



(12)发明专利申请

(10)申请公布号 CN 111909991 A

(43)申请公布日 2020.11.10

(21)申请号 201910384194.2

(22)申请日 2019.05.09

(71)申请人 中国科学院生物物理研究所
地址 100101 北京市朝阳区大屯路15号

(72)发明人 薛愿超 蔡兆奎 曹唱唱

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245

代理人 张立娜

(51)Int.Cl.

C12Q 1/6869(2018.01)

权利要求书2页 说明书22页
序列表1页 附图8页

(54)发明名称

一种捕获RNA原位高级结构及相互作用的方法

(57)摘要

本发明公开了一种捕获RNA原位高级结构及相互作用的方法。该方法包括：固定细胞或组织中蛋白质介导的RNA相互作用；在保持细胞完整的情况下行膜穿孔；降解游离RNA；在RNA3'末端行pCp-biotin标记并原位近端连接；细胞消解后纯化含C-biotin的嵌合体RNA；构建链特异性文库；高通量测序。本发明在不破坏细胞结构、保持细胞完整的情况下，原位处理胞内RNA，捕获生理状态下RNA分子内(间)相互作用；用pCp-biotin标记RNA末端，非变性条件下原位连接，极大提高标记效率，同时降低分子间非特异连接；用C1磁珠能高效富集标记C-biotin的嵌合体RNA，增加测序有效数据比例，降低测序成本。

1. 一种原位捕获RNA高级结构和/或鉴定原位RNA-RNA相互作用的方法,包括如下步骤:

- (1) 对细胞或组织样本进行处理以固定蛋白质介导的RNA-RNA近距离相互作用;
- (2) 在保持细胞完整的情况下进行膜穿孔;
- (3) 降解游离的不受蛋白质保护的RNA;
- (4) 在受蛋白质保护的RNA的3'末端进行“pCp-标记物1”标记并在原位进行近端连接;
- (5) 细胞消解后纯化含有“C-标记物1”的嵌合体RNA;进行链特异性文库构建;
- (6) 进行高通量测序。

2. 根据权利要求1所述的方法,其特征在于:步骤(1)中,所述对细胞或组织样本进行处理为对细胞或组织样本进行甲醛交联;

进一步地,所述步骤(1)是按照包括如下步骤的方法进行的:

(a1) 将所述细胞或组织样本置于甲醛溶液中,室温放置10分钟;

进一步地,所述甲醛溶液为体积百分比为1%的甲醛溶液;

更进一步地,在步骤(a1)之后还包括如下步骤(a2):

(a2) 向步骤(a1)处理过的细胞或组织样本中加入甘氨酸溶液,放置10分钟;

进一步地,所述甘氨酸溶液为浓度为0.125mol/L的甘氨酸溶液。

3. 根据权利要求1或2所述的方法,其特征在于:步骤(2)中,进行所述膜穿孔时采用的穿孔液为Permeabilization溶液;

进一步地,所述步骤(2)是按照包括如下步骤的方法进行的:

(b1) 将经步骤(1)处理过的细胞或组织样本置于所述Permeabilization溶液中,0-4℃放置15分钟;其中,所述Permeabilization溶液的溶剂为10mM pH 7.5的Tris-HCl缓冲液,溶质及浓度如下:10mM NaCl,0.5%NP-40,0.3%Triton X-100,0.1%Tween 20,1×protease inhibitors和2 U/ml SUPERase • InTM RNase Inhibitor;%均表示体积百分含量;

更进一步地,在步骤(b1)之后还包括如下步骤(b2):

(b2) 将步骤(b1)处理过的细胞或组织样本用1×PNK溶液洗涤;其中,所述1×PNK溶液的溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶剂及浓度如下:10mM MgCl₂,0.1mg/ml BSA,0.2%NP-40;%表示体积百分含量。

4. 根据权利要求1-3中任一所述的方法,其特征在于:步骤(3)中,是采用微球菌核酸酶来实现所述“降解游离的不受蛋白质保护的RNA”的。

5. 根据权利要求1-4中任一所述的方法,其特征在于:所述步骤(4)是按照包括如下步骤的方法进行的:

(d1) 将所述受蛋白质保护的RNA的3'末端羟基化;

进一步地,是通过将步骤(3)处理过的样本进行碱性磷酸酶处理,从而实现将所述受蛋白质保护的RNA的3'末端羟基化;

(d2) 将RNA的3'末端标记为“Cp-标记物1”;

进一步地,是通过向步骤(d1)处理过的样本中加入所述“pCp-标记物1”,进行连接反应,从而实现将RNA的3'末端标记为“Cp-标记物1”;

(d3) 将RNA的3'末端的“Cp-标记物1”中的磷酸基团转变为羟基;

进一步地,是通过将步骤(d2)处理过的样本进行碱性磷酸酶处理,从而实现将RNA的3'

末端的“Cp-标记物1”中的磷酸基团转变为羟基；

(d4) 将RNA的5' 末端磷酸化；

进一步地, 是通过将步骤(d3)处理过的样本进行T4 PNK酶处理, 从而实现将RNA的5' 末端磷酸化；

(d5) 在原位进行近端连接；

进一步地, 是通过向步骤(d4)处理过的样本中加入T4 RNA连接酶, 从而实现在原位进行近端连接。

6. 根据权利要求1-5中任一所述的方法, 其特征在于: 所述步骤(5)是按照包括如下步骤的方法进行的:

(e1) 利用蛋白酶K消解细胞；

(e2) 提取总RNA, 并进行片段化处理；

(e3) 利用固定有标记物2的磁珠富集标记有“C-标记物1”的嵌合体RNA; 所述标记物2能够与所述标记物1特异性结合；

(e4) 构建链特异性文库。

7. 根据权利要求1-6中任一所述的方法, 其特征在于: 所述方法中, 所述细胞的最高起始量为 1×10^7 个细胞；

进一步地, 所述细胞为动物细胞, 所述组织为动物组织。

8. 一种文库构建方法, 包括权利要求1-7中任一所述方法的步骤(1)至步骤(5)。

9. 权利要求8所述文库在原位捕获RNA高级结构和/或鉴定原位RNA-RNA相互作用中的应用。

10. 如下任一应用:

(A1) 权利要求1-8中任一所述方法在鉴定活细胞内的lncRNA靶标中的应用；

(A2) pCp-biotin在鉴定RNA-RNA近距离相互作用中应用；

(A3) pCp-biotin在RNA原位近端连接中的应用；

(A4) pCp-biotin在嵌合体RNA富集中的应用。

一种捕获RNA原位高级结构及相互作用的方法

技术领域

[0001] 本发明属于生物技术领域,具体涉及一种捕获RNA原位高级结构及相互作用的方法。

背景技术

[0002] 遗传信息的载体DNA需要先转录成RNA,然后再翻译成蛋白质才能发挥生物学功能。RNA作为遗传信息的传递者,主要作用是编码和指导蛋白质的合成,这一类编码蛋白质的RNA统称为信使RNA(messenger RNA,mRNA)。此外,人类基因组还转录产生了海量的不编码蛋白质的RNA,这类RNA被称为非编码RNA(noncoding RNA,ncRNA)。目前已发现的具有调控作用的非编码RNA包括:tRNA、rRNA、siRNA、miRNA、piRNA、snoRNA、circRNA和lncRNA等,它们的异常表达和突变与癌症发生、发育和生殖缺陷等重大疾病相关。作为遗传信息的关键调控者,RNA往往需要先通过分子内碱基配对形成复杂的高级结构,然后再通过与其他RNA分子相互作用才能发挥重要的生物学调控功能。利用测序技术,我们已经可以获得RNA的详细序列信息,但是对于RNA的结构,尤其是高级结构信息的获取依然是个世界性难题。虽然一些物理手段,比如核磁共振、冷冻电镜和晶体学等方法可以解析RNA的高分辨率结构和分子内及分子间相互作用,但是这些技术的通量太低。目前国际蛋白质数据库PDB中收录的人源RNA的高分辨率结构屈指可数。因此,如何系统和精确的解析RNA的分子内和分子间相互作用依然是我们所面临的巨大挑战。

[0003] 近年来,大量分析RNA二级结构的技术被开发出来,这些技术的特点是先利用化学修饰或者酶切处理RNA,然后再进行建库测序,例如:DMS-seq、Structure-seq、icSHAPE等,他们利用了单链区的RNA容易被化合物DMS(dimethyl sulfoxide,硫酸二甲酯)或NAI-N3等修饰的特点,通过分析反转录酶停止的位置而推导RNA的哪些碱基处于单链区。此外,对于RNA结构中的双链配对区域,目前也有多个方法可以解析,例如PARIS、LIGR-seq和SPLASH等。这三种方法的基本原理是:在培养基中加入Psoralen或AMT,它们进入细胞后会穿过细胞膜并迅速结合到RNA上双链配对的区域,经254nm的紫外线(UV)照射后,细胞内配对的RNA会被psoralen或AMT共价交联起来,之后对富集出的RNA进行片段化处理,并在溶液中进行近端连接。然后对连接后的RNA进行365nm UV照射后可打开psoralen或AMT与双链RNA之间的共价键,之后进行文库构建和测序。以上方法虽然可以高通量的研究RNA的单链区和双链区,但也存在一些缺陷:第一,不能捕获非Watson-Crick配对以及远距离的RNA loop-loop相互作用。第二,这些连接反应都是在溶液中进行,存在非特异性连接,不能反映RNA在细胞内的真实结构,导致大量的假阳性分子间连接。第三,测序所得数据中,嵌合体reads(chimeric reads,即不同RNA片段间连接的产物)比率较低,无用数据太多。RNA近端连接技术(RPL)理论上可以克服以上技术缺陷,但由于缺乏交联和嵌合体RNA富集,因此导致RPL技术只能鉴定分子内相互作用,而不能鉴定分子间的RNA-RNA相互作用。

[0004] 近年来的高通量转录组测序表明超过90%的基因组被转录,产生了海量的非编码RNA,其中一些非编码RNA与染色质紧密结合,如lncRNA(long non-coding RNA)。lncRNA是

一类长度在200nt以上、不编码蛋白质的RNA。目前NONCODE收录的人类lncRNA数目已经超过了160000个,是蛋白质编码基因的8倍,但是绝大多数非编码RNA的功能、靶标及作用机制还不清楚。目前常用的鉴定lncRNA靶标的方法包括:CHIRP、CHART和RAP。这些方法的原理是:在生理状态下,首先利用甲醛处理细胞,以固定RNA及其相互作用的靶标分子;然后裂解细胞,利用超声或酶对染色质进行片段化;之后利用生物素修饰的DNA探针富集与目的RNA相互作用的DNA片段;对DNA片段加接头后,利用PCR进行文库扩增和高通量测序;最后结合生物信息学分析,鉴定与特定lncRNA相互作用的靶标DNA。CHIRP、CHART和RAP方法只关注DNA靶标,忽略了具有重要功能的RNA靶位点,而且每次只能鉴定一个lncRNA在细胞内的所有潜在的DNA靶标(one to all),通量太低。因此,如何系统鉴定细胞内所有lncRNA在基因组上的所有结合位点还是个难题。

发明内容

[0005] 鉴于以上技术存在的问题,本发明开发了RNA原位构象测序新技术(RNA in situ conformation sequencing,简称RIC-seq)。基本原理是对细胞进行甲醛交联以固定蛋白质介导的RNA-RNA近距离相互作用,并在保持细胞完整的情况下进行膜穿孔,进而利用微球菌核酸酶切割以去除游离的不受蛋白质保护的片段,然后在RNA的3'末端上进行pCp-biotin标记并在原位进行近端连接。细胞消解后纯化含有C-biotin的嵌合体RNA,并进行链特异性文库构建,该步骤大大提高了数据中嵌合体reads的比率,减少了无用数据,降低了测序成本。RIC-seq在保持细胞完整性的条件下进行RNA-RNA的原位连接,可一次性捕获所有直接的RNA-RNA近距离接触,并可在原位检测所有的lncRNA在体内的RNA结合靶标(all to all)。最重要的是可根据RNA的近端空间距离信息重构RNA的高级结构。

[0006] 第一方面,本发明要求保护一种原位捕获RNA高级结构和/或鉴定原位RNA-RNA相互作用的方法(即RIC-seq法)。

[0007] 本发明所要求保护的原位捕获RNA高级结构和/或鉴定原位RNA-RNA相互作用的方法(RIC-seq法),可包括如下步骤:

[0008] (1)对细胞或组织样本进行处理以固定蛋白质介导的RNA-RNA近距离相互作用。其中,所述组织样本的体积大小可为1立方毫米;所述近距离可为距离50埃以内。

[0009] (2)在保持细胞完整的情况下进行膜穿孔(细胞膜和核膜穿孔)。

[0010] (3)降解游离的不受蛋白质保护的RNA。

[0011] (4)在受蛋白质保护的RNA的3'末端进行“pCp-标记物1”标记并在原位进行近端连接。其中,所述近端可为距离50埃以内。

[0012] 其中,所述“pCp-标记物1”为两端为磷酸基团且标记有所述标记物1的胞嘧啶核苷酸。相应的,后文出现的“Cp-标记物”为3'末端为磷酸基团且标记有所述标记物1的胞嘧啶核苷酸;“C-标记物1”为标记有所述标记物1的胞嘧啶核苷酸。

[0013] 在本发明的具体实施例方式中,所述“pCp-标记物1”具体为pCp-biotin。相应的,所述“Cp-标记物1”具体为Cp-biotin;所述“C-标记物1”具体为C-biotin。

[0014] 所述pCp-biotin为两端为磷酸基团且标记有生物素的胞嘧啶核苷酸。相应的,后文出现的Cp-biotin为3'末端为磷酸基团且标记有生物素的胞嘧啶核苷酸;C-biotin为标记有生物素的胞嘧啶核苷酸。

[0015] (5) 细胞消解后纯化含有“C-标记物1”的嵌合体RNA (chimeric reads, 即不同RNA片段间连接的产物); 进行链特异性文库构建。

[0016] (6) 进行高通量测序。

[0017] 在所述方法的步骤(1)之前, 还可包括对细胞或组织样本进行洗涤的步骤, 所述洗涤方法具体可按照如下步骤进行; 向细胞或组织样本中加入预冷PBS溶液(pH 7.4)进行洗涤, 在4℃、2500rpm条件下离心10分钟, 去除PBS溶液, 得到洗涤后的细胞样本。

[0018] 在所述方法的步骤(1)中, 所述对细胞或组织样本进行处理为对细胞或组织样本进行甲醛交联。

[0019] 进一步地, 所述步骤(1)可按照包括如下步骤的方法进行:

[0020] (a1) 将所述细胞或组织样本置于甲醛溶液中, 室温放置10分钟。其中, 所述甲醛溶液为体积百分比为1%甲醛溶液(溶剂为PBS溶液)。

[0021] 更进一步地, 在步骤(a1)之后还可包括如下步骤(a2):

[0022] (a2) 向步骤(a1)处理过的细胞或组织样本中加入甘氨酸溶液终止反应, 混匀10分钟。其中, 所述甘氨酸溶液为浓度为0.125mol/L的甘氨酸溶液(溶剂为DEPC水)。

[0023] 在所述方法的步骤(2)中, 进行所述膜穿孔时采用的穿孔液为Permeabilization溶液。

[0024] 进一步地, 所述步骤(2)可按照包括如下步骤的方法进行的:

[0025] (b1) 将经步骤(1)处理过的细胞或组织样本置于所述Permeabilization溶液中, 0℃-4℃(如冰浴)放置15分钟, 每隔2分钟混匀一次。其中, 所述Permeabilization溶液的溶剂为10mM pH 7.5的Tris-HCl缓冲液, 溶质及浓度如下: 10mM NaCl, 0.5%(体积百分含量) NP-40, 0.3%(体积百分含量) Triton X-100, 0.1%(体积百分含量) Tween 20, 1×protease inhibitors和2U/ml SUPERase • In™ RNase Inhibitor。

[0026] 在本发明的具体实施方式中, 所述1×protease inhibitors具体为Sigma产品, 货号为P8340-5ML(其具体组份为AEBSF, Aprotinin, Bestatin, E-64, Leupeptin, Pepstatin A)。当然, 所述1×protease inhibitors也可为组成相同的其他产品。

[0027] 在本发明的具体实施方式中, 所述SUPERase • In™ RNase Inhibitor为Thermo Fisher产品, 货号为AM2694。当然, 所述SUPERase • In™ RNase Inhibitor也可为组成相同的其他产品。

[0028] 更进一步地, 在步骤(b1)之后还可包括如下步骤(b2):

[0029] (b2) 将步骤(b1)处理过的细胞或组织样本用1×PNK溶液洗涤。其中, 所述1×PNK溶液的溶剂为50mM pH 7.4的Tris-Cl缓冲液, 溶质及浓度为: 10mM MgCl₂, 0.1mg/ml BSA, 0.2%(体积百分含量) NP-40。

[0030] 在步骤(b2)中, 所述洗涤可为多次洗涤, 如3次。每次洗涤可包括如下步骤: 4℃、旋转(如20rpm)混匀5分钟, 4℃、3500rpm条件下离心5分钟, 去除洗涤溶液。

[0031] 在所述方法的步骤(3)中, 是采用微球菌核酸酶来实现所述“降解游离的不受蛋白质保护的RNA”的。

[0032] 进一步地, 所述步骤(3)可按照包括如下步骤的方法进行:

[0033] (c1) 将经步骤(2)处理过的样本置于1×微球菌核酸酶溶液中反应。其中, 所述1×微球菌核酸酶溶液中微球菌核酸酶的浓度可为0.03U/μl。所述反应的条件可为: 37℃反应

10分钟,每隔2分钟1000rpm震荡15秒。

[0034] 更进一步地,在步骤(c1)之后还可包括如下步骤(c2):

[0035] (c2)将步骤(c1)处理过的样本先后用1×PNK+EGTA溶液和1×PNK溶液进行洗涤。其中,所述1×PNK+EGTA溶液的溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶质及浓度为20mM EGTA,0.5% (体积百分含量) NP-40。所述1×PNK溶液的溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶剂及浓度如下:10mM MgCl₂,0.1mg/ml BSA,0.2% (体积百分含量) NP-40。

[0036] 在步骤(c2)中,所述洗涤可为多次洗涤,如用所述1×PNK+EGTA溶液洗涤2次,用所述1×PNK溶液洗涤2次。每次洗涤可包括如下步骤:4℃、旋转(如20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,去除洗涤溶液。

[0037] 在所述方法中,所述步骤(4)可按照包括如下步骤的方法进行:

[0038] (d1)将所述受蛋白质保护的RNA的3'末端羟基化。

[0039] 进一步地,可通过将步骤(3)处理过的样本进行碱性磷酸酶处理,从而实现将所述受蛋白质保护的RNA的3'末端羟基化;

[0040] 更进一步地,所述“将步骤(3)处理过的样本进行碱性磷酸酶处理”时,反应体系中碱性磷酸酶的含量可为0.1U/μl。反应条件可为:37℃反应10分钟,每隔3分钟1000rpm震荡15秒。

[0041] 再进一步地,所述步骤(d1)反应结束后,还可包括洗涤的步骤;所述洗涤具体为先后用所述1×PNK+EGTA溶液(配方同前)、High-salt溶液和1×PNK溶液先后对细胞样本进行洗涤。所述High-salt溶液的溶剂为5×PBS(no Mg²⁺,Ca²⁺) (即5×PBS缓冲液(pH 7.4):NaCl 685mmol/L,KCl 13.5mmol/L,Na₂HPO₄ 50mmol/L,KH₂PO₄10mmol/L),溶质及浓度为0.5% (体积百分含量) NP-40。所述1×PNK溶液的溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶剂及浓度如下:10mM MgCl₂,0.1mg/ml BSA,0.05% (体积百分含量) NP-40。其中,所述洗涤可为多次洗涤,如用所述1×PNK+EGTA溶液洗涤2次,用所述High-salt溶液洗涤2次,用所述1×PNK溶液洗涤2次。每次洗涤可包括如下步骤:4℃、旋转(如20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,去除洗涤溶液。

[0042] (d2)将RNA的3'末端标记为Cp-biotin。

[0043] 进一步地,可通过向步骤(d1)处理过的样本中加入所述pCp-biotin,进行连接反应,从而实现将RNA的3'末端标记为Cp-biotin。

[0044] 更进一步地,进行所述连接反应时采用的酶可为T4RNA连接酶。在反应体系中,所述pCp-biotin的终浓度可为40μM;所述T4RNA连接酶的终浓度可为1U/μl。反应条件可为:16℃反应12-16h,每隔3分钟1000rpm震荡15秒。

[0045] 再进一步地,所述步骤(d2)反应结束后,还可包括洗涤的步骤;所述洗涤具体为先后用所述1×PNK溶液(配方见具体实施方式实施例1前)对细胞样本进行洗涤;其中,所述洗涤可为多次洗涤,如3次。每次洗涤可包括如下步骤:4℃、旋转(如20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,去除洗涤溶液。

[0046] (d3)将RNA的3'末端的Cp-biotin中的磷酸基团转变为羟基。

[0047] 进一步地,可通过将步骤(d2)处理过的样本进行碱性磷酸酶处理,从而实现将RNA的3'末端的Cp-biotin中的磷酸基团转变为羟基;

[0048] 更进一步地,所述“将步骤(d2)处理过的样本进行碱性磷酸酶处理”时,反应体系

中碱性磷酸酶的含量可为 $0.1\text{U}/\mu\text{l}$ 。反应条件可为： 37°C 反应10分钟，每隔3分钟1000rpm震荡15秒。

[0049] 再进一步地，所述步骤(d3)反应结束后，还可包括洗涤的步骤；所述洗涤具体可为先后用所述 $1\times\text{PNK}+\text{EGTA}$ 溶液(配方同前)、所述High-salt溶液(配方同前)和所述 $1\times\text{PNK}$ 溶液(配方同步骤(d1))先后对细胞样本进行洗涤。其中，所述洗涤可为多次洗涤，如用所述 $1\times\text{PNK}+\text{EGTA}$ 溶液洗涤2次，用所述High-salt溶液洗涤2次，用所述 $1\times\text{PNK}$ 溶液洗涤2次。每次洗涤可包括如下步骤： 4°C 、旋转(如20rpm)混匀5分钟， 4°C 、3500rpm条件下离心5分钟，去除洗涤溶液。

[0050] (d4)将RNA的5'末端磷酸化。

[0051] 进一步地，可通过将步骤(d3)处理过的样本进行T4PNK酶处理，从而实现将RNA的5'末端磷酸化。

[0052] 更进一步地，所述“将步骤(d3)处理过的样本进行T4PNK酶处理”时，反应体系中T4PNK酶的含量可为 $1\text{U}/\mu\text{l}$ 。反应条件可为： 37°C 反应45分钟，每隔3分钟1000rpm震荡15秒。

[0053] 再进一步地，所述步骤(d4)反应结束后，还可如包括洗涤的步骤；所述洗涤具体可为先后用所述 $1\times\text{PNK}+\text{EGTA}$ 溶液(配方同前)和所述 $1\times\text{PNK}$ 溶液(配方同步骤(d1))进行洗涤。其中，所述洗涤可为多次洗涤，如用所述 $1\times\text{PNK}+\text{EGTA}$ 溶液洗涤2次，用所述 $1\times\text{PNK}$ 溶液洗涤2次。每次洗涤可包括如下步骤： 4°C 、旋转(如20rpm)混匀5分钟， 4°C 、3500rpm条件下离心5分钟，去除洗涤溶液。

[0054] (d5)在原位进行近端连接。其中，所述近端可为距离50埃以内。

[0055] 进一步地，是通过向步骤(d4)处理过的样本中加入T4RNA连接酶，从而实现在原位进行近端连接。

[0056] 更进一步地，所述“向步骤(d4)处理过的样本中加入T4RNA连接酶”时，反应体系中T4RNA连接酶的含量可为 $0.5\text{U}/\mu\text{l}$ 。反应条件可为： 16°C 反应12-16h，每隔3分钟1000rpm震荡15秒。

[0057] 再进一步地，所述步骤(d2)反应结束后，还可包括洗涤的步骤；所述洗涤具体可为先后用所述 $1\times\text{PNK}$ 溶液(配方同前)对细胞样本进行洗涤；其中，所述洗涤可为多次洗涤，如3次。每次洗涤可包括如下步骤： 4°C 、旋转(如20rpm)混匀5分钟， 4°C 、3500rpm条件下离心5分钟，去除洗涤溶液。

[0058] 在所述方法中，所述步骤(5)可按照包括如下步骤的方法进行：

[0059] (e1)利用蛋白酶K消解细胞。

[0060] 进一步地，所述“利用蛋白酶K消解细胞”时，反应体系中蛋白酶K的含量可为 $0.12\text{U}/\mu\text{l}$ 。反应条件可为： 37°C 反应60分钟， 56°C 反应15分钟。

[0061] (e2)提取总RNA，并进行片段化处理。

[0062] 该步骤中，所述提取总RNA可利用TRIzol LS、氯仿进行抽提。另外，在沉淀RNA时可加入 $500\mu\text{l}$ 异丙醇， $15\mu\text{g}$ glycoblue， -20°C 过夜。

[0063] 进一步地，在提取得到所述总RNA后，还可包括去除基因组DNA(如用DNaseI处理)和去除核糖体RNA(如利用与核糖体RNA互补配对的探针去除)的步骤。

[0064] 其中，利用与核糖体RNA互补配对的DNA探针去除核糖体RNA的步骤可如下：向RNA中加入等质量的核糖体RNA探针， 95°C 反应2分钟， $-0.1^{\circ}\text{C}/\text{s}$ 降至 22°C ， 22°C 反应5分钟。(反

应结束后可立即放冰上)。降解DNA:RNA杂合链中的RNA(如加入RNase H),降解DNA探针(如加入Turbo DNA酶)。然后纯化RNA(如用Zymo RNA clean试剂盒)。

[0065] 该步骤中,所述对RNA进行片段化处理具体可采用碱裂解法。在本发明的具体实施方式中,具体是使用1×first strand buffer(配方:50mM Tris-HCl,pH 8.3;75mM KCl;3mM MgCl₂),PCR仪中94°C,5分钟对RNA进行片段化处理。

[0066] (e3) 利用固定有标记物2的磁珠富集标记有“C-标记物1”(如C-biotin)的嵌合体RNA;所述标记物能够与所述标记物1特异性结合。

[0067] 在本发明的具体实施方式中,所述标记物1具体为生物素,所述标记物2具体为链霉亲和素。所述固定有标记物2的磁珠即为链霉亲和素磁珠。

[0068] 该步骤中,在利用链霉亲和素磁珠富集标记有C-biotin的嵌合体RNA前还包括对所述链霉亲和素磁珠进行封闭的步骤。具体步骤可如下:取20μl C1磁珠,把离心管放在磁力架上,待溶液澄清,吸出上清,加20μl溶液A,重悬磁珠,室温放置2分钟,把离心管放在磁力架上,待溶液澄清,吸出上清,重复此步骤1次,加入20μl溶液B,重悬磁珠,把离心管放在磁力架上,待溶液澄清,吸出上清,加入32μl酵母RNA(50μg),68μl DEPC water和100μl 2×TWB溶液,重悬磁珠,将离心管放在旋转混匀仪上,旋转混匀1小时,然后将离心管放在磁力架上,待溶液澄清,吸出上清,加入500μl 1×TWB溶液,重悬磁珠,将离心管放在磁力架上,待溶液澄清,吸出上清,重复此步骤2次。

[0069] 该步骤中,在利用链霉亲和素磁珠富集标记有C-biotin的嵌合体RNA后还包括从磁珠上洗脱RNA的步骤。

[0070] (e4) 构建链特异性文库。

[0071] 该步骤主要包括:合成一链cDNA;合成二链DNA;对dsDNA进行末端修复;对末端修复的DNA进行加‘A’;连接接头;以加完接头的DNA为模板进行PCR扩增,用琼脂糖胶进行回收特定片段大小的PCR产物,得到所述链特异性文库;高通量测序。这些步骤均为本领域常规操作。按照常规流程构建链特异性文库的方法可参照“Levin,J.Z.,Yassour,M.,Adiconis,X.,Nusbaum,C.,Thompson,D.A.,Friedman,N.,Gnirke,A.,and Regev,A.(2010).Comprehensive comparative analysis of strand-specific RNA sequencing methods.Nature methods7,709-715.”一文相关记载进行。

[0072] 在本发明的具体实施方式中,所述合成二链DNA时,使用的是25mM dNTPs与dUTP的混合物,其中dTTP与dUTP摩尔比为4:1。

[0073] 该步骤中,在“合成二链DNA”和“对dsDNA进行末端修复”之间,“对dsDNA进行末端修复”和“对末端修复的DNA进行加‘A’”之间,以及“连接接头”之后均可包括DNA纯化步骤。所述纯化的方法可为磁珠纯化。所述磁珠纯化的具体方法可按照如下步骤进行:将AMPure XP磁珠(简称XP磁珠)提前室温混匀平衡30分钟。然后将XP磁珠加入所述洗脱液中,轻柔混匀。室温静置5分钟,转移至磁力架静置5分钟,去除上清液,并使用新鲜的80%乙醇溶液漂洗磁珠2次。磁珠置于磁力架上晾干2分钟,加入TE缓冲液重悬磁珠,吹打50次。室温静置5分钟,然后放入磁力架静置5分钟,收集上清液即为纯化后的DNA产物。其中,所述“连接接头”之后的DNA纯化步骤(如磁珠纯化)可为两次。

[0074] 在本发明的具体实施方式中,该步骤中进行所述PCR扩增时采用的上下游引物为由SEQ ID No.1和SEQ ID No.2所示的两条单链DNA组成的引物对。具体的,该步骤中进行的

所述PCR扩增的反应体系具体为：上清液（经在“连接接头”之后的磁珠纯化DNA步骤所获得的上清液）15.7 μ l、10 \times Pfx缓冲液（Invitrogen）2.5 μ l、10 μ M的上下游引物（SEQ ID No.1和SEQ ID No.2）各1 μ l、50mM的MgSO₄溶液1 μ l、25mM的dNTP 0.4 μ l、Pfx酶（Invitrogen）0.4 μ l、USER酶（NEB）3 μ l。所述PCR扩增的反应程序具体为：37 $^{\circ}$ C反应15分钟；94 $^{\circ}$ C反应2分钟；94 $^{\circ}$ C变性15秒，62 $^{\circ}$ C退火30秒，72 $^{\circ}$ C延伸30秒，反应12个循环；72 $^{\circ}$ C反应10分钟。

[0075] 在所述方法的步骤（6）中，所述高通量测序可使用Illumina HiSeq X Ten测序仪对步骤（5）所得文库进行测序，可进行PE150双端测序。

[0076] 在所述方法中，所述细胞的最高起始量为 1×10^7 个细胞。

[0077] 进一步地，所述细胞可为动物细胞（如人源细胞），所述组织可为动物组织。在本发明的具体实施方式中，所述细胞具体为HeLa细胞。

[0078] 第二方面，本发明要求保护一种文库构建方法。

[0079] 本发明所要求保护的文库构建方法，包括前文第一方面所述方法的步骤（1）至步骤（5）。

[0080] 第三方面，本发明要求保护利用第二方面所述方法构建得到的所述文库在原位捕获RNA高级结构和/或鉴定原位RNA-RNA相互作用中的应用。

[0081] 第四方面，本发明还要求保护如下任一应用：

[0082] （A1）前文第一方面所述方法在鉴定活细胞内的lncRNA靶标中的应用。

[0083] （A2）pCp-biotin在鉴定RNA-RNA近距离相互作用中应用；其中，所述近距离可为距离50埃以内。

[0084] （A3）pCp-biotin在RNA原位近端连接中的应用；其中，所述近端可为距离50埃以内。

[0085] （A4）pCp-biotin在嵌合体RNA富集中的应用。

[0086] 第五方面，本发明还要求保护如下任一：

[0087] （B1）洗涤剂，为前文所述Permeabilization溶液。

[0088] （B2）步骤（B1）所述洗涤剂在对细胞进行膜穿孔中的辅助性用途。

[0089] （B3）微球菌核酸酶（Micrococcal Nuclease）、碱性磷酸酶（Alkaline Phosphatase）和/或T4多聚核苷酸激酶（T4PNK）在RNA的原位连接（如原位近端连接）中的应用。

[0090] （B4）蛋白酶K和加热连用在甲醛固定的细胞样本或组织样本中提取RNA的用途。其中，所述加热指的是37 $^{\circ}$ C反应60分钟，56 $^{\circ}$ C反应15分钟。

[0091] 在本发明中，所述原位连接为非变性条件下的原位连接。

[0092] 本发明的有益效果：1、能够在不破坏细胞结构、保持细胞完整性的情况下，在原位对细胞内RNA进行处理，捕获生理状态下的RNA分子内和分子间相互作用；2、使用pCp-biotin标记RNA末端，在非变性条件下进行原位连接，极大的提高标记效率，同时降低分子间非特异性连接；使用C1磁珠富集标记了C-biotin的嵌合体RNA，进而用于构建文库，能够高效的富集嵌合体RNA，增加有效数据比例，降低测序成本。

附图说明

[0093] 图1为RIC-seq流程及数据验证。A为RIC-seq技术示意图。原位处理部分包括甲醛

交联、细胞穿孔、RNA消解、pCp-biotin标记和近端连接。体外部分包括富集嵌合体RNA和构建双端链特异性文库。RBP代表结合RNA结合蛋白。B为嵌合体RNA的接合处碱基组成。C为比较RIC-seq嵌合体RNA(浅色部分)与miR-3064已知结构(弧线)。嵌合体RNA之间空白区域代表缺口。一条RIC-seq read构成的茎环结构展示在下方,实心圆圈为插入了pCp-biotin的地方。D为pCp-biotin富集在miRNA前体顶端环状区。E-G为RIC-seq重现了U4,U6,RPPH1和U3snoRNA已知的结构和相互作用。H为RIC-seq鉴定了U1在MALAT1上的结合位点。深色阴影区域是RIC-seq、PARIS和RAP方法共同鉴定的位点。浅色阴影是RIC-seq特异性鉴定的位点。虚线方框区域在I中展示。I为新的U1和MALAT1相互作用位点是保守的,而且有嵌合体RNA簇支持(箭头)。U1基序由实心方框表示。J为RNA图谱展示了HeLa细胞中所有的RNA-RNA相互作用,左下方为+pCp样品,右上方为-pCp样品。NEAT1和MALAT1相互作用被放大后展示在右侧。K为超高分辨成像分析揭示了MALAT1与NEAT1的5'端是共定位的。方框1和方框2放大后展示在中间。方框3和方框4中展示直接相互作用的位点。它们之间的相对距离展示在右侧。

[0094] 图2为RIC-seq分析流程及可重复性。A为RIC-seq数据分析流程。PCR重复、接头序列和含有多聚N的测序片段首先被去除,然后利用STAR软件将reads比对到hg19参考基因组。B为RIC-seq两个重复高度相关。C为细胞混合策略检测到RIC-seq技术的整体假阳性率是0.6%(虚线方框区域)。D为RIC-seq成功捕获snoRNA在28S rRNA上的作用和修饰位点。箭头表示已知的修饰位点,方框区域代表D' box。E为RIC-seq检测到的snoRNA相互作用位点在基因组上的分布。F为SNORD22与SPHK2和BCL2L2基因相互作用的位点。下划线区域表示D-box。G为分子内与分子间RNA-RNA相互作用的统计饼状图。H为小提琴图展示MALAT1和NEAT1的靶基因表达水平高于其他基因。双尾t检验计算显著性差异。I为MALAT1或者NEAT1靶标富集的基序。J为15个细胞中MALAT1的位点总结。K为smFISH在15个细胞中检测到的NEAT1位点及其与MALAT1位点重叠的总结。深色柱代表NEAT1位点与MALAT1位点直接重叠。

[0095] 图3为RIC-seq精确捕获28S rRNA的3D结构。A为根据28S rRNA冷冻电镜结构绘制的物理相互作用图谱。浅灰色区域表示空间距离大于25 Å(远端)。25 Å以内的Watson-Crick,非Watson-Crick碱基相互作用,包含Watson-Crick,非Watson-Crick碱基配对的相互作用以及其他类型的近端相互作用用不同颜色标示。Not available表示没有可用的结构数据。B为根据HeLa细胞RIC-seq数据绘制的28S rRNA3D图谱。方框区域展示了Watson-Crick碱基配对和远距离非Watson-Crick碱基配对的相互作用。C为从28S rRNA冷冻电镜结构中产生真阳性和真阴性数据集。深色代表真阳性,浅色代表真阴性。D为ROC(受试者工作特征曲线)分析RIC-seq在28S rRNA 3D结构预测中的准确性。RIC-seq技术以深色线表示,而PARIS用浅色线表示。虚线为随机分布区线。冷冻电镜结构中缺失的部分没有被用来生成ROC曲线。

[0096] 图4为RNA在体内的拓扑结构域和折叠原则。A为在PDE3A、IMMP2L、FTX和PVT1RNA前体中观测到的拓扑结构域(虚线三角框)。将每个转录本分成100bins(每个bin代表1%的长度=1个像素)然后将总和标准化到1,生成的热图。B为前体mRNA和lncRNA线性距离上的实际接触机率函数。斜率为-1符合分型球体的理论模型。C为成熟mRNA和lncRNA线性距离上的实际接触机率函数。虚线表示的斜率为-1。

[0097] 图5为在不同细胞系中绘制的RNA 3D相互作用图谱。A为GM12878、IMR-90、H1hESC、hNPC和HT29细胞中所有RNA的作用矩阵。下面是chr4:93.2Mb-94.8Mb区域RNA相互作

用的放大。B为A图中5种细胞系共有的RNA-RNA相互作用的放大。C-D分别展示了细胞系特有的和共有的RNA-RNA相互作用。RNA-seq、ChIP-seq和TAD信号来自于H1 hESC ENCODE数据。E为Cas9-KRAB系统敲降LncPRESS2。F为敲降LncPRESS2后,定量GRID2和OCT4的表达水平。 $*P<0.05$, $**P<0.01$ 和 $***P<0.001$,双尾t检验($n=3$)。

[0098] 图6为原位RNA-RNA相互作用的特征。A为6种细胞系中染色体内部和染色体之间的RNA-RNA相互作用所占百分比。B为基因内部和基因之间的RNA-RNA相互作用所占百分比。C为6种细胞系中染色体内部嵌合体reads的百分比和所能跨越的距离。实线和虚线分别表示基因内部和基因之间的相互作用。D为Hi-C、RIC-seq和RNA-seq在染色体1上的区间分布,展示的数据为GM12878。E为来自不同区间的RNA-RNA相互作用分别在基因内部和基因间所占百分比。

[0099] 图7为细胞特异性hub-RNA的鉴定。A为HeLa细胞中每一个RNA根据其嵌合体read的密度和相互作用基因的数目进行分选,其中排名在top 5%的被认为是hub-RNA。GAPDH作为一个阴性对照。B为MALAT1、CCAT1和PDE3A与23条染色体上的RNA相互作用。箭头表示基因的位置。C为Meta分析Hub-RNA与其它RNA的RIC-seq信号强度与分布规律。转录起始位点及转录终止位点周围的RIC-seq信号被展示出来。D为Hub-RNA比其他RNAs更保守。E-H为Hub-RNA与其他RNAs上RNA聚合酶II、H3K4me3、H3K27ac和H3K27me3的ChIP-seq信号分布情况。I为大部分hub-RNA呈现细胞特异性。

[0100] 图8为Hub-RNA CCAT1-5L协同MYC的启动子与增强子RNA正调控MYC基因的表达。A为RIC-seq、RNA-seq和H3K27ac信号在8q24上的分布。5'和3' RACE获得的CCAT1转录本展示在下方。Northern blot探针用黑色的线标记。CCAT1、MYC和PVT1基因放大标示在图中。CCAT1-5L与MYC的嵌合体reads展示在图中。B为Northern blot分析CCAT1-5L在不同细胞系中的表达。5L探针检测显示CCAT1-5L只在HeLa细胞中表达。18S rRNA与28S rRNA作为上样对照。C为smFISH发现CCAT1-5L定位在细胞核中。CCAT1-5L,CCAT1-Exon2 (CCAT1-E),NEAT1 5'端探针用不同颜色标示。标尺:5 μ m。D为5L特异的LNA探针敲降CCAT1后,MYC的表达量显著性下调。5L和Exon2特异引物用来检测CCAT1表达水平。E为smFISH检测CCAT1-5L、MYC启动子和MYC增强子RNA的共定位情况。CCAT1,MYC,PVT1用不同颜色标示。标尺:5 μ m。F为敲降CCAT1-5L或者异位表达CCAT1-5L后对细胞增殖率的影响。G为敲降CCAT1-5L或者异位表达CCAT1-5L影响克隆形成。H为Transwell分析表明CCAT1-5L对细胞侵染与转移很重要。标尺:50 μ m。 $*P<0.05$ 和 $**P<0.01$,双尾t检验($n=3$)。

具体实施方式

[0101] 下述实施例中所使用的实验方法如无特殊说明,均为常规方法。

[0102] 下述实施例中所用的材料、试剂等,如无特殊说明,均可从商业途径得到。

[0103] 下述实施例中所使用的溶液配方如下:

[0104] PBS缓冲液(pH 7.4):溶剂为水,溶质及浓度如下:NaCl 137mmol/L,KCl 2.7mmol/L, Na_2HPO_4 10mmol/L, KH_2PO_4 2mmol/L。

[0105] 1 \times PNK溶液:溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶质及浓度如下:10mM MgCl_2 ,0.1mg/ml BSA,0.2% (体积百分含量)NP-40。

[0106] 1 \times PNK+EGTA 溶液:溶剂为50mM pH 7.4的Tris-Cl缓冲液,溶质及浓度如下:20mM

EGTA, 0.5% (体积百分含量) NP-40。

[0107] High-salt溶液: 溶剂为5×PBS (no Mg^{2+} , Ca^{2+}), 溶质及浓度如下: 0.5% (体积百分含量) NP-40。其中, 所述5×PBS (no Mg^{2+} , Ca^{2+}) 即为5×PBS缓冲液 (pH 7.4): NaCl 685mmol/L, KCl 13.5mmol/L, Na_2HPO_4 50mmol/L, KH_2PO_4 10mmol/L。

[0108] Permeabilization溶液: 10mM Tris-HCl (pH 7.5), 10mM NaCl, 0.5% (体积百分含量) NP-40, 0.3% (体积百分含量) Triton X-100, 0.1% (体积百分含量) Tween 20, 1× protease inhibitors (Sigma, 货号为: P8340-5ML, 具体组份为AEBSF, Aprotinin, Bestatin hydrochloride, E-64, Leupeptin hemisulfate salt and Pepstatin A) 和2U/ml SUPERase • In™ RNase Inhibitor (Thermo Fisher, 货号为: AM2694)。

[0109] 1×MN反应溶液: 溶剂为50mM pH 8.0的Tris-Cl缓冲液, 溶质及浓度如下: 5mM $CaCl_2$ 。

[0110] Proteinase K溶液: 溶剂为10mM pH 7.5的Tris-Cl缓冲液, 溶质及浓度如下: 10mM EDTA, 0.5% (体积百分含量) SDS。

[0111] 5×hybridization溶液: 1M NaCl, 500mM Tris-HCl (pH 7.4)。

[0112] 溶液A: 0.1M NaOH, 0.05M NaCl。

[0113] 溶液B: 0.1M NaCl。

[0114] 2×TWB溶液: 10mM Tris-HCl (pH 7.5), 1mM EDTA, 2M NaCl, 0.02% (体积百分含量) Tween 20。

[0115] PK溶液: 100mM NaCl, 10mM Tris-HCl (pH 7.0), 1mM EDTA, 0.5% (0.5g/100mL) SDS。

[0116] TE缓冲液: 10mM Tris-HCl (pH 8.0), 1mM EDTA。

[0117] 实施例1、RIC-seq文库的制备方法

[0118] 本发明的RIC-seq文库构建流程如图1中A所示。其中包括培养细胞, 经过甲醛交联、细胞膜和核膜穿孔、MNase酶处理、RNA 3' 末端羟基化处理、pCp-biotin连接、RNA 3' 末端羟基化, 5' 末端磷酸化处理、近端连接、总RNA提取、DNase I去除基因组DNA、核糖体RNA去除、RNA片段化、C1磁珠富集及洗脱富集的RNA、cDNA一链合成、合成DNA二链、末端修复、加“A”、连接接头、PCR扩增等步骤。具体步骤如下:

[0119] 1、取密度大约为80~90%的15cm盘细胞, 倒掉培养基, 加10ml预冷PBS (pH 7.4) 洗涤细胞, 倒掉PBS, 重复此步骤3次, 得到洗涤后的细胞。

[0120] 2、完成步骤1后, 向步骤1获得的洗涤后的细胞中加入10ml的1% (体积百分比) 的甲醛溶液 (溶剂为PBS溶液), 室温静置10分钟。然后再加入甘氨酸溶液 (终浓度为0.125mol/L, 溶剂为DEPC水) 终止反应, 室温放置10分钟, 得到甲醛交联并终止后的细胞。

[0121] 3、完成步骤2后, 向步骤2获得的甲醛交联并终止后的细胞中加10ml预冷PBS (pH 7.4), 洗涤3次, 用细胞刮铲将细胞刮下并转移到50ml离心管中, 4℃、2500rpm条件下离心10分钟, 弃上清液, 加入2ml预冷PBS (pH 7.4) 重悬细胞细胞沉淀, 将细胞悬液转移至2个1.5ml离心管中, 每管1ml, 4℃、2500rpm条件下离心10分钟, 弃上清液, 继续下一步实验或者将细胞沉淀放在-80℃冰箱中保存。

[0122] 4、完成步骤3后, 向步骤3获得细胞沉淀中加入1ml Permeabilization溶液, 冰浴15分钟, 每隔2分钟混匀一次。4℃、3500rpm条件下离心5分钟, 弃上清液, 加入600μl 1×PNK

溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤2次。

[0123] 5、完成步骤4后,向步骤4获得细胞沉淀中加入200μl按照1:10000体积比稀释的微生物核酸酶(Micrococcal nuclease)(Thermo Fisher,货号为EN0181)(其中MNase酶使用浓度为0.03U/μl),重悬细胞沉淀,稀释溶液为1×MN反应溶液,在ThermoMixer中37℃反应10分钟,设置反应程序为每隔2分钟1000rpm震荡15秒。反应结束后,4℃、3500rpm条件下离心5分钟,弃上清液,加入600μl 1×PNK+EGTA溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。加入600μl 1×PNK溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。

[0124] 6、完成步骤5后,向步骤5获得细胞沉淀中加入10μl 10×FastAP buffer(Thermo Fisher公司产品),10μl Fast Alkaline Phosphatase(Thermo Fisher公司产品,货号为EF0651;在反应体系中的终浓度为0.1U/μl),80μl DEPC水,重悬细胞沉淀,在ThermoMixer中37℃反应10分钟,设置反应程序为每隔3分钟1000rpm震荡15秒。反应结束后,4℃、3500rpm条件下离心5分钟,弃上清液,加入600μl 1×PNK+EGTA溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。加入600μl High-salt溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。加入600μl 1×PNK溶液(与前述配方相比,将NP-40的含量调整至0.05%体积百分含量,而其他成分及含量均不变)重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。

[0125] 7、完成步骤6后,向步骤6获得细胞沉淀中加入10μl 10×RNA ligase reaction buffer(Thermo Fisher公司产品),6μl RNase inhibitor,4μl Biotinylated Cytidine(Bis)phosphate(即pCp-biotin,Thermo Fisher公司产品,货号为20160)(1mM),10μl T4 RNA ligase(Thermo Fisher公司产品,货号为EL0021);在反应体系中的终浓度为1U/μl),20μl DEPC水,50μl 30%PEG,重悬细胞沉淀,在ThermoMixer中16℃反应过夜,设置反应程序为每隔3分钟1000rpm震荡15秒。反应结束后,4℃、3500rpm条件下离心5分钟,弃上清液,加入600μl 1×PNK溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤2次。

[0126] 8、完成步骤7后,向步骤7获得细胞沉淀中加入10μl 10×FastAP buffer(Thermo Fisher公司产品),10μl Fast Alkaline Phosphatase(Thermo Fisher公司产品,货号为EF0651);在反应体系中的终浓度为0.1U/μl),80μl DEPC水,重悬细胞沉淀,在ThermoMixer中37℃反应10分钟,设置反应程序为每隔3分钟1000rpm震荡15秒。反应结束后,4℃、3500rpm条件下离心5分钟,弃上清液,加入600μl 1×PNK+EGTA溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。加入600μl High-salt溶液重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤1次。加入600μl 1×PNK溶液(与前述配方相比,将NP-40的含量调整至0.05%体积百分含量,而其他成分及含量均不变)重悬细胞沉淀,4℃旋转(20rpm)混匀5分钟,4℃、3500rpm条件下离心5分钟,弃上清,重复此步骤2次。

[0127] 9、完成步骤8后,向步骤8获得细胞沉淀中加入10μl 10×PNK buffer(Thermo Fisher公司产品),15μl 10mM ATP,10μl T4PNK(Thermo Fisher公司产品,货号为EK0032);

在反应体系中的终浓度为 $1\text{U}/\mu\text{l}$), $65\mu\text{l}$ DEPC水, 重悬细胞沉淀, 在ThermoMixer中 37°C 反应45分钟, 设置反应程序为每隔3分钟 1000rpm 震荡15秒。反应结束后, 4°C 、 3500rpm 条件下离心5分钟, 弃上清液, 加入 $600\mu\text{l}$ $1\times\text{PNK}+\text{EGTA}$ 溶液重悬细胞沉淀, 4°C 旋转混匀5分钟, 4°C 、 3500rpm 条件下离心5分钟, 弃上清, 重复此步骤1次。加入 $600\mu\text{l}$ $1\times\text{PNK}$ 溶液(与前述配方相比, 将NP-40的含量调整至 0.05% 体积百分含量, 而其他成分及含量均不变)重悬细胞沉淀, 4°C 旋转(20rpm)混匀5分钟, 4°C 、 3500rpm 条件下离心5分钟, 弃上清, 重复此步骤1次。

[0128] 10、完成步骤9后, 向步骤9获得细胞沉淀中加入 $20\mu\text{l}$ $10\times\text{RNA ligase reaction buffer}$ (Thermo Fisher公司产品), $8\mu\text{l}$ RNase inhibitor, $10\mu\text{l}$ T4RNA ligase (Thermo Fisher公司产品, 货号为EL0021); 在反应体系中的终浓度为 $0.5\text{U}/\mu\text{l}$), $20\mu\text{l}$ BSA ($1\text{mg}/\text{ml}$), $142\mu\text{l}$ DEPC水, 重悬细胞沉淀, 在ThermoMixer中 16°C 反应过夜, 设置反应程序为每隔3分钟 1000rpm 震荡15秒。反应结束后, 4°C 、 3500rpm 条件下离心5分钟, 弃上清液, 加入 $600\mu\text{l}$ $1\times\text{PNK}$ 溶液重悬细胞沉淀, 4°C 旋转(20rpm)混匀5分钟, 4°C 、 3500rpm 条件下离心5分钟, 弃上清, 重复此步骤2次。

[0129] 11、完成步骤10后, 向步骤10获得细胞沉淀中加入 $200\mu\text{l}$ Proteinase K溶液和 $50\mu\text{l}$ proteinase K (Takara公司产品, 货号为9034; 在反应体系中的终浓度为 $0.12\text{U}/\mu\text{l}$), 混匀, 在ThermoMixer中 37°C 反应60分钟, 56°C 反应15分钟, 反应结束后, 待样品冷却至室温, 加入 $750\mu\text{l}$ Trizol LS (Thermo Fisher公司产品, 货号为10296028), 混匀, 室温放置5分钟, 加入 $220\mu\text{l}$ 氯仿, 剧烈混匀15秒, 室温放置3分钟。 4°C 、 13000rpm 条件下离心15分钟, 转移上清至 1.5ml 离心管中, 加入 $500\mu\text{l}$ 异丙醇, $1\mu\text{l}$ glycoblu (浓度为 $15\mu\text{g}/\mu\text{l}$), 混匀, 将离心管放在 -20°C 冰箱, 过夜沉淀。

[0130] 12、完成步骤11后, 将步骤11获得的样品在 4°C 、 13000rpm 条件下离心20分钟, 弃上清液, 加入 $500\mu\text{l}$ 75% 乙醇, 洗涤沉淀, 4°C 、 13000rpm 条件下离心5分钟, 重复此步骤1次, 自然晾干沉淀, 加入 $20\mu\text{l}$ DEPC水溶解沉淀, 取 $1\mu\text{l}$ 样品用NanoDrop定量。

[0131] 13、完成步骤12后, 从步骤12获得的样品中取出 $20\mu\text{g}$ 总RNA, 加入 $10\mu\text{l}$ $10\times\text{RQ1DNase I buffer}$ (Promega公司产品), $3\mu\text{l}$ RNasin (Thermo Fisher公司产品, 货号为E00381)和 $5\mu\text{l}$ DNase I (Promega公司产品, 货号为M6101), 补水至总体积为 $100\mu\text{l}$, 在ThermoMixer中 37°C 反应20分钟, 反应结束后, 加入 $100\mu\text{l}$ DEPC水, 然后加入 $200\mu\text{l}$ 酸性酚氯仿, 混匀, 室温放置3分钟, 4°C 、 13000rpm 条件下离心15分钟, 取上清至 1.5ml 离心管中, 加入 $20\mu\text{l}$ 3M 乙酸钠 ($\text{pH } 5.5$), $1\mu\text{l}$ glycoblu和 $500\mu\text{l}$ 100% 乙醇, 混匀, 将离心管放在 -20°C 冰箱, 过夜沉淀。

[0132] 14、完成步骤13后, 将步骤13获得的样品在 4°C 、 13000rpm 条件下离心20分钟, 弃上清液, 加入 $500\mu\text{l}$ 75% 乙醇, 洗涤沉淀, 4°C 、 13000rpm 条件下离心5分钟, 重复此步骤1次, 自然晾干沉淀, 加入 $6\mu\text{l}$ DEPC水溶解沉淀, 将样品转移至PCR管中。

[0133] 15、完成步骤14后, 向步骤14获得的样品中加入 $10\mu\text{l}$ rRNA probe mix (探针序列的设计与合成参考已发表文献 (Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysocker, A., Fennell, T., et al. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature methods 10, 623-629.)) ($2\mu\text{g}/\mu\text{l}$), $4\mu\text{l}$ $5\times\text{Hybridization buffer}$, 混匀, 将PCR管放入PCR仪, 设置反应程序: 95°C 2分钟, $-0.1^\circ\text{C}/\text{s}$ 降

至22℃,22℃5分钟,反应完立即将样品放在冰上。

[0134] 16、完成步骤15后,向步骤15获得的样品中加入3μl 10×RNase H buffer (Thermo Fisher公司产品),5μl RNase H(Thermo Fisher公司产品,货号为EN0202) (25U)和2μl DEPC水,混匀,将样品放入PCR仪,设置程序:37℃30分钟。反应完立即将样品放在冰上。

[0135] 17、完成步骤16后,向步骤16获得的样品中加入4μl 10×TURBO buffer (Thermo Fisher公司产品),5μl TURBO DNase (Thermo Fisher公司产品,货号为AM2238;在反应体系中的终浓度为0.25U/μl)和1μl DEPC water,混匀,将样品放入PCR仪,设置程序:37℃30分钟。反应完立即将样品放在冰上。

[0136] 18、完成步骤17后,将步骤17获得的样品转移至1.5ml离心管中,加入160μl DEPC水,加入200μl酸性酚氯仿,混匀,室温放置3分钟,4℃、13000rpm条件下离心15分钟,取上清至1.5ml离心管中,加入20μl 3M乙酸钠(pH 5.5),1μl glycoblu和500μl 100%乙醇,混匀,将离心管放在-20℃冰箱,过夜沉淀。

[0137] 19、完成步骤18后,将步骤18获得的样品在4℃、13000rpm条件下离心20分钟,弃上清液,加入500μl 75%乙醇,洗涤沉淀,4℃、13000rpm条件下离心5分钟,重复此步骤1次,自然晾干沉淀,加入16μl DEPC水溶解沉淀,将样品转移至PCR管中,加入4μl 5×First-strand buffer (Thermo Fisher公司产品,货号为18064-014),混匀,放入PCR仪,94℃反应5分钟,反应完立即放冰上。

[0138] 20、取1.5ml离心管,加入20μl C1磁珠,把离心管放在磁力架上,待溶液澄清,吸出上清,加20μl溶液A,重悬磁珠,室温放置2分钟,把离心管放在磁力架上,待溶液澄清,吸出上清,重复此步骤1次,加入20μl溶液B,重悬磁珠,把离心管放在磁力架上,待溶液澄清,吸出上清,加入32μl酵母RNA (Roche公司产品,货号为10109223001) (50μg),68μl DEPC水和100μl 2×TWB溶液,重悬磁珠,将离心管放在旋转混匀仪上,旋转混匀1小时,然后将离心管放在磁力架上,待溶液澄清,吸出上清,加入500μl 1×TWB溶液,重悬磁珠,将离心管放在磁力架上,待溶液澄清,吸出上清,重复此步骤2次。

[0139] 21、取步骤19的样品,加入30μl DEPC水,共50μl样品加入封闭好的磁珠中,混匀,室温旋转混匀30分钟,将离心管放在磁力架上,待溶液澄清,吸出上清,加入500μl 1×TWB溶液,洗涤4次。

[0140] 22、完成步骤21后,向步骤21获得的洗涤后的磁珠中加入100μl PK溶液,混匀,在ThermoMixer中95℃1000rpm震荡反应10分钟,将离心管放在磁力架上,待溶液澄清,吸出上清,加入新的1.5ml离心管中,在原管中加入100μl PK溶液,混匀,在ThermoMixer中95℃1000rpm震荡反应10分钟,将离心管放在磁力架上,待溶液澄清,吸出上清,加入1.5ml离心管中,在原管中加入100μl PK溶液,混匀,将离心管放在磁力架上,待溶液澄清,吸出上清,加入1.5ml离心管中,共300μl洗脱液,加入300μl酸性酚氯仿,混匀,室温放置3分钟,4℃、13000rpm条件下离心15分钟,吸出上清加入新的1.5ml离心管中,加入18μl 5M NaCl,混匀,加入1μl glycoblu和900μl 100%乙醇,混匀,将离心管放在-20℃冰箱,过夜沉淀。

[0141] 23、完成步骤22后,将步骤22获得的样品在4℃、13000rpm条件下离心20分钟,弃上清液,加入500μl 75%乙醇,洗涤沉淀,4℃、13000rpm条件下离心5分钟,重复此步骤1次,自然晾干沉淀,加入10μl DEPC水溶解沉淀,将样品转移至PCR管中,加入0.5μl N6引物(序列为NNNNNN,其中,N表示A或T或C或G) (0.1μg/μl),混匀,将PCR管放入PCR仪中,65℃反应5分

钟,反应完立即放冰上。

[0142] 24、完成步骤23后,向步骤23获得的样品中加入3 μ l 5 \times First-strand buffer (Thermo Fisher公司产品,货号为18064-014),1 μ l dNTP mix (10mM),0.5 μ l 100mM DTT,0.5 μ l RNase Inhibitor (40U/ μ l),0.5 μ l Superscript II (Thermo Fisher公司产品,货号为18064-014) (200U/ μ l),混匀,将PCR管放入PCR仪中,设置程序:25 $^{\circ}$ C 10分钟,42 $^{\circ}$ C 40分钟,70 $^{\circ}$ C 15分钟。反应完成后将样品放冰上。

[0143] 25、完成步骤24后,将步骤24获得的样品转移至新1.5ml离心管中,加入10 μ l 5 \times Second-strand buffer (Thermo Fisher公司产品,货号为10812-014),0.8 μ l dNTP (dUTP) (25mM) (即25mM dNTPs与dUTP的混合物,其中dTTP与dUTP摩尔比为4:1),0.2 μ l RNaseH (Thermo Fisher公司产品,货号为EN0202) (5U/ μ l),2.5 μ l DNA pol I (Enzymatics公司产品,货号为P705-500) (10U/ μ l),将离心管放入ThermoMixer中,设置反应程序:16 $^{\circ}$ C 2小时,300rpm,15s/2min。

[0144] 26、完成步骤25后,将AMPure XP磁珠 (简称XP磁珠,Beckman) 提前室温混匀平衡30分钟。然后将90 μ l (1.8 \times) XP磁珠加入步骤25获得的反应液中,轻柔混匀。室温静置5分钟,转移至磁力架静置5分钟,去除上清液,并使用200 μ l新鲜的80%乙醇溶液漂洗磁珠2次。磁珠置于磁力架上晾干2分钟,加入43 μ l TE缓冲液重悬磁珠,吹打50次。室温静置5分钟,然后放入磁力架静置5分钟,吸取42 μ l上清液,加入1.5ml离心管中。

[0145] 27、完成步骤26后,向步骤26获得的样品中加入5 μ l 10 \times PNK溶液 (T4PNK配套反应溶液),0.4 μ l dNTPs (25mM),1.2 μ l T4DNA polymerase (Enzymatics公司产品,货号为P7080L) (3U/ μ l),0.2 μ l Klenowfragment (Enzymatics公司产品,货号为P7060L) (5U/ μ l),1.2 μ l T4PNK (Enzymatics公司产品,货号为Y9040L) (10U/ μ l),混匀,将离心管放入ThermoMixer中,20 $^{\circ}$ C条件下反应30分钟。反应结束后,加入90 μ l AMPure XP磁珠进行纯化,具体步骤同步骤26,最后用20.5 μ l TE缓冲液洗脱,吸取19.7 μ l上清,加入新的1.5ml离心管中。

[0146] 28、完成步骤27后,向步骤27获得的样品中加入2.3 μ l 10 \times blue buffer (Enzymatics公司产品,货号为B0110L),0.5 μ l dATP (5mM)和0.5 μ l Klenow exo- (3' to 5' exo minus) (Enzymatics公司产品,货号为P7010-LC-L) (5U/ μ l),混匀,将离心管放入ThermoMixer中,37 $^{\circ}$ C条件下反应30分钟。

[0147] 29、完成步骤28后,向步骤28获得的样品中加入1.4 μ l 2 \times Rapid ligation buffer (Enzymatics公司产品,货号为B1010L),0.1 μ l 10mM ATP,1 μ l Adapter (PEI Adapter oligo A:GATCGGAAGAGCACACGTCT,PEI Adapter oligo B:ACACTCTTCCCTAC ACGACGCTCTCCGATCT,反应中的adapter为两条oligo退火形成) (2 μ M)和1 μ l T4quick DNAligase (Enzymatics公司产品,货号为L6030-HC-L) (600U/ μ l),混匀,将离心管放入ThermoMixer中,20 $^{\circ}$ C条件下反应15分钟。反应结束后,加入47.7 μ l AMPure XP磁珠进行纯化,具体步骤同步骤26,最后用26 μ l TE缓冲液洗脱,吸取25 μ l上清,加入新的1.5ml离心管中。加入45 μ l AMPure XP磁珠进行二次纯化,具体步骤同步骤26,最后用16.5 μ l TE缓冲液洗脱,吸取15.7 μ l上清,加入PCR管中。30、完成步骤29后,以步骤29获得的上清液为模板于PCR管中进行PCR反应,得到PCR反应液 (25 μ l)。

[0148] PCR反应体系为25 μ l:上清液15.7 μ l,10 \times Pfx缓冲液 (Invitrogen) 2.5 μ l,10 μ M的

上下游引物各1 μ l, 50mM的MgSO₄溶液1 μ l, 25mM的dNTP 0.4 μ l, Pfx酶 (Invitrogen) 0.4 μ l, USER enzyme (NEB) 3 μ l。

[0149] PCR反应程序为:37 $^{\circ}$ C反应15分钟;94 $^{\circ}$ C反应2分钟;94 $^{\circ}$ C变性15秒,62 $^{\circ}$ C退火30秒,72 $^{\circ}$ C延伸30秒,反应12个循环;72 $^{\circ}$ C反应10分钟。

[0150] 31、完成步骤30后,将步骤30获得的PCR反应液用2%琼脂糖胶进行电泳,对200-450bp范围产物使用Qiagen MinElute试剂盒进行切胶回收,操作步骤参照试剂盒说明书,最后用16 μ l TE缓冲液进行洗脱,得到PCR洗脱液。

[0151] 32、完成步骤31后,吸取1 μ l步骤31获得的PCR洗脱液使用Qubit 3.0定量。定量合格的样品用于测序分析。

[0152] 实施例2、RIC-seq文库的制备方法的应用

[0153] 一、培养HeLa细胞及果蝇S2细胞样本

[0154] 以实验室培养的HeLa细胞为样本,细胞样本起始量为 1×10^7 个细胞,果蝇S2细胞作为spike-in,评估近端连接特异性。

[0155] 二、RIC-seq文库的制备

[0156] 基于步骤一中的细胞样本按照实施例1中的方法构建RIC-seq文库。其中,步骤30中的上下游引物分别如下(NNNNNNN为文库Index序列)

[0157] Primer1.0

[0158] 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' (SEQ ID No.1);

[0159] Index primer

[0160] 5'-CAAGCAGAAGACGGCATAACGAGATANNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3' (SEQ ID No.2)。

[0161] 其中,N表示A或T或C或G。

[0162] 三、测序

[0163] 使用Illumina HiSeq X Ten测序仪对步骤二构建得到的RIC-seq文库进行PE150双端测序。

[0164] 四、数据分析及结果

[0165] 1、数据分析方法

[0166] 数据分析流程如图2中A所示。首先利用Trimmomatic (0.36) 软件过滤RIC-seq原始测序数据中的接头序列和低质量的测序片段,进一步去除冗余片段后利用Cutadapt (v1.15) 剪除低复杂度序列如polyA等,随后利用STAR (2.5.2b) 将高质量数据比对到人类参考基因组 (hg19版本),最后从比对结果中筛选来自于RNA连接产物的测序片段(定义为嵌合体reads)。通过比较每个基因的嵌合体reads个数,计算皮尔森相关系数 (Pearson correlation coefficient) 以评估实验的可重复性。利用IGVtools以及Juicebox对RIC-seq数据进行可视化展示。

[0167] 2、数据分析结果

[0168] 为了捕获由蛋白质介导的RNA近端连接,我们发明了RIC-seq方法 (RNA In situ Conformation Sequencing)。具体流程如图1中A所示。首先,用甲醛处理细胞以固定蛋白与RNA、蛋白与DNA以及蛋白与蛋白之间的相互作用,这样空间位置接近的不同RNA片段便被固

定下来。其次,利用多组合洗涤剂对固定的细胞进行细胞膜和核膜打孔,并进行微球菌核酸酶处理以去除不受蛋白质保护的游离RNA。微球菌核酸酶消解后RNA的3'末端是磷酸基,而5'末端是羟基(图1中A)。为了进行pCp-biotin标记,我们采用碱性磷酸酶先将3'末端磷酸基变成羟基,然后用T4RNA连接酶将pCp-biotin连接到RNA的3'末端。紧接着,对样本分别进行碱性磷酸酶和T4PNK酶处理,将Cp-biotin的3'末端变成羟基,而RNA的5'末端则变成磷酸基(图1中A)。之后,在原位和非变性的条件下,利用T4RNA连接酶将空间上相互靠近的RNA连接起来。然后,利用蛋白酶K消解结合TRIzol抽提以获得总RNA,在去除基因组DNA和rRNA后,进行碱裂解片段化处理。最后,用链霉亲和素磁珠富集含有C-biotin的嵌合体RNA,并按照常规流程构建链特异性文库(Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* 7, 709-715.), 上机测序并分析。

[0169] 我们在HeLa细胞中构建了两个RIC-seq文库,共得到了155M可比对的reads。为了便于数据分析和可视化,我们整合多种算法和软件建立了一套完整的分析流程(图2中A)。分析发现嵌合体reads占有所有测序片段的比例大约是9%,而且超过90%的嵌合体reads在连接处含有一个多余的“C”,说明了pCp-biotin的高效率和链霉素磁珠富集的高特异性(图1中B)。每一个RIC-seq嵌合体read代表两个不同的RNA片段间的近端相互作用(图1中C-G)。多个RIC-seq嵌合体reads可以揭示共同的结构或者特异的反式RNA相互作用。RIC-seq技术的重复性非常好,两次生物学重复之间的Pearson相关系数为0.963(图2中B)。为了确定假阳性率,我们采用了细胞混合的策略(Li, X., Zhou, B., Chen, L., Gou, L. T., Li, H., and Fu, X. D. (2017). GRID-seq reveals the global RNA-chromatin interactome. *Nature biotechnology* 35, 940-950.),即将HeLa细胞和果蝇S2细胞按1:5比例混合在一起,然后构建RIC-seq文库并测序。结果显示只有约0.6%的嵌合体reads是一端来自果蝇的RNA,而另一端是来自人HeLa细胞的RNA(图2中C),这表明RIC-seq技术的假阳性率在1%以下。

[0170] 接下来我们通过与已知的RNA结构和相互作用比较,系统检测了RIC-seq方法的分辨率、敏感度和特异性,这些RNA包括microRNA、snRNA、snoRNA和lncRNA(图1中C-I,图2中D-F)。RIC-seq可以精确的在单碱基分辨率水平捕获经典的miRNA前体茎环结构(图1中C, pCp插入位置在下方),而且这些miRNA的表达水平(RPM, reads per million)从0.05到31,067不等(图1中D),说明了RIC-seq技术的动态检测范围。出乎意料的是, pCp-biotin标记的位置主要富集在miRNA前体的顶端环上(图1中C,图1中D),表明顶端环可能很少被蛋白保护。更进一步的, RIC-seq成功检测到了已知的snRNA、snoRNA、RPPH1 (the RNA component of Ribonuclease P) 和TERC (telomerase RNA, 数据没有展示)的分子间和分子内相互作用(图1中E-I和图2中D-F)。与PARIS和RAP方法相比(Engreitz, J. M., Sirokman, K., McDonel, P., Shishkin, A. A., Surka, C., Russell, P., Grossman, S. R., Chow, A. Y., Guttman, M., Lander, E. S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* 159 (1):188-199.); (Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., et al. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* 165, 1267-1279.), RIC-seq不仅可以捕获

已知的U1-MALAT1相互作用,还可以鉴定一些HeLa细胞中特异的U1和MALAT1相互作用位点(图1中H)。不出所料,这些相互作用位点是保守的、包含U1基序、有嵌合体reads支持(箭头标示)(图1中I),提示了一定的功能性。

[0171] 在充分验证了RIC-seq方法和数据后,我们合并了两次生物学重复的数据,用Juicebox绘制了全基因组的相互作用矩阵(Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* 3, 99-101.),在矩阵中成对的相互作用通过二维的热图进行可视化(IGV/Juicebox),强度表明嵌合体RNA连接的频率(图1中J和图2中A)。与-pCp的对照(检测到极少的嵌合体reads)相比,+pCp文库涵盖了复杂的分子内(~7M)和分子间相互作用(~6M)(图2中G),表明RNA在体内不仅是高度结构化的,而且具有广泛的相互纠缠(图1中J)。有意思的是,一些lncRNA在所有染色体上都具有广泛的结合,比如NEAT1和MALAT1。为了鉴定这些高丰度RNA的真正结合位点,我们对嵌合体reads进行了成簇分析,在HeLa细胞中共鉴定出了0.74M高可信度的RNA-RNA相互作用位点。在这些位点中,MALAT1和NEAT1不仅彼此可以相互作用而且还有数千个其他靶标(图1中J)。与近期的报道一致(West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Molecular cell* 55, 791-802.),我们也发现MALAT1和NEAT1更加倾向于结合转录活跃的基因(图2中H, $p < 2.2e-16$),而且它们的结合基序高度相似(图2中I)。

[0172] RIC-seq揭示MALAT1可结合在NEAT1的5'端,(NEAT1_5',图1中J右)。为了验证这些相互作用,我们采用单分子原位杂交(smFISH)和超高分辨成像(SIM)检测了这两个lncRNA的相互作用。我们发现NEAT1_5'可形成环状结构,而MALAT1则为点状分布(每个HeLa细胞核中平均有248个点)(图1中K和图2中J)。一些NEAT1与MALAT1的荧光信号直接重叠(图1中K box 2-4)。我们发现一个HeLa细胞平均有7.5个核旁斑(paraspeckles),其中有~63.7%与MALAT1共定位(图2中K)。总结起来,以上这些数据表明RIC-seq是一种高特异性、高可重复性和高精度的鉴定原位RNA-RNA相互作用的新方法。

[0173] 为了检查RIC-seq能否捕获RNA的高级结构,我们将RIC-seq检测到的RNA临近位置信息与人80S核糖体冷冻电镜结构得到的数据(Anger, A.M., Armache, J.P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N., and Beckmann, R. (2013). Structures of the human and Drosophila 80S ribosome. *Nature* 497, 80-85.)进行了比较。首先我们基于每对5-nt窗口的相对空间距离绘制了28S rRNA的物理相互作用图谱(图3中A)。我们同样基于RIC-seq数据绘制了28S rRNA的3D图谱(图3中B)。这两个图谱无论在全局还是细微尺度上都是高度相似的(图3中A,图3中B)。出乎意料的是,RIC-seq不仅能捕获WC碱基配对(图3中B, box 1和box 2),还可以检测到远距离的loop-loop相互作用,如图3中B所示28S rRNA的50-200nt与4300-4400nt之间的相互作用(图3中B, box 3)。我们发现28S rRNA的结构中~70%的non-WC碱基配对的相互作用可以被RIC-seq检测到(图3中A,图3中B)。这些数据表明RIC-seq可以如实的捕获RNA的3D结构信息。

[0174] 为了定量RIC-seq在检测RNA高级结构中的表现,我们根据28S rRNA的冷冻电镜结构数据生成了两组数据:真阳性集合(5-nt窗口对应的区域之间的3D距离小于 25 Å)和真阴

性集合(距离大于25 Å)(4,847 vs 369,698)(图3中C)。并利用这两个数据集来评估RIC-seq敏感度(成功检测到真阳性)和特异性(成功排除掉真阴性)。我们将RIC-seq检测到的28S rRNA的临近相互作用与真阳性数据集和真阴性数据集进行了比较,并生成了ROC曲线。ROC分析得到的AUC值是0.89,表明RIC-seq在RNA高级结构鉴定中具有很高的准确度(图3中D,深色线)。作为对照,我们利用同样的数据集评估了PARIS的表现,令人遗憾的是,由于大量28S rRNA中配对区域和远距离相互作用的位点不能被PARIS捕获,导致不能得到完整的曲线,也无法生成AUC值(图3中D,浅色线)。

[0175] RIC-seq技术产生的高质量RNA-RNA分子内相互作用数据使得我们可以在体内检测RNA的折叠规律。为了这个目的,我们重点关注了5179个mRNA前体,这些mRNA前体含有至少100个RNA分子内连接事件。有意思的是,我们发现在mRNA的内含子和外显子区域存在很多独立的拓扑结构域(图4中A),它们的共同特点是在某一区间内存在异常复杂的RNA-RNA相互作用,如在PDE3A和IMMP2L前体RNA中。为了系统鉴定类似的拓扑结构域,我们发明了一种迭代算法,可通过最大化结构域内部与结构域之间RIC-seq密度的比值来鉴定拓扑结构域的边界。与mRNA前体相比,类似的拓扑结构域在初始转录的lncRNA如FTX和PVT1中也很明显(图4中A)。这些数据表明RNA在体内和特定区间是高度结构化的,而且RNA共转录加工过程可能发生在独立的拓扑结构域中。

[0176] 拓扑结构域的发现表明大的RNA分子可能边转录边形成复杂的局部结构,然后按照级联折叠的规律形成特定的高级结构,但是RNA在体内的具体折叠规则目前还不清楚。与DNA聚合物类似,RNA聚合物也可以以无规卷曲、平衡球体或分形球体的形式存在。RNA具体以那种构象存在可以通过计算不同核苷酸距离的RNA片段间的连接概率来推导(Fudenberg,G.,and Mirny,L.A.(2012).Higher-order chromatin structure:bridging physics and biology.Current opinion in genetics&development 22,115-124.)。

[0177] 利用RIC-seq数据和类似的模拟方法(Lieberman-Aiden,E.,van Berkum,N.L.,Williams,L.,Imakaev,M.,Ragoczy,T.,Telling,A.,Amit,I.,Lajoie,B.R.,Sabo,P.J.,Dorschner,M.O.,et al.(2009).Comprehensive mapping of long-range interactions reveals folding principles of the human genome.Science 326,289-293.),我们检查了同一个RNA分子内部任意两个片段间的接触概率(或连接频率)与其线性距离之间的相关性。在聚合物的物理学特征中,如果聚合物以无规卷曲状态存在,则两个基因座间的接触概率随着线性距离的增加而迅速衰减,曲线的斜率预计会是-3/2(Fudenberg,G.,and Mirny,L.A.(2012).Higher-order chromatin structure:bridging physics and biology.Current opinion in genetics&development 22,115-124.)。相反,如果聚合物以平衡球体形式存在,接触概率将首先以与无规卷曲类似的速率下降,但随后达到平衡,最后连接频率变得与线性距离无关(Fudenberg,G.,and Mirny,L.A.(2012).Higher-order chromatin structure:bridging physics and biology.Current opinion in genetics&development 22,115-124.)。然而,无规卷曲和平衡球体模型看起来都不符合RIC-seq得到的实际规律。因为无论内含子是否被统计,RIC-seq数据都显示不同RNA片段间的接触概率随着距离的增加而逐步下降,曲线的斜率接近-1(图4中B和图4中C),该值与分形球体模型能够非常好的吻合(Lieberman-Aiden,E.,van Berkum,N.L.,Williams,L.,Imakaev,M.,Ragoczy,T.,Telling,A.,Amit,I.,Lajoie,B.R.,Sabo,P.J.,Dorschner,M.O.,et al.

(2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.)。因此,基于以上数据我们认为RNA前体可能类似于基因组DNA在体内以分形球体构象进行折叠,该构象在保证RNA缺乏结节的同时,可以保持最大程度的包装,同时保持了RNA随时打开和重折叠其局部结构的能力。

[0178] 我们接下来使用仅来自外显子和非翻译区的嵌合体reads检查了成熟mRNA的构象。通过数据模拟,我们发现成熟mRNA的折叠呈现出幂律依赖性,其曲线斜率也接近-1(图4中C),提示成熟mRNA也被压缩成分形球体状态。类似的,内含子缺失的lncRNA,例如NEAT1和MALAT1,与mRNA前体和含内含子的lncRNA折叠类似(图4中C)。总之,这些结果表明mRNA和ncRNA可能都遵循分形球体的折叠途径以形成复杂的3D结构。

[0179] 高度结构化的RNA进而与其他RNA分子通过分子间相互作用而发挥调节功能。为了探索新的分子间相互作用特征,我们在多种细胞系中构建RNA 3D图谱,包括人神经前体细胞(hNPC)和结肠腺癌细胞系HT29。此外,还包括三种ENCODE常用细胞类型:人B淋巴细胞系GM12878、H1人胚胎干细胞(hESCs)和人胎肺成纤维细胞IMR-90,以便于后期整合ENCODE项目产生的基因组数据。我们在这5种细胞系中也构建了RIC-seq文库,去除PCR重复后一共得到了1,001M唯一比对的read,其中,嵌合体占比8.4%。正如所料,这五种新的细胞类型中的RNA-RNA相互作用也异常复杂(图5中A)。

[0180] 利用这些高质量的数据,我们在六种不同的细胞类型中共鉴定了约3M的细胞特异性相互作用簇(fragment cutoff=2)和大量组成型相互作用位点(图5中B)。如图5中A所示LncPRESS2,一个受P53调控的胚胎干细胞特异性lncRNA(Jain,A.K.,Xi,Y.,McCarthy,R.,Allton,K.,Akdemir,K.C.,Patel,L.R.,Aronow,B.,Lin,C.,Li,W.,Yang,L.,et al.(2016).LncPRESS1 Is a p53-Regulated LncRNA that Safeguards Pluripotency by Disrupting SIRT6-Mediated De-acetylation of Histone H3K56.*Molecular cell* 64, 967–981.),RIC-seq在H1 hESC中检测到LncPRESS2与其邻近基因GRID2具有广泛的相互作用(图5中C)。相反,在ChrX:73.1Mb-73.6Mb基因座中,我们观察到了lncRNA FTX和JPX之间的组成型结合(图5中B,图5中D)。已知这两个lncRNA都在XIST表达中起正调控作用(Carmona,S.,Lin,B.,Chou,T.,Arroyo,K.,and Sun,S.(2018).LncRNA Jpx induces Xist expression in mice using both trans and cis mechanisms.*PLoS genetics* 14, e1007378.;Chureau,C.,Chantalat,S.,Romito,A.,Galvani,A.,Duret,L.,Avner,P.,and Rougeulle,C.(2011).Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region.*Human molecular genetics* 20,705–718.;Sun,S.,Del Rosario,B.C.,Szanto,A.,Ogawa,Y.,Jeon,Y.,and Lee,J.T.(2013).Jpx RNA activates Xist by evicting CTCF.*Cell* 153,1537–1551.;Tian,D.,Sun,S.,and Lee,J.T.(2010).The long noncoding RNA,Jpx,is a molecular switch for X chromosome inactivation.*Cell* 143,390–403.) ,并且对XIST介导的X染色体沉默至关重要。这两个lncRNA之间的紧密结合表明它们有可能以复合物的形式调节XIST。以上这些细胞类型特异性和组成型相互作用进一步凸显了RIC-seq方法的特异性,并表明RIC-seq可用于鉴定活细胞内的lncRNA靶标。

[0181] 为了进一步验证LncPRESS2-GRID2相互作用的功能,我们采用了Cas9-KRAB介导的

lncRNA沉默策略 (Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* 154, 442-451.)，其中，sgRNA可特异性的将Cas9-KRAB直接靶向lncRNA的启动子区域，而KRAB作为RNA聚合酶II的转录抑制因子(图5中E)，可高效阻断lncRNA的转录。我们发现使用三种不同的sgRNA特异性敲降LncPRESS2后，GRID2的表达水平显著性降低(图5中F)，表明尽管线性距离大于25kb，LncPRESS2也能正调控GRID2的表达。意外的是，在LncPRESS2敲低细胞中，干细胞多能性关键因子OCT4的表达显著性降低(图5中F)，意味着LncPRESS2介导的GRID2调控可能与干性密切相关。以上数据表明RIC-seq技术确实可以鉴定功能性的lncRNA靶标。

[0182] 为了揭示不同细胞类型中RNA-RNA相互作用的普遍特征，我们首先分别计算了染色体内和染色体间相互作用的频率。利用在上述六种细胞类型中产生的RIC-seq数据，我们发现~70%的RNA-RNA相互作用是发生在同一染色体内部，而剩余的~30%发生在染色体间(图6中A)。由于RNA既可以顺式作用也可以反式作用于其他RNA分子，我们还分别计算了基因内和基因间RNA相互作用的频率。类似地，约60%的嵌合体read是来源于基因内部的顺式作用，这部分数据可以用来推断RNA的3D结构；剩余40%则表现出反式RNA-RNA相互作用的特性(图6中B)，这说明细胞内大量的RNAs可以跨越很长的距离与相同染色体上的其他RNA或不同染色体上的RNA相互作用。如果我们仅统计同一染色体上的嵌合体reads，这一趋势也很明显，其中我们检测到两个明显的峰：第一个峰对应于基因内部的RNA-RNA相互作用，可跨越数百个核苷酸的距离；而第二个峰对应于基因间的相互作用，其跨越距离超过1Mb(图6中C)。

[0183] 染色质在体内级联折叠并根据转录活性的差异被分成区室A和B(Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289-293.)。与染色质的组织类似，RNA相互作用也似乎是区室化的并且可以在很大程度上重现DNA的区室(图6中D)，表明由于空间接近，同一区室中的RNA可能更倾向于彼此相互作用。我们接下来量化了不同区室之间的相互作用。有趣的是，在基因内部的相互作用中，嵌合体read主要在同一区室中富集，A至A可占嵌合体read总量的~90%(图6中E)，这可能是由于区室A的活跃转录和基因内部的空间距离相对更接近导致的。相反，对于基因间相互作用，区室A至A的相互作用降低至约65%，但区室A至B的相互作用增加至~30%(图6中E)，表明此类反式RNA相互作用可能具有一些未知功能并可能调控区室B内基因的活性。

[0184] 由于反式RNA-RNA相互作用可以跨越超过1Mb甚至跨越不同的染色体，我们接下来基于两个标准对RNA-RNA相互作用进行分类：靶基因的数量和嵌合体reads的密度(嵌合体reads的量除以RNA的表达水平)。有意思的是，该分析在HeLa细胞中意外的揭示了~500个高度丰富的RNA-RNA相互作用中枢(图7中A)，包括众所周知的lncRNA，例如MALAT1、NEAT1、CCAT1和PVT1。出乎意料的是，许多蛋白质编码基因也显示出复杂的RNA-RNA相互作用，例如PDE3A、GPC5和TRIO(图7中A)。MALAT1、CCAT1和PDE3A的相互作用模式和基因组位置被可视

化为Circos图(图7中B)。由于从这些基因座转录的RNA似乎在全基因组范围内起到组织RNA-RNA相互作用的作用,我们将这些RNA命名为hub-RNA,其中包括hub-mRNAs和hub-lncRNAs。

[0185] 为了表征hub-RNA的特征,我们将HeLa细胞中表达的所有RNA分成hub-RNA和其他RNA两组。仅基于RIC-seq信号,我们发现hub-RNA具有更强的反式RNA-RNA相互作用并且作用位点显著的富集在基因体(gene body)(图7中C)。此外,这些hub-RNAs在进化上比其他RNA更保守(图7中D),并且转录也很活跃,在这些基因的TSS(转录起始位点)区域RNA聚合酶II也存在富集(图7中E)。对应的,hub-RNAs对应的基因座活性组蛋白标志物H3K4me3(图7中F)和H3K27ac(图7中G)结合信号稍高。相反,抑制性组蛋白标记H3K27me3信号略低(图7中H)。同时我们发现hub-RNAs是具有细胞特异性的(图7中I)。因此,RIC-seq意外地揭示了一组可能在基因调控中发挥重要作用的组织特异性hub-RNAs。

[0186] 为了研究hub-RNAs的作用,我们选择CCAT1进行进一步分析,因为它具有广泛的反式RNA相互作用(图7中B)和潜在的超级增强子活性(Hnisz,D.,Abraham,B.J.,Lee,T.I.,Lau,A.,Saint-Andre,V.,Sigova,A.A.,Hoke,H.A.,and Young,R.A.(2013).Super-enhancers in the control of cell identity and disease.*Cell* 155,934-947.;Loven,J.,Hoke,H.A.,Lin,C.Y.,Lau,A.,Orlando,D.A.,Vakoc,C.R.,Bradner,J.E.,Lee,T.I.,and Young,R.A.(2013).Selective inhibition of tumor oncogenes by disruption of super-enhancers.*Cell* 153,320-334.)。CCAT1定位于人类8q24基因区,并在多种癌症如结肠直肠癌,前列腺癌和肝癌中异常高表达(Chen,H.,He,Y.,Hou,Y.S.,Chen,D.Q.,He,S.L.,Cao,Y.F.,and Wu,X.M.(2018a).Long non-coding RNA CCAT1 promotes the migration and invasion of prostate cancer PC-3cells.*European review for medical and pharmacological sciences* 22,2991-2996.;Deng,L.,Yang,S.B.,Xu,F.F.,and Zhang,J.H.(2015).Long noncoding RNA CCAT1 promotes hepatocellular carcinoma progression by functioning as let-7 sponge.*Journal of experimental&clinical cancer research:CR* 34,18.;Tseng,Y.Y.,Moriarity,B.S.,Gong,W.,Akiyama,R.,Tiwari,A.,Kawakami,H.,Ronning,P.,Reuland,B.,Guenther,K.,Beadnell,T.C.,et al.(2014).PVT1 dependence in cancer with MYC copy-number increase.*Nature* 512,82-86.;Xiang,J.F.,Yin,Q.F.,Chen,T.,Zhang,Y.,Zhang,X.O.,Wu,Z.,Zhang,S.,Wang,H.B.,Ge,J.,Lu,X.,et al.(2014).Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus.*Cell research* 24,513-531.)。在结肠癌细胞中,研究人员报道了一个具有额外延长的3'末端的CCAT1转录本,并发现该转录本可调节MYC基因的启动子和增强子之间的DNA相互作用(Xiang,J.F.,Yin,Q.F.,Chen,T.,Zhang,Y.,Zhang,X.O.,Wu,Z.,Zhang,S.,Wang,H.B.,Ge,J.,Lu,X.,et al.(2014).Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus.*Cell research* 24,513-531.) ,但确切的机制不明。

[0187] 由于CCAT1与报道的超级增强子部分重叠(Khan,A.,and Zhang,X.(2016).dbSUPER:a database of super-enhancers in mouse and human genome.*Nucleic acids research* 44,D164-171.) ,通过检查HeLa细胞中的RIC-seq和RNA-seq数据,我们意

外地发现CCAT1可能存在5'末端额外延伸的转录本,而不是先前在结肠癌中报道的3'末端延长(图8中A)。在HeLa细胞中的5'和3'RACE也证实了CCAT1的确存在5'末端额外延伸(图8中A,底部标记为褐色),且转录本总长度为~4,700nt,并与超增强子直接重叠。我们将该新发现的lncRNA命名为CCAT1-5L。基于Exon2(E2)探针的Northern印迹(参见图8中A)证实了CCAT1在不同细胞类型中都表达,而5L特异性探针则仅在HeLa细胞中检测到CCAT1-5L(图8中B),表明CCAT1-5L可能是宫颈癌中的特定转录本,这在宫颈癌患者的RNA-seq数据中也得到进一步证实(数据未显示)。

[0188] smFISH显示CCAT1-5L是核定位的lncRNA并且每个细胞核内存在2-3个点(图8中C)。CCAT1-5L似乎是有功能的,因为CCAT1与8q24“基因沙漠”中其他区域的RNA相互作用大多数来自第一个外显子和额外的5'末端延伸区(图8中A)。此外,我们检测到CCAT1-5L、MYC启动子RNA和PVT1之间具有广泛的远距离RNA-RNA相互作用。更重要的是,在PVT1基因座中观察到的CCAT1-5L结合位点主要位于含有MYC增强子的内含子序列中(图8中A)。以上数据提示CCAT1-5L可能作为超增强子RNA与启动子和增强子RNA相互作用进而调控MYC癌基因的表达。

[0189] 我们接下来探讨了CCAT1-5L是否可以调节MYC表达。当用两种靶向5'末端延伸区的LNA寡核苷酸敲低CCAT1-5L后(图8中D),MYC的RNA水平显著降低~40%(图7中D),表明在HeLa细胞中CCAT1-5L确实可以调节MYC的表达。出乎意料的是,MYC的正调节因子PVT1(Tseng,Y.Y.,Moriarity,B.S.,Gong,W.,Akiyama,R.,Tiwari,A.,Kawakami,H.,Ronning,P.,Reuland,B.,Guenther,K.,Beadnell,T.C.,et al.(2014).PVT1 dependence in cancer with MYC copy-number increase.Nature 512,82-86.)的表达水平也显著降低(图8中D)。因此,我们假设CCAT1-5L hub-lncRNA可能与MYC的启动子和增强子RNA协同作用以协调它们的表达水平。

[0190] 为了验证这一假设,我们首先检查了CCAT1-5L、MYC启动子和MYC增强子RNA是否在体内共定位。为此,我们首先合成了smFISH探针,分别靶向RIC-seq检测到的CCAT1-5L部分、MYC的第一外显子和第一内含子、以及位于PVT1内含子中的增强子。结果显示三种RNA完美的定位在一起(图8中E)。从而进一步证实了CCAT1-5L对MYC和PVT1的调节作用。此外,在LNA寡核苷酸敲降CCAT1-5L后,MYC启动子和增强子RNA之间的共定位模式也没有改变(图8中E)。

[0191] 由于CCAT1-5L在宫颈癌患者中高表达,我们接下来检查了CCAT1-5L是否可以促进细胞增殖和转移,这是癌症的两个标志性特征(Hanahan,D.,and Weinberg,R.A.(2011).Hallmarks of cancer:the next generation.Cell 144,646-674.)。与LNA对照相反,5L特异性LNA寡核苷酸对CCAT1-5L的敲低显著降低了HeLa细胞的增殖速率(图8中F);相反,使用慢病毒质粒异位表达CCAT1-5L能够显著增强细胞增殖(图8中F),这与CCAT1-5L的致癌作用一致。集落形成实验进一步佐证了CCAT1-5对细胞增殖的影响(图8中G)。为了检测CCAT1-5L是否可以影响细胞转移和侵袭,我们使用transwell实验测定HeLa细胞的侵袭能力,发现CCAT1-5L的敲降显著降低了HeLa细胞的转移能力,而CCAT1-5L的过表达则显著增加了细胞的侵袭和转移(图8中H)。总之,这些数据表明CCAT1-5L hub-lncRNA可以直接调节MYC表达以促进肿瘤发生。

<110> 中国科学院生物物理研究所
<120> 一种捕获RNA原位高级结构及相互作用的方法
<130> GNCLN190785
<160> 2
<170> PatentIn version 3.5
<210> 1
<211> 58
<212> DNA
<213> Artificial sequence
<400> 1
aatgatacgg cgaccaccga gatctacact ctttcctac acgacgtct tccgatct 58
<210> 2
<211> 66
<212> DNA
<213> Artificial sequence
<220>
<221> misc_feature
<222> (26)..(32)
<223> n i s a o r t o r c o r g
<400> 2
caagcagaag acggcatacg agatannnnn nngtgactgg agttcagacg tgtgctcttc 60
cgatct 66

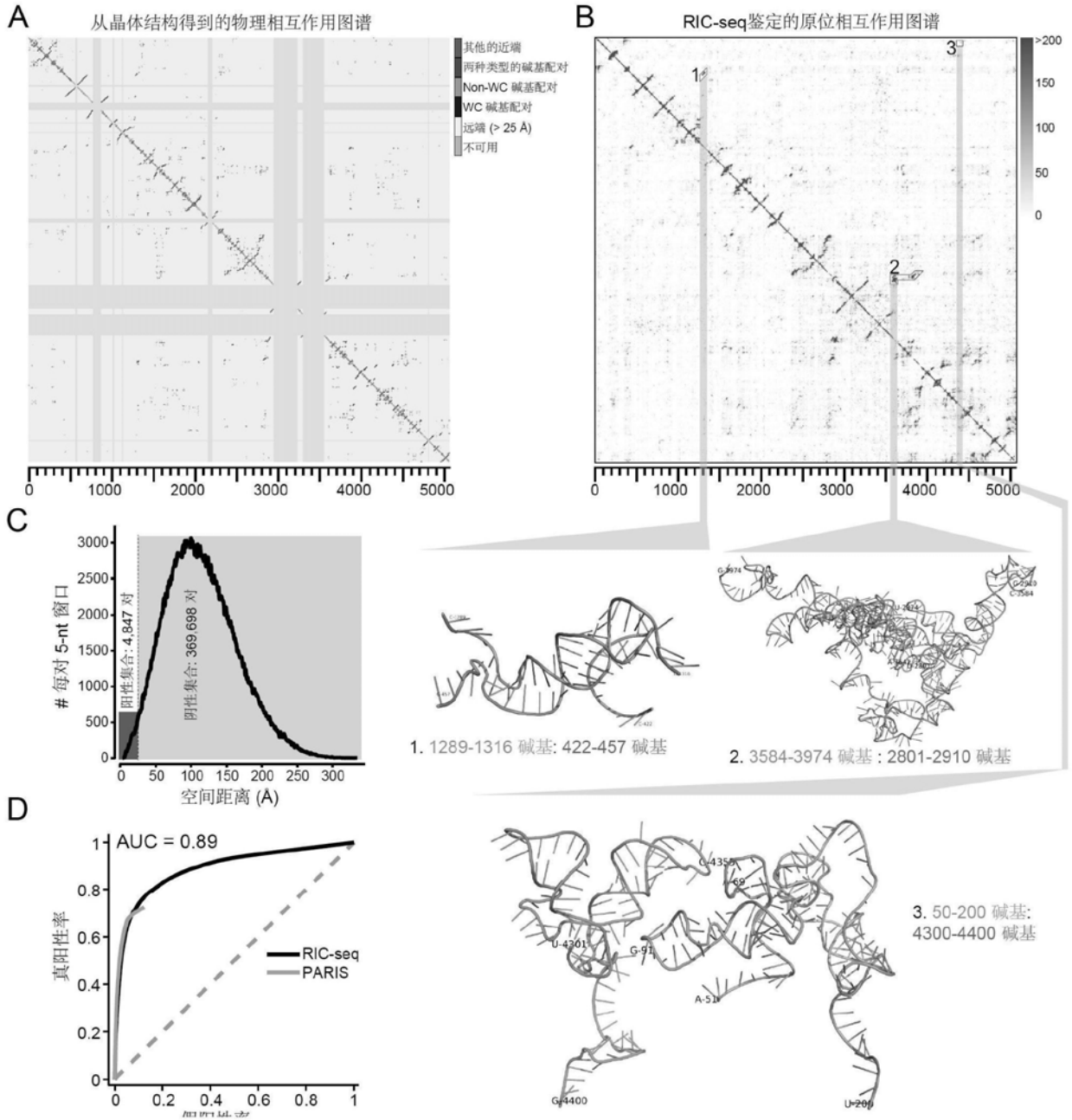


图3

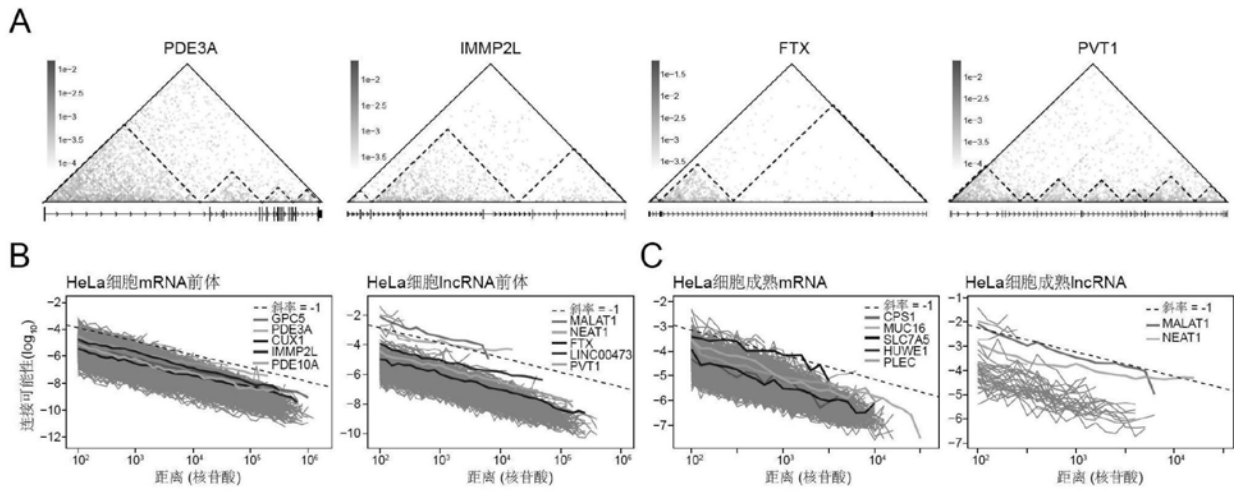


图4

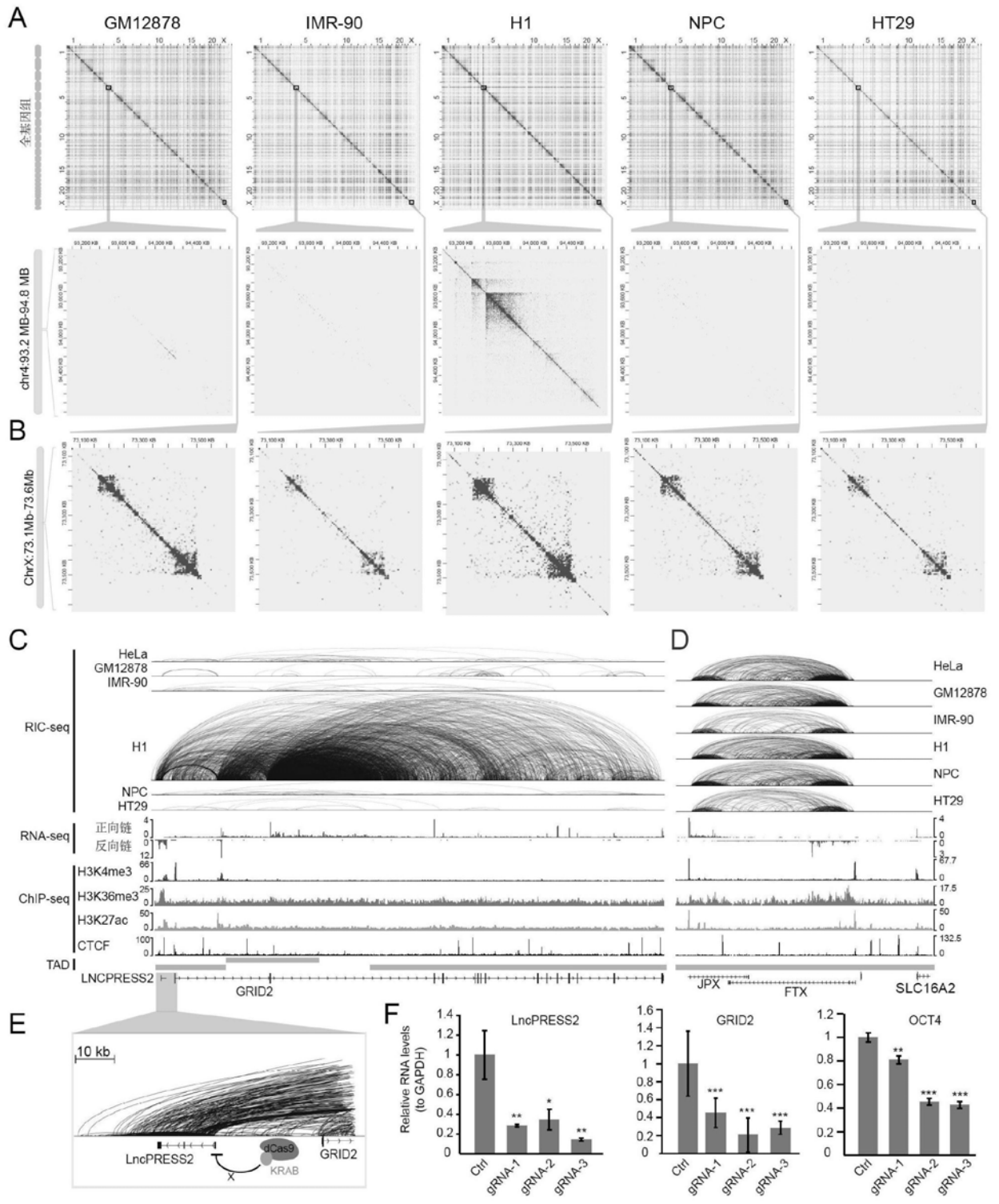


图5

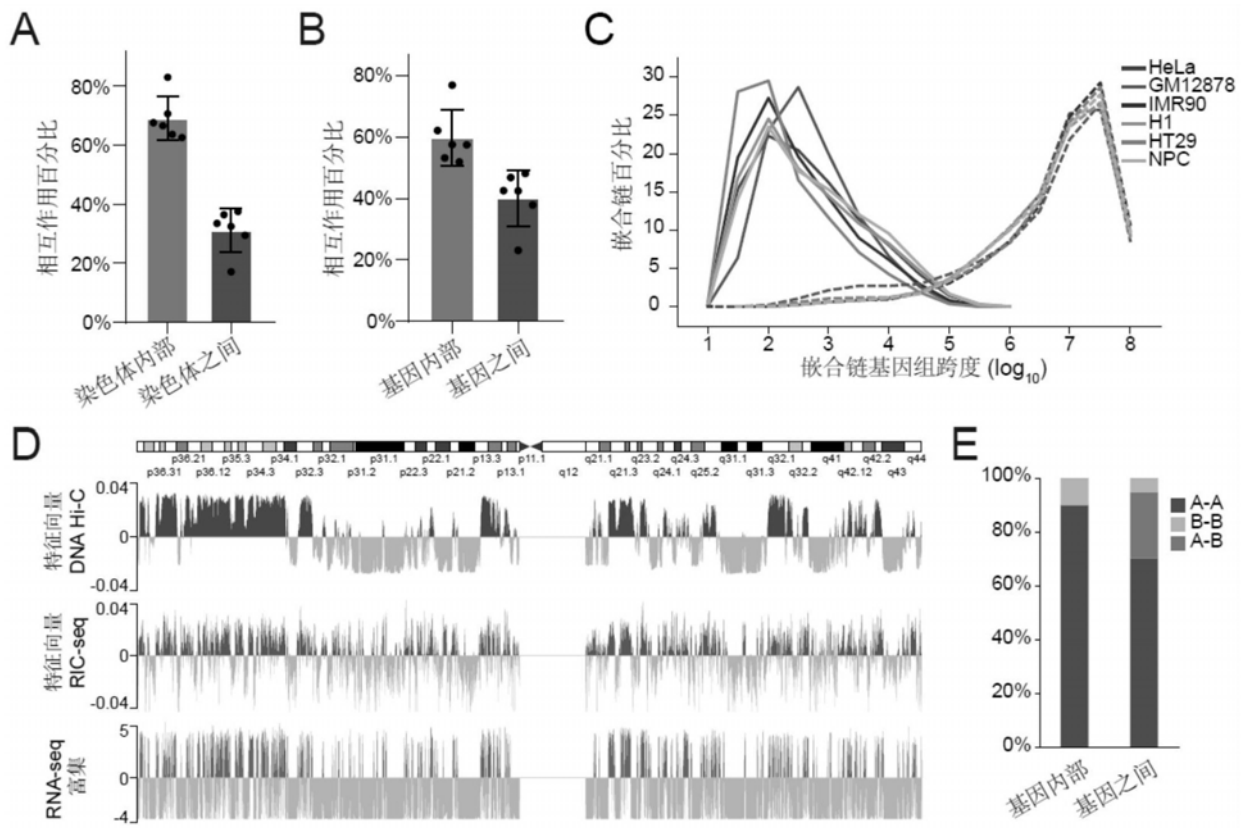


图6

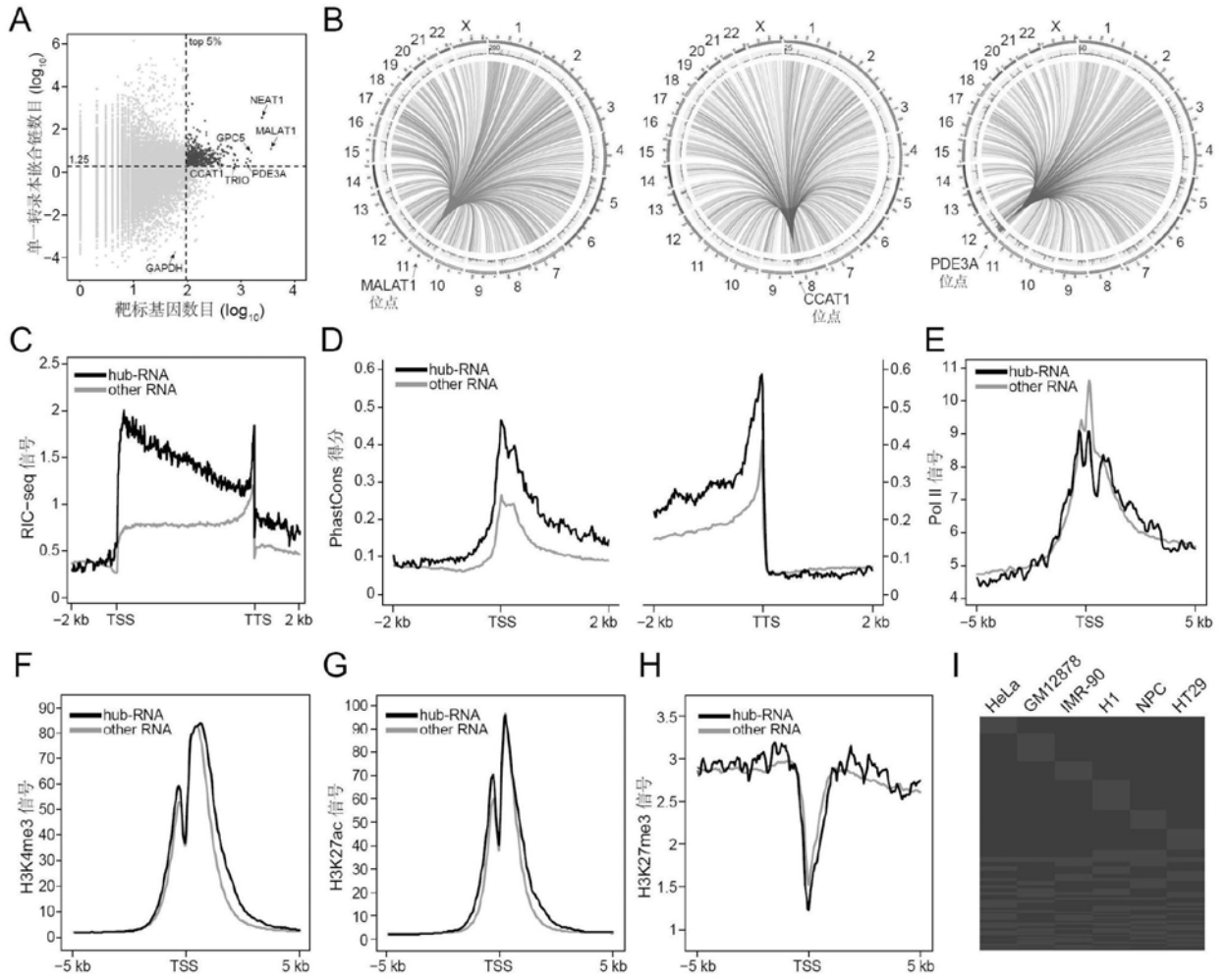


图7

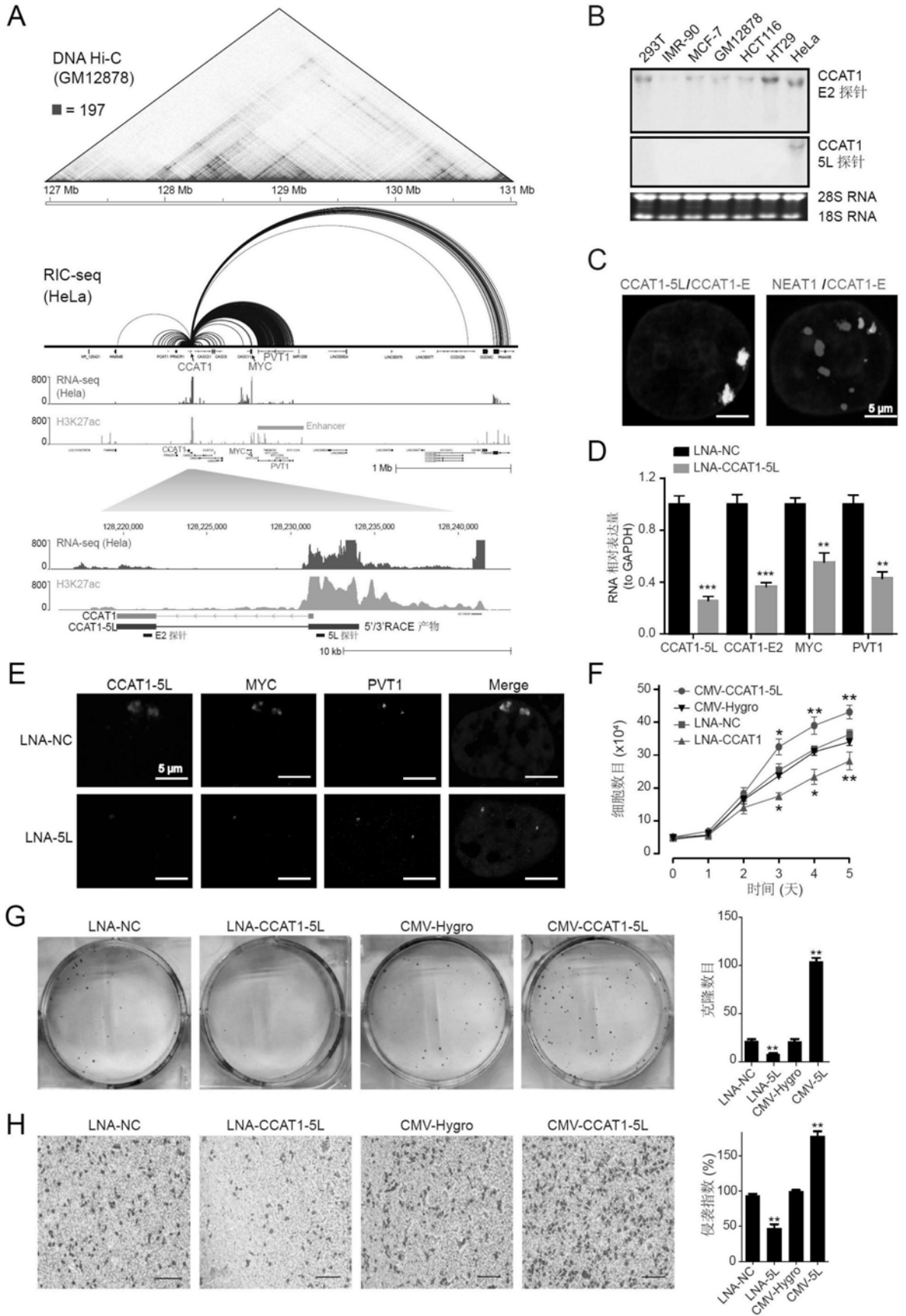


图8