(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103710336 B
(45)授权公告日 2017.02.22

(21)申请号 201210379402.8

(22)申请日 2012.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 103710336 A

(43)申请公布日 2014.04.09

(73)专利权人 深圳华大基因科技服务有限公司
地址 518083 广东省深圳市盐田区北山工业区综合楼科技创业园201

(72)发明人 祝珍珍 黄文潘 章文蔚 陈茂山 张艳艳

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东

(51)Int.Cl.
C12N 15/10(2006.01)
C40B 50/06(2006.01)
C40B 40/08(2006.01)
C12Q 1/68(2006.01)
C40B 60/14(2006.01)
C12M 1/34(2006.01)

(56)对比文件

Elitza Deltcheva等.CRISPRRNAmaturationbytrans-encoded small RNA and host factor RNase III.《NATURE》.2011,第471卷全文.

Carsten Kröger等.The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium.《PNAS》.2012,全文.

Jan Mitschke等.Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120.《PNAS》.2011,第108卷(第50期),全文.

Cynthia M. Sharma等.The primary transcriptome of the major human pathogen Helicobacter pylori.《nature》.2010,第464卷全文.

审查员 张锦广

权利要求书3页 说明书13页
序列表1页 附图9页

(54)发明名称

从RNA样本富集转录本的方法及其用途

(57)摘要

本发明提出了从RNA样本富集转录本的方法及其用途。从RNA样本富集转录本的方法,包括:利用富集试剂对RNA样本进行处理,以便富集转录本,其中,所述富集试剂具有5'-单磷酸外切酶活性,所述转录本为在其5'末端具有帽子结构或三磷酸基团的RNA分子。利用该方法能够有效地富集转录本。

CN 103710336 B

1.一种确定转录起点的方法,其特征在于,包括:

从宿主提取RNA样本,利用富集试剂对所述RNA样本进行处理,以便富集转录本,其中,所述富集试剂具有5'-单磷酸外切酶活性,所述转录本为在其5'末端具有帽子结构或三磷酸基团的RNA分子;

对处理后的RNA样本进行测序文库构建,包括,

去除所述转录本的帽子结构或三磷酸基团,以便获得去除帽子结构或三磷酸基团的转录本,

在去除帽子结构或三磷酸基团的转录本的5'末端连接RNA接头,以便获得连接有RNA接头的转录本,

对连接有RNA接头的转录本进行反转录,以便获得与所述转录本对应的cDNA,所述反转录的反转录引物在其末端具有与所述RNA接头相对应的序列,所述反转录采用序列如SEQ ID NO:1所示的寡核苷酸5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN-3'作为反转录引物,所述反转录引物中至少一个N被硫代修饰,

对所述cDNA进行扩增,以便获得扩增产物,

基于所述扩增产物,构建测序文库;

对所述测序文库进行测序,获得由多个测序序列构成的测序结果;

基于所述测序结果,确定转录起点,包括,

将所述测序结果与参考序列进行比对,所述参考序列中包含预定基因的5'-UTR序列的至少一部分:针对原核宿主,所述参考序列包含所述预定基因的翻译起始位点与该翻译起始位点上游700bp位点之间的核酸序列,针对真核宿主,所述参考序列包含所述预定基因的翻译起始位点与该翻译起始位点上游5000bp位点之间的核酸序列,

基于获得的比对结果,选择能够和所述参考序列对上、且在所述参考序列最上游的测序序列作为阳性序列,对所述阳性序列进行筛选,所述筛选的原则是:所述阳性序列的数目是比对到所述预定基因内部的测序序列数目平均值的N倍以上,其中所述N为至少10的实数,

确定获得的筛选结果中的阳性序列的第一位碱基作为所述转录起始位点。

2.根据权利要求1所述的方法,其特征在于,利用末端修整试剂去除所述转录本的帽子结构或三磷酸基团,其中,

所述末端修整试剂具有烟草酸焦磷酸酶活性。

3.根据权利要求1所述的方法,其特征在于,所述反转录引物中倒数第二个N被硫代修饰。

4.根据权利要求1所述的方法,其特征在于,所述测序利用Illumina Hiseq2000、Genome Analyzer、SOLiD测序系统、Ion Torrent、Ion Proton、454、PacBio RS测序系统、Helicos tSMS技术以及纳米孔测序技术的至少一种进行。

5.根据权利要求1所述的方法,其特征在于,所述RNA样本为宿主的总RNA的至少一部分。

6.根据权利要求1所述的方法,其特征在于,采用SOAP Alignment进行所述比对。

7.根据权利要求1所述的方法,其特征在于,进一步包括对筛选结果进行卡方检验,所述卡方检验的检验值为3.84以上。

8.一种确定转录起点的系统，其特征在于，包括：

样品提取设备，所述样品提取装置用于从宿主提取RNA样本；

核酸样本测序设备，所述核酸样本测序设备与所述样品提取装置相连，所述核酸样本测序设备包括文库构建装置和测序装置，

所述文库构建装置包括，

转录本富集单元，所述转录本富集装置中设置有富集试剂，所述富集试剂具有5’-单磷酸外切酶活性，以便从RNA样本富集转录本，

末端修整单元，所述末端修整单元与所述转录本富集单元相连，并且适于去除所述转录本的5’帽子结构或5’三磷酸，以便获得去除5’帽子结构或5’三磷酸的转录本，

RNA接头连接单元，所述RNA接头连接单元与末端修整单元相连，并且适于在去除5’帽子结构或5’三磷酸的转录本的5’末端连接RNA接头，以便获得连接有RNA接头的转录本，

反转录单元，所述反转录单元与所述RNA接头连接单元相连，并且适于对连接有RNA接头的转录本进行反转录，以便获得与所述转录本对应的cDNA，所述反转录单元中设置有反转录引物，所述反转录引物在其末端具有与所述RNA接头引物相对应的序列，所述反转录引物的序列如SEQ ID NO:1所示：5’-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN-3’，所述反转录引物中至少一个N被硫代修饰，

扩增单元，所述扩增单元与所述反转录单元相连，并且适于对所述cDNA进行扩增，以便获得扩增产物，以及

文库构建单元，所述文库构建单元与所述扩增单元相连，并且适于基于所述扩增产物，构建测序文库，

所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得获得由多个测序序列构成的测序结果；以及

转录起点确定设备，所述转录起点确定装置与所述测序装置相连，并且适于基于所述测序结果，确定转录起点，所述转录起点确定设备包括：

比对装置，所述比对装置用于将所述测序结果与参考序列进行比对，所述参考序列中包含预定基因的5’-UTR序列的至少一部分，

确定装置，所述确定装置适于基于比对结果，确定所述转录起点，所述确定装置适于：选择能够和所述参考序列对上、且在所述参考序列最上游的测序序列作为阳性序列，所述确定装置进一步包括筛选单元，所述筛选单元适于对所述阳性序列进行筛选，其中所述筛选的原则是：所述阳性序列的数目是比对到所述预定基因内部的测序序列数目平均值的N倍以上，其中所述N为大于1的实数，

确定从所述筛选单元中获得的筛选结果中的阳性序列的第一位碱基作为所述转录起始位点。

9.根据权利要求8所述的系统，其特征在于，所述N为至少10的实数。

10.根据权利要求8所述的系统，其特征在于，所述末端修整单元中设置有末端修整试剂，其中，所述末端修整试剂具有烟草酸焦磷酸酶活性。

11.根据权利要求8所述的系统，其特征在于，所述反转录引物中倒数第二个N被硫代修饰。

12.根据权利要求8所述的系统，其特征在于，所述RNA接头连接单元中设置有连接试

剂，其中，所述连接试剂具有T4RNA连接酶活性。

13.根据权利要求8所述的系统，其特征在于，所述测序装置选自Illumina Hiseq2000、Genome Analyzer、SOLiD测序系统、Ion Torrent、Ion Proton、454、PacBio RS测序系统、Helicos tSMS系统以及纳米孔测序系统中的至少一种。

14.根据权利要求8所述的系统，其特征在于，所述比对装置适于采用SOAP Alignment进行所述比对。

15.根据权利要求8所述的系统，其特征在于，所述确定装置进一步包括检验单元，所述检验单元适于对筛选结果进行卡方检验，所述卡方检验的检验值为3.84以上。

从RNA样本富集转录本的方法及其用途

技术领域

[0001] 本发明涉及生物技术领域，具体的，本发明涉及从RNA样本富集转录本的方法及其用途，更具体的，本发明涉及从RNA样本富集转录本的方法、构建测序文库的方法、测序文库、核酸样本测序方法、确定转录起点(transcription start site,TSS)的方法、用于从RNA样本富集转录本的富集试剂、构建测序文库的装置、核酸样本测序设备以及确定TSS的系统。

背景技术

[0002] 基因的转录过程是从RNA聚合酶与DNA模板的启动子位置结合开始，然后从转录起点(transcription start site,在本文中简称为:TSS)进行转录延伸，最终形成完整的RNA。生物体内存在的RNA分子都是从TSS开始的，因此通过高通量测序研究TSS有助于我们从全基因组推测启动子的位置及结构，从而全局了解基因转录调控网络。TSS的研究也有助于修正原有的基因注释或发现新的基因。

[0003] 然而，目前对于TSS的研究，仍有待改进。

发明内容

[0004] 本发明旨在至少在一定程度上解决上述技术问题之一或至少提供一种有用的商业选择。为此，本发明的一个目的在于提出一种能够有效富集转录本，进而可以有效确定TSS的手段。

[0005] 本发明是基于发明人的下列发现而完成的：

[0006] 目前，关于高通量测序研究TSS的方法通常是针对具有帽子结构的RNA，采用CAGE或RACE的方法捕获RNA分子的5’末端。常见的有deepCAGE，PEAT，deep-RACE，nanoCAGE和CAGEscan。其中deepCAGE，PEAT，deep-RACE和CAGEscan需要酶切等繁琐操作，对RNA的要求量很高，而且产生的测序序列(reads)较短(大约20nt)，只适用于具有帽子结构的RNA，不能用于没有帽子结构的原核RNA的TSS的研究。尽管nanoCAGE操作比较简单，对RNA的使用量要求也低，但是也只适用于具有帽子结构的RNA，而且产生的数据中假阳性比较多。发明人发现通过采用5’单磷酸外切酶，能够特异性地降解5’单磷酸的RNA，保留具有5’帽子和5’三磷酸的完整的RNA分子，可以有效地应用于富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0007] 在本发明的第一方面，本发明提出了一种从RNA样本富集转录本的方法。根据本发明的实施例，该从RNA样本富集转录本的方法包括：利用富集试剂对RNA样本进行处理，以便富集转录本，其中，所述富集试剂具有5’-单磷酸外切酶活性，所述转录本为在其5’末端具有帽子结构或5’三磷酸的RNA分子。由于5’单磷酸外切酶，能够特异性地降解5’单磷酸的RNA，而不会降解具有5’帽子和5’三磷酸的完整的RNA分子，从而利用具有该5’单磷酸外切酶活性的富集试剂，可以有效地富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0008] 在本发明的第二方面，本发明提出了一种构建测序文库的方法。根据本发明的实施例，该构建测序文库的方法包括：根据前面所述的方法，从RNA样 本富集转录本；去除所述转录本的5’帽子结构或5’三磷酸，以便获得去除5’帽子结构或5’三磷酸的转录本；在去除5’帽子结构或5’三磷酸的转录本的5’末端连接RNA接头，以便获得连接有RNA接头的转录本；对连接有RNA接头的转录本进行反转录，以便获得与所述转录本对应的cDNA；对所述cDNA进行扩增，以便获得扩增产物；以及基于所述扩增产物，构建测序文库。由此，利用该方法，能够有效地针对核酸样本中所富集的转录本构建测序文库，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0009] 在本发明的第三方面，本发明提出了一种测序文库，其特征在于，是由前面所述的方法构建的。利用该测序文库，能够有效的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0010] 在本发明的第四方面，本发明提出了一种核酸样本测序方法。根据本发明的实施例，该核酸样本测序方法包括：根据前面所述的方法，构建测序文库；以及对所述测序文库进行测序，以便获得测序结果。利用该方法，能够有效的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0011] 在本发明的第五方面，本发明提出了一种用于确定TSS的方法。根据本发明的实施例，该确定TSS的方法包括：从宿主提取RNA样本；利用前面所述的方法，获得由多个测序序列构成的测序结果；以及基于所述测序结果，确定TSS。利用该方法，可以有效地确定转录起始位点，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0012] 在本发明的第六方面，本发明提出了一种用于从RNA样本富集转录本的富集试剂。根据本发明的实施例，富集试剂具有5’-单磷酸外切酶活性。利用该富集试剂，可以有效地富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0013] 在本发明的第七方面，本发明提出了一种构建测序文库的装置。根据本发明的实施例，该构建测序文库的装置包括：转录本富集单元，所述转录本富集装置中设置有前面所述的富集试剂，以便从RNA样本富集转录本；末端修整单元，所述末端修整单元与所述转录本富集单元相连，并且适于去除所述转录本的5’帽子结构或5’三磷酸，以便获得去除5’帽子结构或5’三磷酸的转录本；RNA接头连接单元，所述RNA接头连接单元与末端修整单元相连，并且适于在去除5’帽子结构或5’三磷酸的转录本的5’末端连接RNA接头，以便获得连接有RNA接头的转录本；反转录单元，所述反转录单元与所述RNA接头连接单元相连，并且适于对连接有RNA接头的转录本进行反转录，以便获得与所述转录本对应的cDNA；扩增单元，所述扩增单元与所述反转录单元相连，并且适于对所述cDNA进行扩增，以便获得扩增产物；以及文库构建单元，所述文库构建单元与所述扩增单元相连，并且适于基于所述扩增产物，构建测序文库。利用该装置，能够有效地针对核酸样本中所富集的转录本构建测序文库，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0014] 在本发明的第八方面，本发明提出了一种核酸样本测序设备，其特征在于，包括：文库构建装置，所述文库构建装置为前面所述的装置，以便针对核酸样本构建测序文库；以

及测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得测序结果。利用该装置，能够有效 的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0015] 在本发明的第九方面，本发明提出了一种确定TSS的系统。根据本发明的实施例，该系统包括：样品提取设备，所述样品提取设备用于从宿主提取RNA样本；核酸样本测序设备，所述核酸样本测序设备与所述样品提取设备相连，并且所述测序设备为前面所述的核酸样本测序设备，以便针对所述RNA样本进行测序，从而获得由多个测序序列构成的测序结果；以及TSS确定设备，所述TSS确定装置与所述测序设备相连，并且适于基于所述测序结果，确定TSS。根据本发明的实施例，利用该系统能够有效的确定核酸样本中的TSS。

[0016] 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0017] 本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

[0018] 图1显示了根据本发明一个实施例的构建测序文库的方法的流程示意图；

[0019] 图2显示了根据本发明一个实施例的确定TSS序列的信息学分析流程示意图；

[0020] 图3显示了根据本发明有一个实施例的确定TSS的系统的示意图；

[0021] 图4显示了根据本发明一个实施例的核酸样本测序设备的示意图；

[0022] 图5显示了根据本发明一个实施例的构建测序文库的装置的示意图；

[0023] 图6显示了根据本发明有一个实施例的确定TSS的设备的示意图；

[0024] 图7显示了根据本发明一个实施例，筛选后的TSS在基因组上的分布，上图和下图分别是人RNA和大肠杆菌RNA样品的TSS分布图，其中0是基因编码区的起始位点，其上游就是转录起始的位点，从图中可以看出，大部分的序列都落在基因编码区的上游；

[0025] 图8显示了根据本发明一个实施例，展示了8个人RNA样品的TSS图谱，从图中可以看到不同样品中TSS的分布情况；图9显示了根据本发明一个实施例，TSS上游的碱基分布图形，其中横坐标1对应的就是TSS的位置，以嘌呤为主(A/G)。上图是人RNA样品的TSS上游碱基分布图，有明显的GC富集区，这也是真核生物主要的启动子类型；下图是大肠杆菌RNA样品的TSS上游碱基分布图，在其上游-10区处也能找到典型的TATA盒；

[0026] 图10显示了根据本发明一个实施例，5’UTR的长度分布，也就是TSS到编码区的距离。上图是人RNA样品5’UTR的长度分布，下图是大肠杆菌RNA样品5’UTR的长度分布；

[0027] 图11显示了相关性分析可获得对实验结果可靠性和操作稳定性的评估，如图11所示，上图是人RNA样品的两次重复，下图是大肠杆菌RNA样品的两次重复；以及

[0028] 图12，显示了根据本发明的实施例预测基因的结果示意图。上图是人的两个基因NM_018997和NM_031901的TSS分布，他们是发生了可变剪切的基因，图中红色竖线表示筛选的TSS，黑色的竖线是过滤前得到的序列，蓝色横线代表基因的外显子，黄色横线是基因的内含子；下图是大肠杆菌一个操纵子的TSS分布，原核的不存在内含子，所以只有代表基因的蓝色横线，这个操纵子的4个基因共有一个TSS。

具体实施方式

[0029] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

[0030] 在本发明中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”、“固定”等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本发明中的具体含义。另外，在本文中所使用的术语“上游”“下游”是按照5’端至3’端的方向所确定的。

[0031] 本发明是基于发明人的下列发现而完成的：

[0032] 目前，关于高通量测序研究TSS的方法通常是针对具有帽子结构的RNA，采用CAGE或RACE的方法捕获RNA分子的5’末端。常见的有deepCAGE，PEAT，deep-RACE，nanoCAGE和CAGEscan。其中deepCAGE，PEAT，deep-RACE和CAGEscan需要酶切等繁琐操作，对RNA的要求量很高，而且产生的测序序列(reads)较短(大约20nt)，只适用于具有帽子结构的RNA，不能用于没有帽子结构的原核RNA的TSS的研究。尽管nanoCAGE操作比较简单，对RNA的使用量要求也低，但是也只适用于具有帽子结构的RNA，而且产生的数据中假阳性比较多。发明人发现通过采用5’单磷酸外切酶，能够特异性地降解5’单磷酸的RNA，保留具有5’帽子和5’三磷酸的完整的RNA分子，可以有效地应用于富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0033] 在本发明的第一方面，本发明提出了一种从RNA样本富集转录本的方法。根据本发明的实施例，该从RNA样本富集转录本的方法包括：利用富集试剂对RNA样本进行处理，以便富集转录本，其中，所述富集试剂具有5’-单磷酸外切酶活性，所述转录本为在其5’末端具有帽子结构或三磷酸的RNA分子。根据本发明实施例，具有5’单磷酸外切酶活性的酶的例子可以包括：核糖核酸外切酶XRN-1，Terminator™依赖于5’磷酸的核酸外切酶(Terminator™ 5′-Phosphate-Dependent Exonuclease)或者TAKARA™碱性磷酸酶(TAKARA™ Alkaline Phosphatase)。由于5’单磷酸外切酶，能够特异性地降解5’单磷酸的RNA，而不会降解具有5’帽子和5’三磷酸的完整的RNA分子，从而利用具有该5’单磷酸外切酶活性的富集试剂，可以有效地富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0034] 根据本发明的实施例，该从RNA样本富集转录本的方法可以采用任何具有5’单磷酸外切酶活性的富集试剂。根据本发明实施例，具有5’单磷酸外切酶活性的酶的例子可以包括：核糖核酸外切酶XRN-1，Terminator™依赖于5’磷酸的核酸外切酶或者TAKARA™碱性磷酸酶。根据本发明的一个实施例，所述富集试剂含有DNase I。由此，可以进一步提高降解5’单磷酸的RNA的特异性和效率，从而进一步提高富集转录本方法的效率。根据本发明的一个实施例，所述富集试剂还可以进一步含有缓冲液和可溶性盐，以便进一步提高DNase I的酶活性。根据本发明的一个实施例，所述富集试剂的pH为8.0。根据本发明的一个实施例，所述缓冲液为Tris-HCl，所述可溶性盐为选自氯化钠和氯化镁的至少一种。根据本发明的一

个实施例，在30摄氏度下，利用所述富集试剂对所述RNA样本进行处理。从而，可以进一步提高利用根据本发明实施例的富集试剂进行富集转录本的效率。根据本发明实施例，具有5’单磷酸 外切酶活性的酶的例子可以包括：核糖核酸外切酶XRN-1，Terminator™依赖于5’磷酸的核酸外切酶或者TAKARA™碱性磷酸酶。

[0035] 在本发明的第二方面，本发明提出了一种构建测序文库的方法。参考图1，根据本发明的实施例，该构建测序文库的方法包括：

[0036] S100(富集转录本)：根据前面所述的方法，从RNA样本富集转录本。关于该步骤，前面已经进行了详细描述，在此不再赘述。

[0037] S200(末端修整)：去除所述转录本的5’帽子结构或5’三磷酸，以便获得去除5’帽子结构或5’三磷酸的转录本。根据本发明的一个实施例，利用末端修整试剂去除所述转录本的5’帽子结构或5’三磷酸，其中，所述末端修整试剂具有烟草酸焦磷酸酶活性。根据本发明的一个实施例，所述修整试剂包含：烟草酸焦磷酸酶、可溶性盐、EDTA、β-巯基乙醇和Triton-X 100。根据本发明的一个实施例，所述可溶性盐为醋酸钠。根据本发明的一个实施例，所述修整试剂的pH为7.5。由此，可以进一步提高对RNA进行末端修整的效果，即能够有效的去除转录本的5’帽子结构或5’三磷酸，从而提高构建测序文库的效率。

[0038] S300(连接接头)：在去除5’帽子结构或5’三磷酸的转录本的5’末端连接RNA接头，以便获得连接有RNA接头的转录本。根据本发明的一个实施例，利用连接试剂，在去除5’帽子结构或5’三磷酸的转录本的5’末端连接RNA接头，其中，所述连接试剂具有T4RNA连接酶活性。根据本发明的一个实施例，所述连接试剂包含：T4RNA连接酶，缓冲液、可溶性盐、二硫苏糖醇。根据本发明的一个实施例，所述连接试剂的pH为7.5。根据本发明的一个实施例，所述缓冲液为Tris-HCl。根据本发明的一个实施例，所述可溶性盐为氯化镁。根据本发明的一个实施例，在30摄氏度下，利用连接试剂，在去除5’帽子结 构或5’三磷酸的转录本的5’末端连接RNA接头。由此，可以提高连接接头的效率，从而提高构建测序文库的效率。

[0039] S400(反转录)：对连接有RNA接头的转录本进行反转录，以便获得与所述转录本对应的cDNA。根据本发明的实施例，进行反转录所采用的反转录引物在其末端具有与所述RNA接头相对应的序列，由此，所得到的cDNA在其末端也将具有接头，从而便于后续文库构建和测序。在本文中所使用的术语“与RNA接头相对应”的含义是指，反转录引物中包含的序列能够与RNA接头匹配，并且能够进行扩增反应，从而得到在两个末端具有接头的cDNA。例如，在进行反转录的两条反转录引物之一中包含与RNA接头之一相同的序列，而在另一个反转录引物中，则包含于另一个RNA接头互补的序列。根据本发明的一个实施例，所述反转录采用具有SEQ ID NO:1所示序列的寡核苷酸作为反转录引物。根据本发明的一个实施例，所述反转录引物(SEQ ID NO:1)中至少一个N被硫代修饰，从而可以防止该引物被核酸酶降解。根据本发明的一个实施例，所述反转录引物(SEQ ID NO:1)中倒数第二个N被硫代修饰。

[0040] S500(扩增)：对所述cDNA进行扩增，以便获得扩增产物。本领域技术人员可以通过任何已知的方法进行扩增，例如可以通过常规的PCR方法，只需要采用根据接头的序列进行设计相应的引物即可。

[0041] S600(文库构建)：基于所述扩增产物，构建测序文库。本领域技术人员可以根据所期望使用的测序方法来针对扩增产物，本领域技术人员可以参考制造商所提供的操作说明，在此不在赘述。需要说明的是，利用根据本发明的方法进行处理所得到的扩增产物，可

以应用于Illumina Hiseq2000、Genome Analyzer、SOLiD测序系统、Ion Torrent、Ion Proton、454、PacBio RS测序系 统、Helicos tSMS技术以及纳米孔测序技术，从而可以实现高通量测序。

[0042] 由此，利用该方法，能够有效地针对核酸样本中所富集的转录本构建测序文库，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。另外，需要说明的是，在上述各处理步骤之间，可以任选地包括纯化产物的步骤，根据本发明的实施例，纯化RNA可以采用苯酚/氯仿/异戊醇(体积比为25:24:1)抽提，乙醇沉淀，是为了去除反应混合物中的酶，以免影响下一步骤的反应，而且用乙醇沉淀，还能保留一些小分子的转录本，如microRNA，使这一部分非编码RNA的TSS信息得以获得，从而帮助了解转录调控状态。

[0043] 在本发明的第三方面，本发明提出了一种测序文库，其特征在于，是由前面所述的方法构建的。利用该测序文库，能够有效的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。

[0044] 在本发明的第四方面，本发明提出了一种核酸样本测序方法。根据本发明的实施例，该核酸样本测序方法包括：根据前面所述的方法，构建测序文库；以及对所述测序文库进行测序，以便获得测序结果。利用该方法，能够有效的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。根据本发明的实施例，所述测序利用Illumina Hiseq2000、Genome Analyzer、SOLiD测序系统、Ion Torrent、Ion Proton、454、PacBio RS测序系统、Helicos tSMS技术以及纳米孔测序技术的至少一种进行的。由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高了测序的效率。在本发明的一个实施例中，所述测序是利用Illumina Hiseq2000进行的。

[0045] 根据本发明的第五方面，本发明提出了一种确定TSS的方法。根据本发明的实施例，该确定TSS的方法包括：从宿主提取RNA样本；利用前面所述的方法，获得由多个测序序列构成的测序结果；以及基于所述测序结果，确定TSS。利用该方法，能够有效的确定核酸样本中的TSS。

[0046] 根据本发明的一个实施例，所述RNA样本为宿主的总RNA的至少一部分。根据本发明的实施例，宿主可以为真核生物，例如人，也可以为原核生物，例如大肠杆菌。

[0047] 根据本发明的一个实施例，基于所述测序结果，确定TSS，进一步包括：将所述测序数据与参考序列进行比对；

[0048] 基于比对结果，确定所述转录起点，

[0049] 其中，所述参考序列中包含预定基因的5’-UTR序列的至少一部分，选择能够和所述参考序列对上、且在所述参考序列最上游的测序序列作为阳性序列，并且确定所述阳性序列的第一位碱基作为所述转录起始位点。这里所使用的术语“预定基因”指的是，在参考基因组上，预先设定了一系列基因的可能包括的范围，这些基因可能是已知的，也可以是未知而通过生物信息学推测出来的。根据本发明的实施例，参考序列的长度并不受特别限制，根据本发明的实施例，参考序列至少包含预定基因的翻译起始位点及其上游预定长度的序列。由于转录起始位点在翻译位点的上游，因而，通过选择参考序列的长度，可以将转录起始位点包括在其中。例如，根据本发明的实施例，针对原核宿主，所述参考序列包含所述预

定基因的翻译起始位点与该翻译起始位点上游700bp位点之间的核酸序列，针对真核宿主，所述参考序列包含所述预定基因的翻译起始位点与该翻译起始位点上游5000bp位点之间的核酸序列。

[0050] 根据本发明的一个实施例，可以采用SOAPAlignment进行所述比对。在本发明中，通过一种短序列映射程序soapalignmentv2.2，将高通量测序技术得到的干净的序列片段分别比对到参考基因组和参考基因序列上，不允许碱基的错配。参考基因组序列和参考基因序列可取于公共数据库。

[0051] 根据本发明的一个实施例，进一步包括对所述阳性序列进行筛选，其中所述筛选的原则是：所述阳性序列的数目是比对到所述预定基因内部的测序序列数目平均值的N倍以上，其中所述N为大于1的实数，优选地，所述N为至少10的实数。根据本发明的实施例，比对后，可以首先对比对结果进行筛选，以获得可靠地TSS信息。筛选方法为：假设干净序列比对到基因(与预定基因对应的序段)的第一个位置即为原始的TSS，但是这些序列有可能是比对到基因的内部成为假阳性的TSS，所以需要再进一步进行过滤。该方法可以使获得的序列在基因的5’端富集，因此真实的TSS的序列数会比落在基因内部的序列的平均数要高，于是在他们之间引进一个倍数N过滤TSS，即筛选的TSS的序列数要是落在对应的基因内部序列数平均值的N倍才将其认定为真实的TSS。根据本发明的实施例，N可以是至少为10的实数。

[0052] 根据本发明的一个实施例，进一步包括对筛选结果进行卡方检验。根据本发明的一个实施例，所述卡方检验的检验值为3.84以上时，即置信度大于95%。根据本发明的实施例，过滤后，对于过滤结果本方法使用卡方检验来验证其可靠性，具体地，在上一实施例的基础上，计算所有的TSS对应的倍数的平均值，以及他们的标准差，标准化之后，用下述公式计算卡方值：$\frac{(\text{默认得到的参数或用户设定的参数} - \text{平均值})}{(\text{标准差}/\sqrt{\text{总个数}})}$，根据卡方检验表中查到当置信度为0.95时卡方值为3.84，所以可以获得可靠度大于95%的TSS，根据公式算出的卡方值就必须大于3.84。

[0053] 另外，根据本发明的实施例，还可以在进行获得测序结果之后，还可以包括对测序序列去除不合格的序列，获得干净的测序序列的步骤。根据本发明的实施例，不合格的序列包括：

[0054] 测序质量低于某一阀值的碱基个数超过整条序列碱基个数的50%则认为是不合格序列。低质量阀值由具体测序技术及测序环境而定；

[0055] 序列中测序结果不确定的碱基(如Illumina Hiseq2000测序结果中的N)个数超过整条序列碱基个数的10%则认为是不合格序列；

[0056] 除样本接头序列外，与其它实验引入的外源序列比对，如各种接头序列。若序列中存在外源序列则认为是不合格序列。

[0057] 原始的序列数据经过去除不合格序列处理后得到的序列数据我们称为干净的序列片段(clean reads)，可以作为后续分析的基础，由此，可以提高后续分析的有效性。

[0058] 另外，在卡方检验后，对验证可靠的结果进行一系列的生物信息分析，如：

[0059] 1)TSS(转录起始位点)的分类：根据本发明的实施例，可以将筛选的TSS分为两大类，一类是能比对到基因组上且有对应的基因注释的TSS，称之为有注释的TSS；另一类是能

比对到基因组上但是在其周围没有注释的基因信息，称之为未注释的TSS，可以用于新基因的预测。

[0060] 2)TSS注释：这里主要对落在已知基因的TSS进行注释，包括TSS的表达量，TSS所处的位置，以及对应的基因注释信息。

[0061] 3)构建TSS图谱：根据本发明的实施例，可以将同一物种在该方法中找到的TSS用图片的形式直观的展示出来形成TSS图谱，从图谱上可以很直观的看出每个TSS所在的位置以及他们的表达量。同时也可以看到不同样品中的TSS表达，分布的差异。

[0062] 4)启动子区寻找及5'UTR长度统计。

[0063] 5)实验重复性分析：根据本发明的实施例，对两次平行实验的结果相关性分析可获得对实验结果可靠性和操作稳定性的评估。

[0064] 6)新基因预测：根据本发明的实施例，对于附近没有找到参考基因的TSS，可以将这些TSS附近的序列提取出来进行基因预测。原核生物用glimmer进行预测，真核生物用genscan进行预测。

[0065] 7)数据可视化：根据本发明的实施例，利用分析结果，可以针对感兴趣的基因或者区域的TSS分布作图观察。

[0066] 在本发明的第六方面，本发明提出了一种用于从RNA样本富集转录本的富集试剂。根据本发明的实施例，富集试剂具有5'-单磷酸外切酶活性。利用该富集试剂，可以有效地富集转录本，因而能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。根据本发明的一个实施例，所述富集试剂含有DNase I。由此，可以进一步提高降解5'单磷酸的RNA的特异性和效率，从而进一步提高富集转录本的方法的效率。根据本发明的一个实施例，所述富集试剂还可以进一步含有缓冲液和可溶性盐，以便进一步提高DNase I的酶活性。根据本发明的一个实施例，所述富集试剂的pH为8.0。根据本发明的一个实施例，所述缓冲液为Tris-HCl，所述可溶性盐为选自氯化钠和氯化镁的至少一种。根据本发明的一个实施例，在30摄氏度下，利用所述富集试剂对所述RNA样本进行处理。从而，可以进一步提高利用根据本发明实施例的富集试剂进行富集转录本的效率。根据本发明实施例，具有5'单磷酸外切酶活性的酶的例子可以包括：核糖核酸外切酶XRN-1，Terminator™依赖于5'磷酸的核酸外切酶或者TAKARA™碱性磷酸酶。

[0067] 在本发明的第七方面，本发明提出了一种构建测序文库的装置。参考图5，根据本发明的实施例，该构建测序文库的装置包括：转录本富集单元211、末端修整单元212、RNA接头连接单元213、反转录单元214、扩增单元215以及文库构建单元216。根据本发明的实施例，转录本富集单元211中设置有前面所述的富集试剂，以便从RNA样本富集转录本；末端修整单元212与所述转录本富集单元211相连，并且适于去除所述转录本的5'帽子结构或5'三磷酸，以便获得去除5'帽子结构或5'三磷酸的转录本；RNA接头连接单元213与末端修整单元212相连，并且适于在去除5'帽子结构或5'三磷酸的转录本的5'末端连接RNA接头，以便获得连接有RNA接头的转录本；反转录单元214与所述RNA接头连接单元213相连，并且适于对连接有RNA接头的转录本进行反转录，以便获得与所述转录本对应的cDNA；扩增单元215与所述反转录单元214相连，并且适于对所述cDNA进行扩增，以便获得扩增产物；文库构建单元216与所述扩增单元215相连，并且适于基于所述扩增产物，构建测序文库。利用该装置，能够有效地针对核酸样本中所富集的转录本构建测序文库，因而能够同时应用于真核

和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。根据本发明的一个实施例，所述末端修整单元212中设置有末端修整试剂，其中，所述末端修整试剂具有烟草酸焦磷酸酶活性。据本发明的一个实施例，所述修整试剂包含：烟草酸焦磷酸酶、可溶性盐、EDTA、β-巯基乙醇和Triton-X 100。据本发明的一个实施例，所述可溶性盐为醋酸钠。据本发明的一个实施例，所述修整试剂的pH为7.5。据本发明的一个实施例，所述反转录单元214中设置有具有SEQ ID NO:1所示序列的寡核苷酸作为反转录引物。据本发明的一个实施例，所述反转录引物中至少一个N被硫代修饰。据本发明的一个实施例，所述反转录引物中倒数第 二个N被硫代修饰。据本发明的一个实施例，所述RNA接头连接单元213中设置有连接试剂，其中，所述连接试剂具有T4RNA连接酶活性。据本发明的一个实施例，所述连接试剂包含：T4RNA连接酶，缓冲液、可溶性盐、二硫苏糖醇。据本发明的一个实施例，所述连接试剂的pH为7.5。据本发明的一个实施例，所述缓冲液为Tris-HCl。据本发明的一个实施例，所述可溶性盐为氯化镁。

[0068] 在本发明的第八方面，本发明提出了一种核酸样本测序设备。参考图4，根据本发明的实施例，该设备包括：文库构建装置210，所述文库构建装置210为前面所述的装置，以便针对核酸样本构建测序文库；以及测序装置220，所述测序装置220与所述文库构建装置210相连，并且适于对所述测序文库进行测序，以便获得测序结果。利用该设备，能够有效的对RNA转录本进行测序，能够同时应用于真核和原核的RNA的TSS的高通量测序，具有操作简单，准确率高和成本低的众多优点。根据本发明的实施例，所述测序设备为Illumina Hiseq2000、Genome Analyzer、SOLiD测序系统、Ion Torrent、Ion Proton、454、PacBio RS测序系统、Helicos tSMS系统以及纳米孔测序系统的至少一种。

[0069] 在本发明的第九方面，本发明提出了一种确定TSS的系统。参考图3，根据本发明的实施例，该系统包括：样品提取设备100，所述样品提取设备用于从宿主提取RNA样本；核酸样本测序设备200，所述核酸样本测序设备与所述样品提取设备相连，并且所述测序设备为前面所述的核酸样本测序设备，以便针对所述RNA样本进行测序，从而获得由多个测序序列构成的测序结果；以及TSS确定设备300，所述TSS确定设备300与所述测序设备200相连，并且适于基于所述测序结果，确定TSS。根据本发明的实施例，利用该系统能够有效的确定核酸样本中的TSS。参考图6，根据本发明的一个实施例，所述 TSS确定设备进一步包括：比对装置310，所述比对装置用于将所述测序数据与参考序列进行比对；确定装置320，所述确定装置适于基于比对结果，确定所述TSS，其中，所述参考序列中包含预定基因的5’-UTR序列的至少一部，并且，所述确定装置320适于：选择能够比对到所述与预定基因对应的序段的并且最接近所述与预定基因对应的序段5’端的测序序列作为阳性序列，并且确定所述阳性序列的第一碱基为转录起始位点。根据本发明的一个实施例，所述比对装置适于采用SOAP Alignment进行所述比对。根据本发明的一个实施例，所述确定装置进一步包括筛选单元，所述筛选单元适于对所述阳性序列进行筛选，其中所述筛选的原则是：所述阳性序列的序列数目是所述与预定基因对应的序段内部序列数目平均值的N倍以上，其中所述N为大于1的实数，优选地，N可以是至少为10的实数。根据本发明的一个实施例，所述确定装置进一步包括检验单元，所述检验单元适于对筛选结果进行卡方检验。根据本发明的一个实施例，所述卡方检验的检验值为3.84以上，对应置信度大于95%。

[0070] 在本发明中所使用的术语“预定基因”应做广义理解，其可以指任何已知的基因，

也可以指通过已知的方法，预测可能会编码蛋白的核酸序列。

[0071] 下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件(例如参考J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社)或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自Illumina公司。

[0072] 一般方法

[0073] 在实施例中所采用的方法主要包括TSS文库构建以及测序后分析，其中TSS文库构建方法主要包括下述步骤：

[0074] (1)取总RNA样品，用DNaseI消化后，乙醇沉淀纯化消化后的RNA；

[0075] (2)将(1)得到的RNA与试剂I混匀反应，富集含有5’帽子或5’三磷酸的RNA；

[0076] (3)苯酚/氯仿/异戊醇(25:24:1)抽提纯化(2)得到的RNA；

[0077] (4)将(3)纯化后的RNA与试剂II混匀反应，去除5’帽子或5’三磷酸得到5’单磷酸；

[0078] (5)苯酚/氯仿/异戊醇(25:24:1)抽提纯化(4)得到的RNA；

[0079] (6)将(5)的RNA加入RNA接头，并与试剂III混匀反应，在得到5’端加上接头的RNA；

[0080] (7)用特定的反转录引物将(6)的RNA反转录，得到两端都有特定序列接头的cDNA并用磁珠纯化；

[0081] (8)采用聚合酶链式反应(PCR)扩增(7)所得两端加接头的cDNA片段，使用磁珠纯化PCR产物；

[0082] (9)采用Agilent Bioanalyzer 2100和Q-PCR检测文库浓度及片段大小。

[0083] 步骤(1)中，总RNA的量为5μg。

[0084] 步骤(2)中，试剂I，含有：1μL 5’单磷酸外切酶(1U/μL)，50mM缓冲盐，2mM-100mM可溶性盐，pH 8.0，溶剂为水。试剂I中缓冲盐为Tris-HCl。试剂I中可溶性盐为氯化钠或氯化镁。步骤(2)中所得RNA与试剂I混合温度为30℃。

[0085] 所述步骤(4)中，试剂II含有：0.2μL烟草酸焦磷酸酶(10U/μL)，50mM可溶性盐，pH 6.0，1mM EDTA，0.1%β-巯基乙醇，0.01%Triton X-100，溶剂为 水。试剂II中可溶性盐为醋酸钠。样品与试剂II混合温度为37℃。

[0086] 所述步骤(6)中，试剂III含有：1μL T4 RNA连接酶1，50mM缓冲盐，10mM可溶性盐，1mM二硫苏糖醇，pH 7.5，溶剂为水。试剂III中缓冲盐为Tris-HCl。试剂III中可溶性盐为氯化镁。步骤(6)中所得RNA与试剂III混合温度为20℃。

[0087] 所述步骤(7)中所用特定反转录引物序列为：5-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN-3，其中倒数第二个N做硫代修饰。

[0088] 所述步骤(3)和(5)之后，纯化RNA均采用苯酚/氯仿/异戊醇抽提，乙醇沉淀，是为了去除反应混合物中的酶，以免影响下一步骤的反应，而且用乙醇沉淀，还能保留一些小分子的转录本，如microRNA，使这一部分非编码RNA的TSS信息得以获得，从而帮助了解转录调控状态。

[0089] 参考图2，对TSS文库测序所产生的数据，进行生物信息分析，包括以下步骤：

[0090] (1)过滤测序序列；

[0091] 在本发明中，接收到高通量测序序列后，对测序序列进行过滤，去除不合格的序

列。其中高通量测序技术可以为Illumina Hiseq2000测序技术，也可以为现有的其他高通量测序技术。

[0092] 不合格序列包括：测序质量低于某一阀值的碱基个数超过整条序列碱基个数的50%则认为是不合格序列。低质量阀值由具体测序技术及测序环境而定；序列中测序结果不确定的碱基(如Illumina Hiseq2000测序结果中的N)个数超过整条序列碱基个数的10%则认为是不合格序列；除样本接头序列外，与其它实验引入的外源序列比对，如各种接头序列。若序列中存在外源序列则认为 是不合格序列。原始的序列数据经过去除不合格序列处理后得到的序列数据我们称为干净的序列片段(clean reads)，作为后续分析的基础。

[0093] (2)干净的序列片段与参考序列比对；

[0094] 在本发明中，通过一种短序列映射程序soapalignment v2.2，将高通量测序技术得到的干净的序列片段分别比对到参考基因组和参考基因序列上，不允许碱基的错配。参考基因组序列和参考基因序列可取于公共数据库。

[0095] (3)比对后，首先对比对结果进行筛选，以获得可靠地TSS信息。筛选方法为：假设干净序列比对到基因组的第一个位置即为原始的TSS，但是这些序列有可能是比对到基因的内部成为假阳性的TSS，所以需要再进一步进行过滤。该方法可以使我们获得的序列在基因的5’端富集，因此真实的TSS的序列数会比落在基因内部的序列的平均数要高，于是在他们之间引进一个倍数N过滤TSS，即筛选的TSS的序列数要是落在对应的基因内部序列数平均值的N倍才将其认定为真实的TSS。

[0096] (4)过滤后，对于过滤结果本方法使用卡方检验来验证其可靠性，即卡方检验值应该大于3.84即置信度大于95%。

[0097] (5)在卡方检验后，对验证可靠的结果进行一系列的生物信息分析，如：

[0098] 1)TSS(转录起始位点)的分类：根据本发明的实施例，可以将筛选的TSS分为两大类，一类是能比对到基因组上且有对应的基因注释的TSS，称之为有注释的TSS；另一类是能比对到基因组上但是在其周围没有注释的基因信息，称之为未注释的TSS，可以用于新基因的预测。

[0099] 2)TSS注释：这里主要对落在已知基因的TSS进行注释，包括TSS的表达量，TSS所处的位置，以及对应的基因注释信息。

[0100] 3)构建TSS图谱：根据本发明的实施例，可以将同一物种在该方法中找 到的TSS用图片的形式直观的展示出来形成TSS图谱，从图谱上可以很直观的看出每个TSS所在的位置以及他们的表达量。同时也可以看到不同样品中的TSS表达，分布的差异。

[0101] 4)启动子区寻找及5’UTR长度统计。

[0102] 5)实验重复性分析：根据本发明的实施例，对两次平行实验的结果相关性分析可获得对实验结果可靠性和操作稳定性的评估。

[0103] 6)新基因预测：根据本发明的实施例，对于附近没有找到参考基因的TSS，可以将这些TSS附近的序列提取出来进行基因预测。原核生物用glimmer进行预测，真核生物用genscan进行预测。

[0104] 7)数据可视化：根据本发明的实施例，利用分析结果，可以针对感兴趣的基因或者区域的TSS分布作图观察。

[0105] 实施例1人RNA样本和大肠杆菌RNA样本的转录起始位点序列分析

[0106] 人RNA样本(样本一)购自安捷伦公司,大肠杆菌RNA(样本二)是将大肠杆菌培养至对数生长期后提取的RNA。取1-5μg的总RNA,用DNaseI进行消化,乙醇沉淀纯化,纯化后的RNA与试剂I混匀反应,富集得到含有5'帽子或5'三磷酸的完整RNA,用苯酚/氯仿/异戊醇抽提纯化后,与试剂II混匀反应,去除5'端的帽子或三磷酸使之变成单磷酸,用苯酚/氯仿/异戊醇抽提纯化,将5'单磷酸的RNA与试剂III和RNA接头混匀反应,在RNA 5'端加上接头,用特定的反转录引物将加有5'接头的RNA反转录为两端带有固定序列的cDNA,使用磁珠纯化cDNA产物,采用聚合酶链式反应(PCR)扩增所得cDNA片段,磁珠纯化PCR产物,上机测序。测序使用Illumina Hiseq2000。

[0107] 按照一般方法的信息分析流程,筛选得到了一系列TSS信息,图7是筛选后的TSS在基因组上的分布,上图和下图分别是人RNA和大肠杆菌RNA 样品的TSS分布图,其中0是基因编码区的起始位点,其上游就是转录起始的位点,从图中可以看出,大部分的序列都落在基因编码区的上游。

[0108] 另外,在本实施例中,针对这些TSS信息进行了一系列的分析。

[0109] 首先是TSS的分类,将筛选的TSS分为两大类,一类是能比对到基因组上且有对应的基因注释的TSS,称之为有注释的TSS;另一类是能比对到基因组上但是在其周围没有注释的基因信息,称之为未注释的TSS,可以用于新基因的预测。

[0110] 其次做了TSS的注释,这里主要对落在已知基因的TSS进行注释,包括TSS的表达量,TSS所处的位置,以及对应的基因注释信息。然后构建TSS图谱,发明人将同一物种在该方法中找到的TSS用图片的形式直观的展示出来形成TSS图谱,从图谱上可以很直观的看出每个TSS所在的位置以及他们的表达量。同时也可以看到不同样品中的TSS表达,分布的差异。如图8所示,每个是8个人的样品的TSS图谱,从图中可以看到不同样品中TSS的分布情况。

[0111] 接下来是启动子区的寻找及5'UTR长度统计,图9是TSS上游的碱基分布图,其中横坐标1对应的就是TSS的位置,以嘌呤为主(A/G),上图显示的是人的TSS上游碱基分布图,有明显的GC富集区,这也是真核生物主要的启动子类型,下图显示了大肠杆菌的碱基分布图,在其TSS上游-10区处也能找到典型的TATA盒;图10显示了人(上图)和大肠杆菌(下图)的5'UTR的长度分布,也就是TSS到编码区的距离,5'UTR的长度影响基因功能的发挥,真核的5'UTR比原核的要长。

[0112] 在本实施例中,还对两次平行实验的结果做了相关性分析可获得对实验结果可靠性和操作稳定性的评估,如图11所示,同一样本两次平行实验之间的 相关性越接近1,说明可重复性高。

[0113] 本发明对于附近没有找到参考基因的TSS,提取出这些TSS附近的序列进行基因预测。大肠杆菌的用glimmer进行预测,人的用genscan进行预测。最后,本发明利用分析结果,针对感兴趣的基因或者区域的TSS分布作图观察,如图12所示,上图是人的两个基因NM_018997和NM_031901的TSS分布,他们是发生了可变剪切的基因,图中红色竖线表示筛选的TSS,黑色的竖线是过滤前得到的序列,蓝色横线代表基因的外显子,黄色横线是基因的内含子,下图是大肠杆菌一个操纵子的TSS分布,原核的不存在内含子,所以只有代表基因的蓝色横线,这个操纵子的4个基因共有一个TSS。可以看到发明人筛选出来的TSS是位于基因上游的,也是可靠的。

[0114] 本发明的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本发明的原理和实际应用，并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

[0115] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0116] 尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在不脱离本 发明的原理和宗旨的情况下在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

SEQUENCE LISTING

<110> 深圳华大基因科技服务有限公司

<120> 从 RNA 样本富集转录本的方法及其用途

<130> PIDC121985

<140> 201210379402.8
<141> 2012-09-29

<160> 1

<170> PatentIn version 3.5

<210> 1
[0001] <211> 40
<212> DNA
<213> 人工序列

<220>
<221> misc_feature
<222> (1)..(40)
<223> 反转录引物

<220>
<221> misc_feature
<222> (35)..(40)
<223> n 为 a、t、c 或者 g

<400> 1
gtgactggag ttcagacgtg tgctcttccg atctnnnnnn 40

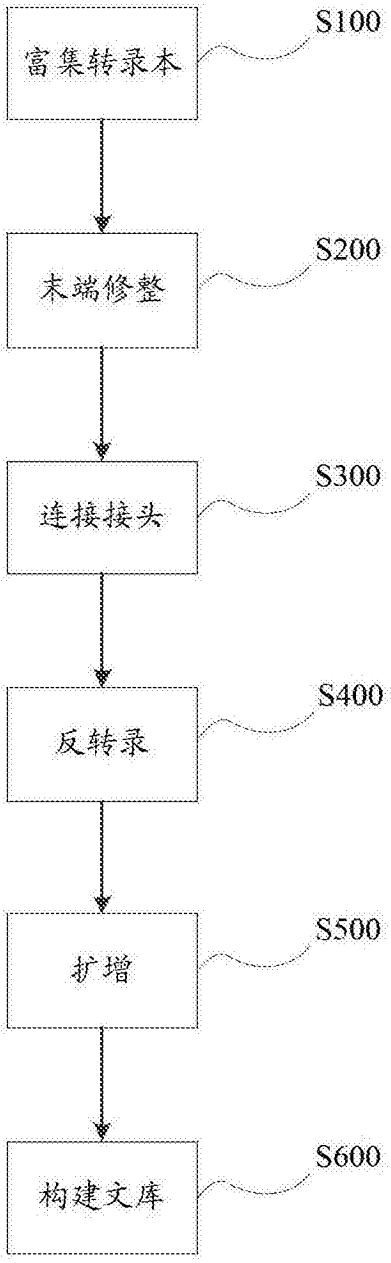

图1

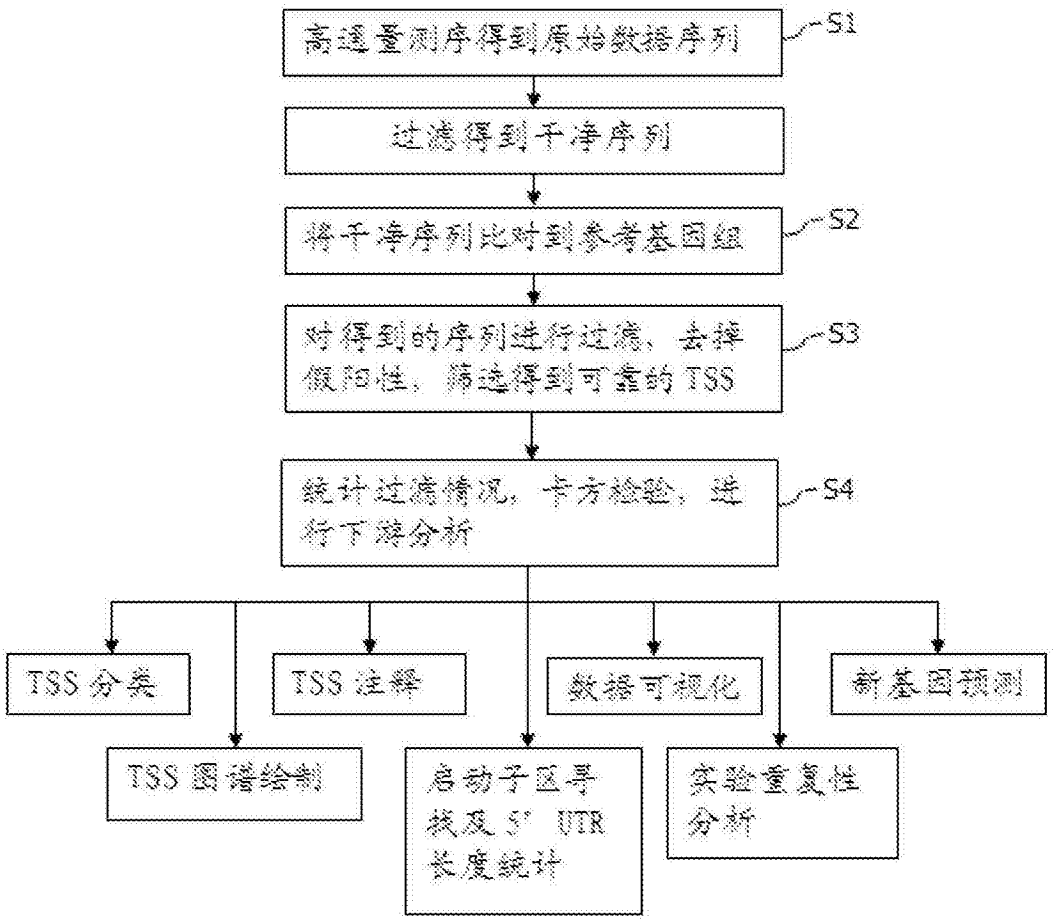
高通量测序得到原始数据序列
S1
过滤得到干净序列
将干净序列比对到参考基因组
S2
对得到的序列进行过滤，去掉假阳性，筛选得到可靠的 TSS
S3
统计过滤情况，卡方检验，进行下游分析
S4
TSS 分类
TSS 注释
数据可视化
新基因预测
TSS 图谱绘制
启动子区寻找及 5' UTR 长度统计
实验重复性分析

图2


1000
100
200
300
样品提取设备
核酸样本测序设备
转录起点确定设备

图3

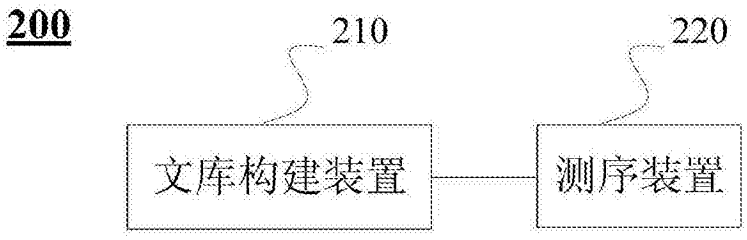
200
210
220
文库构建装置
测序装置

图4

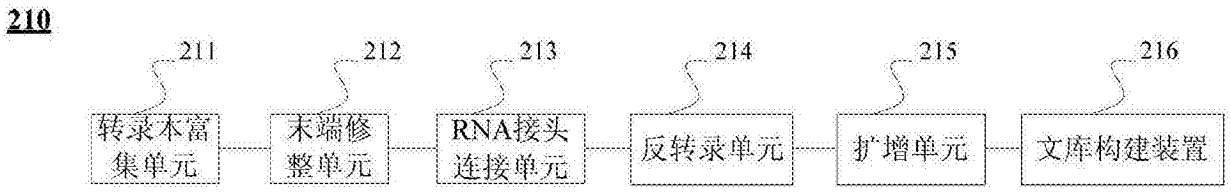
210
211
转录本富集单元
212
末端修整单元
213
RNA接头连接单元
214
反转录单元
215
扩增单元
216
文库构建装置

图5


300
310
比对装置
320
确定装置

图6

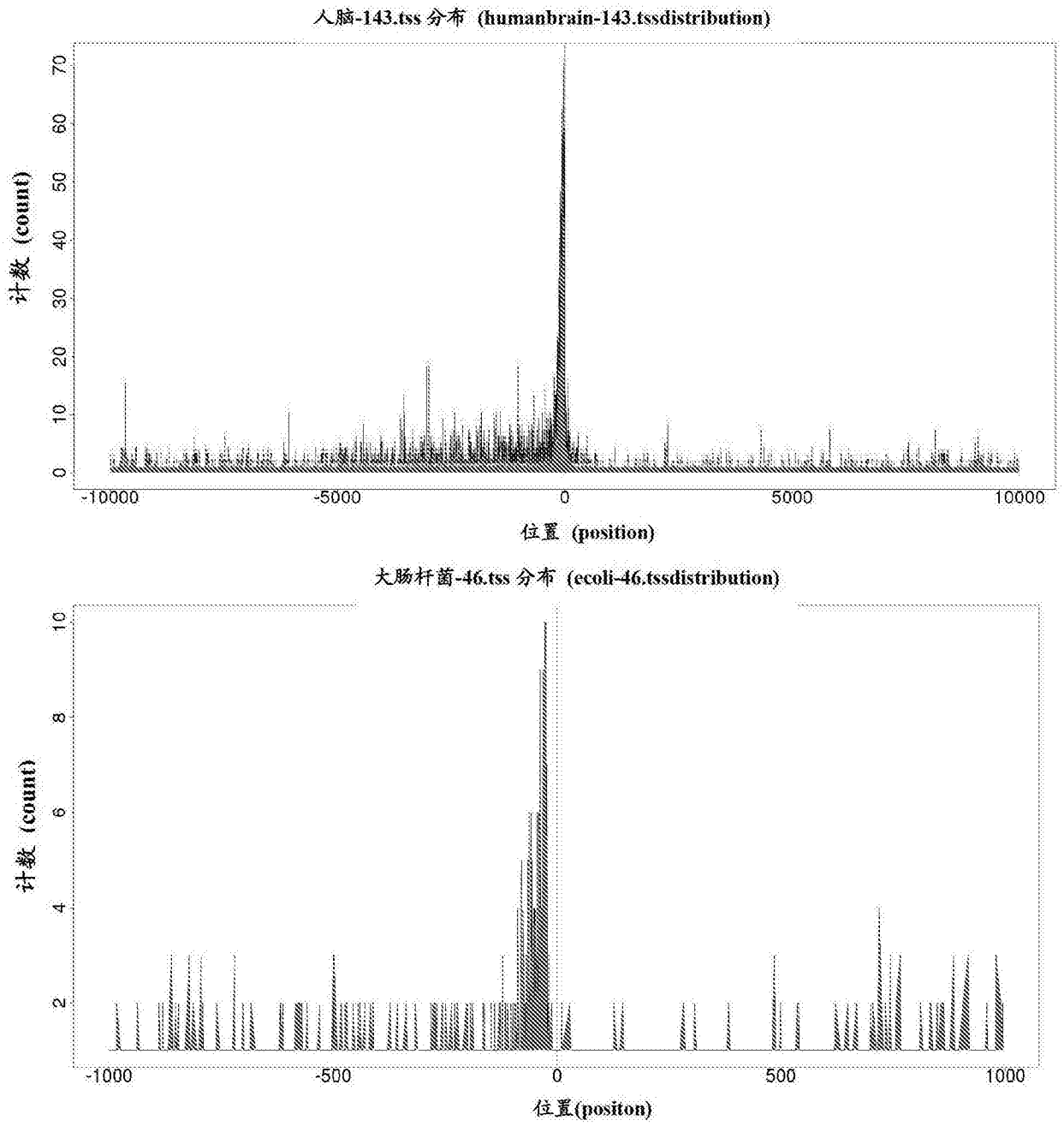
人脑-143.tss 分布 (humanbrain-143.tssdistribution)
计数 (count)
0
10
20
30
40
50
60
70
-10000
-5000
0
5000
10000
位置 (position)
大肠杆菌-46.tss 分布 (ecoli-46.tssdistribution)
计数 (count)
2
4
6
8
10
-1000
-500
0
500
1000
位置(positon)

图7

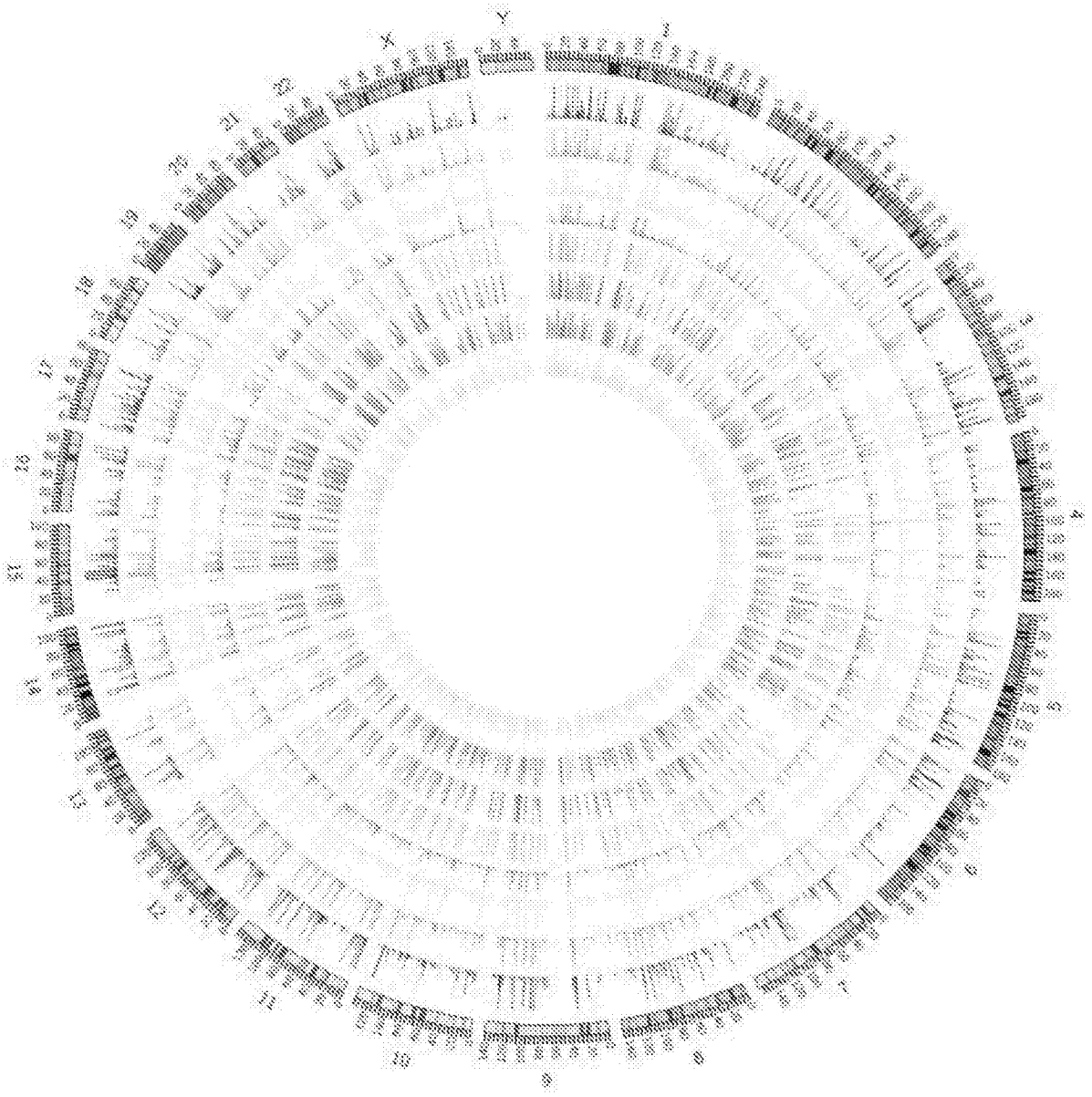

图8

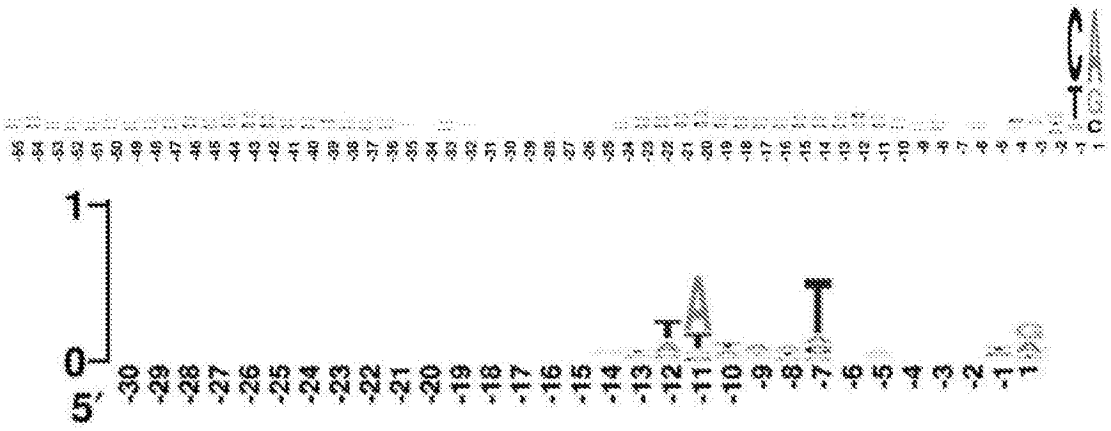

图9

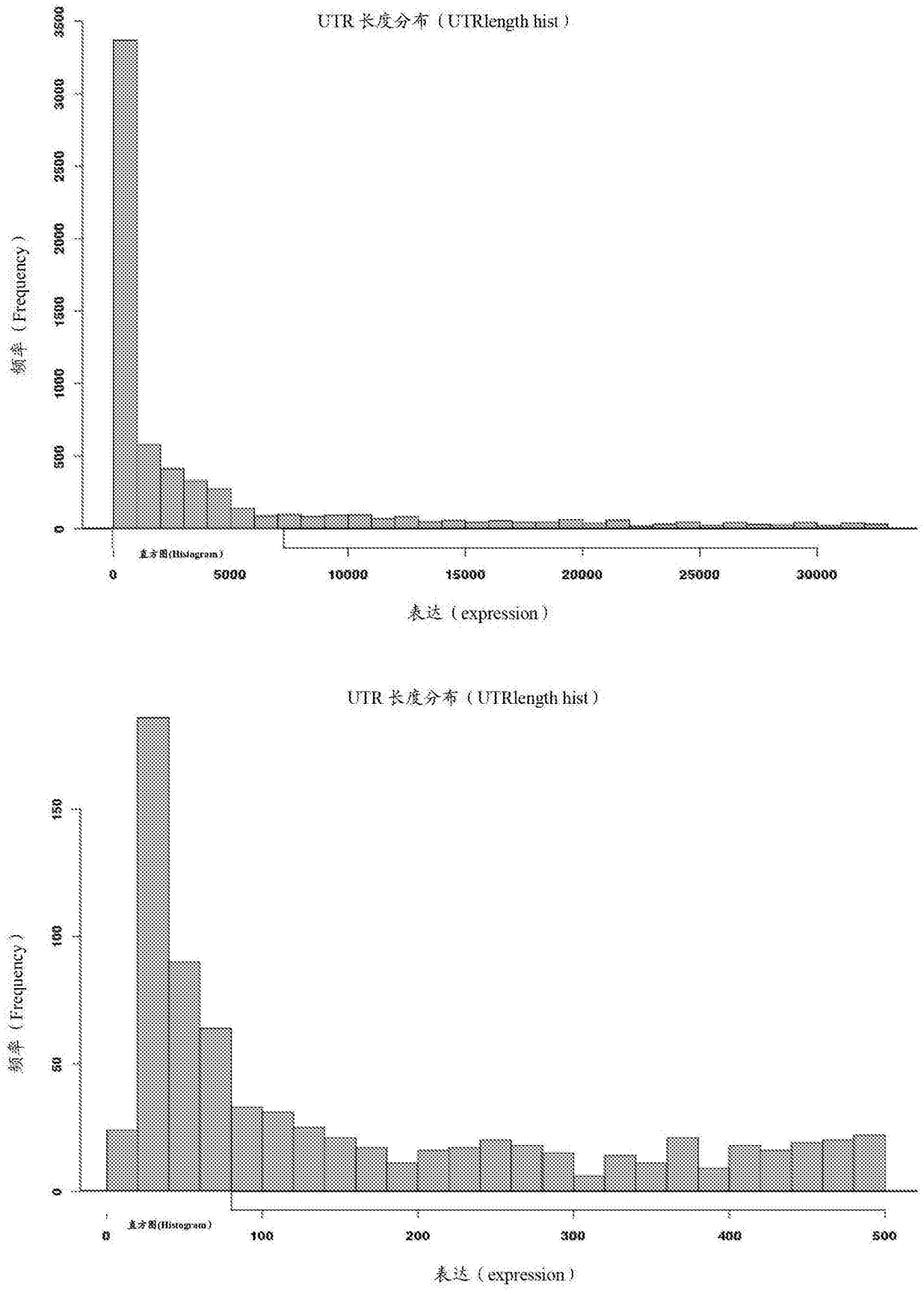
UTR 长度分布（UTRlength hist）
频率（Frequency）
0
500
1000
1500
2000
2500
3000
3500
直方图(Histogram)
0
5000
10000
15000
20000
25000
30000
表达（expression）
UTR 长度分布（UTRlength hist）
频率（Frequency）
0
50
100
150
直方图(Histogram)
0
100
200
300
400
500
表达（expression）

图10

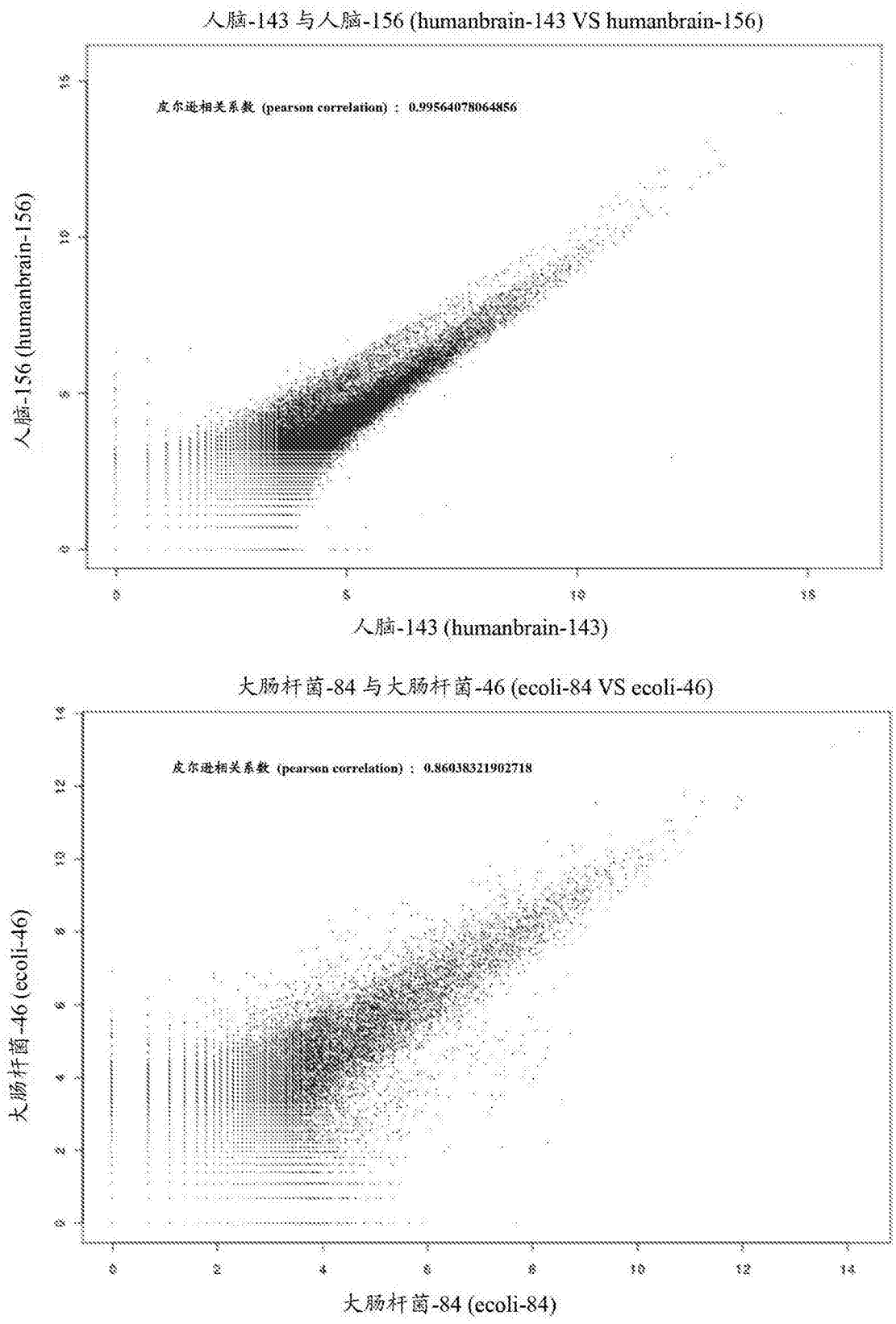
人脑-143 与人脑-156 (humanbrain-143 VS humanbrain-156)
皮尔逊相关系数 (pearson correlation) : 0.99564078064856
人脑-156 (humanbrain-156)
人脑-143 (humanbrain-143)
大肠杆菌-84 与大肠杆菌-46 (ecoli-84 VS ecoli-46)
皮尔逊相关系数 (pearson correlation) : 0.86038321902718
大肠杆菌-46 (ecoli-46)
大肠杆菌-84 (ecoli-84)

图11

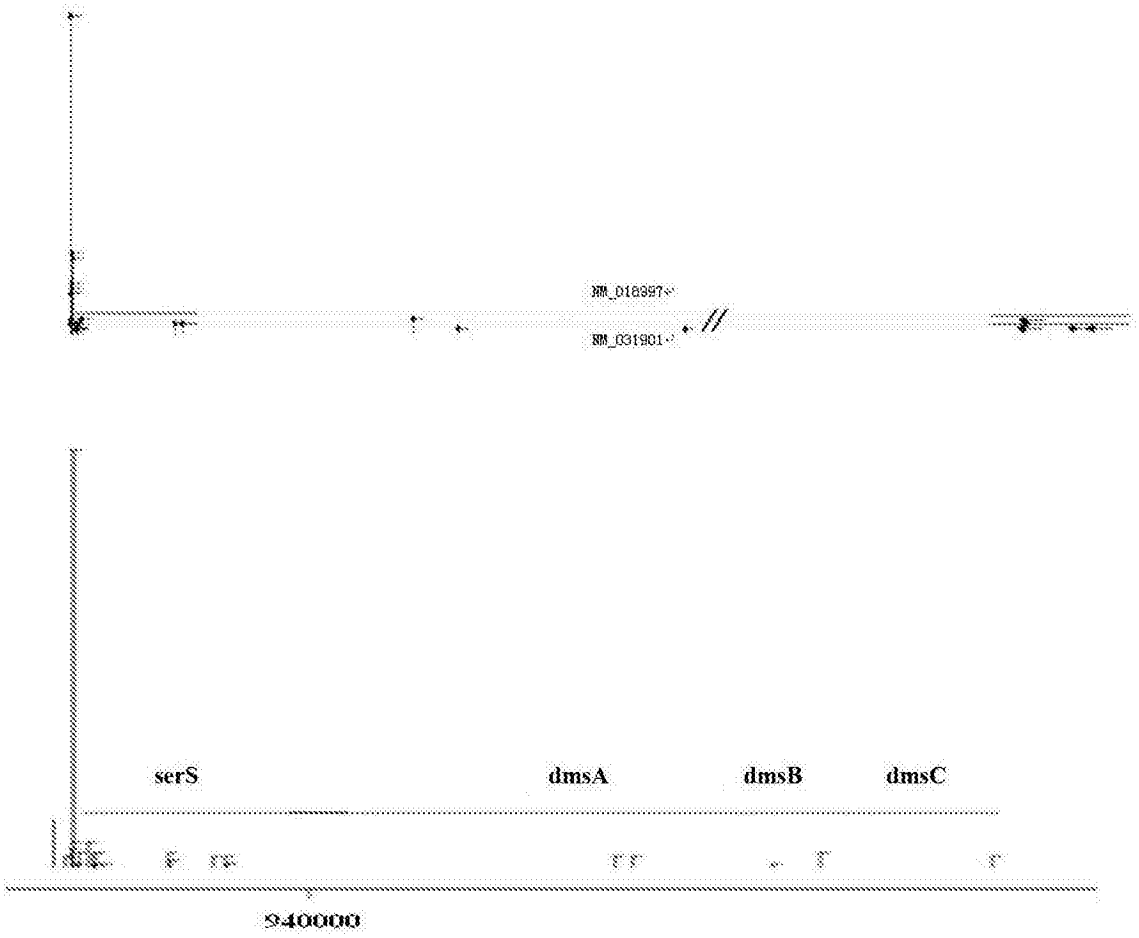

图12