**eCLIP测序报告**

**客户单位：**

**客户名称：**

**合同编号：**

**项目名称：**

# 一、介绍

eCLIP（enhanced Crosslinking and Immunoprecipitation followed by high-throughput sequencing）是一项先进的分子生物学技术，专门设计来精确绘制RNA结合蛋白（RBPs）或DNA结合蛋白与其对应的核酸（RNA或DNA）互作的全基因组图谱。作为CLIP-seq技术的升级版，eCLIP通过优化实验流程，显著增强了检测的特异性和灵敏度，为科学家们提供了更为详尽和准确的蛋白质结合位点信息。

该技术的核心步骤包括：首先，利用紫外线照射或其他交联手段，在细胞内促使蛋白质与它们结合的核酸分子形成共价连接，这一过程有效地冻结了生理状态下的分子互作。随后，细胞被裂解，而这些交联的复合物经过特定抗体的富集，这些抗体针对性地捕获目标蛋白质及其绑定的核酸片段。与标准CLIP-seq相比，eCLIP在此阶段通常引入了更多的质量控制措施，比如增加洗涤步骤以减少非特异性结合，以及采用更精细的片段化策略来提升序列读取的效率和准确性。

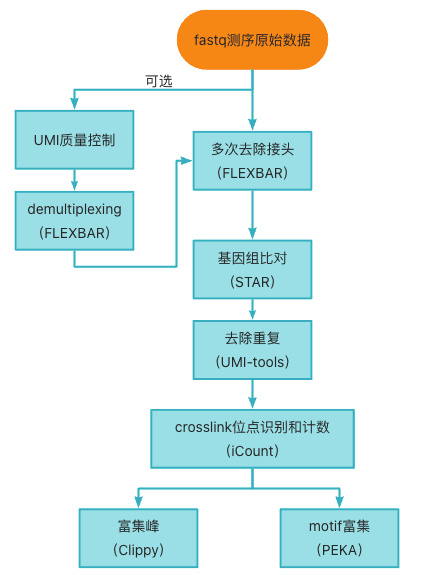
接着，从免疫沉淀下来的复合体中回收的核酸片段经历反转录和PCR扩增，最终通过高通量测序技术进行分析。测序结果能够揭示蛋白质在基因组上的结合偏好和模式，帮助科研人员理解蛋白质如何调控基因表达、RNA剪接、稳定性等关键生物过程。eCLIP数据为探索疾病机制、药物靶标发现及基本生物医学研究提供了宝贵的资源和深刻见解。

# 流程

2.1 样本信息

|  |  |  |  |
| --- | --- | --- | --- |
| 样品名\* | **样品描述** | \*分析名称 | 组名 |
| ENCFF111LIY |  | ENCFF111LIY |  |
| ENCFF217LJP |  | ENCFF217LJP |  |
| ENCFF319MIE |  | ENCFF319MIE |  |
| ENCFF647ULQ |  | ENCFF647ULQ |  |
|  |  |  |  |

2.2 分析流程



# 三、结果

（1）数据质控：

测序得到的原始数据会存在一定比例的低质量数据，为了保证后续信息分析结果的准确可靠，首先要对原始数据进行质控，得到有效数据。Flexbar程序高效地预处理高通量测序数据。它对条形码运行进行解复用，并删除适配器序列。包含Illumina库的几个适配器移除预设。Flexbar使用SIMD和多核并行计算精确的重叠对齐。此外，还提供了修整和过滤功能，例如在读取端修整均聚物。Flexbar提高了读取映射率，改善了基因组和转录组组装。可以灵活地提取独特的分子标识符。该软件支持来自多个测序平台的fasta和fastq格式的数据。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| file | num\_seqs | sum\_len | min\_len | avg\_len | max\_len | Q1 | Q2 | Q3 | sum\_gap | N50 | N50\_num | Q20(%) | Q30(%) | AvgQual | GC(%) |
| ENCFF111LIY | 9,695,975 | 436,318,875 | 45 | 45 | 45 | 45 | 45 | 45 | 0 | 45 | 1 | 97.26 | 95.82 | 19.97 | 62.28 |
| ENCFF217LJP | 10,673,135 | 480,291,075 | 45 | 45 | 45 | 45 | 45 | 45 | 0 | 45 | 1 | 97.33 | 95.97 | 20.08 | 59.25 |
| ENCFF319MIE | 10,673,135 | 416,252,265 | 39 | 39 | 39 | 39 | 39 | 39 | 0 | 39 | 1 | 99.14 | 98.2 | 27.44 | 59.15 |
| ENCFF647ULQ | 9,695,975 | 378,143,025 | 39 | 39 | 39 | 39 | 39 | 39 | 0 | 39 | 1 | 99.04 | 97.97 | 27.1 | 61.03 |

|--01.fastq /  
| |--\*.fastq.gz [fastq测序原始序列文件]  
| |--fastq\_stat.tsv [fastq测序原始序列文件统计汇总结果]

（2）基因组对比

质控处理后的序列数据通过STAR软件比对到参考基因组得到BAM文件。

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | Number of input reads | Average input read length | Uniquely mapped reads number | Uniquely mapped reads % | Average mapped length | Mismatch rate per base, % | Deletion rate per base | Deletion average length | Insertion rate per base | Insertion average length | Number of reads mapped to multiple loci | % of reads mapped to multiple loci |
| ENCFF217LJP | 10323570 | 38 | 4446749 | 43.07% | 35.43 | 0.38% | 0.05% | 1.59 | 0.00% | 1.15 | 0 | 0.00% |
| ENCFF647ULQ | 9619686 | 33 | 2408905 | 25.04% | 32.64 | 0.47% | 0.05% | 1.54 | 0.00% | 1.1 | 0 | 0.00% |
| ENCFF111LIY | 9314737 | 39 | 2792403 | 29.98% | 36.23 | 0.38% | 0.06% | 1.61 | 0.00% | 1.17 | 0 | 0.00% |
| ENCFF319MIE | 10592006 | 32 | 4070615 | 38.43% | 31.99 | 0.44% | 0.04% | 1.51 | 0.00% | 1.1 | 0 | 0.00% |

文件说明：

|--02.align/  
| |--\*.Aligned.sortedByCoord.out.duprm.sort.bam [fastq对比到参考基因组的结果]  
| |--align\_stat.xls [每个样本对比基因组结果统计汇总]

1. 去除重复

使用umitools去除重复序列，去重是至关重要的预处理步骤，它通过排除由于测序错误或复杂基因组区域导致的非生物性差异序列，确保数据分析的准确性和高效性，进而精确识别蛋白和RNA的真实结合位点，提升结果的可信度与可比性，同时优化计算资源利用，为深入探索基因调控机制提供坚实的基础。

1. 蛋白和RNA结合位点识别

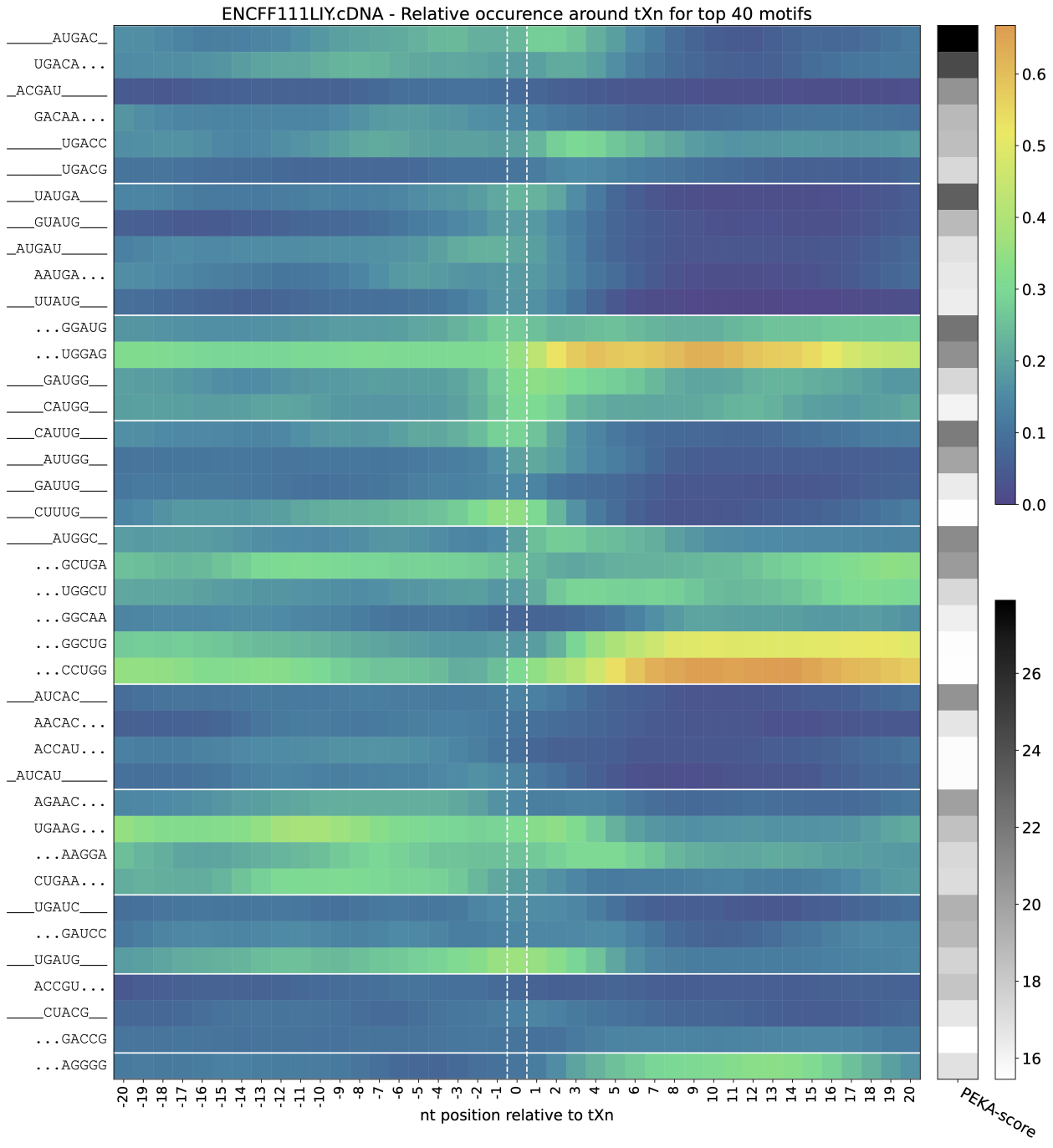
使用iCount软件识别蛋白和RNA结合位点并计数结果保存到BED格式文件中用于后续分析。

文件说明：

|--03.iCount/  
| |--\*.cDNA\_unique.bed [蛋白与RNA结合位点识别和计数]

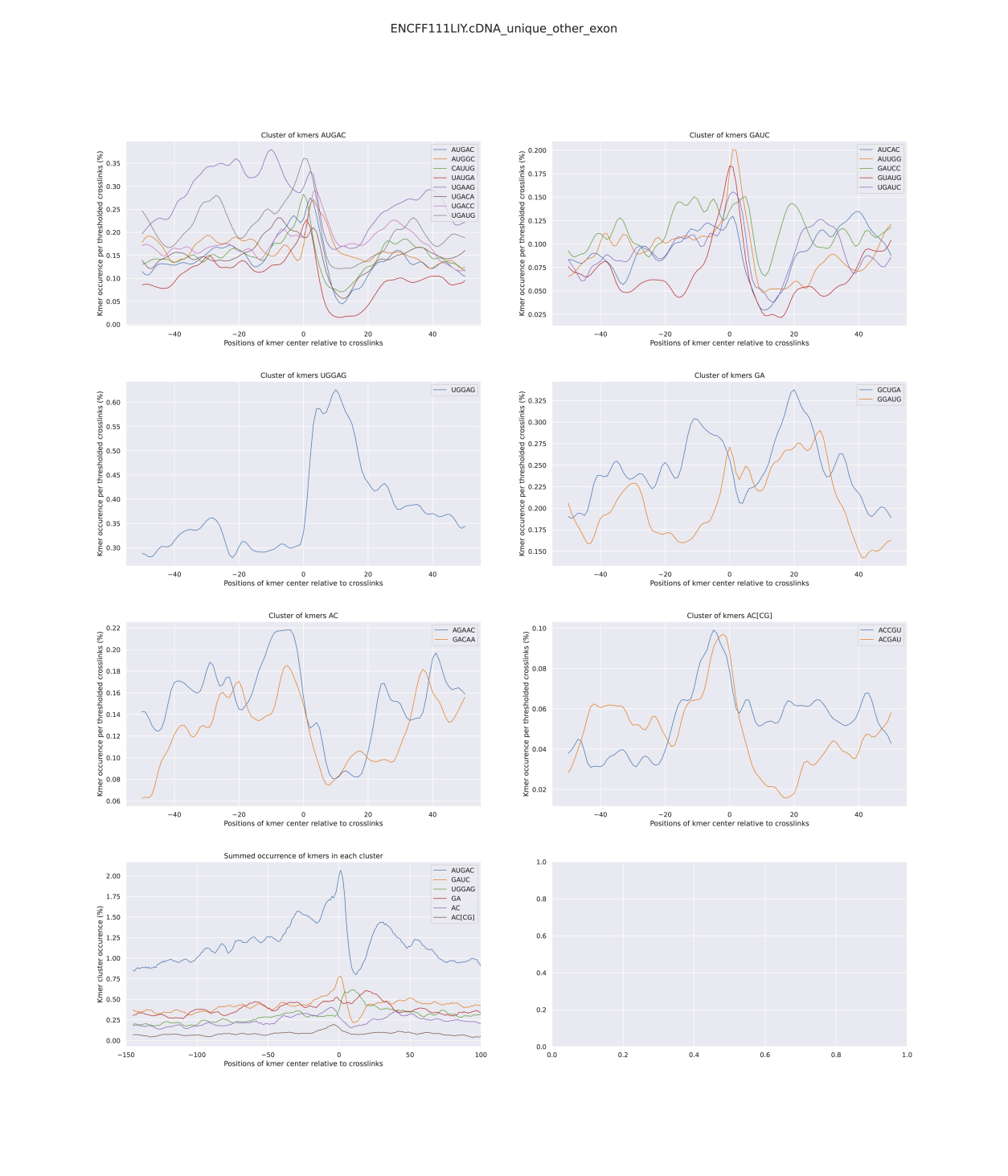
（5）motif富集分析

位置富集k-mer分析（Positionally-enriched k-mer analysis，简称PEKA）是一款软件工具包，旨在从CLIP（Crosslinking and ImmunoPrecipitation）数据集中识别富集的蛋白质-RNA结合基序。PEKA通过比较靠近高置信度交联位点（tXn，即阈值化的交联点）附近的k-mer富集情况，这些位点位于交联峰内部且具有高cDNA计数，与位于峰外且计数较低的交联位点（oXn，外部交联）进行对比。这种方法能有效减少技术偏差的影响，比如紫外交联过程中对尿嘧啶的偏好性。每个k-mer都会被赋予一个PEKA分数，用于按富集程度从高到低对k-mer进行排序。此外，PEKA还提供了围绕高置信度交联位点的基序富集轮廓的全面可视化，并对显示相似轮廓的基序进行聚类。PEKA还支持在特定转录组区域中进行基序发现，包括或排除重复元件，从而增强了分析的灵活性和针对性。



结合位点富集到的前40个motif

Heatmap showing relative occurrences (RtXn) and PEKA scores for top 40 k-mers identified by PEKA for TIA1 eCLIP in HepG2 cell line. K-mers are clustered based on their sequences, which are aligned to the position of their maximal RtXn (on the left), and the most frequently crosslinked nucleotide is highlighted in blue. When the maximal occurrence position is located further than 3 nt from the crosslink site, three dots are shown



k-mer occurrence distributions around thresholded crosslink sites for top n most enriched k-mers in the region spannig -50...50 nt around thersholded crosslink sites. K-mers are clustered based on their sequence and distributions.y-axis on the plots denotes a % of thresholded crosslinks for which a particular k-mer occurs at a specified position (y-axis).

文件说明：

|--04.PEKA/  
| |--\*.cDNA\_heatmap\_top40\_clustered\_rtxn\_underscores.pdf [结合位点富集到的前40个motif热图]

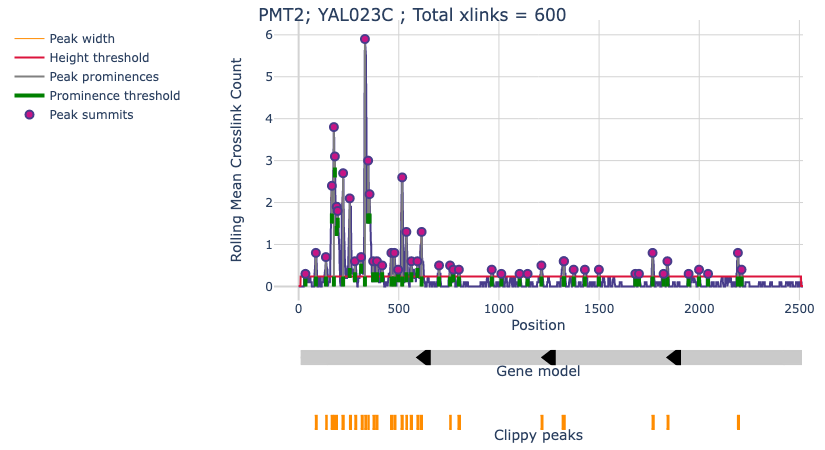
| |--\*.cDNA\_unique\_5mer\_other\_exon.pdf [k-mer分布图]

| |--\*.cDNA\_unique\_5mer\_cluster\_distribution\_other\_exon.tsv[summed occurrence distributions of k-mers within defined clusters. Distributions in this file correspond to the curves on the last plot in the .pdf file]

| |--\*..cDNA\_unique\_5mer\_distribution\_other\_exon.tsv [calculated PEKA score and occurrence distribution for all possible k-mers in the window -48 to +50 around thresholded crosslinks.]

（6）交联峰识别

clippy软件根据提供的注释信息，每个基因上的交联信号会通过移动平均（rolling mean）的方法进行平滑处理，窗口大小可由用户自行设定。对于每个基因，首先计算平滑信号的均值（以红线表示），然后计算均值加上（均值乘以调整因子）的结果（以绿线表示）。这里的均值用于定义峰值的最小高度，而均值加上（均值乘以调整因子）则用来确定峰值的最小突出度。简而言之，这个参数的作用是限制在存在明显更突出峰值的区域中，将那些浅显的峰值也识别出来。通过这样的设置，可以更准确地区分和识别出具有显著意义的交联峰。



文件说明：

|--05.clippy/  
| |--\*\_intergenic\_regions.gtf [基因间隔区注释信息]

| |--\*\_Peaks.bed [结合位点区域]

| |--\*\_Summits.bed [Peaks中的最高点]

# 四、软件版本

|  |  |  |
| --- | --- | --- |
| 软件 | 版本 | 链接 |
| racoon-clip | 1.1.3 | https://github.com/ZarnackGroup/racoon\_clip |
| umitools | 0.3.4 | https://github.com/weng-lab/umitools |
| STAR | 2.7.11b | https://github.com/alexdobin/STAR |
| Flexbar | 3.5 | https://github.com/seqan/flexbar |
| clippy | 1.5.0 | https://github.com/ulelab/clippy |
| PEKA | 1.0.2 | https://github.com/ulelab/peka |
| iCount | 2.0.0 | https://github.com/tomazc/iCount |
|  |  |  |
|  |  |  |

# 五、参考文献

Kuret, K., Amalietti, A. G., Jones, D. M., Capitanchik, C., & Ule, J. (2022). Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP. Genome biology, 23(1), 1-34.

Varier, R. A., Sideri, T., Capitanchik, C., Manova, Z., Calvani, E., Rossi, A., ... & van Werven, F. (2022). m6A reader Pho92 is recruited co-transcriptionally and couples translation efficacy to mRNA decay to promote meiotic fitness in yeast. eLife

Kuret K, Amalietti AG, Jones DM, Capitanchik C, Ule J. Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP. Genome Biol. 2022 Sep 9;23(1):191. doi: 10.1186/s13059-022-02755-2IF: 10.1 Q1 . PMID: 36085079; PMCID: PMC9461102.

Curk et al. (2019) iCount: protein-RNA interaction iCLIP data analysis (in preparation).