(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108319814 A
(43)申请公布日 2018.07.24

(21)申请号 201810082985.5

(22)申请日 2018.01.29

(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 范珍 陈小伟 陈润生

(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 关畅

(51)Int.Cl.
G06F 19/18(2011.01)

权利要求书2页 说明书18页
序列表4页

(54)发明名称

基于染色体空间相互作用预测长非编码RNA生物学功能的方法

(57)摘要

本发明公开了基于染色体空间相互作用预测长非编码RNA生物学功能的方法。本发明的方法包括如下步骤:1)选择候选长非编码RNA;2)确定长非编码RNA在全基因组范围内的结合位点;3)确定组织细胞中染色体精细的空间相互作用数据;4)预测长非编码RNA的靶基因;5)GO功能富集分析,预测长非编码RNA的生物学功能。本发明的方法能够结合最新的染色体空间相互作用数据,提高长非编码RNA生物学功能预测的准确性。

CN 108319814 A

1.一种预测长非编码RNA生物学功能的方法，包括如下步骤：

(1)确定细胞中长非编码RNA在全基因组范围内结合位点，根据所述长非编码RNA结合位点的基因组定位信息，以所述长非编码RNA结合位点的中心位置为准，将所述长非编码RNA结合位点的基因组定位向上下游进行扩展，得到扩展后结合位点区域内的基因，并将其作为所述长非编码RNA的候选靶基因；

(2)确定所述细胞中染色体空间相互作用数据，得到与所述长非编码RNA结合位点在空间上有相互作用的基因组区域，将与所述长非编码RNA结合位点在空间上有相互作用的基因组区域向上下游进行扩展，得到扩展后基因组区域内的基因，并将其作为所述长非编码RNA远程调控的候选靶基因；

(3)分别计算所述长非编码RNA与步骤(1)和(2)中获得的所述候选靶基因表达水平的皮尔森相关性，得到所述长非编码RNA与所述候选靶基因表达水平的皮尔森相关系数，根据所述皮尔森相关系数的大小选择候选靶基因作为所述长非编码RNA的靶基因；

(4)对步骤(3)获得的所述长非编码RNA的靶基因进行GO功能富集分析，得到所述长非编码RNA的生物学功能。

2.根据权利要求1所述的方法，其特征在于：所述GO功能富集分析的方法包括如下步骤：将所述长非编码RNA的靶基因与GO term中的基因进行比较，通过超几何分布检验基因富集的显著性，并且按照FDR排序，选取靶基因富集最显著的15个GO term作为所述长非编码RNA的生物学功能。

3.根据权利要求1或2所述的方法，其特征在于：所述步骤(1)中，将所述长非编码RNA结合位点的基因组定位向上下游各扩展5kb；

或，所述步骤(2)中，将与所述长非编码RNA结合位点在空间上有相互作用的基因组区域向上下游各扩展5kb。

4.根据权利要求1-4任一所述的方法，其特征在于：所述步骤(3)中，选择皮尔森相关系数绝对值大于0.3的候选靶基因作为长非编码RNA的靶基因。

5.根据权利要求1-4任一所述的方法，其特征在于：所述步骤(1)前还包括选择长非编码RNA的步骤；

或，所述选择长非编码RNA的方法包括如下步骤：

1)收集长非编码RNA的注释数据，得到长非编码RNA数据集；

2)去除所述长非编码RNA数据集之间的冗余，得到完备的长非编码RNA数据集；从所述完备的长非编码RNA数据集中过滤掉没有实验证据支持和注释数据不一致的长非编码RNA，得到过滤后的长非编码RNA数据集；

3)从所述过滤后的长非编码RNA数据集中选取表达量高且位于蛋白质编码基因间区域独立转录的长非编码RNA，并确定其细胞核与细胞质定位信息，选择至少90％定位在细胞核的长非编码RNA，即为步骤(1)中所述的长非编码RNA。

6.根据权利要求5所述的方法，其特征在于：所述步骤1)中，所述注释数据包括名称、基因组定位、序列和表达水平。

7.根据权利要求5所述的方法，其特征在于：所述步骤2)中，利用所述长非编码RNA数据集间的交叉注释、序列相似性比对和基因组定位的方法去除所述长非编码RNA数据集之间的冗余，使序列相似度大于95％、基因组位置重叠度大于95％，得到完备的长非编码RNA数

据集。

8.根据权利要求5所述的方法，其特征在于：所述步骤3)中，所述表达量高为在至少1个组织中FPKM>1。

基于染色体空间相互作用预测长非编码RNA生物学功能的方法

技术领域

[0001] 本发明涉及分子生物学、功能基因组学和生物信息学领域，具体涉及一种基于染色体空间相互作用预测长非编码RNA生物学功能的方法。

背景技术

[0002] 人类基因组计划已经完成，但是对基因组还有待于进一步认识，蕴含在其中的大量功能元件仍然未被发现。ENCODE计划最新公布的数据显示，人类基因组74.7％的区域能够发生转录，而蛋白质编码基因的外显子区只占人类基因组的2.94％。说明人类基因组大部分的转录产物不能编码蛋白质。转录组可以分为能够编码蛋白质的信使RNA(mRNA)和不能够编码蛋白质的非编码RNA。发现较早的非编码RNA有参与蛋白质合成的转运RNA(tRNA)和核糖体RNA(rRNA)等。上世纪90年代初，研究人员发现了一种新的非编码RNA--长非编码RNA。长非编码RNA(long noncoding RNA，lncRNA)是一类长度大于200个核苷酸并且不具有编码蛋白质能力的核糖核酸。1989年，科学家们发现了第一个长非编码RNA H19，并发现该RNA分子能够参与基因印记过程。1990年，科学家找到了参与X染色体失活的lncRNA并将其命名为Xist。之后，随着高通量检测技术(如基因芯片和高通量测序技术)地不断发展，数以万计的长非编码RNA已被科学家们发现。根据长非编码RNA相对于附近蛋白编码基因所在的基因组位置关系，大致可以将其分为以下五类：Exonic lncRNA(外显子型lncRNA)、Intronic lncRNA(内含子型lncRNA)、Antisense lncRNA(反义型lncRNA)、Divergent lncRNA(反向型lncRNA)和Intergenic lncRNA(基因间型lncRNA)。

[0003] 随着lncRNA大量地被发现，针对其功能进行的研究也逐渐增多。目前已知的lncRNA的作用机制大致可以分为以下几种：(一)lncRNA作为诱饵分子通过与蛋白质或者其他分子相结合，阻断其与其靶向目标物的结合，从而影响所结合分子的原有功能。一个比较经典的例子是lncRNA作为内源RNA分子与mRNA竞争性结合miRNA，影响了miRNA与mRNA的结合，进而间接上调了mRNA的表达。例如长非编码RNAlinc-RoR能够跟胚胎干细胞分化相关核心转录因子Oct4、Sox2和Nanog竞争性结合miR-145，从而阻止miR-145对Oct4等基因的抑制，linc-RoR、转录因子和miR-145共同构成一个调控环路调控胚胎干细胞的干性维持与分化。(二)lncRNA作为脚手架，促使生物大分子之间的相互作用以及蛋白质复合物的形成，如HOTAIR。HOTAIR的5’端能够跟PRC2蛋白结合，3’端能够跟LSD1/CoREST/REST复合物结合，PRC2具有组蛋白甲基转移酶活性，能够使H3组蛋白第27位的赖氨酸发生三甲基化，从而沉默基因的转录，而LSD1具有去甲基化酶的活性，能够使H3组蛋白第4位的赖氨酸去甲基化。HOTAIR作为脚手架分子将两种不同的染色质修饰复合物联系起来共同沉默基因的表达。(三)lncRNA作为向导，指引蛋白质复合物到特定的地点或者基因组区域行使功能。受p53调控的长非编码RNAlincRNA-p21是一个很好的例子。在小鼠中lincRNA-p21能够抑制p53依赖的转录应答。lincRNA-p21能够跟hnRNP-K相互作用把hnRNP-K引导到特定的基因组位置去抑制基因的表达。

[0004] 伴随着lncRNA的系统发现和lncRNA功能机制研究的显著进展，人们也开始探讨lncRNA与疾病的关系。lncRNA与代谢疾病、神经退行性疾病、精神疾病、心血管疾病和自身免疫疾病的关系都有明确的报道，但是还是主要集中在肿瘤的研究上。HOTAIR是从HOX基因位点转录出来的一个lncRNA，其作用机制已经有所了解，同时HOTAIR与很多种肿瘤密切相关。在2010年，Howard Y.Chang实验室发现HOTAIR在乳腺癌的原发灶和转移灶中表达显著上调，因此，HOTAIR在肿瘤组织中的表达水平可以作为预测肿瘤转移的分子标识物。在上皮肿瘤细胞中过表达HOTAIR，导致PRC2靶向目标的改变，进而影响H3K27的甲基化、相应基因的表达，最终增强了肿瘤细胞侵袭和转移的能力。2011年，研究人员在结肠癌中发现了同样的结果，HOTAIR在癌组织中的表达水平要高于癌旁组织，而且HOTAIR的高表达与结肠癌的肝转移显著相关。结合患者的随访信息，还发现HOTAIR表达水平高的患者预后较差。研究人员在前列腺癌组织中发现了很多组织特异性表达的长非编码RNA，例如PCA3/DD3、PCGEM1、PCAT-1、PRNCR1等。除了以上列举的一些癌症相关的lncRNA，还有一些诸如aHIF、ANRIL、Oct4-pg、PTENP1和BC200等在神经母细胞瘤、乳腺癌、胶质瘤、结直肠癌、神经退行性等疾病中有功能的长非编码RNA。近些年来在几乎各种已知的各种肿瘤中都发现了lncRNA的存在以及两者间的显著关联，表明了lncRNA在肿瘤发生发展中起到了至关重要的作用。

[0005] GENCODE最新公布的第27版的数据中包括了27,908条长非编码RNA，其中却仅有一小部分的lncRNA的功能被报道。由于lncRNA在生物体中发挥着重要作用以及其与许多疾病密切相关，因此对其进行的研究日益增多。然而，lncRNA自身结构比较复杂，对于它们如何发挥功能还需进一步深入地研究。目前，对lncRNA功能进行预测的方法主要是通过基因表达量的信息来确定的。最早的关于lncRNA功能预测的方法是由在2009年提出的关联推定(Gulit by association)的方法。该方法的假定是共表达的RNA更有可能会受到同样的调控，并倾向于具有相似的功能或者参与相同的生物过程。通过分析lncRNA和mRNA的共表达水平，得到与所研究lncRNA显著相关的mRNA。由于mRNA的功能大都是已知的，可以通过将富集出的mRNA的功能或参与的通路推定给该lncRNA。通过这一方法，John L.Rinn等发现TUG1能够与PRC2结合并且参与p53依赖型细胞周期的调控过程。随后又衍生出一些相似的lncRNA功能的预测方法，如ncFANs和lnc-GFP。ncFANs和lnc-GFP主要基于长非编码RNA与蛋白质编码基因表达的相关性以及蛋白质之间的相互作用来预测长非编码RNA的生物学功能。由于长非编码RNA的表达水平通常低于蛋白质编码基因，目前的预测往往不能为长非编码RNA的生物学功能研究提供有效的线索。

发明内容

[0006] 本发明的目的在于提供一种基于染色体空间相互作用预测长非编码RNA生物学功能的方法，能够结合最新的染色体空间相互作用数据，提高长非编码RNA生物学功能预测的准确性。

[0007] 为了解决上述技术问题，本发明提供了一种预测长非编码RNA生物学功能的方法。

[0008] 本发明提供的预测长非编码RNA生物学功能的方法包括如下步骤：

[0009] (1)确定细胞中长非编码RNA在全基因组范围内结合位点，根据所述长非编码RNA结合位点的基因组定位信息，以所述长非编码RNA结合位点的中心位置为准，将所述长非编码RNA结合位点的基因组定位向上下游进行扩展，得到扩展后结合位点区域内的基因，并将

其作为所述长非编码RNA的候选靶基因；

[0010] (2)确定所述细胞中染色体空间相互作用数据，得到与所述长非编码RNA结合位点在空间上有相互作用的基因组区域，将与所述长非编码RNA结合位点在空间上有相互作用的基因组区域向上下游进行扩展，得到扩展后基因组区域内的基因，并将其作为所述长非编码RNA远程调控的候选靶基因；

[0011] (3)分别计算所述长非编码RNA与步骤(1)和(2)中获得的所述候选靶基因表达水平的皮尔森相关性，得到所述长非编码RNA与所述候选靶基因表达水平的皮尔森相关系数，根据所述皮尔森相关系数的大小选择候选靶基因作为所述长非编码RNA的靶基因；

[0012] (4)对步骤(3)获得的所述长非编码RNA的靶基因进行GO功能富集分析，得到所述长非编码RNA的生物学功能。

[0013] 上述方法中，所述确定细胞中长非编码RNA在全基因组范围内结合位点的方法为现有技术中公知方法，该方法在文献“Simon等，The genomic binding sites of a noncoding RNA.PNAS.108:20497-20502.”中公开过。本领域技术人员可根据现有技术中公知方法来确定长非编码RNA在全基因组范围内的结合位点。具体方法包括如下步骤：

[0014] 步骤S21、收集细胞并用1％甲醛交联，然后加入裂解液，得到交联的细胞核。

[0015] 步骤S22、设计靶标长非编码RNA的捕获寡核苷酸，并对其进行生物素标记，得到标记后的寡核苷酸。

[0016] 步骤S23、加入超声缓冲液进行超声处理，将其打断到约300bp的片段，得到超声后的细胞核提取物。

[0017] 步骤S24、将所述标记后的寡核苷酸与所述超声后的细胞核提取物混匀，室温孵育过夜。

[0018] 步骤S25、加入链霉素磁珠孵育，得到结合产物。因为链霉素可以与寡核苷酸上所带的生物素结合从而拉下靶标RNA，同时与靶标RNA相结合的DNA片段也被捕获到。

[0019] 步骤S26、用洗涤液清洗所述结合产物几次，以除去非特异性的结合。

[0020] 步骤S27、将清洗后的结合产物从珠子上洗脱下来构建文库并进行测序分析，确定长非编码RNA在全基因组范围内的结合位点。该结合位点是指长非编码RNA在全基因组范围内的具体结合位置，如某染色体的第几位至第几位。

[0021] 上述方法中，所述确定细胞中染色体空间相互作用数据的方法为现有技术中公知方法，该方法在文献“Goh等，Chromatin Interaction Analysis with Paired-End Tag Sequencing(ChIAPET)for Mapping Chromatin Interactions and Understanding Transcription Regulation.JOVE.62.”中公开过。本领域技术人员可根据现有技术中公知方法来确定细胞中染色体空间相互作用数据。具体方法包括如下步骤：

[0022] 步骤S31、收集细胞并用1％甲醛交联，然后加入细胞质裂解液和细胞核裂解液，获得交联的染色质。

[0023] 步骤S32、将交联的染色质进行超声处理，将其打断到约300bp的片段，之后用IgG磁珠孵育过夜，以除去非特异性结合的DNA，得到预纯化后的染色质。与此同时，用RNA聚合酶Ⅱ的抗体孵育IgG磁珠过夜，使抗体结合在磁珠表面。

[0024] 步骤S33、次日，将预纯化后的染色质与用抗体包被后的磁珠混匀，孵育过夜，使磁珠与所需的目标染色质相结合，得到结合产物。

[0025] 步骤S34、用洗涤液清洗结合产物几次，以除去非特异性的结合。

[0026] 步骤S35、将结合产物从珠子上洗脱下来并测定浓度。

[0027] 步骤S36、将洗脱下来的染色质DNA碎片分为两等分，分别用不同DNA半连接子(A/B)连接，两个连接子除了中间的两个核苷酸不一样之外(连接子A是CG；连接子B是AT)，其他部分的核苷酸序列完全相同。

[0028] 步骤S37、在连接子进行连接后去除多余的序列，将两部分混合，两等分又会重新结合到一起发生邻近式连接。在邻近连接时，如果同一个染色质复合物内的DNA碎片被相同的连接子连接在一起，那么则会产生同源二聚体形式的连接产物(即AA或BB)。然而，如果连接反应发生在不同染色质的DNA碎片之间，那么这样非特异性连接的产物将有50％的几率形成异源二聚体的形式(AB或者BA)。这些异源二聚体的连接子可以作为非特异性连接的标志，用来评估每一次建立ChIA-PET文库发生非特异性连接概率的大小。

[0029] 步骤S38、在邻近连接之后，获得的连接产物可以用来提取配对的末端标签(PET)，这些末端标签的模板将被用来构建文库并进行测序分析，得到细胞中染色体空间相互作用数据。

[0030] 上述方法中，所述GO功能富集分析的方法包括如下步骤：将所述长非编码RNA的靶基因与GO term中的基因进行比较，通过超几何分布检验基因富集的显著性，并且按照FDR排序，选取靶基因富集最显著的15个GO term作为所述长非编码RNA的生物学功能。

[0031] 上述方法中，所述步骤(1)中，将所述长非编码RNA结合位点的基因组定位向上下游各扩展5kb；所述步骤(2)中，将与所述长非编码RNA结合位点在空间上有相互作用的基因组区域向上下游各扩展5kb。

[0032] 上述方法中，所述步骤(3)中，选择皮尔森相关系数绝对值大于0.3的候选靶基因作为长非编码RNA的靶基因。

[0033] 上述方法中，所述步骤(1)前还包括选择长非编码RNA的步骤；

[0034] 所述选择长非编码RNA的方法包括如下步骤：

[0035] 1)收集长非编码RNA的注释数据，得到长非编码RNA数据集；

[0036] 2)去除所述长非编码RNA数据集之间的冗余，得到完备的长非编码RNA数据集；从所述完备的长非编码RNA数据集中过滤掉没有实验证据支持和注释数据不一致的长非编码RNA，得到过滤后的长非编码RNA数据集；

[0037] 3)从所述过滤后的长非编码RNA数据集中选取表达量高且位于蛋白质编码基因间区域独立转录的长非编码RNA，并确定其细胞核与细胞质定位信息，选择至少90％定位在细胞核的长非编码RNA，即为步骤(1)中所述的长非编码RNA。

[0038] 上述方法中，步骤1)中，所述注释数据包括名称、基因组定位、序列和表达水平。在本发明中，所述注释数据收集自公开发表的文献：Cabili等，Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.2011.Genes Dev 25:1915-27和GENCODE公共数据库(公开GENCODE公共数据库的文献如下：GENCODE.Harrow等，GENCODE:the reference human genome annotation for The ENCODE Project.Genome Research.2012.22:1760-74.，GENCODE公共数据库的查询网址如下：http://www.gencodegenes.org/)。

[0039] 上述方法中，步骤2)中，利用所述长非编码RNA数据集间的交叉注释、序列相似性

比对和基因组定位的方法去除所述长非编码RNA数据集之间的冗余，使序列相似度大于95%、基因组位置重叠度大于95%，得到完备的长非编码RNA数据集。

[0040] 上述方法中，步骤3)中，所述表达量高为在至少1个组织中FPKM>1。

[0041] 上述方法中，所述细胞可为常见细胞系，如HCT116细胞系、HeLa细胞系、K562细胞系等。在本发明中，所述细胞具体为MCF-7细胞系。

[0042] 上述方法中，所述长非编码RNA为MALAT1。利用上述方法预测其生物学功能如下：1)参与mRNA、rRNA等转录后加工代谢过程；2)mRNA翻译调控；3)与蛋白质结合；4)与具有多聚A尾的RNA结合；5)基于SRP的膜靶向共翻译蛋白；6)病毒转录。本发明预测的功能与文献“Hutchinson等，A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35splicing domains.2007.BMC Genomics 8:39；Bernard等，A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression.2010.EMBO J.29:3082-3093”中已经证实的MALAT1在细胞核内能够与其他蛋白质结合，参与mRNA的转录后加工代谢过程的结果一致。

[0043] 本发明基于染色体空间相互作用提供了一种预测长非编码RNA生物学功能的方法。本发明的方法包括如下步骤：1)选择候选长非编码RNA；2)确定细胞中长非编码RNA在全基因组范围内的结合位点；3)确定细胞中染色体精细的空间相互作用数据；4)预测长非编码RNA的靶基因；5)GO功能富集分析，预测长非编码RNA的生物学功能。本发明的预测方法能够结合最新的染色体空间相互作用数据，提高长非编码RNA生物学功能预测的准确性。

具体实施方式

[0044] 为了使本发明的技术方案和优点更加清楚明白，以下结合实施例对本发明进行进一步说明。此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0045] 实施例1、基于染色体空间相互作用预测长非编码RNA生物学功能的方法

[0046] 一、选择候选长非编码RNA

[0047] 1、构建完备的长非编码RNA数据集

[0048] 从公开发表的文献：Cabili等，Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.2011.Genes Dev25:1915-27和GENCODE公共数据库(公开GENCODE公共数据库的文献如下：GENCODE.Harrow等，GENCODE:the reference human genome annotation for The ENCODE Project.Genome Research.2012.22:1760-74.，GENCODE公共数据库的查询网址如下：http://www.gencodegenes.org/)中收集长非编码RNA的注释数据，包括名称、基因组定位、序列和表达水平等信息，得到长非编码RNA数据集。

[0049] 利用数据集间的交叉注释、序列相似性比对和基因组定位的方法去除数据集之间的冗余，使序列相似度大于95%、基因组位置重叠度大于95%，得到完备的长非编码RNA数据集。

[0050] 2、对长非编码RNA数据集进行过滤

[0051] 从完备的长非编码RNA数据集中过滤掉没有实验证据支持、注释数据不一致的长非编码RNA，得到过滤后的长非编码RNA数据集。

[0052] 3、确定长非编码RNA的核质定位比例

[0053] 从过滤后的长非编码RNA数据集中选取表达量高(在至少1个组织中FPKM>1)且位于蛋白质编码基因间区域独立转录的长非编码RNA,利用细胞核质分提实验和RNA原位杂交技术确定这些长非编码RNA的细胞核与细胞质定位信息,选择至少90%定位在细胞核的长非编码RNA作为候选长非编码RNA。

[0054] 二、确定细胞中长非编码RNA在全基因组范围内的结合位点

[0055] 1、收集细胞并用1%(体积分数)甲醛交联,然后加入裂解液,得到交联的细胞核。

[0056] 2、设计靶标RNA的捕获寡核苷酸,并对其进行生物素标记,得到生物素标记的捕获寡核苷酸。

[0057] 3、向步骤1中的交联的细胞核中加入超声缓冲液进行超声处理,将其打断到约300bp的片段,得到超声后的细胞核提取物。

[0058] 4、将步骤2中的生物素标记的捕获寡核苷酸与步骤3中的超声后的细胞核提取物混匀,室温孵育过夜,得到捕获反应产物。

[0059] 5、向步骤4中的捕获反应产物中加入链霉素磁珠,孵育,得到结合后产物。因为链霉素可以与寡核苷酸上所带的生物素结合,从而捕获得到与寡核苷酸相结合的靶标RNA,同时与靶标RNA相结合的DNA片段也被捕获到。

[0060] 6、用洗涤液清洗步骤5中的结合后产物几次,以除去非特异性的结合。

[0061] 7、将结合到的CHART-DNA(清洗后的结合后产物)从珠子上洗脱下来构建文库,并进行测序分析,得到长非编码RNA在全基因组范围内的结合位点的基因组定位信息。

[0062] 三、确定细胞中染色体精细的空间相互作用数据

[0063] 1、收集细胞并用1%(体积分数)甲醛交联,然后加入细胞质裂解液和细胞核裂解液,获得交联的染色质。

[0064] 2、向步骤1获得的交联的染色质中加入超声缓冲液进行超声处理,将其打断到约300bp的片段,然后加入IgG磁珠,孵育过夜,以除去非特异性结合的DNA,得到预纯化后的染色质;与此同时,用RNA聚合酶Ⅱ的抗体孵育IgG磁珠过夜,使抗体结合在磁珠表面,得到用抗体包被后的磁珠。

[0065] 3、次日,将步骤2中的预纯化后的染色质与用抗体包被后的磁珠混匀,孵育过夜,使磁珠与所需的目标染色质相结合,得到结合后产物。

[0066] 4、用洗涤液清洗步骤3中的结合后产物几次,以除去非特异性的结合。

[0067] 5、将结合到的ChIP-DNA(清洗后的结合后产物)从珠子上洗脱下来,得到染色质DNA碎片并测定其浓度。

[0068] 6、将染色质DNA碎片分为两等分,分别用不同DNA半连接子:连接子A和连接子B连接,分别得到连接产物A和连接产物B。两个连接子除了中间的两个核苷酸不一样之外(连接子A是CG;连接子B是AT),其他部分的核苷酸序列完全相同。

[0069] 连接子A:5'-GGCCGCGAT(biotin)ATCTTATCCAAC-3';

[0070] 5'-GTTGGATAAGATATCGC-3';

[0071] 连接子B:

[0072] 5'-GGCCGCGAT(biotin)ATACATTCCAAC-3';

[0073] 5'-GTTGGAATGTATATCGC-3'。

[0074] 7、在连接子连接后,去除连接产物中多余的序列,然后将两部分连接产物混合,两

等分又会重新结合到一起发生邻近式连接。在邻近连接时，如果同一个染色质复合物内的DNA碎片被相同的连接子连接在一起，那么则会产生同源二聚体形式的连接产物(即AA或BB)。然而，如果连接反应发生在不同染色质的DNA碎片之间，那么这样非特异性连接的产物将有50％的几率形成异源二聚体的形式(AB或者BA)。这些异源二聚体的连接子可以作为非特异性连接的标志，用来评估每一次建立ChIA-PET文库发生非特异性连接概率的大小。

[0075] 8、在邻近连接之后，获得的连接产物可以用来提取配对的末端标签(PET)，这些末端标签的模板将被用来构建文库并进行测序分析，根据分析结果确定组织细胞中染色体精细的空间相互作用数据。

[0076] 四、预测长非编码RNA的靶基因

[0077] 1、根据步骤二获得的长非编码RNA在全基因组范围内结合位点的基因组定位信息，以长非编码RNA结合位点的中心位置为准，将结合位点的基因组定位向上下游各扩展5kb，寻找扩展后结合位点区域内的基因，作为长非编码RNA的候选靶基因。

[0078] 2、结合步骤三中获得的染色体空间相互作用数据，得到与长非编码RNA结合位点在空间上有相互作用的基因组区域，将与长非编码RNA结合位点在空间上有相互作用的基因组区域向上下游各扩展5kb，寻找扩展后基因组区域内的基因，作为长非编码RNA远程调控的候选靶基因。

[0079] 3、分别计算长非编码RNA与步骤1和步骤2获得的候选靶基因表达水平的皮尔森相关性，选择皮尔森相关系数绝对值大于0.3的基因作为长非编码RNA的靶基因。

[0080] 五、GO功能富集分析

[0081] 基于步骤四预测到的长非编码RNA的靶基因，准备Gene Ontology进行GO功能富集分析，预测长非编码RNA的生物学功能。具体方法如下：将预测的长非编码RNA的靶基因与GO term中的基因进行比较，通过超几何分布检验基因富集的显著性，并且按照FDR排序，选取靶基因富集最显著的15个GO term作为预测的长非编码RNA的生物学功能。

[0082] 实施例2、基于染色体空间相互作用预测长非编码RNA的生物学功能的方法的应用

[0083] 一、选择候选长非编码RNA

[0084] 按照实施例1步骤一中的方法，从完备的长非编码RNA数据集中选取长非编码RNA----MALAT1(NR_144568.1)作为靶标RNA，其序列如序列1所示。

[0085] 二、确定长非编码RNA在全基因组范围内的结合位点

[0086] 参照文献“Simon等，The genomic binding sites of a noncoding RNA.PNAS.108:20497-20502.”中的方法确定长非编码RNA----MALAT1在全基因组范围内的结合位点，具体步骤如下：

[0087] 1、收集MCF-7细胞(购自ATCC，ATCC编号为HTB-22)并用1％(体积分数)甲醛交联，然后加入裂解液，得到交联的细胞核。

[0088] 上述裂解液由溶质和溶剂组成，溶剂为水，溶质及其浓度分别如下：0.3M蔗糖，1％(体积分数)Triton X-100，10mM Hepes(pH7.5)，100mM KOAc，0.1mM EGTA，0.5mM spermidine，0.15mM spermine，Roche protease inhibitor tablet(终浓度为1×)，1mM DTT，10U/mL SUPERasIN。

[0089] 2、设计靶标RNA的捕获寡核苷酸，并对其进行生物素标记。序列如下：

[0090] MALAT1CO1：5’-CCTCAGTCCTAGCTTCATCAAACAC-3’；

[0091] MALAT1CO2:5'-GTCTTTCCTGCCTTAAAGTTACATTCG-3',

[0092] 3、向步骤1中的交联的细胞核中加入超声缓冲液进行超声处理,将其打断到约300bp的片段,得到超声后的细胞核提取物。

[0093] 上述超声缓冲液由溶质和溶剂组成,溶剂为水,溶质及其浓度分别如下:50mM HEPES(pH7.5),75mM NaCl,0.5%(体积分数)N-lauroylsarcosine,0.1%(质量分数)Sodium deoxycholate,0.1mM EGTA,10U/mL RNase inhibitor(Promega),1mM DTT,EDTA-free protease inhibitors(Roche)(终浓度为1×)。

[0094] 4、分别将步骤2中的捕获寡核苷酸MALAT1CO1和MALAT1CO2与上述超声后的细胞核提取物混匀,使其在体系中的浓度为800nM,室温孵育过夜,得到捕获反应产物。

[0095] 5、向步骤4中的捕获反应产物中加入链霉素磁珠(Thermo Fisher),孵育,得到结合后产物。因为链霉素可以与寡核苷酸上所带的生物素结合,从而捕获得到与寡核苷酸相结合的靶标RNA,同时与靶标RNA相结合的DNA片段也被捕获到。

[0096] 6、用洗涤液清洗步骤5中的结合后产物5次,以除去非特异性的结合,将结合到的CHART-DNA(清洗后的结合后产物)从珠子上洗脱下来,得到洗脱后产物。

[0097] 上述洗涤液由溶质和溶剂组成,溶剂为水,溶质及其浓度分别如下:250mM NaCl,10mM Hepes(pH7.5),2mM EDTA,1mM EGTA,0.2%(质量分数)SDS,0.1%(体积分数)N-lauroylsarcosine。

[0098] 7、用NEBNext®Ultra™II DNA文库试剂盒(E7645,NEB)基于步骤6中的洗脱后产物构建文库并在HiSeq测序仪上进行双端测序,读长为150bp,得到长非编码RNA----MALAT1在全基因组范围内的结合位点的基因组定位信息。长非编码RNA----MALAT1在全基因组范围内的部分结合位点的基因组定位信息如表1所示。

[0099] 表1、长非编码RNA在全基因组范围内的部分结合位点的基因组定位信息

[0100]

染色体	起始位置	终止位置	染色体	起始位置	终止位置
chrX	487470	489494	chr12	53877191	53883847
chrX	47087417	47092798	chr12	54676535	54682708
chrX	47429400	47436909	chr12	56523830	56540021
chrX	53221822	53227414	chr12	56546142	56575584
chrX	102863193	102868966	chr12	57482722	57495331
chrX	148596815	148622801	chr12	57888888	57900750
chrX	149100310	149105408	chr12	57902032	57914198
chrX	149107501	149119517	chr12	57916740	57922190
chr13	21720418	21727979	chr12	58090234	58104267
chr13	31028903	31036337	chr12	58120340	58130244
chr13	45902467	45912415	chr12	120652615	120660612
chr12	72333	94851	chr12	122263587	122269628
chr12	6496679	6501905	chr11	402881	416346
chr12	7052459	7059633	chr11	1752753	1777796
chr12	49214454	49223861	chr11	45921504	45934624

chr12	49393087	49395788	chr11	46801591	46805630
chr12	49949005	49952902	chr11	47258926	47270661
chr12	49992379	49995639	chr11	47433173	47442251
chr12	50169950	50180651	chr11	47853358	47870210
chr12	50489090	50493535	chr11	60654475	60661177
chr12	50525386	50534964	chr11	62335455	62346170
chr12	51764786	51769143	chr11	62389184	62401223
chr12	52573154	52584121	chr11	62576999	62580230
chr12	53280250	53297672	chr11	62647911	62661477
chr12	53331957	53349613	chr11	63972636	63980148
chr12	53429418	53441113	chr11	63989874	63994357
chr12	53448638	53462892	chr11	64521128	64533637
chr12	53594214	53602269	chr11	64571906	64604174
chr12	53607614	53625986	chr11	64809536	64815174
chr12	53691423	53698792	chr11	64864004	64906084

[0101] 三、确定组织细胞中染色体精细的空间相互作用数据

[0102] 参考文献“Goh等,Chromatin Interaction Analysis with Paired-End Tag Sequencing(ChIAPET)for Mapping Chromatin Interactions and Understanding Transcription Regulation.JOVE.62.”中的方法确定细胞中染色体精细的空间相互作用情况,具体步骤如下:

[0103] 1、收集1×10^8个MCF-7细胞(购自ATCC)并用1%(体积分数)甲醛交联,然后加入15mL细胞质裂解液裂解细胞,得到细胞核提取物,再向细胞核提取物中加入15mL细胞核裂解液,获得交联的染色质。

[0104] 上述细胞质裂解液由溶质和溶剂组成,溶剂为水,溶质及其浓度分别如下:50mM HEPES(pH7.5),150mM NaCl,1mM EDTA,1%(体积分数)Triton X-100,0.1%(体积分数)Sodium Deoxycholate,0.1%(质量分数)SDS,Protease inhibitor(Roche)(终浓度为1×)。

[0105] 上述细胞核裂解液由溶质和溶剂组成,溶剂为水,溶质及其浓度分别如下:50mM HEPES(pH7.5),150mM NaCl,1mM EDTA,1%Triton X-100,0.1%Sodium Deoxycholate,1%(质量分数)SDS,Protease inhibitor(Roche)(终浓度为1×)。

[0106] 2、向步骤1获得的交联的染色质中加入超声缓冲液进行超声处理,将其打断到约300bp的片段,然后加入IgG磁珠(Thermo Fisher),孵育过夜,以除去非特异性结合的DNA,得到预纯化后的染色质;与此同时,用RNA聚合酶Ⅱ的抗体(Covance,MMS-126R)孵育IgG磁珠过夜,使抗体结合在磁珠表面,得到抗体包被后的磁珠。

[0107] 3、次日,将步骤2中的预纯化后的染色质与抗体包被后的磁珠混匀,孵育过夜,使磁珠与所需的目标染色质相结合,得到结合后产物。

[0108] 4、用洗涤液清洗步骤3中的结合后产物5次,以除去非特异性的结合。

[0109] 5、将结合到的ChIP-DNA(清洗后的结合后产物)从珠子上洗脱下来,得到染色质DNA碎片并测定其浓度。

[0110] 6、将染色质DNA碎片分为两等分，分别用不同DNA半连接子：连接子A和连接子B连接，分别得到连接产物A和连接产物B。两个连接子除了中间的两个核苷酸不一样之外(连接子A是CG；连接子B是AT)，其他部分的核苷酸序列完全相同。

[0111] 7、在连接子连接后，去除连接产物中多余的序列，然后将两部分连接产物混合，两等分又会重新结合到一起发生邻近式连接。在邻近连接时，如果同一个染色质复合物内的DNA碎片被相同的连接子连接在一起，那么则会产生同源二聚体形式的连接产物(即AA或BB)。然而，如果连接反应发生在不同染色质的DNA碎片之间，那么这样非特异性连接的产物将有50％的几率形成异源二聚体的形式(AB或者BA)。这些异源二聚体的连接子可以作为非特异性连接的标志，用来评估每一次建立ChIA-PET文库发生非特异性连接概率的大小。

[0112] 8、在邻近连接之后，获得的连接产物可以用来提取配对的末端标签(PET)，基于末端标签的模板用NEBNext®Ultra™II DNA文库试剂盒(E7645，NEB)构建文库并在HiSeq测序仪上进行双端测序，读长为150bp，得到染色体精细的空间相互作用数据。染色体精细的空间相互作用部分数据结果如表2所示。

[0113] 表2、染色体精细的空间相互作用部分数据结果

[0114]

染色体_1	起始位置_1	终止位置_1	染色体_2	起始位置_2	终止位置_2
chr1	113270429	113271459	chr22	25568160	25569141
chr1	113270394	113271449	chr22	25568163	25569143
chr1	145382435	145383000	chr22	26907921	26908514
chr1	145382392	145383011	chr22	26907868	26908487
chr1	227489479	227490368	chr22	46990265	46991152
chr1	227489492	227490367	chr22	46990213	46991146
chr10	56848913	56849893	chr22	31212898	31213611
chr10	56848913	56849878	chr22	31212874	31213606
chr10	103124018	103124703	chr22	23442342	23443022
chr10	103124015	103124665	chr22	23442325	23443009
chr11	789117	789688	chr22	43582474	43583406
chr11	789079	789719	chr22	43582480	43583461
chr11	62608621	62609243	chr22	23442357	23443002
chr11	62608624	62609274	chr22	23442393	23442959
chr12	120728890	120730294	chr22	23520439	23521488
chr12	120728946	120730253	chr22	23520426	23521481
chr13	31774085	31774726	chr22	31742614	31743338
chr13	31774107	31774700	chr22	31742637	31743404
chr13	36327147	36327700	chr22	43010869	43011817
chr13	36327143	36327656	chr22	43010940	43011818
chr13	51666786	51667498	chr22	17257840	17258793
chr13	51666735	51667501	chr22	17257844	17258850
chr14	51706548	51707225	chr22	33757718	33758311
chr14	51706516	51707181	chr22	33757666	33758342
chr14	62978950	62979716	chr22	30519395	30520383
chr14	62978957	62979727	chr22	30519396	30520421
chr16	33962908	33963791	chr22	30819233	30819877
chr16	33962872	33963822	chr22	30819246	30819903
chr17	31001623	31002678	chr22	33884431	33885269
chr17	31001637	31002627	chr22	33884425	33885331
chr17	41465594	41466197	chr22	42707227	42707929
chr17	41465531	41466172	chr22	42707260	42707857
chr17	56736170	56737026	chr22	51021143	51022104
chr17	56736200	56737073	chr22	51021131	51022179
chr18	34089403	34090424	chr22	18483937	18484577
chr18	34089409	34090447	chr22	18483914	18484632
chr19	752333	752990	chr22	38694197	38694869
chr19	752334	752986	chr22	38694244	38694786
chr2	12039060	12039712	chr22	39239752	39240394
chr2	12039079	12039644	chr22	39239689	39240416
chr2	2.19E+08	2.19E+08	chr22	46465611	46466427
chr2	2.19E+08	2.19E+08	chr22	46465683	46466475
chr20	52207703	52208646	chr22	38791587	38792252
chr20	52207684	52208651	chr22	38791556	38792279
chr20	52210324	52211077	chr22	38794653	38795342
chr20	52210290	52211017	chr22	38794651	38795299
chr20	52211456	52212672	chr22	38795714	38796829

[0115]

chr20	52211459	52212637	chr22	38795722	38796799
chr20	52407363	52408325	chr22	24236168	24237194
chr20	52407430	52408371	chr22	24236148	24237206

[0116] 四、预测长非编码RNA的靶基因

[0117] 1、根据步骤二获得的长非编码RNA----MALAT1在全基因组范围内结合位点的基因组定位信息，以长非编码RNA----MALAT1结合位点的中心位置为准，将结合位点的基因组定位向上下游各扩展5kb，寻找扩展后结合位点区域内的基因，作为长非编码RNA的候选靶基因。

[0118] 2、结合步骤三中获得的染色体空间相互作用数据，得到与长非编码RNA结合位点在空间上有相互作用的基因组区域，将与长非编码RNA----MALAT1结合位点在空间上有相互作用的基因组区域向上下游各扩展5kb，寻找扩展后基因组区域内的基因，作为长非编码RNA远程调控的候选靶基因。

[0119] 3、分别计算长非编码RNA----MALAT1与步骤1和步骤2获得的候选靶基因表达水平的皮尔森相关系数，选择皮尔森相关系数绝对值大于0.3的候选靶基因作为长非编码RNA----MALAT1的靶基因。

[0120] 最终预测到的长非编码RNA----MALAT1的靶基因共有477个，具体信息如表3所示。顺式调控类型是指MALAT1直接与靶基因所在的基因组区域结合；远程调控类型是指与MALAT1结合的基因组区域在空间上与靶基因相互作用。

[0121] 表3、预测MALAT1的靶基因

[0122]

靶基因	调控类型	相关系数	靶基因	调控类型	相关系数
AAGAB	顺式,远程	0.500161981	MTMR10	远程	0.412743161
ACBD5	远程	0.495230782	MUC1	顺式,远程	0.505907258
ACD	远程	-0.344970944	MVD	顺式,远程	-0.382255974
ACER3	远程	0.30127658	MYLK4	远程	0.353052785
ACP1	远程	0.483004398	MYO5A	远程	0.303020655
ACP7	远程	0.406176878	NAALADL1	顺式,远程	-0.391271537
ADCY1	顺式,	0.374031421	NAT16	远程	0.502569459
ADGRG1	远程	0.434192597	NAXE	顺式,	0.415708305
AGAP1	远程	0.455775123	NCOA7	远程	0.324270754
AHCYL2	远程	0.403620018	NDUFS6	远程	0.348805492
AK6	远程	0.310370135	NECAB1	远程	0.394754137
ALG3	远程	-0.308066198	NEDD8	顺式,远程	0.362844298
ALKBH1	顺式,远程	-0.342560566	NELFB	顺式,远程	-0.340994039
ANAPC11	远程	-0.331667257	NEMF	远程	0.71645178
ANAPC13	远程	0.631123741	NEMP1	远程	-0.309015479
ANAPC15	顺式,远程	0.306898328	NFASC	远程	0.413179102
ANGEL1	远程	-0.333232776	NGRN	远程	0.339440872
ANGPTL6	顺式,远程	-0.309349364	NKAIN4	顺式,远程	0.350807671
ANKIB1	远程	0.479237019	NME1	远程	0.342391296
ANKRD17	远程	0.403113576	NME1-NME2	远程	0.342391296
ANKS1B	远程	0.335241858	NMRAL1	远程	0.318729452
ANP32E	顺式,远程	0.428228886	NR1H2	顺式,远程	-0.357581002

	ANXA9	远程	0.416325205	NR4A1	远程	-0.30575213
	AP3B1	远程	0.320240949	NRBP2	顺式,远程	0.312114672
	API5	远程	0.376313634	NRCAM	远程	0.413846961
	ARAP1	顺式,远程	-0.40243574	NRG3	远程	0.379766512
	ARF3	远程	0.394249985	NTNG1	远程	0.36712777
	ARFGEF2	远程	0.465475172	NUDT16	远程	0.734691295
	ARHGAP12	远程	0.617216537	NUDT4	远程	0.513692951
	ARHGAP9	远程	-0.306077716	NUP62	顺式,远程	-0.3337404
	ARID3B	远程	-0.34433058	OCIAD1	远程	0.614698813
	ARID4A	远程	0.443521357	OTOL1	远程	0.458220796
	ARL3	远程	0.305536466	OTUD7B	远程	0.374024227
	ARRDC1	远程	-0.345683829	PABPN1	远程	0.345142855
	ARSG	远程	-0.315523071	PAFAH1B1	远程	0.434671433
	ATG16L2	远程	-0.333583556	PAM16	远程	-0.383926298
	ATP1A1	顺式,远程	0.655119815	PARD6B	远程	0.50379384
	ATP1B1	远程	0.697534104	PAX8	顺式,远程	0.608195837
	ATP5S	远程	0.393052475	PDE6B	远程	0.382566619
	ATP9A	远程	0.400632879	PDE8A	顺式,远程	0.414377765
	B3GALNT1	远程	0.422392559	PDZK1	远程	0.562147755
	B4GALT3	远程	-0.330438921	PEX11B	顺式,远程	0.460460262
	B4GAT1	顺式,远程	0.325123942	PEX6	远程	-0.321916337
	BACE1	远程	0.407835838	PHF19	远程	-0.394131841
	BANP	远程	-0.466801028	PI4K2A	远程	0.39851227
	BARD1	远程	-0.345375515	PIP4K2C	远程	0.47576243
	BBC3	顺式,远程	-0.33067529	PLAGL2	远程	-0.308249797
[0123]	BBS1	顺式,远程	0.376260375	PLEKHA6	顺式,远程	0.489016455
	BCAP29	远程	0.317423543	PLEKHO1	远程	-0.39632858
	BCAS2	顺式,远程	0.477332897	PNPLA8	远程	0.398185945
	BCL2L2	远程	0.345142855	POLD1	顺式,远程	-0.322245939
	BCL7A	远程	0.427325823	POLE	远程	-0.327920151
	BICRA	顺式,远程	-0.371688843	POLL	远程	-0.321364947
	BLOC1S2	远程	0.465720257	PPIAL4D	顺式,	0.382748806
	BLVRB	顺式,远程	-0.345744262	PPIAL4E	顺式,	0.327294859
	BRI3	远程	-0.381294477	PPIAL4F	顺式,	0.382748806
	BRINP3	远程	0.308187877	PPM1E	远程	0.380536123
	BRMS1L	远程	0.382465291	PPM1H	远程	0.414258707
	BRWD1	远程	0.333425543	PPP1R15A	远程	-0.318825045
	C11orf68	远程	-0.363641117	PPP1R9A	远程	0.433416022
	C22orf46	远程	0.400874523	PPP2R5B	远程	0.315963398
	C2CD4D	远程	0.337719604	PRDX2	远程	0.304887691
	C3orf14	远程	0.309975604	PRDX5	远程	0.354796032
	C6orf89	远程	0.355869989	PRELID3B	远程	0.48522698
	C8orf44	远程	0.331135744	PRMT3	远程	0.432323068
	C8orf46	远程	0.333718595	PRPF31	顺式,远程	0.340565688
	C9orf66	远程	0.570050476	PRR15L	远程	0.570783493
	CABLES1	远程	0.557963584	PRR19	远程	-0.307382627
	CACNG8	顺式,远程	0.358286234	PRR35	远程	0.409984453
	CADPS	远程	0.319552213	PRRG2	远程	0.404484462
	CALM2	远程	0.382325042	PSMB5	顺式,远程	0.318466965
	CAMSAP2	远程	0.463712492	PSMD1	远程	0.442236812
	CANX	远程	0.343184753	PTER	远程	0.490620596

[0124]

CAPN12	顺式,远程	0.420898959	PTGES3	远程	-0.364406986
CBX5	顺式,远程	0.40884555	PTPRG	远程	0.458511168
CCDC117	顺式,远程	-0.334112085	PTPRT	远程	0.346719197
CCDC47	远程	0.416522029	PTRH2	远程	0.302023748
CCNT1	远程	-0.314164848	PTRHD1	远程	-0.407041769
CD14	远程	-0.30602677	PWP1	远程	0.361254709
CD55	远程	-0.351953352	PWWP2B	远程	-0.3070857
CDC123	远程	0.304523031	RAB11FIP3	远程	0.49660099
CDCA4	远程	-0.328388211	RAB13	远程	0.327709724
CDH4	远程	0.300889248	RAB1A	远程	0.4840911
CDK2AP1	远程	0.348514308	RAB22A	远程	0.382361204
CEBPB	远程	-0.382322864	RAB3IP	远程	0.430645745
CHD6	顺式,远程	0.322094324	RAB4B	顺式,远程	-0.377990772
CHMP2A	顺式,远程	0.510550994	RABL3	远程	0.316443203
CIRBP	远程	0.401004265	RACK1	顺式,	-0.392800441
CLCN3	顺式,	0.525289984	RANBP3	远程	-0.359441626
CLEC18B	顺式,	0.482999448	RASGRF1	远程	0.300130915
CLK3	顺式,远程	-0.448196641	RBM25	顺式,远程	0.445458602
CLN8	远程	0.463121682	RBM3	远程	0.314778074
CLTC	远程	0.57366491	RBP5	远程	0.505939762
CNNM4	远程	-0.311232693	RENBP	远程	0.389482125
CNTNAP4	远程	0.301730844	REXO4	远程	-0.362495114
COLGALT2	远程	0.34441083	RFX1	远程	-0.35988632
COPS6	顺式,远程	0.313171355	RGS7BP	远程	0.43943578
CPE	远程	0.33949655	RHBG	远程	0.512262513
CPSF1	远程	-0.309327898	RIMS2	远程	0.380411457
CPT1A	远程	0.331725061	RNMT	远程	0.374082805
CXCL16	远程	-0.321490728	ROM1	远程	0.364934597
CXorf40B	顺式,远程	0.350288702	RPA3	远程	0.390231664
CXXC5	远程	0.327344065	RPL13	顺式,远程	-0.383870606
CYCS	远程	0.332946495	RPL18	顺式,远程	-0.329186989
CYP24A1	远程	0.549616423	RPL18A	远程	-0.409732895
CYP27B1	远程	0.600946335	RPL28	远程	-0.394712035
DDX42	顺式,远程	0.332770957	RPL30	远程	-0.449656589
DDX5	顺式,远程	0.455926618	RPL31	远程	-0.486242844
DEDD2	顺式,远程	-0.350755692	RPL35A	远程	-0.484585632
DENND2C	顺式,远程	0.477332897	RPL37	顺式,	-0.422587783
DNMT3L	远程	0.438040708	RPL37A	远程	-0.495394864
DOCK1	远程	0.437767456	RPL41	远程	-0.449783528
DOK5	远程	0.37680989	RPL7L1	远程	0.395487535
DOK6	远程	0.352255093	RPLP1	远程	-0.463445449
DPF2	顺式,远程	-0.405757828	RPS12	远程	-0.384329782
DPM1	远程	0.311472103	RPS19	远程	-0.43414249
DPP3	顺式,远程	0.376260375	RPS2	远程	-0.461698727
DSCAM	顺式,远程	0.320098784	RPS21	顺式,远程	-0.505599698
DUS4L	远程	0.68872245	RPS24	顺式,远程	-0.305839927
DYNLL2	远程	0.42253938	RPS26	远程	-0.443361575
EDC4	远程	-0.31908803	RPS5	远程	-0.527856745
EEF1AKMT3	远程	0.351733945	RPS6KB2	顺式,	-0.309725666
EEF2	远程	-0.343234903	RPS9	远程	-0.423771303
EGLN2	远程	-0.377990772	RPUSD3	远程	-0.325361249

[0125]

EID2	远程	0.480401275	RRP36	顺式,远程	0.499989958
EIF1AD	顺式,远程	-0.315841713	RTKN	远程	0.413571477
EMC2	远程	0.370481216	RTN4	远程	0.411382607
EMX1	远程	0.624647381	S100A2	远程	0.451833547
ENOPH1	远程	0.444265864	S100A5	远程	0.380193208
ENSA	顺式,远程	0.473832664	SAP18	顺式,远程	0.359681336
ERLEC1	远程	0.449996502	SCN4B	远程	0.405207559
ERN1	远程	-0.380323899	SCN8A	远程	0.370092127
ESF1	远程	0.408906826	SCNN1A	远程	0.569361244
EWSR1	顺式,远程	0.318082098	SCRIB	顺式,远程	-0.30333378
EXOSC5	远程	-0.300898555	SDC4	远程	0.350569053
FAM117A	远程	-0.383870821	SDCCAG8	远程	0.437472652
FAM136A	远程	0.326884648	SEC14L2	远程	0.372146792
FAM168A	顺式,远程	0.306074869	SEC31B	远程	0.498611896
FAM219A	顺式,	0.344514426	SERPINF1	远程	-0.402517609
FAM84B	远程	0.599081505	SFXN2	顺式,远程	0.580135854
FAM89B	顺式,远程	-0.448158247	SGK1	远程	0.386972018
FANCA	顺式,远程	-0.316878903	SGK3	远程	0.331135744
FANCG	远程	-0.325498751	SH3YL1	远程	0.311439121
FAU	顺式,远程	-0.554464194	SIAH2	远程	-0.482180507
FBXO46	远程	-0.516666385	SIKE1	顺式,远程	0.319198906
FCGRT	远程	-0.324952891	SIVA1	远程	-0.512981269
FKBP2	远程	0.336554971	SKA2	远程	0.592522612
FRMD4A	远程	0.325375609	SLC1A2	远程	0.367175952
FUS	远程	0.391711484	SLC22A20	远程	0.395852053
G6PC3	顺式,远程	0.454780457	SLC22A6	远程	0.620154812
GABBR2	远程	0.347738018	SLC25A36	远程	0.589753822
GAS8	顺式,远程	0.36785769	SLC25A44	远程	0.39586848
GEMIN8	远程	0.574064886	SLC2A8	顺式,	-0.353396075
GFAP	远程	0.390951273	SLC34A3	顺式,远程	0.483047358
GGT6	远程	0.434880997	SLC35G2	顺式,远程	0.395610137
GHITM	远程	0.432493673	SLC39A4	远程	0.427290538
GLS	远程	0.50281162	SLC39A5	远程	0.354747103
GLYR1	远程	0.367459264	SLC4A1	远程	0.482726072
GMPR2	顺式,远程	0.348677806	SMARCD2	顺式,远程	-0.305331738
GNB2	顺式,远程	-0.341644141	SMIM13	远程	0.32330902
GOLM1	顺式,远程	0.711977588	SMIM5	顺式,远程	0.528992597
GPATCH4	顺式,	0.445773799	SNAI3	顺式,远程	-0.302743304
GPI	顺式,远程	0.328211468	SNRPA	远程	-0.300463399
GRAMD1A	顺式,远程	-0.314550612	SNRPD2	远程	-0.329331857
GTF2I	远程	0.372348923	SNX30	远程	0.306803619
HARS	远程	0.448411228	SORT1	远程	0.522769691
HARS2	远程	0.394211246	SPDYC	顺式,远程	-0.382859237
HDAC5	顺式,远程	-0.449483792	SPTLC1	顺式,远程	0.374472235
HDDC2	远程	0.562923228	SRP9	顺式,远程	0.587109499
HELZ2	顺式,远程	-0.308196544	SRSF5	顺式,	0.457276128
HES4	远程	-0.316397794	STARD10	顺式,远程	-0.40243574
HIP1R	远程	0.523338087	STARD8	远程	0.369307134
HNRNPDL	远程	0.372558303	STK11	远程	-0.354252279
HOMER1	远程	0.319301244	STXBP3	远程	0.4126237
HPF1	顺式,	0.508969091	SV2C	远程	0.347057003

[0126]

HSD11B2	远程	0.561582403	SYAP1	远程	0.349578326
HSD17B14	远程	0.354829569	SYT6	远程	0.59944465
HSF1	顺式,远程	-0.385344758	TADA3	远程	-0.334504497
IBTK	远程	0.322784458	TAF4	远程	-0.315944247
ICOSLG	远程	0.437174236	TANC1	远程	0.301139399
IDS	远程	0.30976769	TATDN1	远程	0.326738829
IGSF3	远程	0.372995452	TCIRG1	远程	-0.360779315
IL4I1	顺式,远程	-0.3337404	TERF2IP	远程	0.320354925
INAFM1	远程	-0.335686976	TESK2	远程	-0.303183708
INPP4B	远程	-0.356150905	THEM4	远程	0.358535287
INTS8	远程	-0.369009209	TMCO6	远程	-0.503824483
IRF7	远程	-0.327677395	TMEM139	远程	0.348796265
KCNC3	顺式,远程	0.558741913	TNRC6A	远程	0.392074697
KCNH6	远程	0.371259258	TPRN	顺式,远程	0.404076034
KCNK6	顺式,远程	-0.399557341	TRAPPC6B	顺式,远程	0.535307262
KCNQ3	远程	0.362358866	TREH	远程	0.591048996
KDM6B	远程	-0.418942118	TRIM28	顺式,远程	-0.340674557
KIAA0100	远程	0.404460748	TRIR	远程	-0.328384379
KIF5A	远程	0.381204208	TRPM6	远程	0.36652241
KLHL32	远程	0.318706388	TRPM7	顺式,远程	0.387609931
KLK1	顺式,远程	0.523702183	TSEN34	远程	-0.325360772
KLK11	顺式,远程	-0.315972429	TSFM	远程	0.348210036
L2HGDH	远程	0.488753513	TTC33	远程	0.388905824
LAG3	远程	-0.305570733	TTC39C	远程	-0.33298613
LAMTOR5	远程	0.529993121	UBE3D	远程	0.416786153
LENG9	顺式,远程	-0.366832567	UBN1	远程	-0.384041735
LEO1	远程	0.394372236	UCKL1	顺式,远程	-0.348926872
LIX1L	顺式,远程	-0.384090598	ULK1	远程	-0.317498448
LRCH4	顺式,远程	-0.322276564	UNC13D	远程	-0.310605629
LRRC3	远程	-0.439278729	UPF1	远程	-0.335187692
LRRN2	远程	0.545492921	UXT	远程	-0.36444732
LYL1	远程	-0.321649634	VAMP1	远程	0.35414018
LYRM7	远程	0.425699364	VAPB	远程	0.587136701
MACROD2	远程	0.368413948	VAV3	远程	0.464214114
MAGI3	远程	0.352340418	VPS13B	远程	0.30411564
MAML3	远程	-0.393683878	VPS33B	远程	0.356911864
MAP2K2	远程	-0.313272068	VPS36	远程	0.373359435
MAP3K13	远程	0.448267549	VPS45	远程	0.513388609
MAP3K3	顺式,远程	-0.34907523	VTI1A	远程	0.382881359
MAP7	远程	0.453021806	WASF3	远程	0.416369617
MAT2A	远程	0.345536902	WDR33	远程	-0.327020004
MAZ	顺式,远程	-0.381393295	WDR72	远程	0.599648318
MBD6	顺式,远程	-0.380175587	WNT1	远程	-0.318976046
MCM7	顺式,远程	-0.327684376	YKT6	远程	0.366872434
MECOM	远程	0.476178239	YPEL3	远程	-0.339393287
MED29	顺式,远程	0.338534513	YTHDF1	远程	0.376575164
METTL1	远程	0.341261944	YWHAG	远程	0.365465951
METTL2A	顺式,远程	0.481716088	ZC3H10	远程	-0.322372496
METTL2B	远程	0.394987442	ZCCHC6	远程	-0.300733735
MFAP1	远程	0.360952155	ZFPM1	远程	-0.316791483
MFSD3	顺式,远程	0.345677406	ZKSCAN1	远程	0.535941641

[0127]

MGEA5	远程	0.396980213	ZMAT2	远程	0.366802588
MIA	顺式,远程	-0.377990772	ZMYND8	远程	0.53152269
MLF2	远程	0.35662845	ZNF148	远程	0.415538207
MLLT11	远程	0.321143079	ZNF480	远程	0.435262716
MOB1B	远程	0.694843371	ZNF517	远程	-0.305665982
MPPE1	远程	0.361516046	ZNF524	远程	-0.392913453
MPPED2	远程	0.404896451	ZNF526	远程	-0.432606181
MRPL14	远程	-0.360235161	ZNF620	远程	0.367849774
MRPL9	远程	0.324927619	ZNF621	远程	0.381775623
MRPS23	远程	0.51062024	ZNF629	顺式,远程	0.57593556
MSH2	远程	0.316300393	ZNF672	顺式,远程	-0.379530818
MSI2	顺式,远程	0.370781654	ZNRF3	远程	0.498344975
MTFP1	远程	0.372146792			

[0128] 五、GO功能富集分析

[0129] 将表3中预测的MALAT1的靶基因与GO term中的基因进行比较，通过超几何分布检验基因富集的显著性，并且按照FDR排序，得到靶基因富集最显著的15个GO term(表4)。通过本发明的方法预测MALAT1具有如下功能：1)参与mRNA、rRNA等转录后加工代谢过程；2)mRNA翻译调控；3)与蛋白质结合；4)与具有多聚A尾的RNA结合；5)基于SRP的膜靶向共翻译蛋白；6)病毒转录。文献“Hutchinson等,A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35splicing domains.2007.BMC Genomics 8:39;Bernard等,A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression.2010.EMBO J.29:3082-3093”中已经证实MALAT1在细胞核内能够与其他蛋白质结合参与mRNA的转录后加工代谢过程。与本发明的预测结果一致，说明本发明基于长非编码RNA结合位点和染色体空间结构信息来预测长非编码RNA生物学功能的方法准确、可靠。

[0130] 表4、靶基因富集最显著的10个GO term

[0131]

GO条目	功能描述	P值	FDR值
GO:0006614	基于SRP的膜靶向共翻译蛋白	1.58E-13	2.69E-10
GO:0019083	病毒转录	5.04E-12	8.56E-09
GO:0000184	核转录mRNA代谢过程	1.61E-11	2.74E-08
GO:0005840	核糖体	1.87E-11	2.63E-08
GO:0005654	核质	1.36E-10	1.92E-07
GO:0006413	翻译起始	2.25E-10	3.82E-07
GO:0006412	翻译	2.55E-10	4.34E-07
GO:0044822	多聚A尾RNA结合	4.09E-10	6.08E-07
GO:0003735	核糖体结构性组成	4.84E-10	7.21E-07
GO:0005829	细胞溶质	1.02E-07	1.44E-04
GO:0006364	rRNA加工	1.11E-07	1.89E-04
GO:0022625	细胞溶质核糖体大亚基	7.44E-07	0.001048601
GO:0015935	核糖体小亚基	2.96E-06	0.004167096
GO:0005515	与蛋白质结合	4.16E-06	0.00619447

GO:0022627	细胞溶质核糖体小亚基	1.96E-05	0.027647567

序列表

<110>中国科学院生物物理研究所

<120>基于染色体空间相互作用预测长非编码RNA生物学功能的方法

<160>1

<170>PatentIn version 3.5

<210>1

<211>8302

<212>DNA

<213>人工序列(Artificial Sequence)

<400>1

cgcagcctgc agcccgagac ttctgtaaag gactggggcc ccgcaactgg cctctcctgc 60
cctcttaagc gcagcgccat tttagcaacg cagaagcccg gcgccgggaa gcctcagctc 120
gcctgaaggc aggtcccctc tgacgcctcc gggagcccag gtttcccaga gtccttggga 180
cgcagcgacg agttgtgctg ctatcttagc tgtccttata ggctggccat tccaggtggt 240
ggtatttaga taaaaccact caaactctgc agtttggtct tggggtttgg aggaaagctt 300
ttatttttct tcctgctccg gttcagaagg tctgaagctc atacctaacc aggcataaca 360
cagaatctgc aaaacaaaaa cccctaaaaa agcagaccca gagcagtgta aacacttctg 420
ggtgtgtccc tgactggctg cccaaggtct ctgtgtcttc ggagacaaag ccattcgctt 480
agttggtcta ctttaaaagg ccacttgaac tcgctttcca tggcgatttg ccttgtgagc 540
actttcagga gagcctggaa gctgaaaaac ggtagaaaaa tttccgtgcg ggccgtgggg 600
ggctggcggc aactgggggg ccgcagatca gagtgggcca ctggcagcca acggcccccg 660
gggctcaggc ggggagcagc tctgtggtgt gggattgagg cgttttccaa gagtgggttt 720
tcacgtttct aagatttccc aagcagacag cccgtgctgc tccgatttct cgaacaaaaa 780
agcaaaacgt gtggctgtct tgggagcaag tcgcaggact gcaagcagtt gggggagaaa 840
gtccgccatt ttgccacttc tcaaccgtcc ctgcaaggct ggggctcagt tgcgtaatgg 900
aaagtaaagc cctgaactat cacactttaa tcttccttca aaaggtggta aactatacct 960
actgtccctc aagagaacac aagaagtgct ttaagaggcg gcggaaggtg atcgaattcc 1020
ggtgatgcga gttgttctcc gtctataaat acgcctcgcc cgagctgtgc ggtaggcatt 1080
gaggcagcca gcgcaggggc ttctgctgag ggggcaggcg gagcttgagg aaaccgcaga 1140
taagtttttt tctctttgaa agatagagat taatacaact acttaaaaaa tatagtcaat 1200
aggttactaa gatattgctt agcgttaagt ttttaacgta attttaatag cttaagattt 1260
taagagaaaa tatgaagact tagaagagta gcatgaggaa ggaaaagata aaaggtttct 1320
aaaacatgac ggaggttgag atgaagcttc ttcatggagt aaaaaatgta tttaaaagaa 1380
aattgagaga aaggactaca gagccccgaa ttaataccaa tagaagggca atgcttttag 1440
attaaaatga aggtgactta aacagcttaa agtttagttt aaaagttgta ggtgattaaa 1500
ataatttgaa ggcgatcttt taaaaagaga ttaaaccgaa ggtgattaaa agaccttgaa 1560
atccatgacg cagggagaat tgcgtcattt aaagcctagt taacgcattt actaaacgca 1620
gacgaaaatg gaaagattaa ttgggagtgg taggatgaaa caatttggag aagatagaag 1680

tttgaagtgg aaaactggaa gacagaagta cgggaaggcg aagaaaagaa tagagaagat 1740
agggaaatta gaagataaaa acatactttt agaagaaaaa agataaattt aaacctgaaa 1800
agtaggaagc agaagaaaaa agacaagcta ggaaacaaaa agctaagggc aaaatgtaca 1860
aacttagaag aaaattggaa gatagaaaca agatagaaaa tgaaaatatt gtcaagagtt 1920
tcagatagaa aatgaaaaac aagctaagac aagtattgga gaagtataga agatagaaaa 1980
atataaagcc aaaaattgga taaaatagca ctgaaaaaat gaggaaatta ttggtaacca 2040
atttatttta aaagcccatc aatttaattt ctggtggtgc agaagttaga aggtaaagct 2100
tgagaagatg agggtgttta cgtagaccag aaccaattta gaagaatact tgaagctaga 2160
aggggaagtt ggttaaaaat cacatcaaaa agctactaaa aggactggtg taatttaaaa 2220
aaaactaagg cagaaggctt ttggaagagt tagaagaatt tggaaggcct taaatatagt 2280
agcttagttt gaaaaatgtg aaggactttc gtaacggaag taattcaaga tcaagagtaa 2340
ttaccaactt aatgtttttg cattggactt tgagttaaga ttatttttta aatcctgagg 2400
actagcatta attgacagct gacccaggtg ctacacagaa gtggattcag tgaatctagg 2460
aagacagcag cagacaggat tccaggaacc agtgtttgat gaagctagga ctgaggagca 2520
agcgagcaag cagcagttcg tggtgaagat aggaaaagag tccaggagcc agtgcgattt 2580
ggtgaaggaa gctaggaaga aggaaggagc gctaacgatt tggtggtgaa gctaggaaaa 2640
aggattccag gaaggagcga gtgcaatttg gtgatgaagg tagcaggcgg cttggcttgg 2700
caaccacacg gaggaggcga gcaggcgttg tgcgtagagg atcctagacc agcatgccag 2760
tgtgccaagg ccacagggaa agcgagtggt tggtaaaaat ccgtgaggtc ggcaatatgt 2820
tgtttttctg gaacttactt atggtaacct tttatttatt ttctaatata atgggggagt 2880
ttcgtactga ggtgtaaagg gatttatatg gggacgtagg ccgatttccg ggtgttgtag 2940
gtttctcttt ttcaggctta tactcatgaa tcttgtctga agcttttgag ggcagactgc 3000
caagtcctgg agaaatagta gatggcaagt ttgtgggttt ttttttttta cacgaatttg 3060
aggaaaacca aatgaatttg atagccaaat tgagacaatt tcagcaaatc tgtaagcagt 3120
ttgtatgttt agttggggta atgaagtatt tcagttttgt gaatagatga cctgttttta 3180
cttcctcacc ctgaattcgt tttgtaaatg tagagtttgg atgtgtaact gaggcggggg 3240
ggagttttca gtatttttt ttgtgggggt gggggcaaaa tatgttttca gttctttttc 3300
ccttaggtct gtctagaatc ctaaaggcaa atgactcaag gtgtaacaga aaacaagaaa 3360
atccaatatc aggataatca gaccaccaca ggtttacagt ttatagaaac tagagcagtt 3420
ctcacgttga ggtctgtgga agagatgtcc attggagaaa tggctggtag ttactctttt 3480
ttccccccac ccccttaatc agactttaaa agtgcttaac cccttaaact tgttatttt 3540
tacttgaagc attttgggat ggtcttaaca gggaagagag agggtggggg agaaaatgtt 3600
tttttctaag attttccaca gatgctatag tactattgac aaactgggtt agagaaggag 3660
tgtaccgctg tgctgttggc acgaacacct tcagggactg gagctgcttt tatccttgga 3720
agagtattcc cagttgaagc tgaaaagtac agcacagtgc agctttggtt catattcagt 3780
catctcagga gaacttcaga agagcttgag taggccaaat gttgaagtta agttttccaa 3840
taatgtgact tcttaaaagt tttattaaag gggaggggca aatattggca attagttggc 3900
agtggcctgt tacggttggg attggtgggg tgggtttagg taattgttta gtttatgatt 3960
gcagataaac tcatgccaga gaacttaaag tcttagaatg gaaaaagtaa agaaatatca 4020

acttccaagt tggcaagtaa ctcccaatga tttagttttt ttccccccag tttgaattgg 4080
gaagctgggg gaagttaaat atgagccact gggtgtacca gtgcattaat ttgggcaagg 4140
aaagtgtcat aatttgatac tgtatctgtt ttccttcaaa gtatagagct tttggggaag 4200
gaaagtattg aactgggggt tggtctggcc tactgggctg acattaacta caattatggg 4260
aaatgcaaaa gttgtttgga tatggtagtg tgtggttctc ttttggaatt tttttcaggt 4320
gatttaataa taatttaaaa ctactataga aactgcagag caaaggaagt ggcttaatga 4380
tcctgaaggg atttcttctg atggtagctt ttgtattatc aaacttttt cagataacat 4440
cttctgagtc ataaccagcc tggcagtatg atggcctaga tgcagagaaa acagctcctt 4500
ggtgaattga taagtaaagg cagaaaagat tatatgtcat acctccattg gggaataagc 4560
ataaccctga gattcttact actgatgaga acattatctg catatgccaa aaaatttaa 4620
gcaaatgaaa gctaccaatt taaagttacg gaatctacca ttttaaagtt aattgcttgt 4680
caagctataa ccacaaaaat aatgaattga tgagaaatac aatgaagagg caatgtccat 4740
ctcaaaatac tgcttttaca aaagcagaat aaaagcgaaa agaaatgaaa atgttacact 4800
acattaatcc tggaataaaa gaagccgaaa taaatgagag atgagttggg atcaagtgga 4860
ttgaggaggc tgtgctgtgt gccaatgttt cgtttgcctc agacaggtat ctcttcgtta 4920
tcagaagagt tgcttcattt catctgggag cagaaaacag caggcagctg ttaacagata 4980
agtttaactt gcatctgcag tattgcatgt tagggataag tgcttatttt taagagctgt 5040
ggagttctta aatatcaacc atggcacttt ctcctgaccc cttccctagg ggatttcagg 5100
attgagaaat ttttccatcg agccttttta aaattgtagg acttgttcct gtgggcttca 5160
gtgatgggat agtacacttc actcagaggc atttgcatct ttaaataatt tcttaaaagc 5220
ctctaaagtg atcagtgcct tgatgccaac taaggaaatt tgtttagcat tgaatctctg 5280
aaggctctat gaaaggaata gcatgatgtg ctgttagaat cagatgttac tgctaaaatt 5340
tacatgttgt gatgtaaatt gtgtagaaaa ccattaaatc attcaaaata ataaactatt 5400
tttattagag aatgtatact tttagaaagc tgtctcctta tttaaataaa atagtgtttg 5460
tctgtagttc agtgttgggg caatcttggg ggggattctt ctctaatctt tcagaaactt 5520
tgtctgcgaa cactctttaa tggaccagat caggatttga gcggaagaac gaatgtaact 5580
ttaaggcagg aaagacaaat tttattcttc ataaagtgat gagcatataa taattccagg 5640
cacatggcaa tagaggccct ctaaataagg aataaataac ctcttagaca ggtgggagat 5700
tatgatcaga gtaaaaggta attacacatt ttatttccag aaagtcaggg gtctataaat 5760
tgacagtgat tagagtaata ctttttcaca tttccaaagt ttgcatgtta actttaaatg 5820
cttacaatct tagagtggta ggcaatgttt tacactattg accttatata gggaagggag 5880
ggggtgcctg tggggtttta aagaattttc ctttgcagag gcatttcatc cttcatgaag 5940
ccattcagga ttttgaattg catatgagtg cttggctctt ccttctgttc tagtgagtgt 6000
atgagacctt gcagtgagtt tatcagcata ctcaaaattt ttttcctgga atttggaggg 6060
atgggaggag ggggtggggc ttacttgttg tagctttttt ttttttaca gacttcacag 6120
agaatgcagt tgtcttgact tcaggtctgt ctgttctgtt ggcaagtaaa tgcagtactg 6180
ttctgatccc gctgctatta gaatgcattg tgaaacgact ggagtatgat taaaagttgt 6240
gttccccaat gcttggagta gtgattgttg aaggaaaaaa tccagctgag tgataaaggc 6300
tgagtgttga ggaaatttct gcagttttaa gcagtcgtat ttgtgattga agctgagtac 6360

attttgctgg tgtattttta ggtaaaatgc tttttgttca tttctggtgg tgggagggga 6420
ctgaagcctt tagtcttttc cagatgcaac cttaaaatca gtgacaagaa acattccaaa 6480
caagcaacag tcttcaagaa attaaactgg caagtggaaa tgtttaaaca gttcagtgat 6540
ctttagtgca ttgtttatgt gtgggtttct ctctcccctc ccttggtctt aattcttaca 6600
tgcaggaaca ctcagcagac acacgtatgc gaagggccag agaagccaga cccagtaaga 6660
aaaaatagcc tatttacttt aaataaacca aacattccat tttaaatgtg gggattggga 6720
accactagtt ctttcagatg gtattcttca gactatagaa ggagcttcca gttgaattca 6780
ccagtggaca aaatgaggaa aacaggtgaa caagcttttt ctgtatttac atacaaagtc 6840
agatcagtta tgggacaata gtattgaata gatttcagct ttatgctgga gtaactggca 6900
tgtgagcaaa ctgtgttggc gtgggggtgg aggggtgagg tgggcgctaa gccttttttt 6960
aagatttttc aggtacccct cactaaaggc accgaaggct taaagtagga caaccatgga 7020
gccttcctgt ggcaggagag acaacaaagc gctattatcc taaggtcaag agaagtgtca 7080
gcctcacctg atttttatta gtaatgagga cttgcctcaa ctccctcttt ctggagtgaa 7140
gcatccgaag gaatgcttga agtacccctg ggcttctctt aacatttaag caagctgttt 7200
ttatagcagc tcttaataat aaagcccaaa tctcaagcgg tgcttgaagg ggagggaaag 7260
ggggaaagcg ggcaaccact tttccctagc ttttccagaa gcctgttaaa agcaaggtct 7320
ccccacaagc aacttctctg ccacatcgcc accccgtgcc ttttgatcta gcacagaccc 7380
ttcacccctc acctcgatgc agccagtagc ttggatcctt gtgggcatga tccataatcg 7440
gtttcaaggt aacgatggtg tcgaggtctt tggtgggttg aactatgtta gaaaaggcca 7500
ttaatttgcc tgcaaattgt taacagaagg gtattaaaac cacagctaag tagctctatt 7560
ataatactta tccagtgact aaaaccaact taaaccagta agtggagaaa taacatgttc 7620
aagaactgta atgctgggtg ggaacatgta acttgtagac tggagaagat aggcatttga 7680
gtggctgaga gggcttttgg gtgggaatgc aaaaattctc tgctaagact ttttcaggtg 7740
aacataacag acttggccaa gctagcatct tagcggaagc tgatctccaa tgctcttcag 7800
tagggtcatg aaggtttttc ttttcctgag aaaacaacac gtattgtttt ctcaggtttt 7860
gctttttggc ctttttctag cttaaaaaaa aaaaaagcaa aagatgctgg tggttggcac 7920
tcctggtttc caggacgggg ttcaaatccc tgcggcgtct ttgctttgac tactaatctg 7980
tcttcaggac tctttctgta tttctccttt tctctgcagg tgctagttct tggagttttg 8040
gggaggtggg aggtaacagc acaatatctt tgaactatat acatccttga tgtataattt 8100
gtcaggagct tgacttgatt gtatattcat atttacacga gaacctaata taactgcctt 8160
gtcttttca ggtaatagcc tgcagctggt gttttgagaa gccctactgc tgaaaactta 8220
acaattttgt gtaataaaaa tggagaagct ctaaattgtt gtggttcttt tgtgaataaa 8280
aaaatcttga ttggggaaaa aa 8302