

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(43) 国际公布日
2012 年 1 月 26 日 (26.01.2012)

PCT

(10) 国际公布号
WO 2012/009952 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01)
- (21) 国际申请号: PCT/CN2011/001158
- (22) 国际申请日: 2011 年 7 月 13 日 (13.07.2011)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201010236176.9 2010 年 7 月 22 日 (22.07.2010) CN
- (71) 申请人 (对除美国外的所有指定国): 深圳华大基因科技有限公司 (BGI SHENZHEN CO., LIMITED) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): 彭智宇 (PENG, Zhiyu) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。 韩祖晶 (HAN, Zujing) [CN/CN]; 中国广东省深圳市

盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。 李英睿 (LI, Yingrui) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。 王俊 (WANG, Jun) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: 北京北翔知识产权代理有限公司 (PEKSUNG INTELLECTUAL PROPERTY LTD.); 中国北京市海淀区学院路 35 号世宁大厦 908 室, Beijing 100191 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV,

[见续页]

(54) Title: QUALITY CONTROL METHOD AND APPARATUS FOR RNA SEQUENCING OF GENE EXPRESSION

(54) 发明名称: 关于基因表达的 RNA 测序质控方法及装置

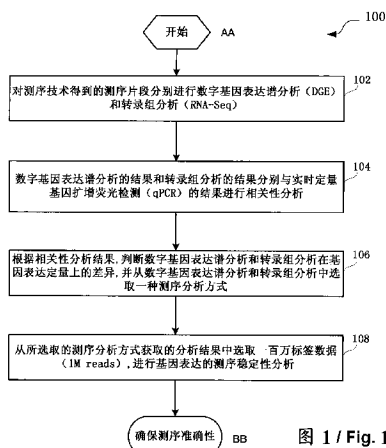


图 1 / Fig. 1

(57) Abstract: Disclosed is a quality control method and apparatus for RNA sequencing of gene expression, which method comprises: performing respectively DGE and RNA-Seq analyses on sequencing fragments obtained by the sequencing technology; performing respectively correlative analyses on the results from the DGE analysis and that from the transcriptome analysis with the results from the qPCR; judging the quantitative differences in the gene expression between the DGE analysis and the transcriptome analysis according to the results of the correlative analysis, then selecting a sequencing analysis mode from among the DGE analysis and the transcriptome analysis, selecting one million lines of tag data (1M reads) from the analysis results obtained by the selected sequencing analysis mode, and performing a sequencing stability analysis of the gene expression.

[见续页]

AA START
102 PERFORMING RESPECTIVELY DIGITAL GENE EXPRESSION PROFILE ANALYSIS (DGE) AND TRANSCRIPTOME ANALYSIS (RNA-SEQ) ON THE SEQUENCING FRAGMENTS OBTAINED BY THE SEQUENCING TECHNOLOGY
104 PERFORMING RESPECTIVELY CORRELATIVE ANALYSES ON THE RESULTS FROM THE DIGITAL GENE EXPRESSION PROFILE ANALYSIS AND THAT FROM THE TRANSCRIPTOME ANALYSIS WITH THE RESULTS FROM THE REAL TIME QUANTITATIVE GENE AMPLIFICATION FLUORESCENCE DETECTION (qPCR)
106 JUDGING THE QUANTITATIVE DIFFERENCES IN THE GENE EXPRESSION BETWEEN THE DIGITAL GENE EXPRESSION PROFILE ANALYSIS AND THE TRANSCRIPTOME ANALYSIS ACCORDING TO THE RESULTS OF THE CORRELATIVE ANALYSIS, THEN SELECTING A SEQUENCING ANALYSIS MODE FROM AMONG THE DIGITAL GENE EXPRESSION PROFILE ANALYSIS AND THE TRANSCRIPTOME ANALYSIS
108 SELECTING ONE MILLION LINES OF TAG DATA (1M READS) FROM THE ANALYSIS RESULTS OBTAINED BY THE SELECTED SEQUENCING ANALYSIS MODE, AND PERFORMING A SEQUENCING STABILITY ANALYSIS OF THE GENE EXPRESSION
BB ENSURING THE SEQUENCING ACCURACY

WO 2012/009952 A1



SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

- (84) **指定国** (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL,

本国际公布:

- 包括国际检索报告(条约第 21 条(3))。
- 包括说明书序列表部分(细则 5.2(a))。

(57) **摘要:**

本发明公开了一种关于基因表达的 RNA 测序的质控方法及装置, 该方法包括: 对测序技术得到的测序片段分别进行 DGE 和 RNA-Seq 分析; 将 DGE 分析的结果和转录组分析的结果分别与 qPCR 的结果进行相关性分析; 根据相关性分析结果, 判断 DGE 分析和转录组分析在基因表达定量上的差异, 并从 DGE 分析和转录组分析中选取一种测序分析方式; 从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析。

关于基因表达的 RNA 测序质控方法及装置

技术领域

本发明涉及生物技术领域，尤其涉及一种关于基因表达的核糖核酸（RNA，RiboNucleic Acid）测序质控方法及装置。

背景技术

基因表达是指基因片段脱氧核糖核酸（DNA，Deoxyribonucleic acid）转录成信使核糖核酸（mRNA，Messenger RNA）及 mRNA 翻译成蛋白质的过程。随着人类基因组计划（HGP，Human Genome Project）全部核苷酸测序的完成，人类基因组研究的重心逐渐进入后基因组时代（Postgenome Era）向基因的功能及基因的多样性倾斜。通过对个体在不同生长发育阶段或不同生理状态下大量基因表达的平行分析，研究相应基因在生物体内的功能，阐明不同层次多基因协同作用的机理，进而在人类重大疾病如癌症、心血管疾病的发病机理、诊断治疗、药物开发等方面的研究发挥巨大的作用。它将大大推动人类结构基因组及功能基因组的各项基因组研究计划。

一直以来，基于分子杂交的方法对基因表达进行分析，从经典的核酸分子杂交方法（southern、northern blotting）到目前的基因芯片技术，都是应用已知核酸序列作为探针与互补的靶核苷酸序列杂交，通过随后的信号检测进行定性与定量分析。

新一代高通量测序技术的发明对生物学特别是基因组学的研究来说具有划时代的意义，它的高通量性使得对一个物种的转录本和基因组进行细致全貌的分析成为可能。随着 Solexa 测序技术的出现，使得高通量，低成本测序成为可能，并且与芯片技术的模拟信号相比，基于 Solexa 测序技术的表达分析避免了芯片技术中的交叉杂交、分析模型复杂以及灵敏度低等缺点；但是，由于高通量测序读长的限制，使其在对未知基因组进行从头测序（de novo sequencing）的应用受到限制，这部分工作仍然需要传统测序手段（读取长度可达到 850 碱基）的协助。而这并不影响高通量测序技术在 mRNA 表达谱、microRNA 表达谱、转录组测序、染色体免疫共沉淀（ChIP-chip，Chromatin Immunoprecipitation）以及 DNA 甲基化等方面的应用。

数字基因表达谱 (DGE, Digital Gene Expression Profiling) 和转录组分析 (RNA-Seq) 是利用新一代高通量测序技术和高性能计算分析技术对某一物种特定组织和状态下的基因表达情况进行序列捕捉和精确解析的新方法。随着新一代测序技术的不断发展, 对基因表达的研究也会更加深入, 因此, 需要对基因表达的分析手段进行相关性评估, 从而排除由于分析手段自身的不准确或不稳定性所造成的分析误差, 从而选取具有较高可靠性的基因表达分析手段, 以便真实反映基因测序的准确性, 确保评估可靠, 从而保证产业可行性和生产的稳定性。

发明内容

本发明要解决的一个技术问题是提供一种关于基因表达的 RNA 测序质控方法及装置, 通过对基因表达的分析为基因测序提供质量控制方案。

本发明的一个方面提供了一种关于基因表达的 RNA 测序的质控方法, 该方法包括: 对测序技术得到的测序片段分别进行数字基因表达谱分析 (DGE) 和转录组分析 (RNA-Seq); 数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测 (qPCR, Real-time Quantitative PCR Detecting System; 其中 PCR, Polymerase Chain Reaction, 聚合酶链式反应) 的结果进行相关性分析; 根据相关性分析结果, 判断数字基因表达谱分析和转录组分析在基因表达定量上的差异, 并从数字基因表达谱分析和转录组分析中选取一种测序分析方式; 从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中, 该方法还包括: 采用高通量测序技术进行关于基因表达的 RNA 测序; 对数字基因表达谱分析的结果和转录组分析的结果分别进行去接头序列和去低质量序列的处理。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中, 通过高通量测序技术对样品片段的基因表达进行多次测序, 并对多次测序的数据取平均值以获得实时定量基因扩增荧光检测的结果。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中, 数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测 (qPCR) 的结果进行相关性分析进一步包括: 当参考基

因不全时，将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测的结果进行相关性分析；和/或 在相同测序量的情况下，比较数字基因表达谱分析的结果和转录组分析的结果检测到的基因数。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中，当参考基因不全时，将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测的结果进行相关性分析的步骤进一步包括：将参考基因从 3'端到 5'端平均切成三份；对三份参考基因分别进行数字基因表达谱分析和转录组分析；将所获得的分析结果分别与实时定量基因扩增荧光检测的结果进行相关性分析。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中，在相同测序量的情况下，比较数字基因表达谱分析的结果和转录组分析的结果检测到的基因数的步骤进一步包括：从高通量测序得到的测序片段中取出三百万标签数据（3M reads）分别进行数字基因表达谱分析和转录组分析；从高通量测序得到的测序片段中取出两百万标签数据（2M reads）分别进行数字基因表达谱分析和转录组分析；和/或 从高通量测序得到的测序片段中取出一百万标签数据（1M reads）分别进行数字基因表达谱分析和转录组分析；在相同测序量的情况下，分别比较数字基因表达谱分析和转录组分析方法能够检测到的基因数。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中，从所选取的测序分析方式获取的分析结果中选取一百万标签数据（1M reads），进行基因表达的测序稳定性分析的步骤进一步包括：从数字基因表达谱分析结果中取出一百万标签数据（1M reads），并将其与全部的数字基因表达谱分析结果进行相关性分析；和/或 从转录组分析结果中取出一百万标签数据（1M reads），并将其与全部的转录组分析结果进行相关性分析。

本发明的另一个方面提供了一种关于基因表达的 RNA 测序的质控装置，该装置包括：基因表达测算模块，用于对测序技术得到的测序片段分别进行数字基因表达谱分析（DGE）和转录组分析（RNA-Seq）；相关性分析模块，用于将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测（qPCR）的结果进行相关性分析；测序分析方式选取模块，用于根据相关性分析结果，判断数字基因表达谱分析和转录组分析在基因表达定量上的差异，并从数字基因表达谱分析和

转录组分析中选取一种测序分析方式；测序稳定性分析模块，用于从所选取的测序分析方式获取的分析结果中选取一百万标签数据（1M reads），进行基因表达的测序稳定性分析。

本发明提供的关于基因表达的 RNA 测序的质控装置的一个实施例中，相关性分析模块进一步包括：第一相关性分析子模块，用于当参考基因不全时，将参考基因从 3'端到 5'端平均切成三份；对三份参考基因分别进行数字基因表达谱分析和转录组分析；将所获得的分析结果分别与实时定量基因扩增荧光检测的结果进行相关性分析；第二相关性分析子模块，用于在相同测序量的情况下，从高通量测序得到的测序片段中取出三百万标签数据（3M reads）分别进行数字基因表达谱分析和转录组分析，从高通量测序得到的测序片段中取出两百万标签数据（2M reads）分别进行数字基因表达谱分析和转录组分析；和/或从高通量测序得到的测序片段中取出一百万标签数据（1M reads）分别进行数字基因表达谱分析和转录组分析；以及在相同测序量的情况下，分别比较数字基因表达谱分析和转录组分析方法能够检测到的基因数。

本发明提供的关于基因表达的 RNA 测序的质控装置的一个实施例中，测序稳定性分析模块进一步包括：第一测序稳定性分析子模块，用于从数字基因表达谱分析结果中取出一百万标签数据（1M reads），并将其与全部的数字基因表达谱分析结果进行相关性分析；第二测序稳定性分析子模块，用于从转录组分析结果中取出一百万标签数据（1M reads），并将其与全部的转录组分析结果进行相关性分析。

本发明提供了一种关于基因表达的 RNA 测序质控方法及装置，通过对基因表达的分析手段进行相关性分析和综合性评估，从而选取具有较高可靠性的基因表达分析手段，真实反映基因测序的准确性，保证产业可行性，为生产的稳定性提供质量控制方案。

附图说明

图 1 示出本发明实施例提供的一种关于基因表达的 RNA 测序的质控方法的流程图；

图 2 示出了本发明两样品的 DGE 分析的结果与 qPCR 结果的相关性分析结果的示意图，其中图 2（a）示出了样品 UHRR 的 DGE 分析的结果与 qPCR 结果的分析结果示意图，图 2（b）示出了样品 HBRR 的 DGE 分析的结果与 qPCR 结果的分析结果示意图；

图 3 示出了本发明两样品的 RNA-Seq 分析的结果与 qPCR 结果的相关性分析结果的示意图, 其中图 3 (a) 示出了样品 UHRR 的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 3 (b) 示出了样品 HBRR 的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图;

图 4 示出本发明提供的关于基因表达的 RNA 测序的质控方法的另一个实施例的流程图;

图 5 示出了本发明样品 UHRR 三等分参考基因序列的 DGE 分析的结果与 qPCR 结果的相关性分析结果的示意图, 其中图 5 (a) 示出了样品 UHRR 第一段的 DGE 分析的结果与 qPCR 结果的分析结果示意图, 图 5 (b) 示出了样品 UHRR 第二段的 DGE 分析的结果与 qPCR 结果的分析结果示意图, 图 5 (c) 示出了样品 UHRR 第三段的 DGE 分析的结果与 qPCR 结果的分析结果示意图;

图 6 示出了本发明样品 UHRR 三等分参考基因序列的 RNA-Seq 分析的结果与 qPCR 结果的相关性分析结果的示意图, 其中图 6 (a) 示出了样品 UHRR 第一段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 6 (b) 示出了样品 UHRR 第二段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 6 (c) 示出了样品 UHRR 第三段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图;

图 7 是本发明样品 UHRR 在相同测序量下, DGE 和 RNA-Seq 检测到的基因数的示意图;

图 8 示出本发明提供的关于基因表达的 RNA 测序的质控方法的另一个实施例的流程图;

图 9 示出本发明实施例提供的一种关于基因表达的 RNA 测序的质控装置的结构示意图;

图 10 示出本发明提供的关于基因表达的 RNA 测序的质控装置的另一个实施例的结构示意图;

图 11 示出本发明提供的关于基因表达的 RNA 测序的质控装置的另一个实施例的结构示意图。

具体实施方式

下面参照附图用本发明的示例性实施例对本发明进行更全面的描述及说明。

图 1 示出本发明实施例提供的一种关于基因表达的 RNA 测序的质控方

法的流程图。

如图 1 所示, 关于基因表达的 RNA 测序的质控方法 100 包括步骤 102, 对测序技术得到的测序片段分别进行数字基因表达谱分析 (DGE) 和转录组分析 (RNA-Seq)。本发明实施例中, 测序方法可以采用高通量测序技术, 例如采用 Illumina GA Solexa 测序技术; Solexa 是一种基于边合成边测序技术 (SBS, Sequencing-By-Synthesis) 的新型测序方法, 通过利用单分子阵列实现在小型芯片 (Flow Cell) 上进行桥式 PCR 反应。新的可逆阻断技术可实现每次只合成一个碱基, 不需要标记荧光基团, 再利用相应的激光激发荧光基团捕获激发光, 从而读取碱基信息。实验可以采用 36 Single End 测序平台, 对 RNA 标准品/实验样品分别进行双酶切测序和随机打断测序。

步骤 104, 数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测 (qPCR) 的结果进行相关性分析。稍后对关于 DGE 和 RNA-Seq 分析的结果与 qPCR 的结果的相关性分析方法做进一步的详细介绍。

步骤 106, 根据相关性分析结果, 判断数字基因表达谱分析和转录组分析在基因表达定量上的差异, 并从数字基因表达谱分析和转录组分析中选取一种测序分析方式。例如, 综合分析数字基因表达谱和 RNA-Seq 在基因表达定量 (涉及基因数和基因表达量) 上的差异, 具体来说, 可以包括分析正常测序量时比较数字基因表达谱和 RNA-Seq 分析的结果与 qPCR 结果的相关性, 分析参考基因不全时比较数字基因表达谱和 RNA-Seq 分析的结果与 qPCR 结果的相关性, 以及在相同测序量下比较数字基因表达谱和 RNA-Seq 能检测到的基因数中的至少任意一种方式。根据前述综合分析结果, 得出 DGE 和 RNA-Seq 在基因表达定量上的差异, 从而选取合适的测序分析方式。

步骤 108, 从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析。例如, 根据前述综合分析, 如果 RNA-Seq 分析方式所获取的基因表达定量更为准确 (即 RNA-Seq 得到的基因表达量更接近于 qPCR 得到的基因表达量), 则从 RNA-Seq 分析方式所获取的分析结果中随机选取 1M reads, 并将其与全部的转录组分析结果进行相关性分析; 所述随机的选取方式可以是将所有测序得到的 reads 完全打乱, 再从中任意取出 1M 的 reads; 如果 DGE 和 RNA-Seq 分析方式所获取的基因表达定量相当, 则可以从中任选一种, 以所选取的方式所

获得的分析结果中选取 1M reads，并将其与全部的转录组分析结果进行相关性分析；从而根据分析结果对生产测序的稳定性进行检测和评估以确保测序工作的准确性（其中关于“检测和评估”主要是通过分析测序结果的重复性，由于 1M reads 的基因数目和表达量是确定的，如果某次测序与确定结果重复性不好就说明该次测序不稳定不正确）。

数字基因表达谱分析（DGE）实验部分主要包括：样本制备实验和测序实验。主要试剂耗材为 Illumina Gene Expression Sample Prep Kit 和 Solexa 测序芯片（flowcell），主要仪器是 Illumina Cluster Station（Illumina 公司）和 Illumina Genome Analyzer（Illumina 公司）系统。具体实验流程：提取 6μg 总 RNA，利用 Oligo(dT)磁珠吸附纯化 mRNA，并以 Oligo(dT)引导反转录合成双链 cDNA。标签 5'末端的产生可用两种内切酶实现：NlaIII 或者 DpnII，通常我们使用 NlaIII，它识别并切断 cDNA 上的 CATG 位点，利用磁珠沉淀纯化带有 cDNA 3'端的片段，将其 5'末端连接 Illumina 接头 1（即序列：ACAGGTTTCAGAGTTCTACAGTCCGACATG）。Illumina 接头 1 与 CATG 位点的结合处是 MmeI 的识别位点，MmeI 是一种识别位点与酶切位点分离的内切酶，酶切 CATG 位点下游 17bp 处，这样就产生了带有接头 1 的 Tag。通过磁珠沉淀去除 3'片段后，在 Tag 3'末端连接 Illumina 接头 2（即序列：CAAGCAGAAGACGGCATACGANN），从而获得两端连有不同接头序列的 21bp 标签 library。经过 15 个循环的 PCR 线性扩增后，通过 6%TBE PAGE 胶电泳纯化 85 碱基条带，解链后，单链分子被加到 Solexa 测序芯片（flowcell）上并固定，每条分子经过原位扩增成为一个单分子簇（cluster）测序模板，加入 4 色荧光标记的 4 种核苷酸，采用边合成边测序法（sequencing by synthesis, SBS）测序。每个通道将产生数百万条原始 Read，Read 的测序读长为 35bp。利用 OligodT 的 beads 富集总 RNA 中 mRNA，并逆转录为双链 cDNA，采用 4 碱基识别酶 NlaIII、酶切双链 cDNA，链接 Illumina 接头 1，利用 MmeI 酶切 3'端 CATG 下游 17bp 碱基，并在 3'端链接 Illumina 接头 2。再加入 Primer GX1 和 Primer GX2 进行 PCR 扩增。扩增后样本通过 6% TBE PAGE 胶回收 85 碱基条带，纯化后通过 Illumina 基因表达测序法测序。转录组分析（RNA-Seq）实验部分测序基本过程包括：提取样品总 RNA 后，用带有 Oligo(dT)的磁珠富集真核生物 mRNA（若为原核生物，则用试剂盒去除 rRNA 后进入下一步）。加入 fragmentation buffer 将 mRNA 打断成短片段，以 mRNA 为模板，用六碱基随机引物（random hexamers）合成第一条 cDNA 链，然后加入缓冲液、

dNTPs、RNase H 和 DNA polymerase I 合成第二条 cDNA 链，再经过 QiaQuick PCR 试剂盒（Qiagen 公司生产）纯化并加 EB 缓冲液洗脱之后做末端修复并连接测序接头，然后用琼脂糖凝胶电泳进行片段大小选择，最后进行 PCR 扩增，使用建好的测序文库进行测序。

接下来对 DGE 和 RNA-Seq 分析的结果与 qPCR 的结果的相关性分析方法做详细介绍：

数字基因表达谱分析（DGE）的结果与 qPCR 的结果的相关性分析，主要涉及 DGE 标准分析中表达量 TPM（Transcripts Per Million clean reads）的计算方式，具体来说：TPM = 每个基因包含的原始 Clean Tags 数/该样本中总 clean Tags 数 $\times 1,000,000$ （参见 Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms, Peter A.C. 't Hoen, Yavuz Ariyurek, *et al.*, Nucleic Acids Research, 15 October 2008, Vol.36, No.21）。

图 2 示出了本发明两样品的 DGE 分析的结果与 qPCR 结果的相关性分析结果的示意图。通常来说，如 DGE 数据产量为 3M reads，可以随机从样品测序数据中取 3M reads 来进行 DGE 结果准确性的分析；所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出 3M 的 reads。由于 UHRR 和 HBRR 是 RNA 标准样品，能下载获取的是该 RNA 标准样品的 qPCR 结果，而样品测序数据是不能下载的，需要自行进行测序。图 2（a）示出了样品 UHRR 的 DGE 分析结果与 qPCR 结果的分析结果示意图，图 2（b）示出了样品 HBRR 的 DGE 分析结果与 qPCR 结果的分析结果示意图；其中本发明使用的 UHRR 是 Stratagene 公司的 Universal Human Reference RNA（UHRR）标准品，HBRR 是 Ambion 公司的 Human Brain Reference RNA（HBRR）标准品。如图 2（a）所示，样品 UHRR 的 DGE 分析的结果与 qPCR 结果的相关系数约为 0.3，如图 2（b）所示，样品 HBRR 的 DGE 分析的结果与 qPCR 结果的相关系数约为 0.53（其中 UHRR 和 HBRR 样品在 DGE 分析中能检测到的基因数都是 716，UHRR 和 HBRR 样品在 qPCR 中能检测到的基因数都是 687）。

转录组分析（RNA-Seq）的结果与 qPCR 结果的相关性分析，主要涉及 RNA-Seq 标准分析中表达量 RPKM（Reads Per Kb per Million reads）的计算方式，具体来说：RNA-Seq 标准分析中表达量 RPKM 的算法（参见 Mapping and quantifying mammalian transcriptomes by RNA-Seq, Ali Mortazavi *et al.*, 30 May 2008, Nature Methods| Advance Online

Publication) 如下所示:

$$RPKM = \frac{10^5 C}{NL / 10^3}$$

其中, RPKM(A)是关于基因 A 的表达量, C 为唯一比对到基因 A 的 reads 数, N 为唯一比对到基因组的总 reads 数, L 为基因 A 编码区的碱基数。

图 3 示出了本发明两样品的 RNA-Seq 分析的结果与 qPCR 结果的相关性分析结果的示意图。通常来说, 如 RNA-Seq 数据产量为 3M reads, 可以从样品测序数据中随机取 3M reads 来进行 DGE 结果准确性的分析, 所述随机的选取方式可以是将所有测序得到的 reads 完全打乱, 再从中任意取出 3M 的 reads。图 3 (a) 示出了样品 UHRR 的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 3 (b) 示出了样品 HBRR 的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图; 其中样品 UHRR 的 RNA-Seq 分析的结果与 qPCR 结果的相关系数约为 0.91, 样品 HBRR 的 RNA-Seq 分析的结果与 qPCR 结果的相关系数约为 0.86 (其中 UHRR 和 HBRR 样品在 RNA-Seq 分析中能检测到的基因数都是 872, UHRR 和 HBRR 样品在 qPCR 中能检测到的基因数都是 851)。此外, 需要说明的是: 从样品 UHRR 和 HBRR 中抽取 3M reads 进行 RNA-Seq 与 qPCR 相关性分析, 与用全部数据计算得到的 RNA-Seq 和 qPCR 的相关系数相同, 都分别是 0.91 和 0.86。由此可见, 对于基因测序的数据量来说, 其对 RNA-Seq 的定量分析几无影响或者说影响甚微。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中, 在对测序技术得到的测序片段分别进行 DGE 和 RNA-Seq 分析步骤之前, 对数字基因表达谱分析的结果和转录组分析的结果分别进行去接头序列; 进一步地, 也可以对去接头序列的结果再进行去低质量序列的处理, 从而获取能够用于标签的数据 (clean tag) 以进行后续分析。

本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例中, 通过高通量测序技术对样品片段的基因表达进行多次测序, 并对多次测序的数据取平均值以获得实时定量基因扩增荧光检测的结果。例如, UHRR 和 HBRR 样品的 qPCR 数据是从 GEO (高通量基因表达, Gene Expression Omnibus) 上下载的, 具体来说其下载路径: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5350>, 其中

UHRR 的 accession number 是 GSM129638, 公开日是 2006 年 9 月 8 日; HBRR 的 accession number 是 GSM129645, 公开日是 2006 年 9 月 8 日。对 UHRR 和 HBRR 样品分别进行多次测序(如 4 次)的平行实验, 并对该 4 次平行实验的关于基因数和基因表达量的结果取平均值以作为 qPCR 定量结果。

本发明提供的关于基因表达的 RNA 测序的质控方法, 基于对测序片段进行 DGE 和 RNA-Seq 分析, 并对 DGE 和 RNA-Seq 分析的结果与 qPCR 结果进行相关性的综合分析, 从而选取适宜的测序分析方式进行基因表达的测序稳定性分析。通过本发明提供的关于基因表达的 RNA 测序的质控方法的一个实施例, 其能够真实反映基因测序的准确性, 保证产业可行性, 为生产的稳定性提供质量控制方案。

图 4 示出本发明提供的关于基因表达的 RNA 测序的质控方法的另一个实施例的流程图。

如图 4 所示, 关于基因表达的 RNA 测序的质控方法 400 包括步骤 402、404、405、406、408, 其中步骤 402、406 和 408 可以分别执行与可以分别执行与图 1 所示的步骤 102、106 和 108 相同或相似的技术内容, 为简洁起见, 这里不再赘述其技术内容。

如图 4 所示, 在步骤 402 后, 执行步骤 404, 当参考基因不全时, 将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测的结果进行相关性分析。参考基因都是现有数据库中已经拼接好的核酸序列 (<http://www.ncbi.nlm.nih.gov/>), 这些核酸序列有很多版本(由不同的研究机构, 数据中心等单位发布的), 每个机构由于其技术水平的限制, 所以发布的结果与基因的真实情况是有不同的, 因此可能存在参考基因不全或不完整的情形。例如, 在用 DGE 分析的结果和 RNA-Seq 分析的结果分别与实时定量基因扩增荧光检测(qPCR)的结果进行相关性分析时, 当参考基因不全/不完整时, 可以采用如下方式进行相关性的分析。

具体来说, 在非模式生物中, 有理由怀疑参考基因序列的不全会造成 DGE 定量不准确; 首先, 将完整的参考基因序列(如 NCBI 中人的 refseq 基因)从 3'端开始进行三等分, 然后将三等分的基因序列当作完整的参考基因序列, 分别进行 DGE 分析的结果与 qPCR 结果的相关性分析。图 5 示出了本发明样品 UHRR 三等分参考基因序列的 DGE 分析的结果与 qPCR 结果的相关性分析结果的示意图, 其中图 5(a)示出了样品 UHRR 第一段的 DGE 分析的结果与 qPCR 结果的分析结果示意图,

图 5 (b) 示出了样品 UHRR 第二段的 DGE 分析的结果与 qPCR 结果的分析结果示意图, 图 5 (c) 示出了样品 UHRR 第三段的 DGE 分析的结果与 qPCR 结果的分析结果示意图; 分析发现样品 UHRR 的这三部分序列, 其 DGE 分析结果与 qPCR 结果的相关系数分别约为 0.71, 0.39 和 0.33 (用完整基因序列作分析时 DGE 分析结果与 qPCR 结果的相关系数为 0.76), 其在 DGE 分析中能检测到的基因数分别是 774, 596 和 435。同样地, 对于采用 RNA-Seq 分析方式进行基因序列表达的, 也是先将完整的参考基因序列从 3'端开始进行三等分, 然后将三等分的基因序列当作完整的参考基因序列, 分别进行 RNA-Seq 与 qPCR 相关性分析。图 6 示出了本发明样品 UHRR 三等分参考基因序列的 RNA-Seq 分析的结果与 qPCR 结果的相关性分析结果的示意图, 其中图 6 (a) 示出了样品 UHRR 第一段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 6 (b) 示出了样品 UHRR 第二段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图, 图 6 (c) 示出了样品 UHRR 第三段的 RNA-Seq 分析的结果与 qPCR 结果的分析结果示意图; 分析发现样品 UHRR 的这三部分序列, 其 RNA-Seq 分析结果与 qPCR 结果的相关系数分别约为 0.85、0.91 和 0.84 (用完整基因序列作分析时 RNA-Seq 与 qPCR 的相关系数为 0.91), 其在 RNA-Seq 分析中能检测到的基因数分别是 917、911 和 896。

DGE 由于其自身的缺点, 它无法检测出不含 CATG (或 GATC) 位点的基因, 倾向于得到每条 mRNA 最靠近 3'端的 Tag 作为该 mRNA 的标签, 因此它对参考基因的要求比较严格, 参考基因序列不完整对 DGE 结果的影响很大; 而 RNA-Seq 是对 mRNA 进行随机打断, 所以每条 mRNA 能够得到很多标签, 对参考基因的依赖性不是很强, 在参考基因不完整的情况下也能够得到比较准确的表达量信息。由此可见, 基因序列不完整对 DGE 分析结果影响较大, 而对于 RNA-Seq 的分析结果影响不大; 也就是说, 对于参考基因不全时, 如果使用 DGE 进行分析, 则优选采用基因表达 3'端开始的第一段; 而进一步地, 优选采用 RNA-Seq 分析方式对基因片段进行表达分析。

步骤 405, 在相同测序量的情况下, 比较数字基因表达谱分析的结果和转录组分析的结果检测到的基因数。例如, 在相同测序量下, 比较 DGE 和 RNA-Seq 两种分析方式检测到的基因数具体可以包括: 从高通量测序得到的测序片段中随机取出 3M 的 reads 数分别进行数字基因表达谱

和 RNA-Seq 分析，从高通量测序得到的测序片段中随机取出 2M 的 reads 数分别进行数字基因表达谱和 RNA-Seq 分析，从高通量测序得到的测序片段中随机取出 1M 的 reads 数分别进行 DGE 和 RNA-Seq 分析；所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出对应数量的 reads。在相同测序量的情况下，可以从前述三种方式中任选至少一种方式来分别比较数字基因表达谱和 RNA-Seq 能检测到的基因数。图 7 是本发明样品 UHRR 在相同测序量下，DGE 和 RNA-Seq 检测到的基因数的示意图。如图 7 所示，在相同测序量时 RNA-Seq 能检测到的基因比 DGE 多约 1000 个基因。

图 8 示出本发明提供的关于基因表达的 RNA 测序的质控方法的另一个实施例的流程图。

如图 8 所示，关于基因表达的 RNA 测序的质控方法 800 包括步骤 802、804、806、808、809，其中步骤 802、804、806 可以分别执行与可以分别执行与图 1 所示的步骤 102、104、106 相同或相似的技术内容，为简洁起见，这里不再赘述其技术内容。

如图 8 所示，在步骤 806 后，执行步骤 808，从数字基因表达谱分析结果中随机取出一百万标签数据（1M reads），并将其与全部的数字基因表达谱分析结果进行相关性分析。所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出 1M 的 reads。

步骤 809，从转录组分析结果中随机取出一百万标签数据（1M reads），并将其与全部的转录组分析结果进行相关性分析。所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出 1M 的 reads。

关于步骤 808 和步骤 809 中的相关性分析方法，可以采用如下方式：每个 lane 加入一个已知库（只测 1M 左右 reads），通过比较已知库的测序结果检测测序稳定性。例如，本发明实施例中采用 1M 测序量标准品相关性分析方法，两次 1M 测序量标准品 UHRR 与重复实验样品 UHRR 相关性对照表，如表 1 所示。

样品	Gene 相关性 (Spearman)	Gene 相关性 (Pearson)
标准品 UHRR / 重复实验样品 UHRR	0.99	1.00

表 1 1M 测序量标准品与重复实验的数据相关性对照表

表 1 表明, 对 1M 测序量标准品与重复实验的数据进行分析比对发现, Gene 相关性无论是 Spearman 相关系数还是 Pearson 相关系数都很高; 由此可以说明测序结果是正常、可信任的, 使用标准品检测测序稳定性的方法具有可行性的, 能够通过对基因表达的分析为基因测序提供质量控制方案。用 1M reads 进行基因表达质控的方法在 RNA 测序中的应用能够评估生产稳定性。

此外, 需要说明的是, 对于本发明的质控方法和装置而言, 不论是从定量准确性的角度来说, 还是从检测到的基因数目来说, 又或者对参考基因的依赖性 etc 角度来比较, 在质控方案中采用 RNA-Seq 分析方法具有比 DGE 更准确地反映基因表达的优势。

图 9 示出本发明实施例提供的一种关于基因表达的 RNA 测序的质控装置的结构示意图。

如图 9 所示, 一种关于基因表达的 RNA 测序的质控装置 900 包括: 基因表达测算模块 902、相关性分析模块 904、测序分析方式选取模块 906 和测序稳定性分析模块 908。

其中, 基因表达测算模块 902, 用于对测序技术得到的测序片段分别进行数字基因表达谱分析 (DGE) 和转录组分析 (RNA-Seq);

相关性分析模块 904, 用于将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测 (qPCR) 的结果进行相关性分析。

测序分析方式选取模块 906, 用于根据相关性分析结果, 判断数字基因表达谱分析和转录组分析在基因表达定量上的差异, 并从数字基因表达谱分析和转录组分析中选取一种测序分析方式。

测序稳定性分析模块 908, 用于从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析。

图 10 示出本发明提供的关于基因表达的 RNA 测序的质控装置的另一个实施例的结构示意图。

如图 10 所示, 一种关于基因表达的 RNA 测序的质控装置 1000 包括: 基因表达测算模块 1002、相关性分析模块 1004、测序分析方式选取模块 1006 和测序稳定性分析模块 1008, 其中基因表达测算模块 1002、测序分析方式选取模块 1006 和测序稳定性分析模块 1008 可以是与图 9

所示基因表达测算模块 902、测序分析方式选取模块 906 和测序稳定性分析模块 908 相同或相似的功能模块。为简洁起见，这里不再赘述。

如图 10 所示，相关性分析模块 1004 进一步包括：第一相关性分析子模块和第二相关性分析子模块；其中

第一相关性分析子模块 10041，用于当参考基因不全时，将参考基因从 3'端到 5'端平均切成三份；对三份参考基因分别进行数字基因表达谱分析和转录组分析；将所获得的分析结果分别与实时定量基因扩增荧光检测的结果进行相关性分析。

第二相关性分析子模块 10042，用于在相同测序量的情况下，从高通量测序得到的测序片段中随机取出三百万标签数据（3M reads）分别进行数字基因表达谱分析和转录组分析，从高通量测序得到的测序片段中随机取出两百万标签数据（2M reads）分别进行数字基因表达谱分析和转录组分析；和/或 从高通量测序得到的测序片段中随机取出一百万标签数据（1M reads）分别进行数字基因表达谱分析和转录组分析；所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出对应数量的 reads。以及在相同测序量的情况下，分别比较数字基因表达谱分析和转录组分析方法能够检测到的基因数。

图 11 示出本发明提供的关于基因表达的 RNA 测序的质控装置的另一个实施例的结构示意图。

如图 11 所示，一种关于基因表达的 RNA 测序的质控装置 1100 包括：基因表达测算模块 1102、相关性分析模块 1104、测序分析方式选取模块 1106 和测序稳定性分析模块 1108。其中基因表达测算模块 1102、相关性分析模块 1104、测序分析方式选取模块 1106 可以是与图 9 所示基因表达测算模块 902、相关性分析模块 904、测序分析方式选取模块 906 相同或相似的功能模块。为简洁起见，这里不再赘述。

如图 11 所示，测序稳定性分析模块 1108 进一步包括：第一测序稳定性分析子模块 11081 和第二测序稳定性分析子模块 11082，其中

第一测序稳定性分析子模块 11081，用于从数字基因表达谱分析结果中随机取出一百万标签数据（1M reads），并将其与全部的数字基因表达谱分析结果进行相关性分析。所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出 1M 的 reads。

第二测序稳定性分析子模块 11082，用于从转录组分析结果中随机取出一百万标签数据（1M reads），并将其与全部的转录组分析结果进行相

关性分析。所述随机的选取方式可以是将所有测序得到的 reads 完全打乱，再从中任意取出 1M 的 reads。

本发明提供的关于基因表达的 RNA 测序的质控装置，通过基因表达测算模块对基因片段进行分析，并通过相关性分析模块和测序分析方式选取模块进行相关性分析和综合性评估，从而选取具有较高可靠性的基因表达分析手段，真实反映基因测序的准确性，为生产的稳定性提供质量控制方案。

参考前述本发明示例性的描述，本领域技术人员可以清楚的知晓本发明提供的关于基因表达的 RNA 测序的质控方法及装置所具有的前述优点，本发明提供的质控方案适用于高通量测序技术，能够有效地评估 RNA 测序的稳定性，确保测序工作的准确性。

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。本发明中描述的功能模块以及功能模块的划分方式仅为说明本发明的思想，本领域技术人员根据本发明的教导以及实际应用的需要可以自由改变功能模块的划分方式及其模块构造以实现相同的功能；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。

权利要求书

1. 一种关于基因表达的 RNA 测序的质控方法，其特征在于，所述方法包括：

对测序技术得到的测序片段分别进行数字基因表达谱分析（DGE）和转录组分析（RNA-Seq）；

所述数字基因表达谱分析的结果和所述转录组分析的结果分别与实时定量基因扩增荧光检测（qPCR）的结果进行相关性分析；

根据相关性分析结果，判断数字基因表达谱分析和转录组分析在基因表达定量上的差异，并从所述数字基因表达谱分析和转录组分析中选取一种测序分析方式；

从所选取的测序分析方式获取的分析结果中选取一百万标签数据（1M reads），进行基因表达的测序稳定性分析。

2. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

采用高通量测序技术进行关于基因表达的 RNA 测序；

对所述数字基因表达谱分析的结果和所述转录组分析的结果分别进行去接头序列和去低质量序列的处理。

3. 如权利要求 1 所述的方法，其特征在于，通过高通量测序技术对样品片段的基因表达进行多次测序，并对多次测序的数据取平均值以获得所述实时定量基因扩增荧光检测的结果。

4. 如权利要求 1 所述的方法，其特征在于，所述数字基因表达谱分析的结果和所述转录组分析的结果分别与实时定量基因扩增荧光检测（qPCR）的结果进行相关性分析进一步包括：

当参考基因不全时，将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测的结果进行相关性分析；和/或

在相同测序量的情况下，比较数字基因表达谱分析的结果和转录组分析的结果检测到的基因数。

5. 如权利要求 4 所述的方法，其特征在于，所述当参考基因不全时，将数字基因表达谱分析的结果和转录组分析的结果分别与实时定量基因扩增荧光检测的结果进行相关性分析的步骤进一步包括：

将所述参考基因从 3'端到 5'端平均切成三份；

对所述三份参考基因分别进行数字基因表达谱分析和转录组分析；

将所获得的分析结果分别与实时定量基因扩增荧光检测的结果进行相关性分析。

6. 如权利要求 4 所述的方法, 其特征在于, 所述在相同测序量的情况下, 比较数字基因表达谱分析的结果和转录组分析的结果检测到的基因数的步骤进一步包括:

从高通量测序得到的测序片段中取出三百万标签数据 (3M reads) 分别进行数字基因表达谱分析和转录组分析, 从高通量测序得到的测序片段中取出两百万标签数据 (2M reads) 分别进行数字基因表达谱分析和转录组分析; 和/或从高通量测序得到的测序片段中取出一百万标签数据 (1M reads) 分别进行数字基因表达谱分析和转录组分析;

在相同测序量的情况下, 分别比较数字基因表达谱分析和转录组分析方法能够检测到的基因数。

7. 如权利要求 1 所述的方法, 其特征在于, 所述从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析的步骤进一步包括:

从数字基因表达谱分析结果中取出一百万标签数据 (1M reads), 并将其与全部的数字基因表达谱分析结果进行相关性分析; 和/或

从转录组分析结果中取出一百万标签数据 (1M reads), 并将其与全部的转录组分析结果进行相关性分析。

8. 一种关于基因表达的 RNA 测序的质控装置, 其特征在于, 所述装置包括:

基因表达测算模块, 用于对测序技术得到的测序片段分别进行数字基因表达谱分析 (DGE) 和转录组分析 (RNA-Seq);

相关性分析模块, 用于将所述数字基因表达谱分析的结果和所述转录组分析的结果分别与实时定量基因扩增荧光检测 (qPCR) 的结果进行相关性分析;

测序分析方式选取模块, 用于根据相关性分析结果, 判断数字基因表达谱分析和转录组分析在基因表达定量上的差异, 并从所述数字基因表达谱分析和转录组分析中选取一种测序分析方式;

测序稳定性分析模块, 用于从所选取的测序分析方式获取的分析结果中选取一百万标签数据 (1M reads), 进行基因表达的测序稳定性分析。

9. 如权利要求 8 所述的装置, 其特征在于, 所述相关性分析模块进一步包括:

第一相关性分析子模块，用于当参考基因不全时，将所述参考基因从 3' 端到 5' 端平均切成三份；对所述三份参考基因分别进行数字基因表达谱分析和转录组分析；将所获得的分析结果分别与实时定量基因扩增荧光检测的结果进行相关性分析；

第二相关性分析子模块，用于在相同测序量的情况下，从高通量测序得到的测序片段中取出三百万标签数据（3M reads）分别进行数字基因表达谱分析和转录组分析；从高通量测序得到的测序片段中取出两百万标签数据（2M reads）分别进行数字基因表达谱分析和转录组分析；和/或从高通量测序得到的测序片段中取出一百万标签数据（1M reads）分别进行数字基因表达谱分析和转录组分析；以及在相同测序量的情况下，分别比较数字基因表达谱分析和转录组分析方法能够检测到的基因数。

10. 如权利要求 8 所述的装置，其特征在于，所述测序稳定性分析模块进一步包括：

第一测序稳定性分析子模块，用于从数字基因表达谱分析结果中取出一百万标签数据（1M reads），并将其与全部的数字基因表达谱分析结果进行相关性分析；

第二测序稳定性分析子模块，用于从转录组分析结果中取出一百万标签数据（1M reads），并将其与全部的转录组分析结果进行相关性分析。

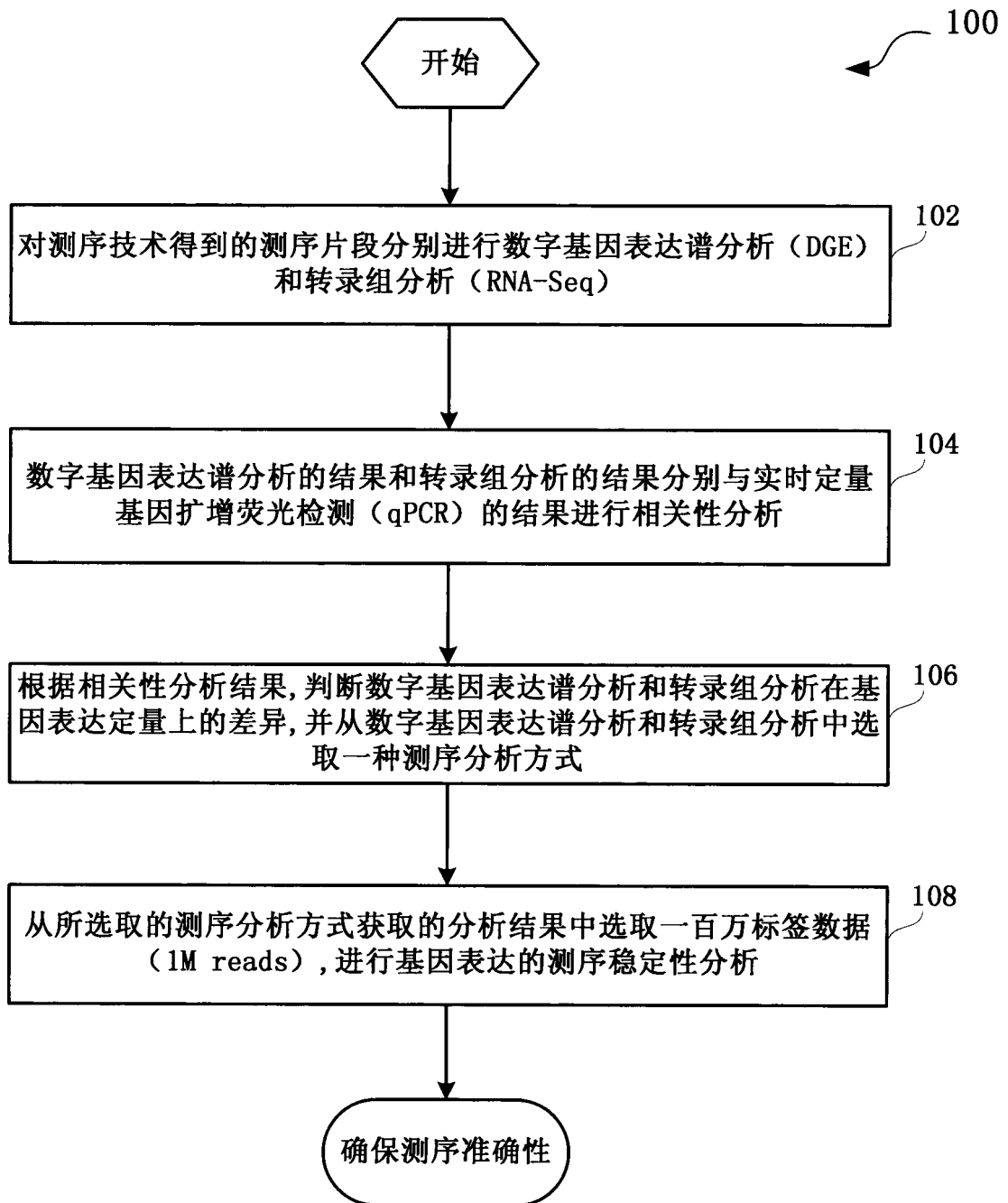


图 1

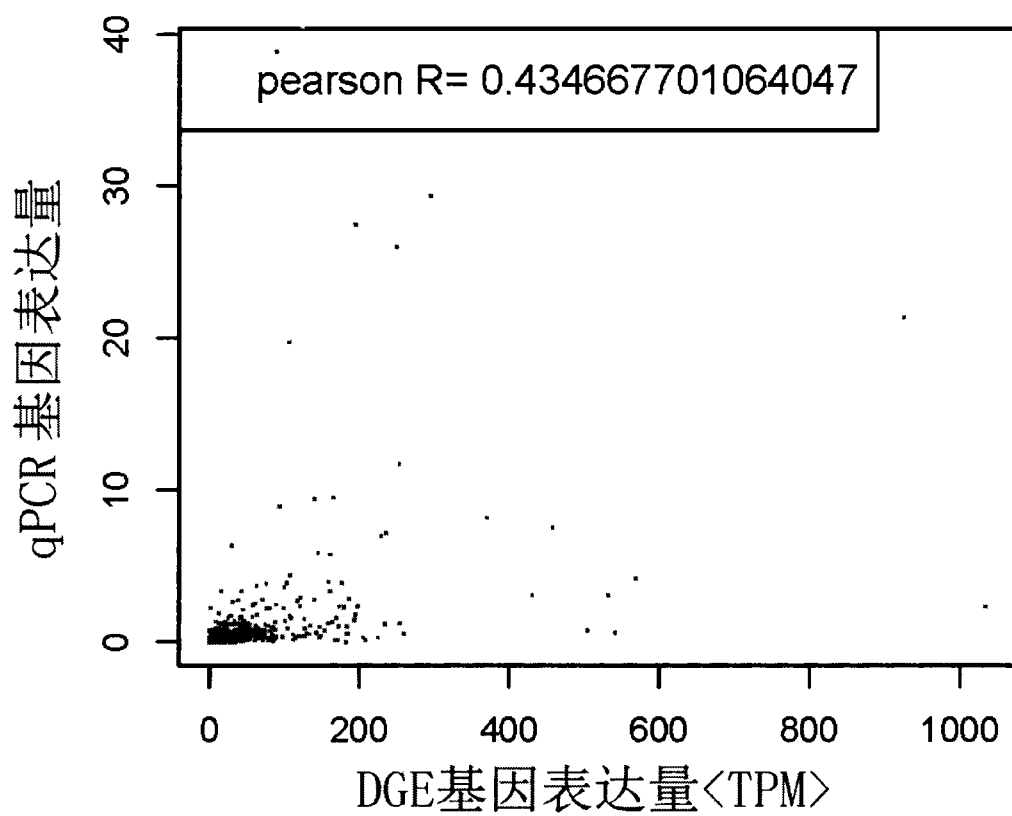


图 2 (a)

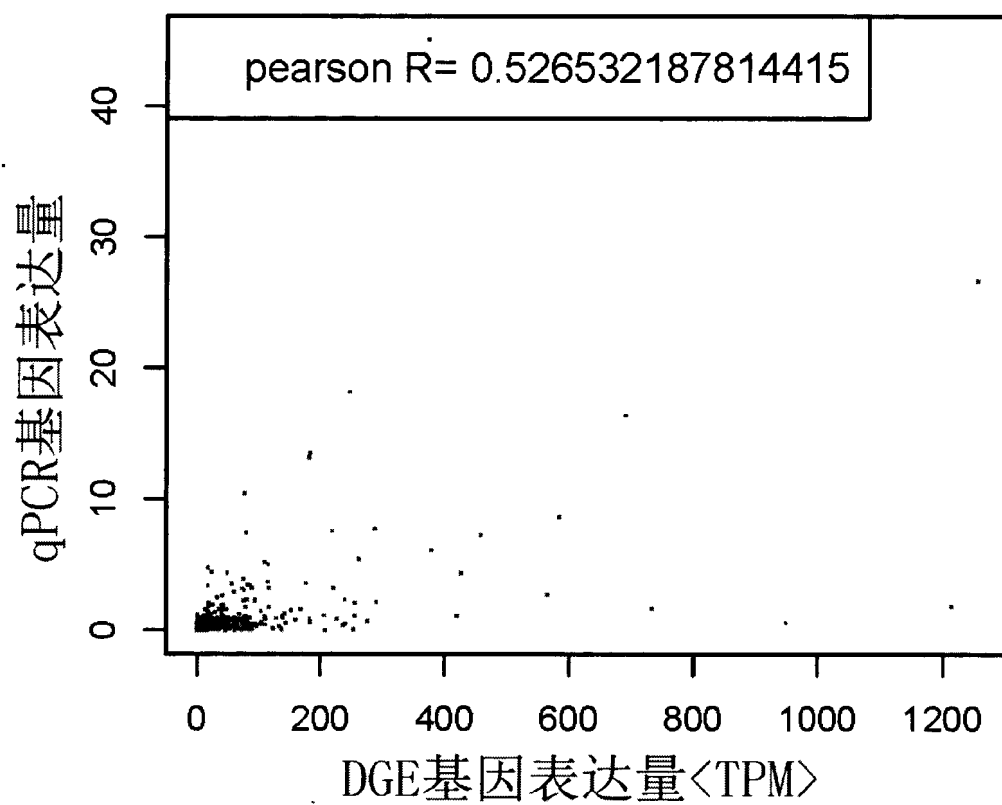


图 2 (b)

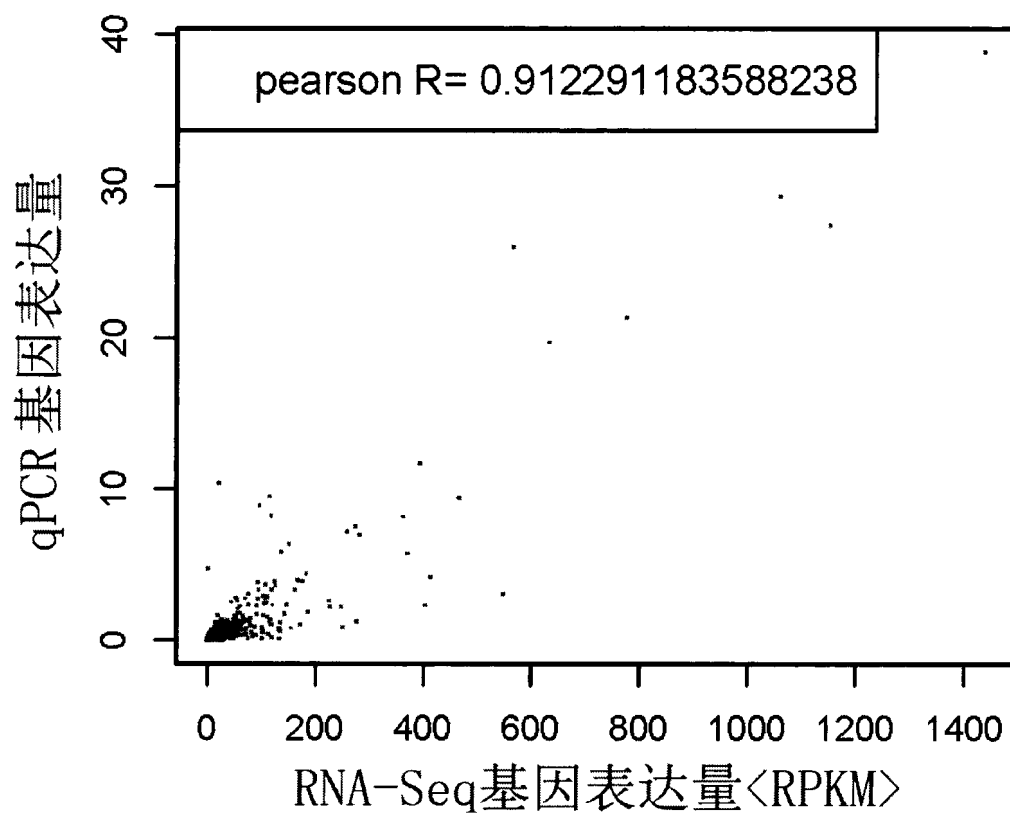


图 3 (a)

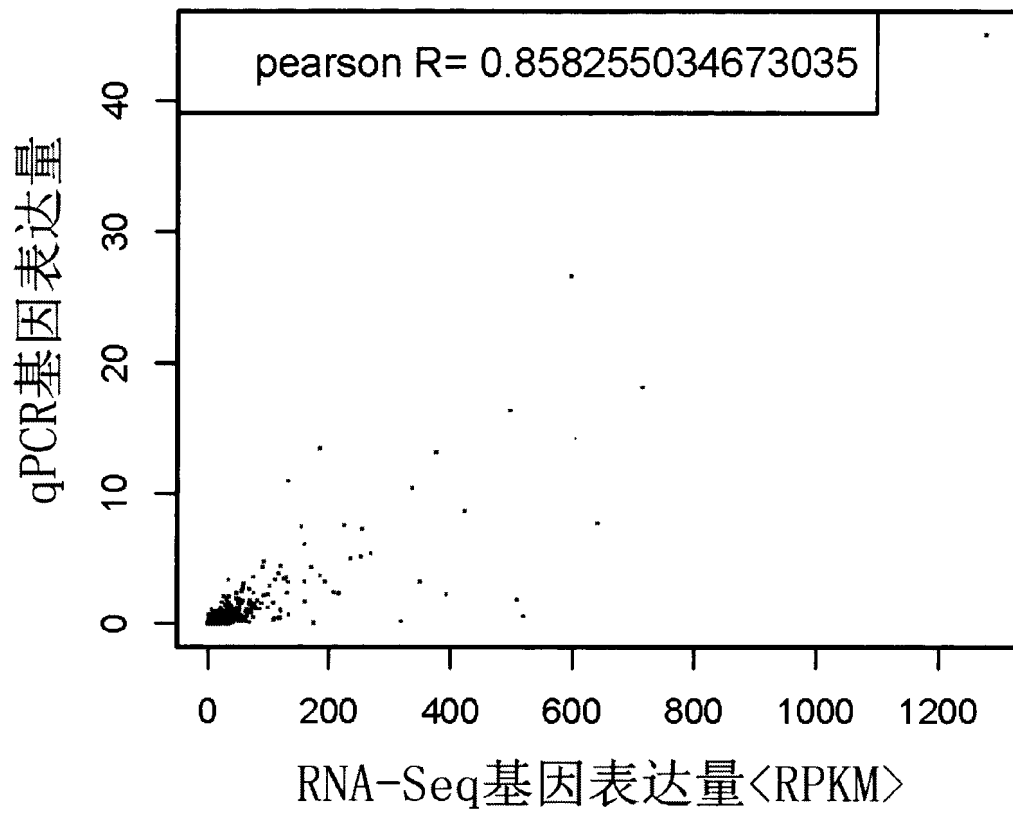


图 3 (b)

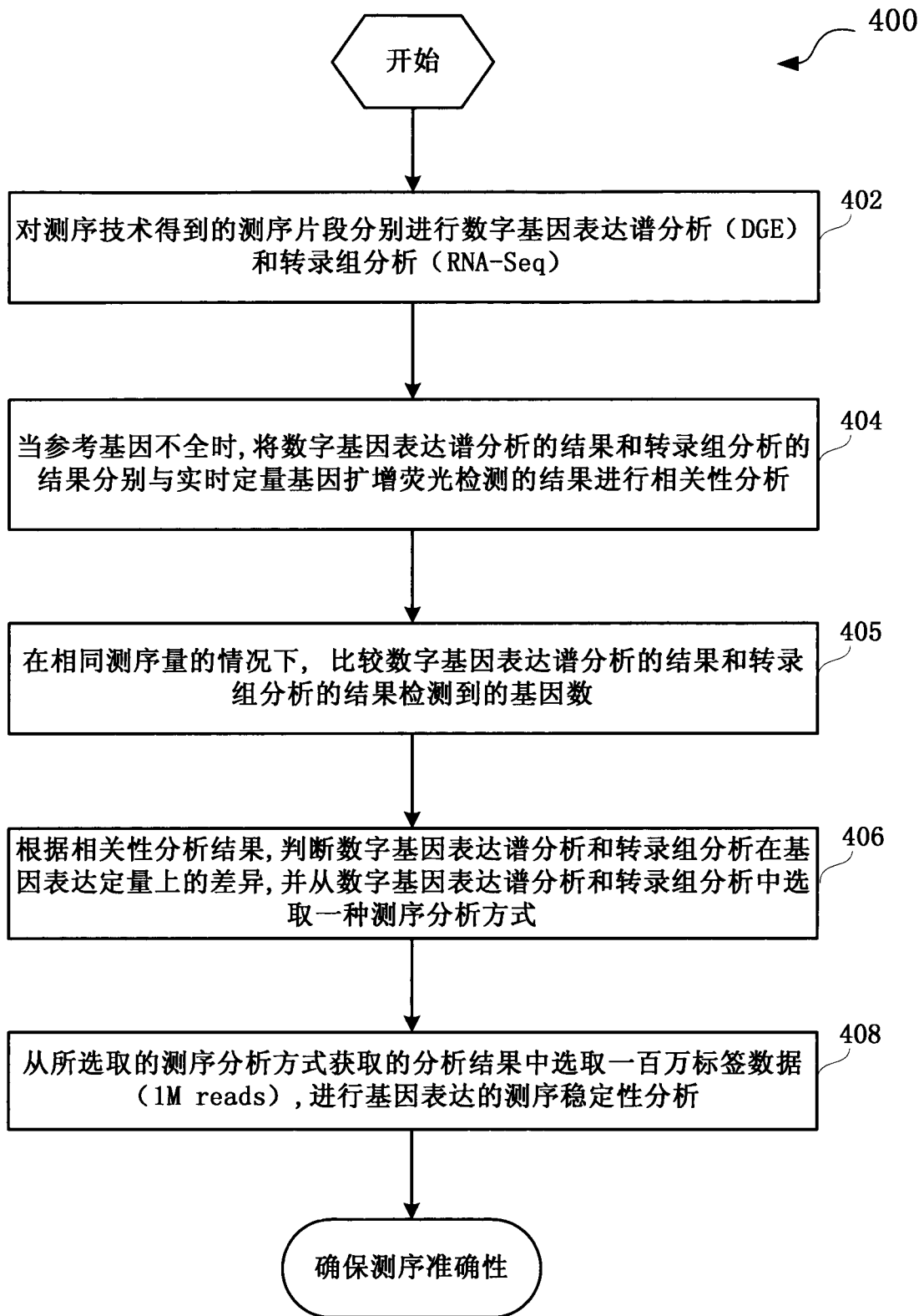


图 4

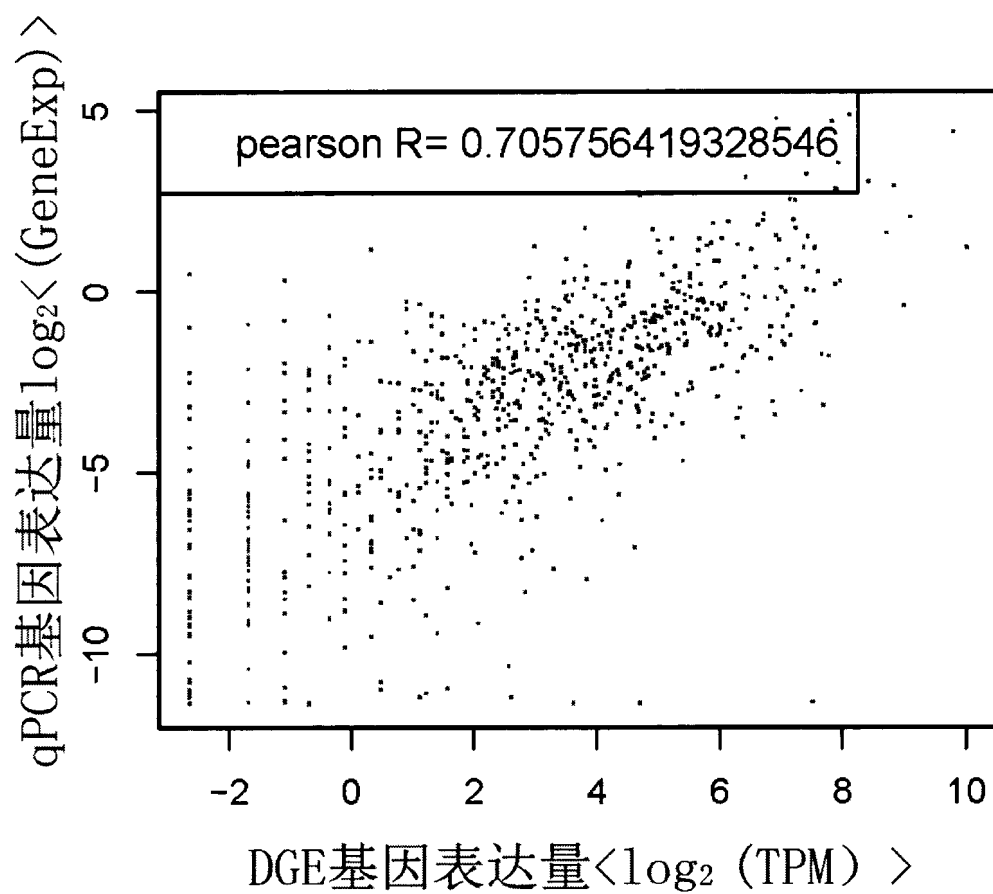


图 5 (a)

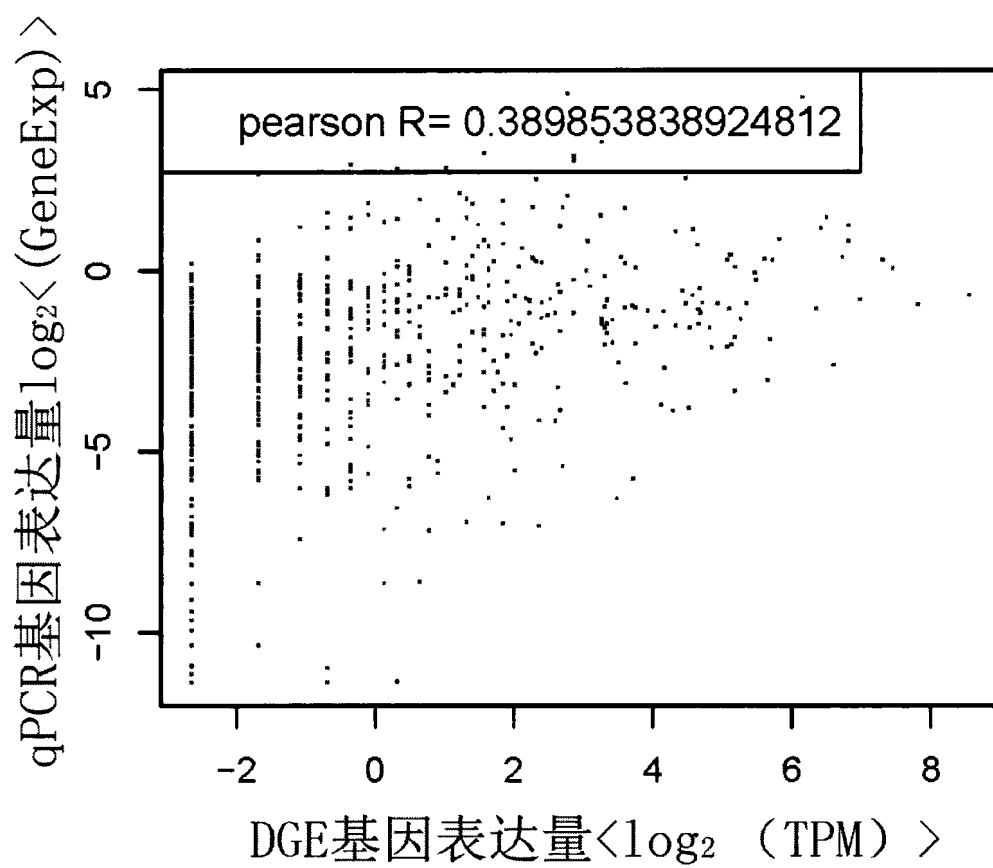


图 5 (b)

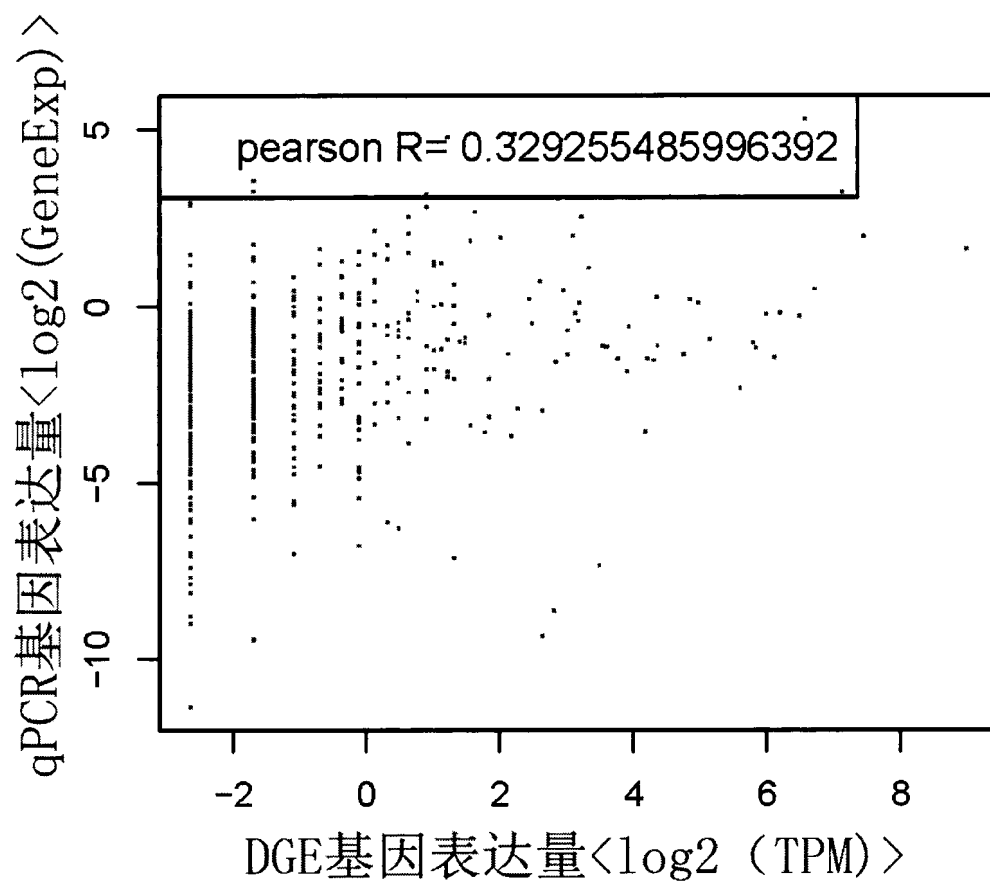


图 5 (c)

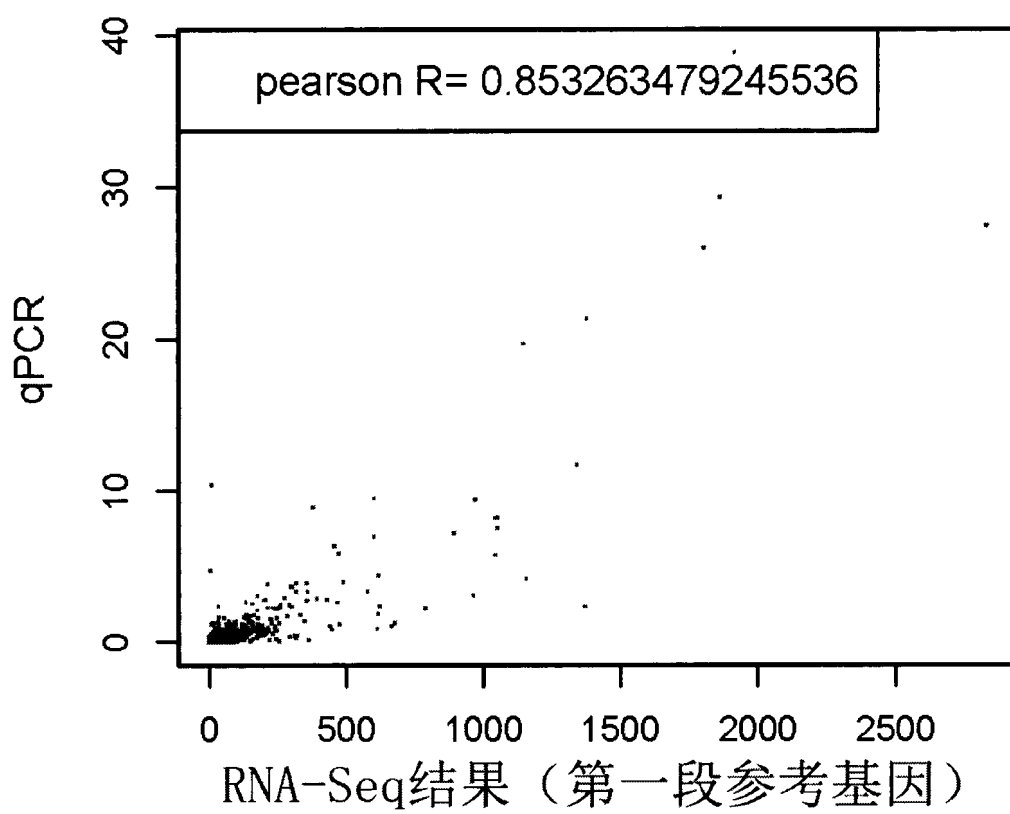


图 6 (a)

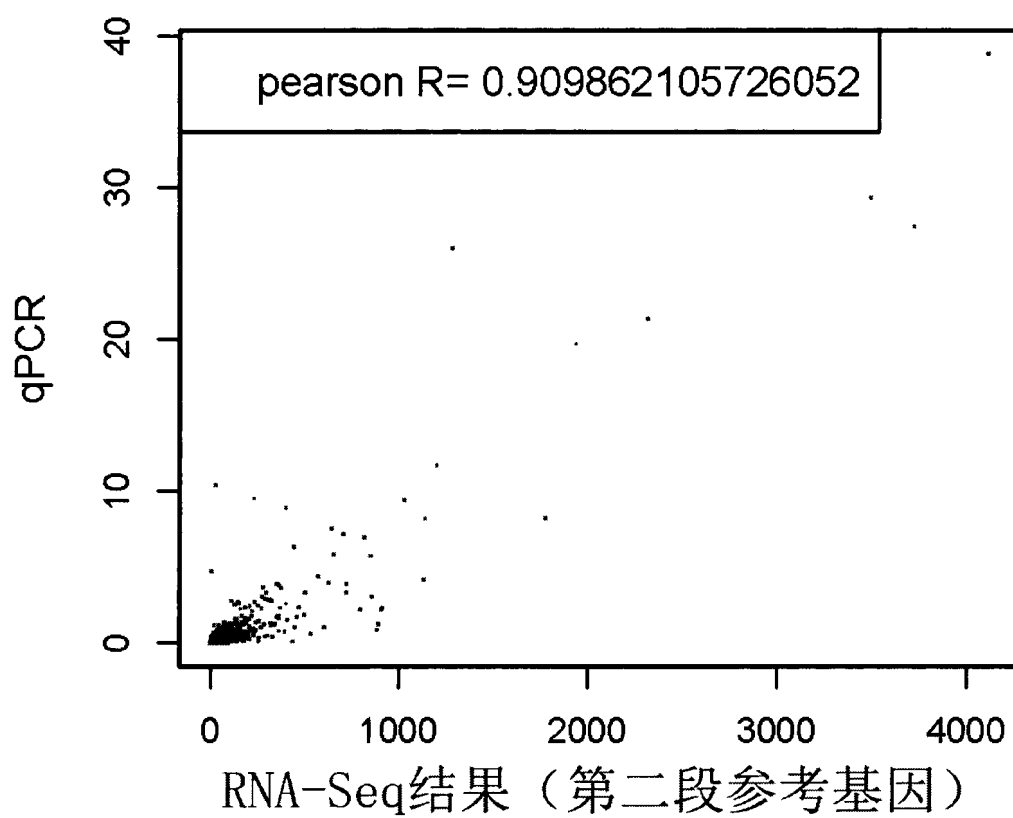


图 6 (b)

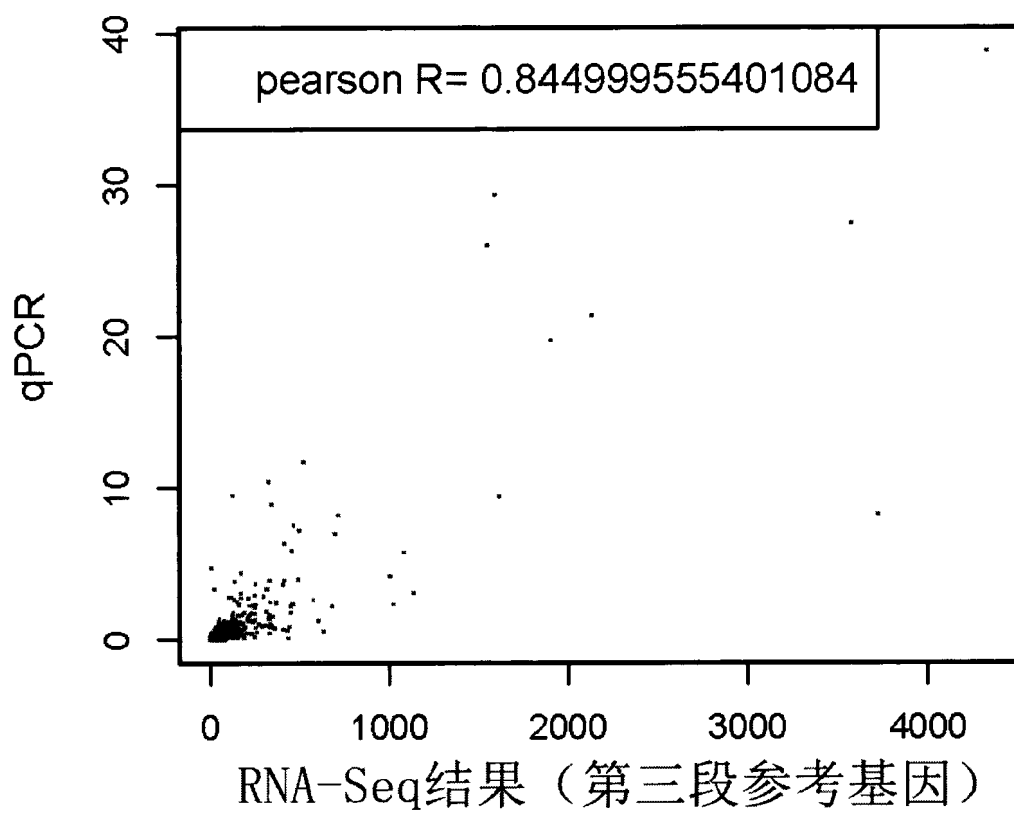


图 6 (c)

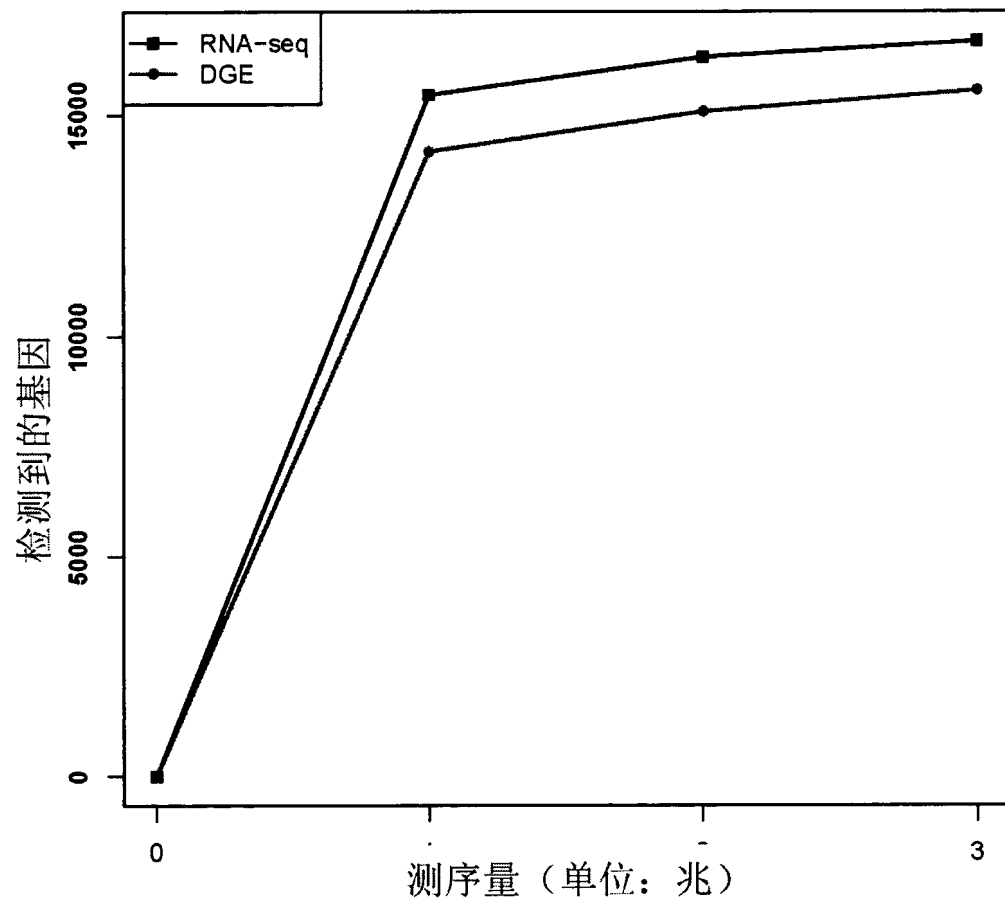


图 7

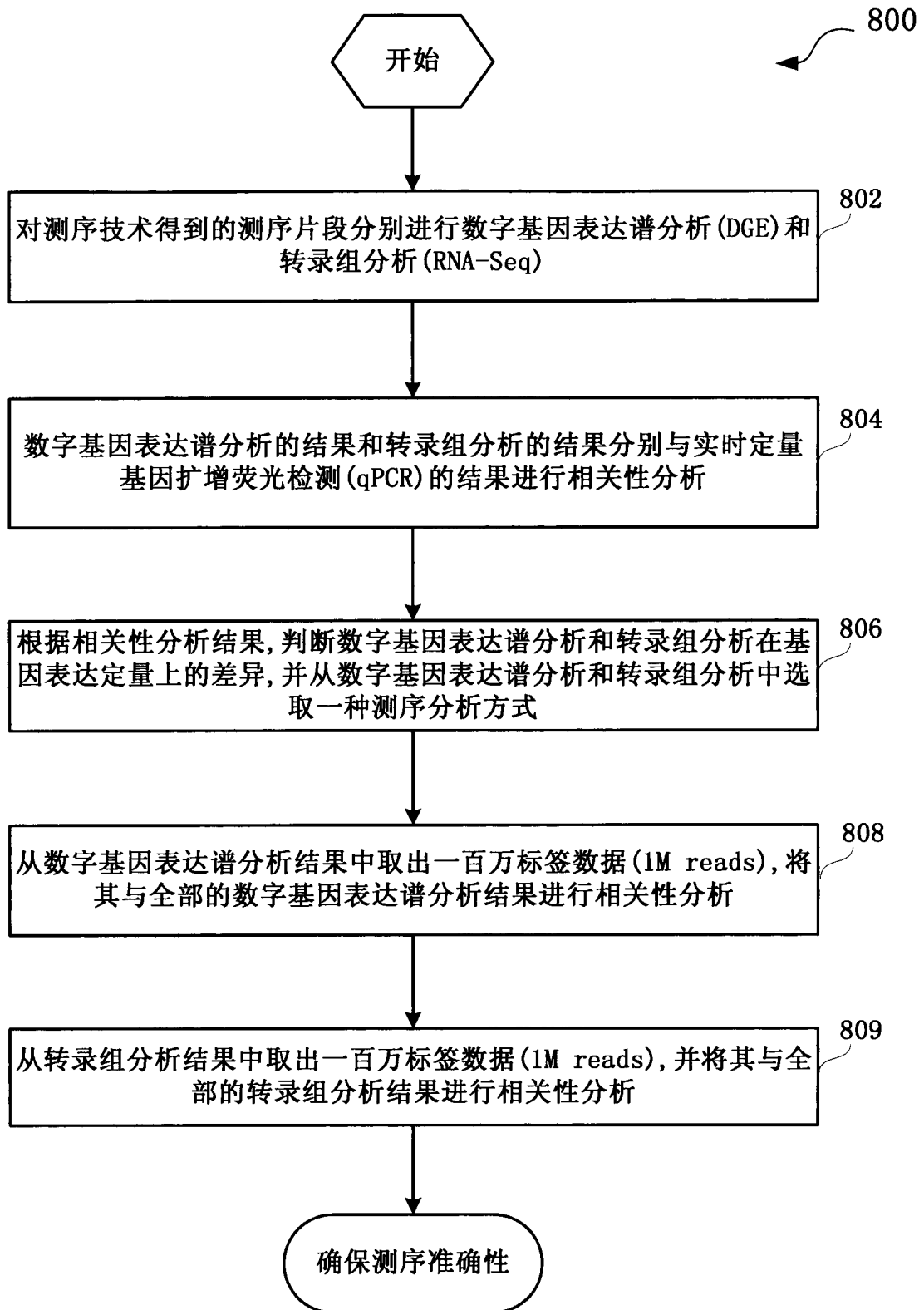


图 8

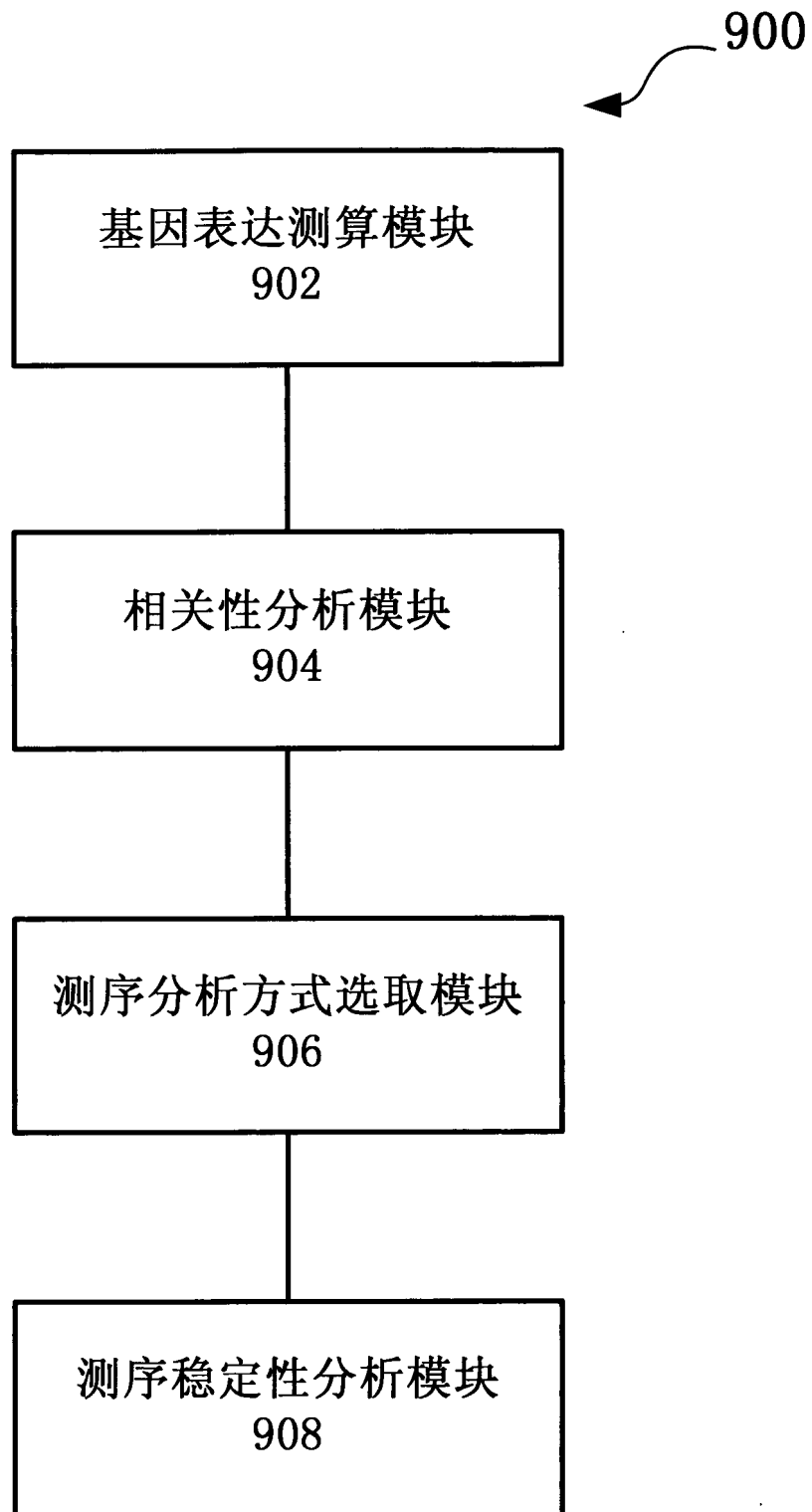


图 9

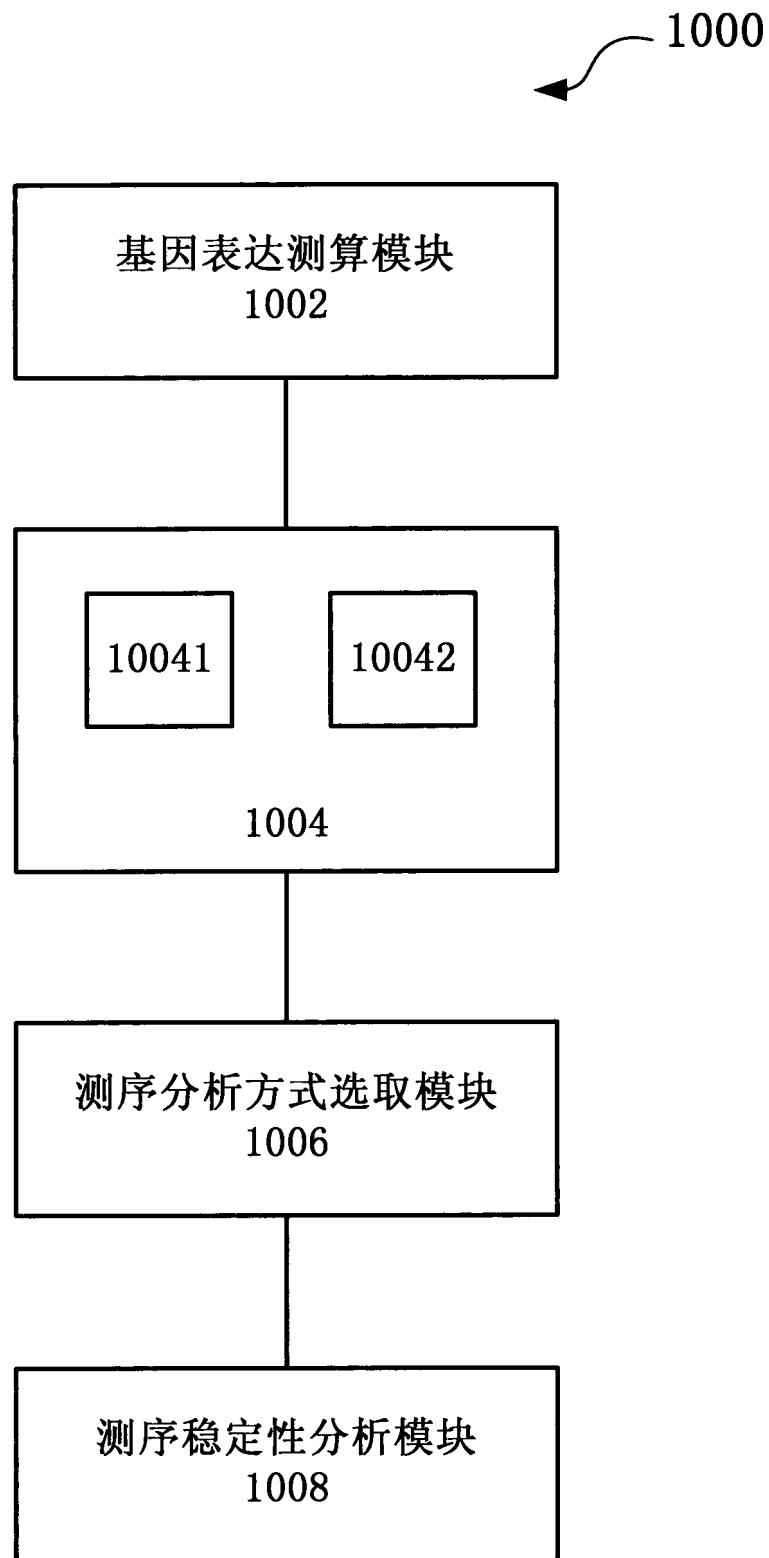


图 10

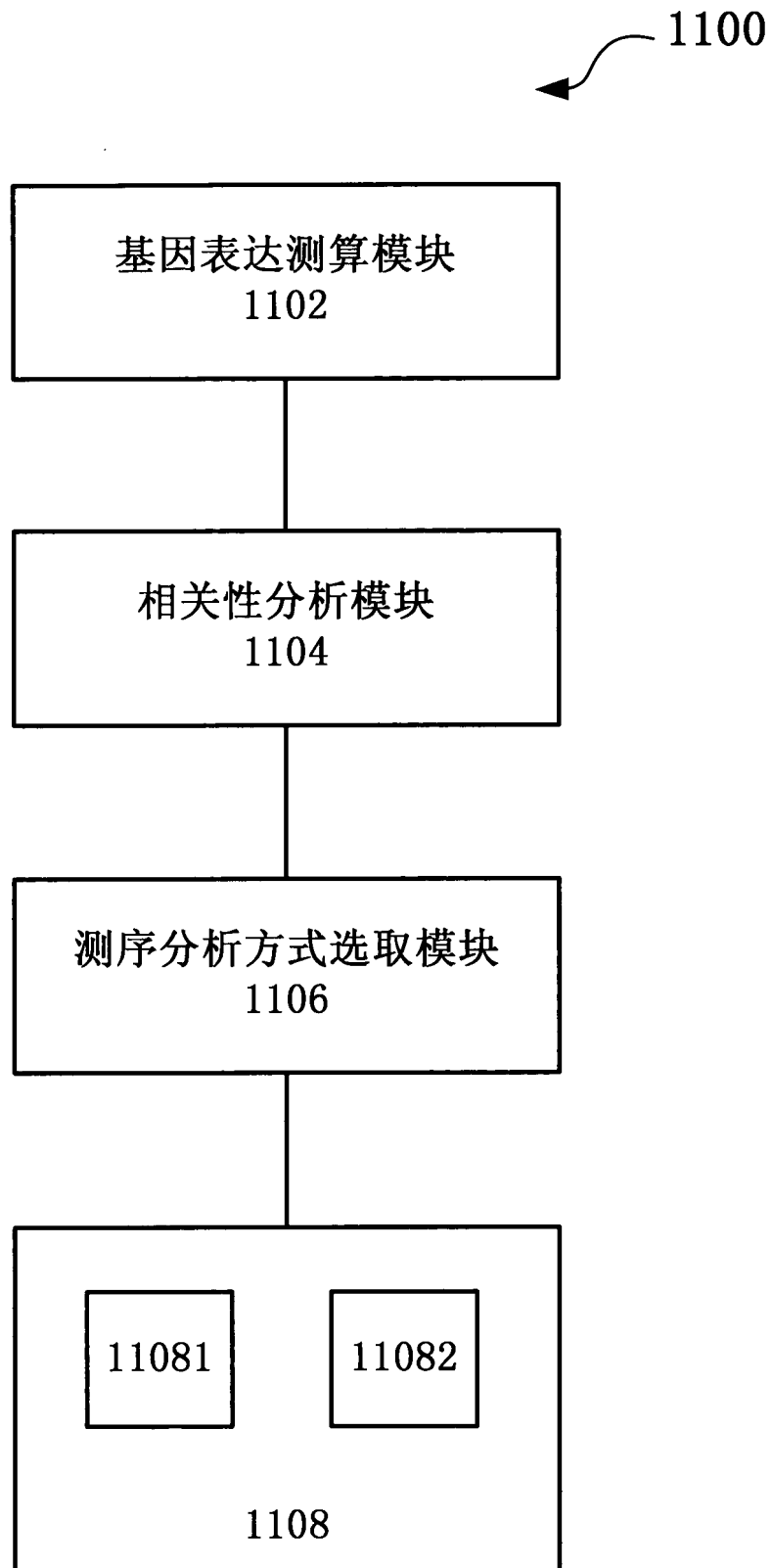


图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/001158**A.CLASSIFICATION OF SUBJECT MATTER**

C12Q 1/68 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B.FIELDS SEARCHEDMinimum documentation searched (classification system followed by classification symbols)
IPC: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
VEN, CNKI, CNABS, CPRSABS, CNTXT, EPTXT, USTXT, WOTXT, JPTXT, ISI web of Knowledge, EMBASE: sequencing, RNA-Seq, DGE, digital gene expression profile, transcription group, qPCR, real-time quantitative gene amplification fluoroscopic examination, stability**C.DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claims
PX	CN101914619 A (SHENZHEN HUADA GENOMIC TECHNOLOGY CO., LTD.), 15 December 2010 (15.12.2010), see claims 1-10.	1-10
A	Yan W Asmann, 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics, pp.531 to 541, vol.10, 16 November 2009.	1-10
A	Doron Lipson, Quantification of the yeast transcriptome by single-molecule sequencing. Nature Biotechnology, pp.652 to 659, vol.27, no.7, 5 July 2009.	1-10
A	WO2009124255 A2 (HELI-N) 8 October 2009 (08.10.2009), see the whole document.	1-10

☐ Further documents are listed in the continuation of Box C.☐, ☒ See patent family annex.

*Special categories of cited documents:

“A” document defining the general state of the prior art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search
8 October 2011 (08.10.2011)Date of mailing of the international search report
27 October 2011 (27.10.2011)Name and mailing address of the ISA/CN:
State Intellectual Property Office of the People's Republic of China
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Authorized officer

WU LiTelephone No.: (86-10) **62411046**

Facsimile No.: (86-10) 62019451

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/CN2011/001158

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN101914619 A	15.12.2010	None	
WO2009124255 A2	08.10.2009	WO2009124255 A8	19.11.2009
		WO2009124255 A3	14.01.2010
		US20110129827 A1	02.06.2011

国际检索报告

国际申请号
PCT/CN2011/001158

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, CNKI, CNABS, CPRSABS, CNTXT, EPTXT, USTXT, WOTXT, JPTXT, ISI web of Knowledge, EMBASE: 测序, RNA-Seq, DGE, 数字基因表达谱, 转录组, qPCR, 实时定量基因扩增荧光检测, 稳定性, stability

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN101914619 A(深圳华大基因科技有限公司), 15.12 月 2010(15.12.2010), 参见权利要求 1-10。	1-10
A	Yan W Asmann, 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics, 2009 年 11 月 16 日, 第 10 卷, 第 531-541 页。	1-10
A	Doron Lipson, Quantification of the yeast transcriptome by single-molecule sequencing. Nature Biotechnology, 2009 年 7 月 5 日, 第 27 卷第 7 期, 第 652-659 页。	1-10
A	WO2009124255 A2(HELI-N), 8.10 月 2009(08.10.2009), 参见全文。	1-10

☐ 其余文件在 C 栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期
8.10 月 2011(08.10.2011)

国际检索报告邮寄日期
27.10 月 2011 (27.10.2011)

ISA/CN 的名称和邮寄地址:
中华人民共和国国家知识产权局
中国北京市海淀区蓟门桥西土城路 6 号 100088
传真号: (86-10)62019451

授权官员

吴立
电话号码: (86-10) **62411046**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2011/001158

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
CN101914619 A	15.12.2010	无	
WO2009124255 A2	08.10.2009	WO2009124255 A8	19.11.2009
		WO2009124255 A3	14.01.2010
		US20110129827 A1	02.06.2011