysis to determine RNA-protein associations. *Methods Mol Biol*, 2015, **1269**: 293-303
- [98] Song J, Gooding A R, Hemphill W O, *et al.* Structural basis for inactivation of PRC2 by G-quadruplex RNA. *Science*, 2023, **381**(6664): 1331-1337
- [99] Kung J T, Kesner B, An J Y, *et al.* Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol Cell*, 2015, **57**(2): 361-375
- [100] Jeon Y, Lee J T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 2011, **146**(1): 119-133

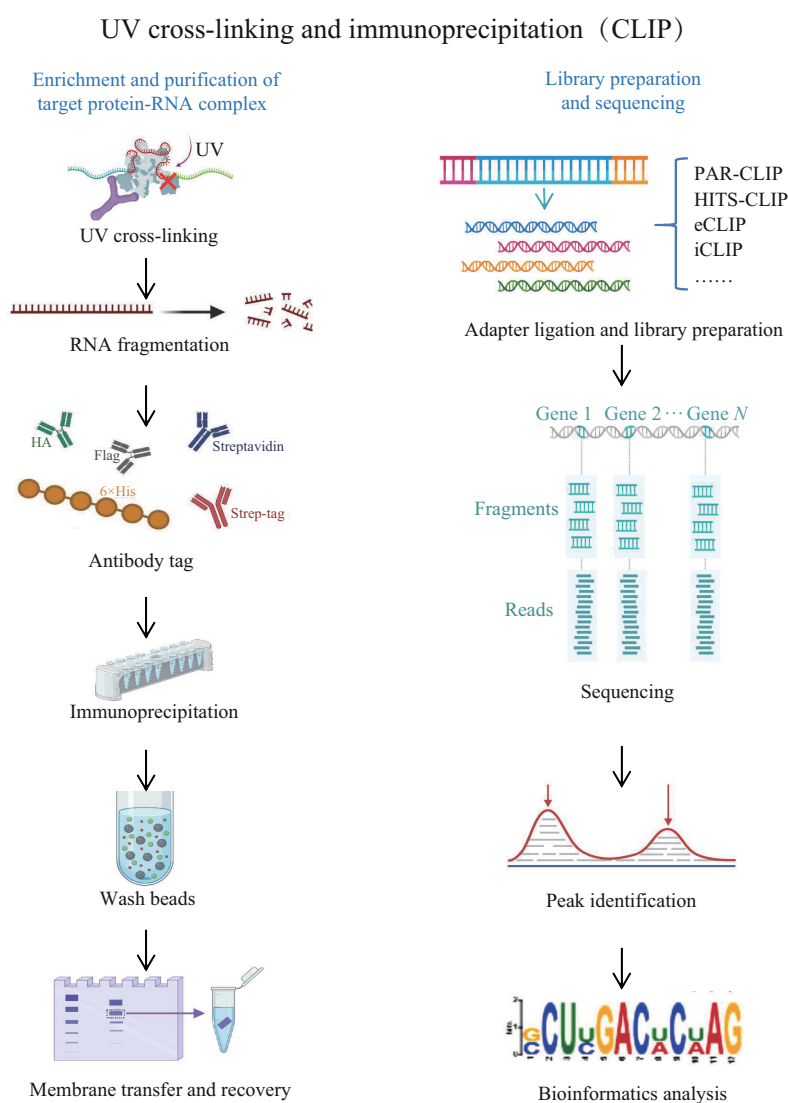
The Refinement and Innovation of The UV Cross-linking and Immunoprecipitation *

ZHAO Jia-Min¹⁾, LU Cheng-Jiang²⁾, YANG Ming²⁾, Nashun Buhe^{1)**}, WANG Gang^{2)**}

¹⁾*School of Life Sciences, Inner Mongolia University, Hohhot 010021, China;*

²⁾*School of Life Sciences, Fudan University, Shanghai 200438, China)*

Graphical abstract



* This work was supported by grants from National Key Research and Development Program of China (2022YFA0806200) and The National Natural Science Foundation of China (32030028).

** Corresponding author.

Nashun Buhe. Tel: 86-471-4996885, E-mail: bnashun@imu.edu.cn

WANG Gang. Tel: 86-13916396401, E-mail: gwang_fd@fudan.edu.cn

Received: September 11, 2024 Accepted: November 26, 2024

Abstract RNA-binding proteins (RBPs) are ubiquitous components within cells, fulfilling essential functions in a myriad of biological processes. These proteins interact with RNA molecules to regulate gene expression at various levels, including transcription, splicing, transport, localization, translation, and degradation. Understanding the intricate network of RBP-RNA interactions is crucial for deciphering the complex regulatory mechanisms that govern cellular function and organismal development. Ultraviolet (UV) cross-linking and immunoprecipitation (CLIP) stands out as a powerful approach designed to map the precise locations where RBPs bind to RNA. By using UV light to create covalent bonds between proteins and RNA, followed by immunoprecipitation to isolate the protein-RNA complexes, researchers can identify the direct targets of specific RBPs. The advent of high-throughput sequencing technologies has revolutionized CLIP, enabling the identification of not only the types but also the exact sequences of RNA bound by RBPs on a genome-wide scale. The evolution of CLIP has led to the development of specialized variants, each with unique features that address specific challenges and expand the scope of what can be studied. High-throughput sequencing CLIP (HITS-CLIP) was one of the first advancements, significantly increasing the throughput and resolution of RNA-protein interaction mapping. Photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) introduced the use of photoactivatable ribonucleosides to enhance cross-linking efficiency and specificity, reducing background noise and improving the detection of low-abundance RNA-protein interactions. Individual-nucleotide resolution CLIP (iCLIP) further refined the technique, achieving unprecedented precision by resolving individual nucleotides involved in RBP binding, which is particularly valuable for studying the fine details of RNA structure and function. Despite the remarkable progress, there remains room for improvement in CLIP technology. Researchers continue to seek methods to increase sensitivity, reduce technical variability, and improve the reproducibility of results. Advances in sample preparation, data analysis algorithms, and computational tools are critical for addressing these challenges. Moreover, the application of CLIP to more diverse biological systems, including non-model organisms and clinical samples, requires the development of tailored protocols and the optimization of existing ones. Looking forward, the field of RNA biology is poised to benefit greatly from ongoing innovations in CLIP technology. The exploration of non-canonical RNA-protein interactions, such as those involving long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs), promises to reveal new layers of cellular regulation and may lead to the discovery of novel therapeutic targets. Furthermore, integrating CLIP data with other omics approaches, such as proteomics and metabolomics, will provide a more comprehensive understanding of the dynamic interplay between RNA and its binding partners within the cell. In conclusion, the continuous refinement and expansion of CLIP techniques have not only deepened our knowledge of RNA biology but have also opened up new avenues for investigating the molecular underpinnings of health and disease. As the technology matures, it is expected to play an increasingly pivotal role in both basic and applied research, contributing to the advancement of medical science and biotechnology.

Key words CLIP, RNA-binding proteins, RNA, sequencing

DOI: 10.16476/j.pibb.2024.0400

CSTR: 32369.14.pibb.20240400