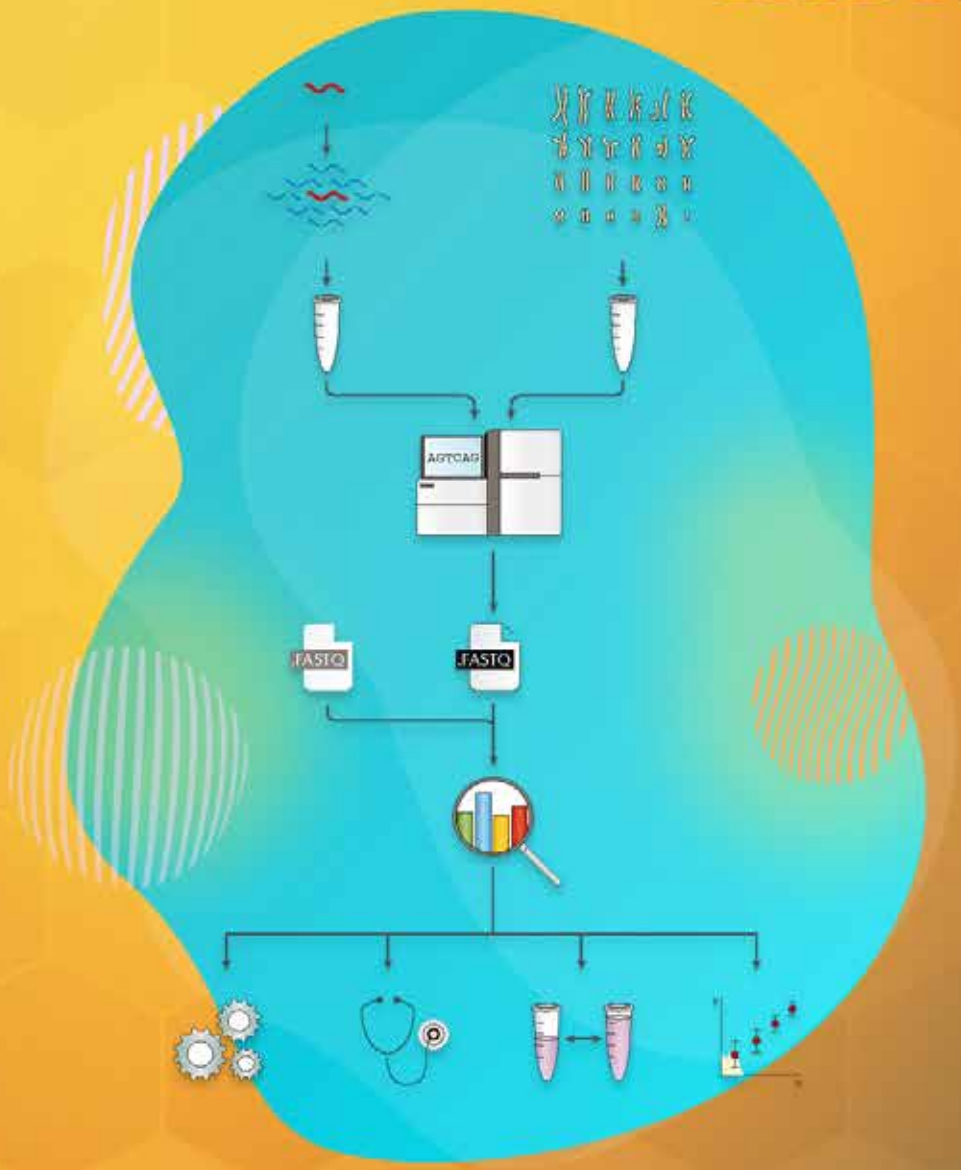
生命奥秘

总127期 2020/7
LIFEOMICS



新一代测序技术的“参照标准”（PART-II）

无奇不有
生命世界
你读生命
走进科学

目录 CONTENTS

专题:新一代测序技术的“参照标准”（PART-II）

前言 01

一、使用人造DNA片段检测多种基因组变异 02

1. 结果 04

1.1 设计及准备DNA sequin 04

1.2 检测遗传变异 06

1.3 检测体细胞突变 08

1.4 检测结构突变 09

1.5 检测基因拷贝数变异 12

2. 讨论 14

二、用于RNA测序试验的新型人工拼接基因内部参照物 16

1. 结果 17

1.1 设计人工染色体及RNA sequin 17

1.2 制备人工RNA sequin分子并对其进行验证 19

1.3 使用RNA sequin分子检测同型物组装 21

1.4 使用RNA sequin了解基因表达及可变剪接情况 21

1.5 使用RNA sequin评估不同样品之间的基因表达情况 24

1.6 使用sequin来分析融合基因 26

2. 讨论 28

本刊文章主要由国外网站文章编译而成，如有版权问题，请版权所有人与本刊联系。
凡本刊所载文章，版权归作者本人和本刊所有，如需转载，请注明作者及出处“生命奥秘”。
本刊提供的任何信息都不能作为医疗凭证和依据，仅供科研参考。

专题

新一代测序技术的“参照标准”（PART-II）

前言

使用NGS来发现遗传变异（genetic variation）是非常复杂的，这是因为人类基因组序列非常复杂，而我们在制备测序文库、进行基因组测序，以及后续的数据分析时，也存在着一定的偏倚。若能开发新型DNA标准品，将对这方面工作起到很大的促进作用。

一、使用人造DNA片段检测多种基因组变异

我们开发了一套人工合成DNA标准品（synthetic DNA standard），目前，称之为“sequins”，以此来模仿人类基因，同时还可以用作基因组测序工作中的定性和定量spike-in对照（spike-in control）。Sequin的测序结果（sequencing read）只能与人造的计算机参照染色体（in silico reference chromosome）匹配，而不能与人类参照基因组匹配，因此它们也可以用于不同样品间的平行分析（parallel analysis）。下文将介绍用这种方法来分析常见的、有临床意义的遗传变异，包括了单核苷酸变异（single nucleotide variants, SNV）、大规模结构重排（large structural rearrangement），以及基因拷贝数变异（copy-number variation）等。我们通过NA12878参照基因组序列，对sequin进行了验证，证实它具有发现遗传变异，并对其进行定量研究的能力。因此，sequin是一种标准化的、定量的方法，可用于检测人类遗传变异，并进行临床诊断。

因为NGS可通过进行个体化基因组测序来发现与疾病相关的突变，因此成为了生物医学研究和临床诊断工作中的重要工具。虽然，还存在大量的可变因素，会影响NGS的敏感性和准确性，比如在样品扩增和文库制备时存在的试验偏倚；测序覆盖不够，或者异质性覆盖问题；测序误差问题等。人类基因组序列的复杂性给测序工作中后续的比对和分析工作带来了极大的挑战，较短的测序读长只适合对重复序列，或者低复杂程度的序列进行测序。而且，可变等位基因频率（variable allele frequency）以及从单核苷酸替换至大规模基因组重排都存在的遗传变异的结构多样性，进一步增加了我们准确识别遗传变异的难度。

我们开发了一套人工合成的测序标准品sequin，来模拟人类基因组的一些特征，比如突变位点等。Sequin是一些长度可达10kb的人造DNA分子，可以模拟各种我们感兴趣的遗传特征，比如总体结构（gross architecture）、核苷酸构成（nucleotide composition），以及可比对性（alignability）都能被保留下来，但是它们的序列却和人类基因组完全不同。虽然，这些sequin在序列上与任何天然的基因组序列都不一样，但是完全符合人造的计算机参照基因组序列，而这个人造的计算机序列又包含了所有人类染色体的特征。

Sequin可以呈现人造参照染色体里最重要区域的特征。我们在构建DNA测序文库时也可以在其中加入sequin，并且在后续的测序和数据分析工作中同时对这些sequin分子进行测序和分析。由于这些sequin的序列只能与人造染色体进行比对，因此，它们可以当作对照，

对不同的样品进行平行比对分析（图1）。

Sequin还可以用作质量控制，进行定量分析，对新一代测序工作流程中的不同阶段进行评估，为生物信息学分析提供信息，构建内部参照标准，以用于基因组生物学（genome biology）的定量分析等（图1）。接下来，我们将介绍使用sequin来模拟各种常见的、与疾病相关的遗传变异，比如单核苷酸突变、小插入及缺失突变（small insertions and deletions, indel）、大规模结构变异和基因拷贝数变异等，同时我们也将介绍如何在各种基因组试验中使用sequin。

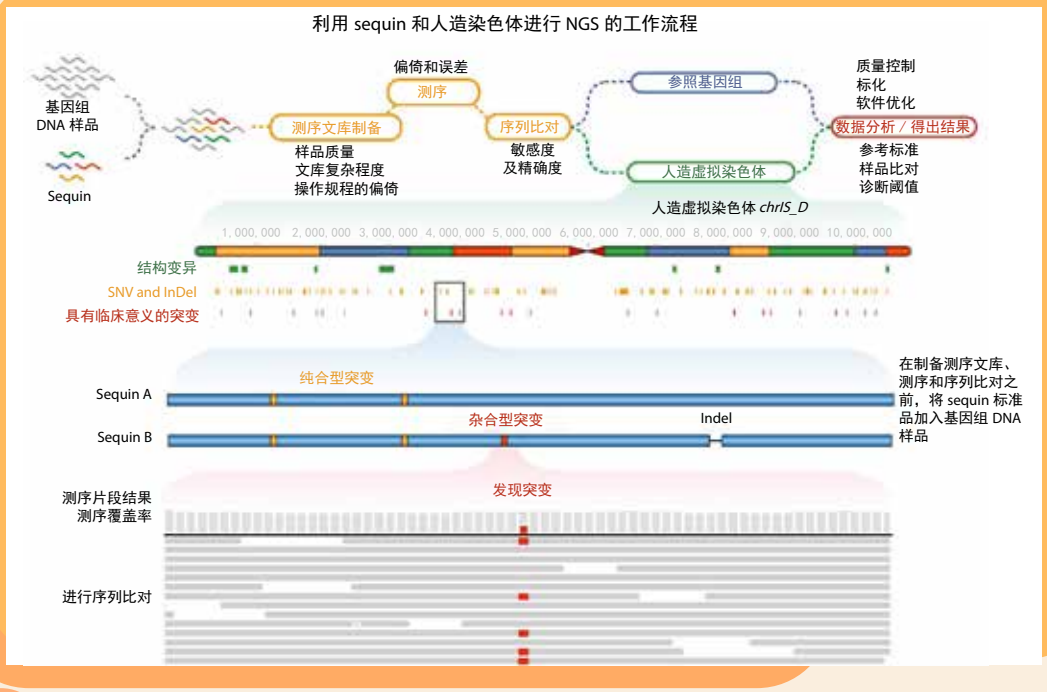


图1 使用sequin标准品代表遗传变异。在制备测序文库、测序和序列比对之前，将sequin标准品加入基因组DNA样品里。待测DNA样品的测序结果与参考基因组（例如图中蓝色表示的Hg38）序列进行比对，sequin的测序结果与人工参考染色体（图中绿色表示）序列进行比对。可以使用sequin对NGS的工作流程里的多个步骤进行评估，比如图中红色所示。人工染色体*chrIS*_D里含有多种遗传突变，比如黄色表示的常见突变、红色表示的与疾病相关的SNV和indel突变，以及绿色表示的大规模结构突变等。也可以使用一对sequin序列来表示遗传变异。其中每一条sequin代表二倍体基因型里的一个等位基因，因此，*chrIS*_D也可以像Hg38等参照基因组一样，构成一个人工参照基因组。通过调整sequin的相对丰度，就可以模拟出纯合型（黄色表示）和杂合型（红色表示）的突变。图中下方灰色表示的是通过和sequin的序列进行比对，在*chrIS*_D中发现了杂合型突变位点。

1. 结果

1.1 设计及准备DNA sequin

我们首先根据人基因组Hg38采样区域的大小、GC含量和重复度，按照比例在计算机上设计了一段长约11 Mb的人工染色体——*chrIS*_D（图1）。人造染色体的序列经过反转，在一些局部区域上如果还保留了与天然基因组序列的同源性，则还需要进行核苷酸替换，或者序列打乱等操作。

人群中常见的遗传变异都在*chrIS*_D中有体现（图1）。首先，我们根据突变的类型、大小、和核酸含量，对常见的人类遗传变异进行排序，然后纳入了一共223个SNV和176个indel突变。这些被选出的突变，以及它们周围的天然基因组序列被插入*chrIS*_D中。由于合成方面的难度，以及与人类基因组之间存在较高的交叉比对风险，所以没有将高度重复的序列，或者低复杂性的序列纳入人造参照序列里。

在电脑里设计人造染色体时，就根据参考序列，在sequin里设计那些有代表性的突变序列。一对sequin就可以代表二倍体基因型里的两个等位基因，而*chrIS*_D则构成了一个类似于Hg38的标准参考序列（图1）。如果要模拟杂合型突变（heterozygous variant），则可以在一条sequin上植入一种突变，在另一条sequin上植入另外一种突变。我们一共设计了36对标准sequin，共编码167种纯合突变和245种杂合突变。这些sequin的长度大约为1kb，其中包含150bp的两端区域，由于测序时存在边缘效应（edge effect），所以这些两端序列能够避免末端突变（terminal variant）的低覆盖问题（low coverage）。

为了验证这种设计是否有效，我们使用sequin和人类基因组模拟组建了末端配对测序文库（paired-end sequencing library）。然后利用BWA-mem，将sequin的测序结果与*chrIS*_D和Hg38的组合序列进行了比对。结果发现，sequin的测序结果只能与*chrIS*_D的序列匹配，而人类基因组的测序结果只能与Hg38的序列匹配。为了更进一步验证交叉比对的效果，还将sequin的测序结果与Hg38序列进行了比对，同时使用了更低的错配惩罚机制（mismatch penalty）——此时会出现更多的错配结果。

这些sequin在合成之后，会被装入模板载体里，可后续进行限制性酶切（restriction digestion）反应。我们将所有的sequin分子根据相同的摩尔浓度进行两两匹配混合，然后用标准的测序方法对其进行测序。和之前一样，这些sequin分子的测序结果和Hg38的序列完全不匹配，同样，NA12878个体的全基因组测序结果也完全不能与*chrIS*_D匹配。这更进一步说明，天然的序列和人造的序列可以根据它们各自的匹配结果，安全地进行区分。

接下来我们又在可测量的丰度范围内，检测了这些sequin标准品在强度（magnitude）和方向（direction）方面的误差情况。这些误差的产生可能是因为我们在制备测序文库、测序和序列比对时那些无法避免的、与序列本身有关的偏倚；也可能是因为在制备sequin混合样品时出现的均一变异性（aliquot

variability）。为了分辨带来误差的不同原因，我们选择了5种不同的sequin混合样品来进行实验。我们假设，序列本身带来的误差会存在于不同的sequin混合样品里，那么只要这些样品里都混入了同样序列的sequin标准品，均一变异性就与sequin标准品的序列无关，可能会随机出现在任何一个sequin混合样品里（图2a）。结果发现，在可检测丰度范围内（丰度的s.d.为0.284），sequin混合样品里每一个sequin标准品的最大误差波动都是可重复的，其它的误差（比如用吸液器吸液带来的变异）则不到10%（图2b），系统校正后的s.d.为0.023。对独立误差的检测可以为我们提供更多有价值的信息，制备出更合适的sequin混合样品。而对可重复误差的检测则可以帮助我们了解序列本身导致的偏倚情况，这对于NGS也非常重要。

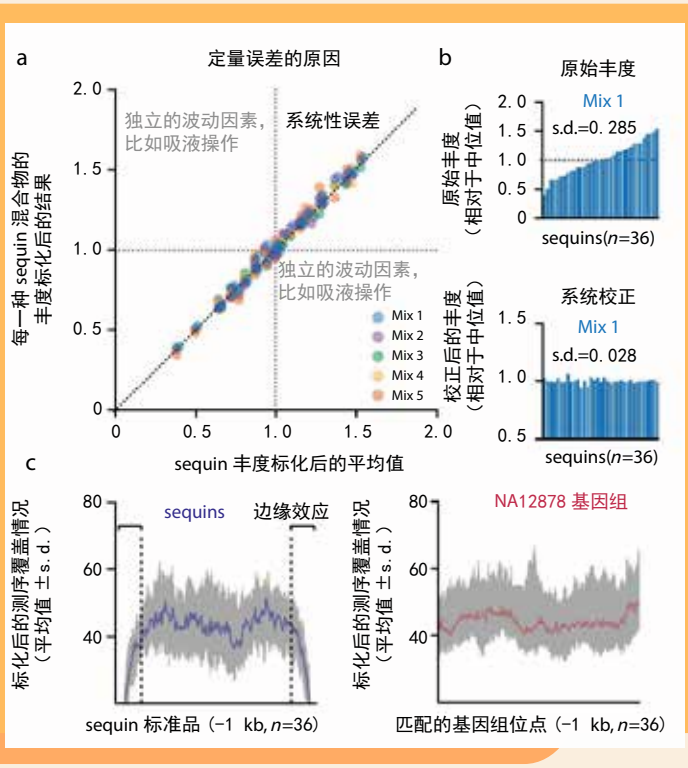


图2 利用同一sequin混合品和不同sequin混合品，了解其在新一代测序操作里的定量误差（quantitative variability）。（a）比较5种sequin混合品里同一种sequin标准品的检测丰度——每碱基中位覆盖值（median per-base coverage），这些sequin标准品都是按照相同的摩尔数加入sequin混合品的，

其丰度已经根据它们在混合品里的中位丰度进行了标化。与系统误差和独立误差有关的载体方向已经标明。（b）直方图表示在一种sequin混合品里的sequin标准品的检测丰度（根据中位数经过了标化），在进行系统校正前后的情况。（c）左图表示每kb sequin标准品的平均测序覆盖情况。右图表示NA12878基因组中同等（即sequin标准品根据相应的人类基因组校正后的数据）每kb序列的平均测序覆盖情况。在sequin两端的150bp范围内存在边缘效应，这会影响覆盖，如图中虚线所示。因此，这部分数据在校正和下游的分析工作中会被剔除。

1.2 检测遗传变异

为了验证sequin在基因组测序中的作用，我们又在NA12878基因组DNA样品里加入了sequin混合物，其终浓度约为1%。然后对这种混合DNA样品进行测序，测序结果与Hg38和*chrIS*_D参考序列进行比对。结果发现，sequin的测序结果在质量上与NA12878基因组DNA的测序结果完全不同，这些sequin在人工染色体和人类染色体之间表现出了非常好的映射置信度（mapping confidence）。

Sequin的序列比对结果还会被二次采样，保证与NA12878基因组达到同样的测序覆盖频次（图2c）。在这部分工作中，sequin在测序深度（平均每碱基43次）和异质性方面（四分位距为19倍）也都达到了足以匹配人类基因组位点的水平。该工作证明，不论其浓度有多低，sequin的测序覆盖度已经确保可以对基因组中的任何目标区域，进行精确的校准。

接下来我们又比较了其辨别sequin中突变位点和NA12878人类基因组中高信度确证突变的能力。我们使用基因组分析工具包（Genome Analysis Toolkit, GATK）和*chrIS*_D及Hg38参考序列索引来发现突变位点信息。最大匹配深度为每碱基平均43次，结果分别在sequin和NA12878人类基因组中发现了95%和99%的杂合型单碱基变异位点，在二者中都发现了99%的纯合型单碱基变异位点（图3a、c），分别发现了95%和93%的杂合型小插入缺失位点，分别发现了100%和98%的纯和型小插入缺失位点（图3b、d）。根据基因组分析工具包推荐的硬过滤阈值（hard-filtering threshold），我们在与*chrIS*_D比对时发现了3个假阳性小插入缺失突变，但是没有发现假阳性单碱基突变。进一步分析之后发现，减少测序深度也会对发现突变有影响（图3a-d）。这说明sequin代表的突变与NA12878基因组里的真实突变具有同等的效力，而且很适合用作基因组测序里的内参和阳性对照。

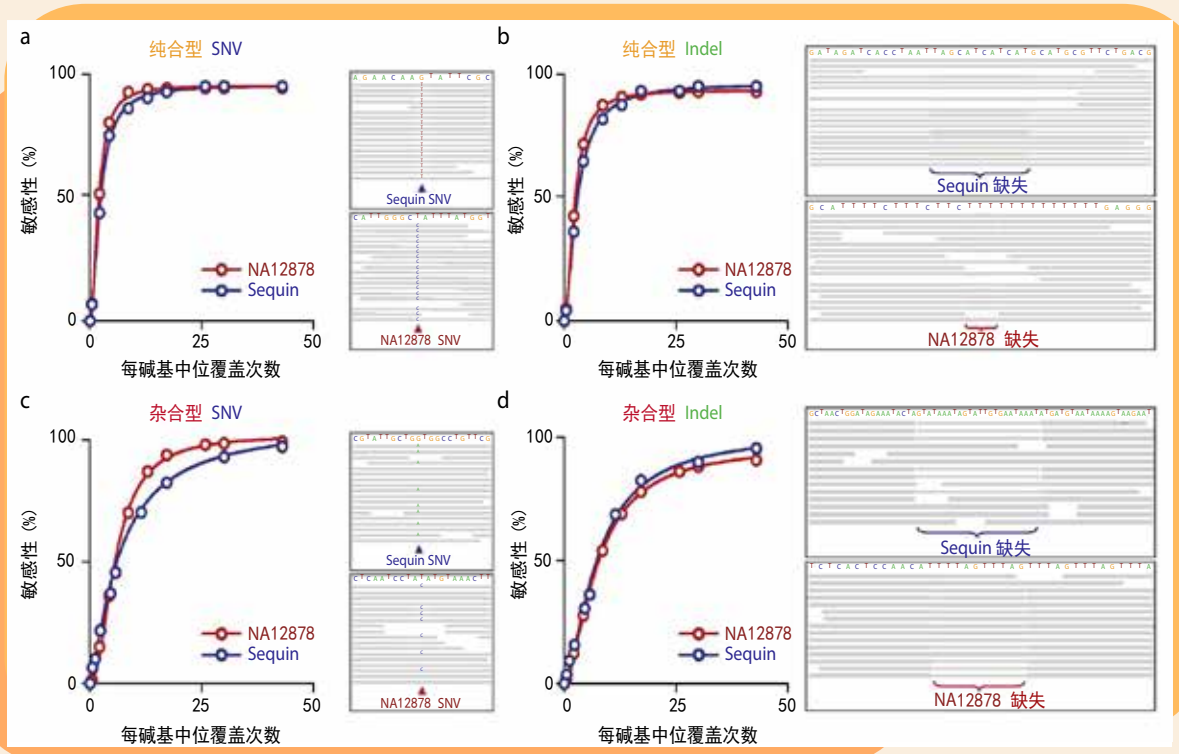


图3 利用NA12878参照基因组对sequin进行验证。（a-d）通过大量群体分布来描述突变检测技术的敏感性，图中蓝色所示的是sequin代表的突变，红色表示的是NA12878基因组代表的突变，这也可作为增加测序深度（每碱基中位覆盖数）的一种方法。（a）（b）分别表示纯合型单核苷酸突变和小插入缺失突变，（c）（d）分别表示杂合型单核苷酸突变和小插入缺失突变。右侧图显示的都是相应的基因组范例，表示序列比对之后发现的突变位点。

1.3检测体细胞突变

基因拷贝数变异、肿瘤异质性和样品不纯等原因导致的与肿瘤相关的体细胞突变（somatic mutation），涉及各种不同的等位基因频率（allele frequency）。将不同浓度的sequin对（sequin pair）进行组合，就可以构建出一个参照系，可用来检测不同的等位基因频率。我们用二倍稀释法制成了sequin对标准品系列，覆盖的等位基因频率从1:1（杂合型）至1:4096，每一个频率都由3个sequin对标准品组成。我们将这些标准品加入NA12878基因组DNA样品里，对这种混合样品进行测序，然后使用VarScan2软件寻找突变位点，并进行定量分析。我们使用这种sequin混合物作为内部参照，对测序技术寻找体细胞突变的敏感性和准确性进行了检测。

我们使用的sequin混合物浓度比较高，达到了10%，以保证在对sequin混合物进行定量分析时可以达到饱和覆盖。在>25,000倍覆盖的饱和覆盖下，我们至少能够得到一个支持性的测序结果（supporting read），但有两种合成的单核苷酸突变是例外，而且这两个例外都属于最罕见的等位基因频率（图4a）。我们还在检测出的突变数量和预计的突变（包括单核苷酸突变和小插入缺失突变）等位基因频率之间发现了非常密切的关系，单核苷酸突变和小插入缺失突变的R^2分别等于0.88和0.45。不过，在真实的试验过程中，测序深度是不可能达到这么多的，所以在发现突变的灵敏性方面也会有所牺牲。为了证实这一点，我们又对比对结果进行了二次分析，来模拟不同的测序深度，以及各自相应的敏感性。比如，测序深度在100倍时，就很难发现各种常见的，但是等位基因频率不超过12.5%的突变。

在用sequin比对*chrIS*_D参照序列寻找突变的工作中出现的假阳性结果，可以帮助我们确定检测的精确度。在充分覆盖的情况下，敏感性可以达到最大，但是在比对*chrIS*_D参照序列之后可以发现1254个潜在的错误单核苷酸突变和64个错误的小插入缺失突变（图4b）。这些错误的结果可能是因为测序本身的问题而导致的（不过这种几率比较低），但是都必须剔除掉，以确保结果的准确性。我们可以利用sequin来对不同的剔除策略进行评估，比如评价支持性测序结果的数量、等位基因频率、支持性测序结果与每条链的比值、突变支持核苷酸结果（variant-supporting nucleotide call）的质量和突变可信评分（variant confidence score）等，这些指标都是可以剔除假阳性突变结果的评价标准。

为了详细验证这种方法是否可行，我们选择突变可信评分进行了试验，根据这个评价标准对真阳性和假阳性结果进行了排序，对支持性测序结果的数量进行了费希尔精确检验分析，生成了受试者工作特征曲线，明确了在每一个预期的等位基因上检出突变的精确度（图4c、d）。比如，在等位基因频率小于1/128时（敏感性 = 75%，精确性 = 94%），就无法准确的区分真阳性单核苷酸突变和假阳性单核苷酸突变（图4c、d）。试验者可以凭借经验做出判断，在敏感性和准确性之间取得所需的平衡。

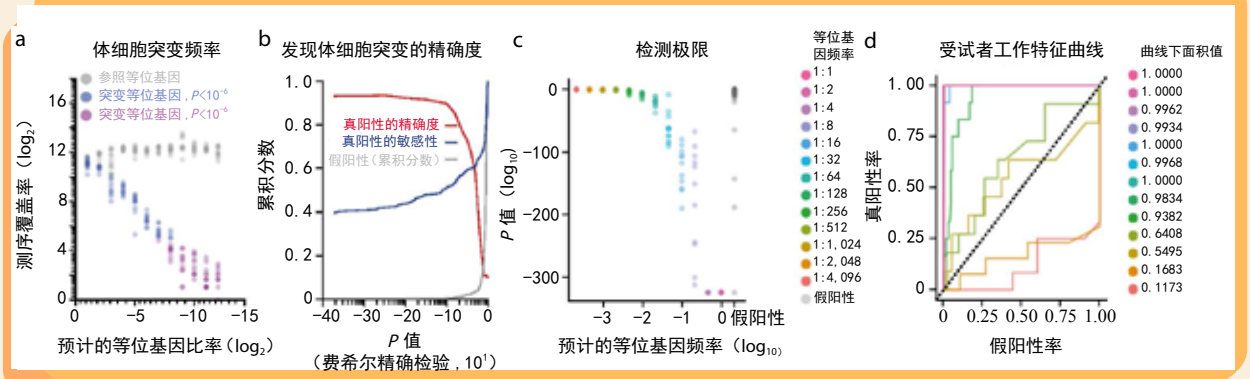


图4 使用sequin来检测体细胞突变的等位基因频率。（a）灰色表示参照序列的测序结果数量，蓝色和紫色则表示人工合成单核苷酸突变等位基因频率与预期突变等位基因的比例。费希尔精确检验结果显示，蓝色的最小可信度p值<10^{-6}、绿色突变未能超出阈值。（b）在检测人工合成单核苷酸突变的试验中，蓝色表示检测敏感度、红色表示检测精确度、灰色表示假阳性累积值（作为可信阈值）。（c）检测极限图，表示每一种单核苷酸突变相比各自预期等位基因频率的可信度。灰色表示的是假阳性结果，以供比较。（d）受试者工作特征曲线表示，根据变异信度（variant confidence）排列的真阳性与假阳性的比值。每一个等位比例都给出了相应的曲线下面积值，可用于比较诊断效力。

1.4 检测结构突变

结构突变（Structural variant, SV）包括了较大范围（>200 bp）的序列缺失、插入、转位、倒位和重排等情况，是一种非常常见的突变，也与多种疾病有关。不过，SV很难被发现，因为现有测序技术只能得到较短的测序结果，无法发现这些规模较大、结构复杂、同时又存在重复序列和低复杂度序列的突变。我们通常借助不对称测序结果配对比对（discordant read-pair alignment）、分隔测序结果比对（split-read alignment）和测序深度异常（read-depth anomalies）等信息来发现SV。

我们根据*chrIS*_D参考序列，设计了一系列sequin标准品来模拟缺失突变

（deletion）、倒位突变（inversion）、新片段插入突变（novel sequence insertion）和移动元件插入突变（mobile element insertion）（图5a）。这些代表性的SV序列都来自基因组变异数据库（Database of Genome Variants），但是我们对这些序列都做了改动，比如倒位或打乱，保证在维持原始结构、核酸组成和重复性的前提下，消除了这些序列的同源性。这些改造后的序列都被加入*chrIS*_D参考序列，同时也根据这些序列设计了相应的sequin标准品（图5a）。Sequin标准品制备好之后，相互混合而成为混合测序样品，然后进行测序，并与*chrIS*_D参考序列进行比对。比对之后，我们使用Lumpy软件，并结合测序深度、不一致配对比对结果和分隔比对结果，来识别SV（图5）。

在测序覆盖分别达到20和30倍时，可以准确地识别出缺失突变和倒位突变（图5b-d）。在倒位突变位点，分隔序列（slope = 0.31, R^2 = 0.75）和不一致配对序列（slope = 0.26, R^2 = 0.96）的累积均具有明显的线性关系，并且随着测序覆盖的加深而更加明显。不过在缺失突变位点，分隔序列（slope = 0.87, R^2 = 0.85）的累积速度要比不一致配对序列（slope = 0.32, R^2 = 0.79）快2倍。在使用NA12878人类基因组高信度缺失突变参照序列做类似分析时，也看到了同样的情况。使用分隔序列比对末端（split-read alignment termini），以及能够清晰区分缺失突变及倒位突变的相对比对极性（opposing alignment polarity），也都能清晰地区分出缺失突变及倒位突变左侧和右侧核苷酸突破点（图5c）。

大于600bp的新片段插入突变已经超过了现有测序技术的读长，因此，无法获得分隔比对结果。不过我们可以观察到不一致配对数量的增加，即在插入位点，一段测序结果无法与参考序列比对上（图5b）。这些悬挂起来，无处安放的序列都是有方向的，其3’末端都会向前或向后汇聚在插入突破位点（insertion breakpoint）处，这些序列的数量也与测序深度呈线性关系（slope = 0.19, 0.25, R^2 = 0.66, 0.25）（图5c、d）。这种序列特异性提供了一个非常好的信号，帮助我们发现新序列插入图表，这一点也在NA12878参照序列实验里得到了验证。

转移元件插入突变也是一种在人类基因组中非常常见的突变，活性转座子（active transposon）也是一种非常常见的突变元件。我们也在人工染色体里设计了人造转座子，同时也在非注释位点（nonannotated site）设计了相应的sequin标准品。测序时插入移动元件会产生特征性的不一致配对比对信号，将多次作图（作图质量mapQ = 0）测序结果与可信的比对序列（作图质量mapQ > 10）匹配，就能够锁定每一个插入位点（图5b-d）。

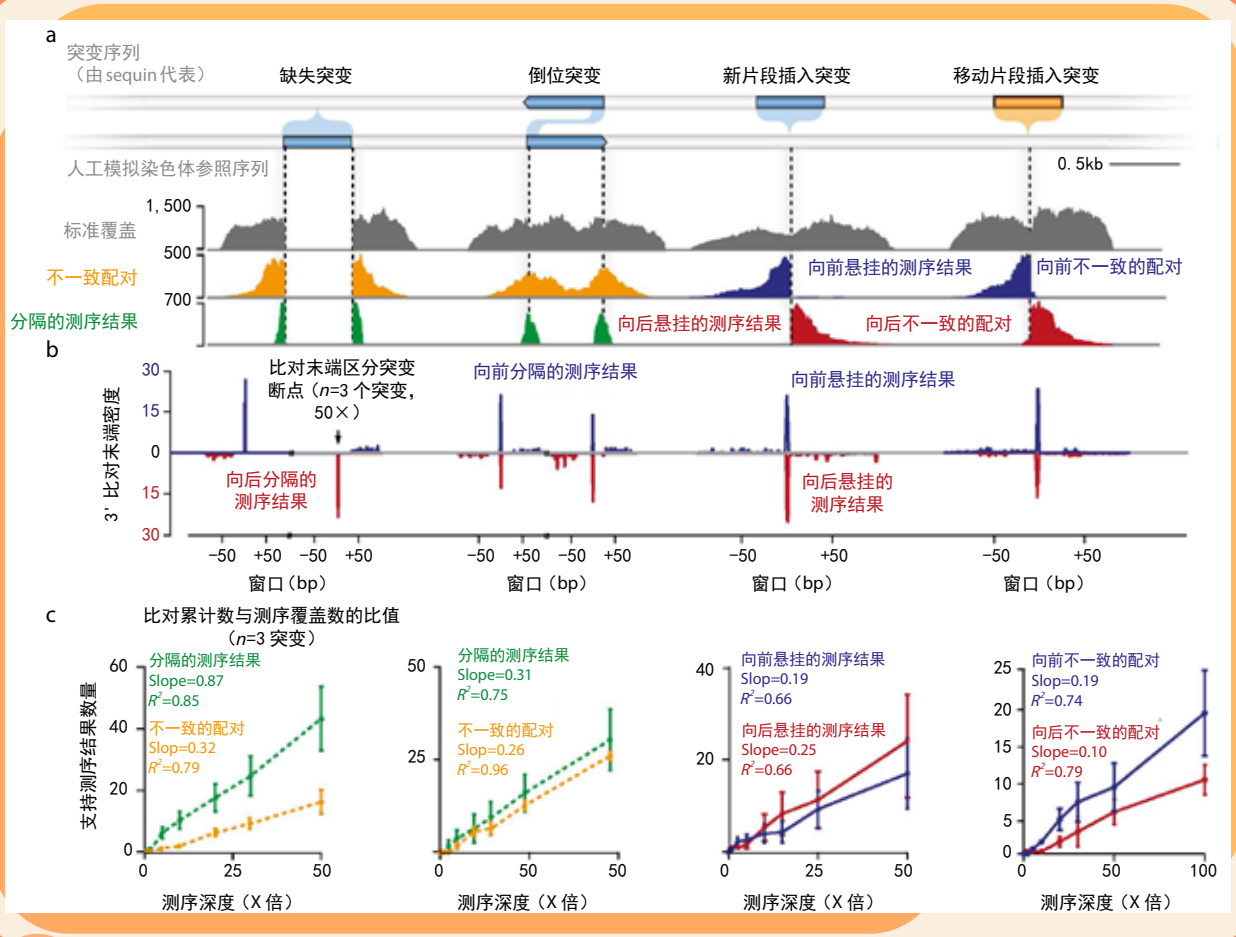


图5 使用人工sequin来发现结构变异。（a）使用人工sequin和*chrIS*_D参照序列发现缺失突变、倒位突变、新片段插入突变和移动片段插入突变。图中展示的都是每一种突变的特征性比对结果。（b）标化后的3’比对末端密度，蓝色表示向前，红色表示向后进行作图。（c）绿色表示分隔测序结果累计，黄色表示不一致配对测序结果累计，分别与各自的测序深度进行作图，图中表示的是平均值，误差条表示标准差，*n*=3。插入突变的不一致配对和悬挂测序结果都分别显示了向前和向后的方向。

1.5 检测基因拷贝数变异

大段基因的数量增加或者减少，就会导致基因拷贝数变异。我们也设计了相应的sequin标准品，来模拟不同大小、不同拷贝数的情况，比如2-copy × 621 bp, 4-copy × 202 bp和6-copy × 96 bp等（图6a）。使用Lumpy软件和分隔测序结果、不一致配对结果，可以成功发现两种较大规模的基因拷贝数变异（图6a），但是不能发现较小的（96bp）基因拷贝数复制情况。在每一个位点，我们都在扩增的和两侧的区域里发现了测序深度漂移（read-depth shift）与基因拷贝数成比例的情况，在拷贝数增加6倍、4倍和2倍时，斜率（slope ratio）分别为5.47、4.62和1.75。

测序深度漂移可以帮助我们发现人类基因组中的基因拷贝数变异，并且进行定量分析，这种方法的准确性主要依赖测序深度和序列覆盖（sequencing coverage）的同质性（homogeneity）。我们设计了一系列sequin标准品，构成了一个定量的梯度，来检测不同拷贝数的基因变异。每一个单独的sequin标准品都由四种不同的、600bp的DNA元件组成，这些DNA元件可以按照1-copy、2-copy、4-copy和8-copy来混合（图6b）。由于这个梯度系列是使用这些DNA元件按照不同的拷贝数混合而成，而不是按照不同的浓度混合而成，因此，就可以保证不受混合样品制备时试验操作的影响。

构建好基因拷贝数变异sequin标准品，对其完成测序、校准测序深度之后，这些标准品就能够以2-copy的方法与二倍体NA12878人类基因组序列匹配（图6b）。我们观察到，在定量范围内，预期丰度（expected abundance）与测序深度之间存在明显的线性关系（R^2 = 0.81, slope = 1.15）（图6b）。可以据此评价检测基因拷贝数变异技术的分辨力（图6c），因为拷贝数也通过平均漂移方法（mean-shift approach）得到了计算。由于在基因组内的任何拷贝数元件都可以进行校准，因此，整个比例可以放大至32倍（图6c）。

我们利用这种放大比例，对之前已经在NA12878基因组里得到过注释的基因拷贝数变异进行了检验，结果发现78.7%的缺失突变都与一个拷贝丢失相吻合，而背景对照的比例是2.7%（图6c）。检测复制的分辨力则稍差一些，只能发现52.8%的串联重复突变，而背景对照的比例是3.5%，这可能是因为平均漂移有所增加。

在人类基因组里，微卫星DNA、rDNA、病毒或线粒体DNA等重复的DNA元件可能有成千上万个拷贝。为了解决这个问题，我们对上面用到的检测基因拷贝数变异的sequin标准品进行了倍比稀释处理，制成了10^6倍的参考品梯度系列，其中每一种DNA元件的丰度都是由其在sequin标准品中的拷贝数，以及加入了多少sequin标准品决定的。

然后我们将这些标准品加入NA12878基因组DNA样品，使其终浓度为8%，然后对这些混合样品进行测序。结果发现，这些DNA元件的丰度（平均覆盖）与初始加入浓度（amol μL^{-1}）之间存在很好的线性关系，在初始浓度为4.88 amol μL^{-1}处才失去线性关系，呈现分段的线性关系。而这个转折点就是定量检测的极限，如果初始浓度比这个更低，则与DNA元件的丰度之间不再具有线性关系。利用在这个极限之上的线性曲线，可以进行定量分析，而且准确度也非常高，在这个浓度范围内的sequin

标准品就是一个非常好的对照，可以对各种DNA进行重复检测。为了证明我们的sequin标准品能够展示人类基因组里的DNA重复特点，我们选择了多个样品进行了检测，比如线粒体染色体、rRNA操纵子、EB病毒基因组和多种微卫星DNA等，所有这些样品的拷贝数都是之前已经确定的，而我们的检测结果也与之前的结果完全吻合。

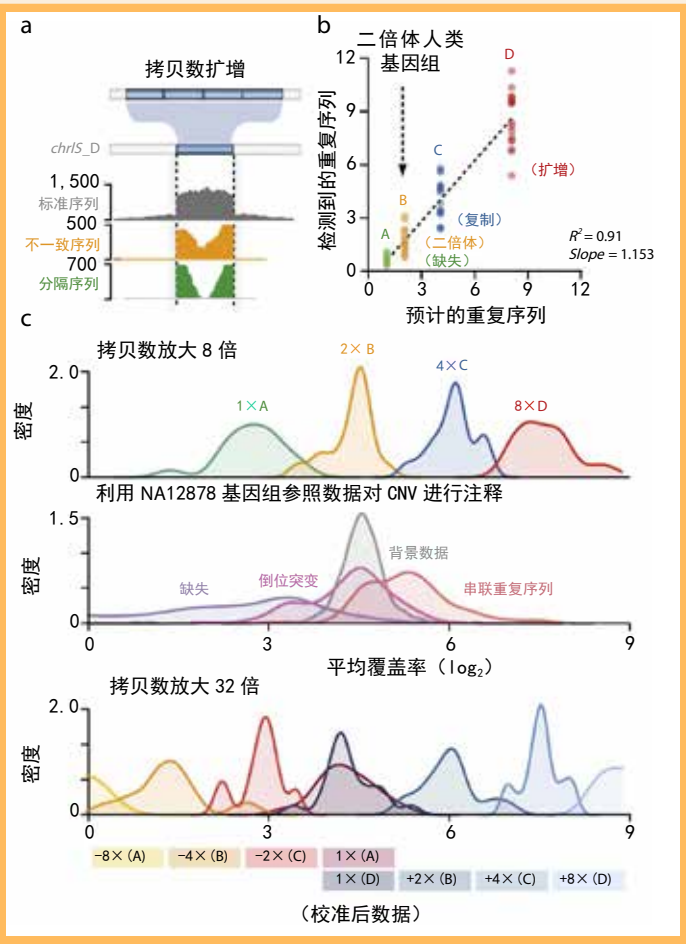


图6 利用sequin来检测基因拷贝数变异和重复序列。（a）利用sequin和*chrIS*_D参考序列来检测基因

拷贝数扩增。（b）散点图表示重复拷贝数变异。横轴是预计的重复数，纵轴是实际检测到的重复数。（c）频率分布图表示各种基因拷贝数变异的平均比对覆盖率。上图标明了参考标尺。中图表示所有已注释拷贝数变异的比对覆盖率与拷贝数sequin比例尺的情况。下图表示校准后的拷贝数可变元件可用于上调或下调基因组的平均比例尺，最多可达32倍。

2. 讨论

NGS已经成为了生物医学研究领域里的核心技术，而与之相配套的参照物开发，则成为影响测序结果准确性的重要因素，同时也是相对滞后的一个方面。包括人工改造细胞系和各种参照基因组在内的参照物都是非常重要的测序参照物。但是这些参照物虽然在大小和复杂程度方面与人类基因组比较相似，可是却不能直接加到测序样品里，也无法评价因样品本身特异性而带来的差异，更无法充当一个标尺，对不同的样品测序结果进行标化。随着NGS不断走进临床，我们也更加迫切地需要这样一个内部参照物（internal standard），帮助我们对不同的样品、不同实验室的结果进行比对，对测序结果有更好的解读和分析，制定出一个有临床价值的定量参照标准和诊断阈值。

Sequin标准品就能满足上述需求。首先，sequin标准品可以直接添加到测序样品里，充当内部参照物使用，既能够对sequin标准品进行定量检测，了解样品制备过程中的误差情况，也能够对不同样品进行标化处理。其次，sequin标准品能够以精确的比例相互混合，制成一系列不同浓度梯度的标准品，对等位基因频率等多种基因组特征进行定量检测。最后，sequin标准品还可以同时模拟各种遗传特征，并且能够很好地区分假阳性结果。综上所述，sequin标准品在测序试验中能够同时满足灵敏度和准确性两方面的需求，可以进一步提高NGS的检测能力。

我们在本文中介绍的各种例子只是为了向读者展示sequin标准品的作用，而不是全面介绍sequin标准品的各种应用，相信读者也可以体会到，其实在新一代测序操作的各个阶段都可以用到sequin标准品。在Hardwick等人的工作中，还介绍了一种RNA sequin标准品，用于RNA测序工作。这些“扩展的”sequin标准品进一步展现了sequin标准品的全方位功用，这也是sequin标准品在DNA或RNA测序中的最

大优势。从原理上说，sequin标准品可以模拟任何基因组或转录组序列。因此，sequin标准品足以成为一大类新兴的分析工具，帮助我们进一步拓展NGS的应用范围。

资讯 · 频道

www.LifeOmics.com

二、用于RNA测序试验的新型人工拼接基因内部参照物

RNA-seq可用于组装各种拼接异构体（spliced isoform），对基因表达情况进行定量研究，对整个转录组（transcriptome）进行宏观研究等。然而，转录组的大小和多样性、基因表达量的差异，以及RNA-seq技术本身的技术限制等因素，都会影响RNA测序结果的质量。为此，我们开发了一套人工RNA标准品，名为“测序spike-in（sequencing spike-ins, sequins）”。这些sequin能够代表各种全长的mRNA拼接异构体，以期能够对RNA测序工作有所帮助。Sequin是100%人造的核酸分子，其序列与天然的参考序列（natural reference genome）完全没有任何同源性，只能与人造染色体（*in silico* chromosome）序列里的基因匹配上。将多种不同浓度的sequin分子混合到一起，就可以模拟可变间接（alternative splicing）和多种不同的基因表达情况（gene expression），也可以为我们提供一把“标尺”，用于对不同的试验样品进行标化。我们已经将这些sequin分子应用于RNA测序试验，验证了功用并检测了样品特异性的偏倚（sample-specifc bias），也确定了对人RNA样品进行RNA测序时，转录子组装和定量分析的可靠极限范围。此外，我们还设计了一套补充sequin分子，来代表人工染色体基因重排时形成的融合基因，这套系统主要用于癌症的诊断。RNA sequin分子可以给我们提供一套定性及定量的参照标准，帮助我们更加深入地了解复杂的人类转录组。

人类基因组能够转录出各种编码及非编码RNA分子，这些RNA分子经过剪接之后，又能变成更加复杂的、相互重叠的同系物。所有这些RNA分子在一起，就被称作转录组（transcriptome）。它们向细胞传递各种遗传信息，决定细胞的发育和表型。

通过RNA测序，我们就可以了解整个转录组的情况，包括高通量决定转录体序列、新型同系物结构组装以及定量检测基因表达情况等。由于RNA测序具有上述优势，因此成为了各个研究领域和临床上的重要实验技术。不过，其准确性还是受到了转录组大小及复杂程度的影响，而文库构建、测序以及后续的生物信息学分析等各个实验步骤本身的偏倚，均使情况变得更加复杂。

Spike-in对照是外源性RNA分子，它们可以被直接添加到待研究的RNA样品中，对各种实验内在的偏倚进行客观评价，同时也可以对不同实验条件的结果进行精确比较。在此之前，RNA外部参照联盟组织（External RNA Controls Consortium, ERCC）曾经开发过一套spike-in标准品，当时主要是用于芯片分析，不过后来也被应用于RNA测序实验。ERCC spike-in可以对RNA测序实验的准确性

进行定量评估，但是作为单个外显子的转录产物，它们并不能反映真核生物基因表达和剪接的复杂情况。

为此，我们又开发了一套人工RNA分子，名为sequin。这些人工RNA分子都能够很好地与人工染色体里的人工基因位点相匹配（图1）。人工染色体的原始序列与相应的sequin分子，在序列上跟任何已知的天然基因组毫无同源性，因此可以共同用于测序结果的比对。这是因为绝对不会有序列交叉的情况出现，所以不用担心会影响测序结果。

这些sequin内部对照可应用于RNA测序工作的所有流程，包括文库构建、测序、测序结果比对、转录体组装、基因表达及可变剪接等。Sequin最适合对下游的生物信息学步骤进行评估，对各种实验参数进行优化，同时也可以对不同的实验样品进行标化处理。下文将介绍这些sequin分子的设计、验证、使用，以及在RNA测序实验中的各种优势。

1. 结果

1.1 设计人工染色体及RNA sequin

我们首先根据Hg38人类基因组的数据，按照大小、GC含量和重复程度等指标，设计出了一个大约有11 Mb的人工染色体*chrIS_R*。为了去除同源性片段，我们将序列倒转（inversion），同时将每一段序列打乱（shuffling），这样既保留了原始序列的核酸组成和重复性，同时又避免了同源性。

我们使用了GENCODE注释数据库的信息，根据可变转录子的大小、外显子的数量，以及GC含量等指标，在人工染色体上同时设计了人工基因位点，来模拟人体内源性的基因。我们一共设计了78个人工基因位点，这些基因共编码164个可变RNA同系物，包含了869个不同的外显子和754个内含子。这些基因有些只含有一个外显子，有些则包括好几个外显子，每一个同系物的大小则从280 bp到7 kb不等，其中最大的同系物里包含36个外显子之多。

人体基因可变剪接机制涉及如此复杂的外显子和内含子结构，同时又有重叠（overlapping）、反义（antisense）和双向（bidirectional）等不同的位点。内含子在可变剪接位点供体和受体里也是非常常见的，由于每一个RNA sequin分子都代表一个成熟的mRNA分子，所以将同一基因位点的不同同系物混合在一起，便可以模拟可变剪接过程（图1）。

为了评价RNA sequin分子在RNA测序过程里序列比对及组装方面的能力（排除了在文库构建和测序过程中可能存在的偏倚），我们使用RNA sequin分子以及相应的人类蛋白质编码基因，构建了测序片段模拟文库（read library）。结果发现，RNA sequin分子的总体匹配率达到了99.5%，人类蛋白质编码基因的

总体匹配率达到了99.4%。同时还发现，没有RNA sequin分子能够与Hg38或者其他人体基因组参照序列匹配，也没有人体基因序列能够与人工染色体匹配。由此可见，RNA sequin分子完全可应用于多种模式生物的测序。在多种不同测序深度的试验中也发现，人工染色体及人类基因组的同系物也有类似的表现。

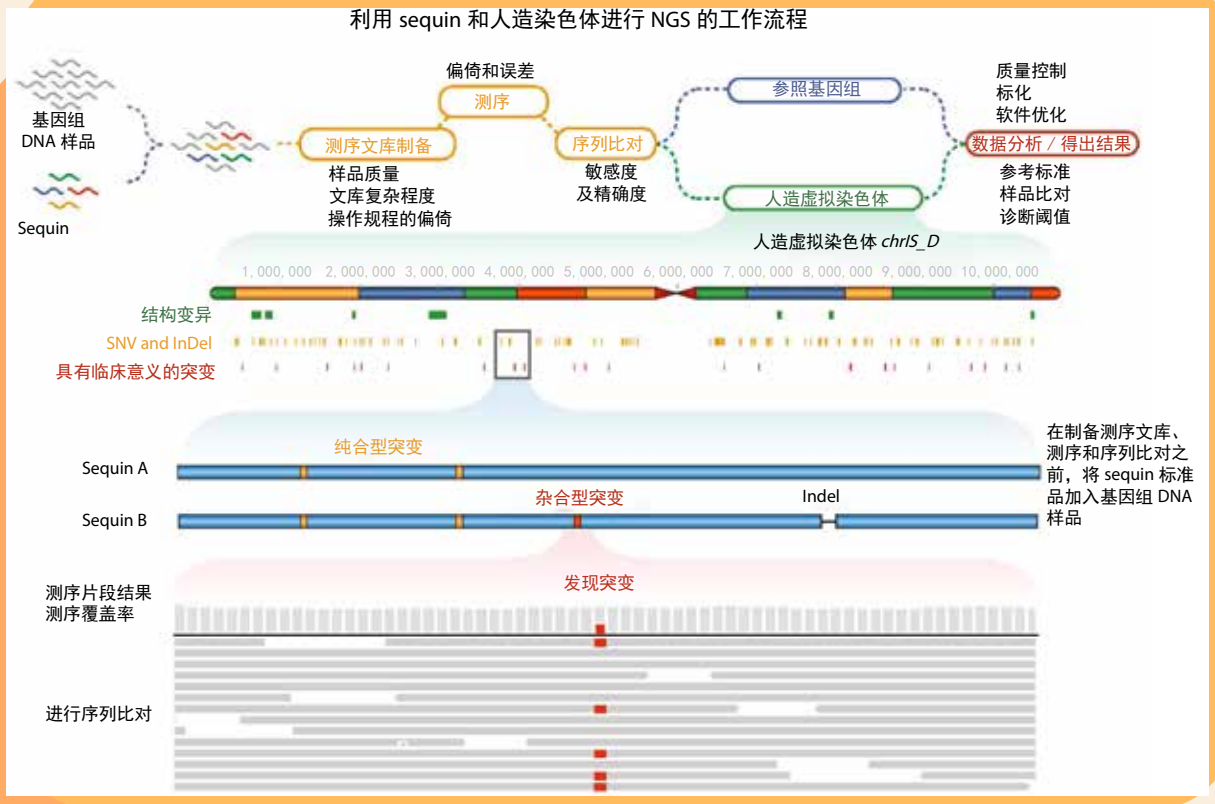


图1 上图整体展现了RNA sequin分子的设计及用途。（a） RNA sequin分子及人工染色体应用于RNA测序示意图。（b）展示了（a）图中人工染色体里的78个人工基因位点。这其中大多数基因都能编码多个不同的mRNA同系物分子，其中每一个同系物RNA都有一个相应的RNA sequin分子与之匹配。（c）具体展示了*R1_82*人工基因，该基因通过可变剪接机制，编码两个可变的同系物RNA。通过将不同浓度的两种RNA sequin分子进行混合，获得A、B两种混合物，就可以模拟3个内部外显子（internal exon）的可变剪接过程。图中灰色的直方图表示RNA测序的覆盖率。

1.2 制备人工RNA sequin分子并对其进行验证

人工合成出包含末端的多聚腺苷酸尾部结构的RNA sequin分子之后，会将其构建到表达载体里，进行体外转录。得到全长RNA分子之后，进行纯化，然后稀释成储存液保存，用于后续操作。

我们起初会对不含天然RNA成分的纯人工RNA sequin分子等摩尔浓度混合物进行测序，来了解多大的测序深度才比较合适。我们也考虑了这些测序结果与人类基因组参照序列存在交叉比对的可能性。结果发现，只有不到0.0075%的测序结果能够同时与人工染色体及Hg38基因组序列匹配，即在4300万个测序结果中，只有3223个能够同时与人工染色体及Hg38基因组序列匹配，这些序列主要来自软件对多聚腺苷酸尾部的剪辑。与其它模式生物的参照序列进行比对，得到的结果也差不多。同样，大型的K562 RNA-seq文库里的序列与*chrIS*_R人工染色体序列的匹配度也只有大约0.03%。

我们也选择了一些人类蛋白质编码基因，以及这些基因在K562文库中表达的外显子数量等指标，对相应长度的RNA sequin分子测序结果的比对和组装进行了评估。结果发现，RNA sequin分子和人体内源基因在多个不同的测序深度情况下，只要两者的测序深度相同，序列组装方面的表现就都是一样的（图2）。值得注意的是，即使是在充分覆盖时，我们使用TopHat2和StringTie软件，按照默认的参数设定条件，都没有组装出完整的人工转录组。随着测序深度的加大，真阳性和假阳性组装都会增多，组装的敏感性和特异性则反向相关。

将不同浓度的RNA sequin分子混合在一起，可以制成一个梯度参照系（reference ladder），对基因表达情况进行检测。我们按照倍比稀释的方法（two-fold serial dilution），制备了一个梯度参照系，最终的浓度梯度范围达到了10^6——足以反映人体转录组里基因表达的动态范围，每份稀释品里最少只有3个基因。接下来，我们又将不同的可变RNA sequin同型分子（alternative RNA sequin isoform）按照不同的浓度进行了混合，用来对基因可变剪接进行分析，最终的浓度梯度也达到了32倍。我们使用这套系统成功模拟出各种基因可变剪接事件。

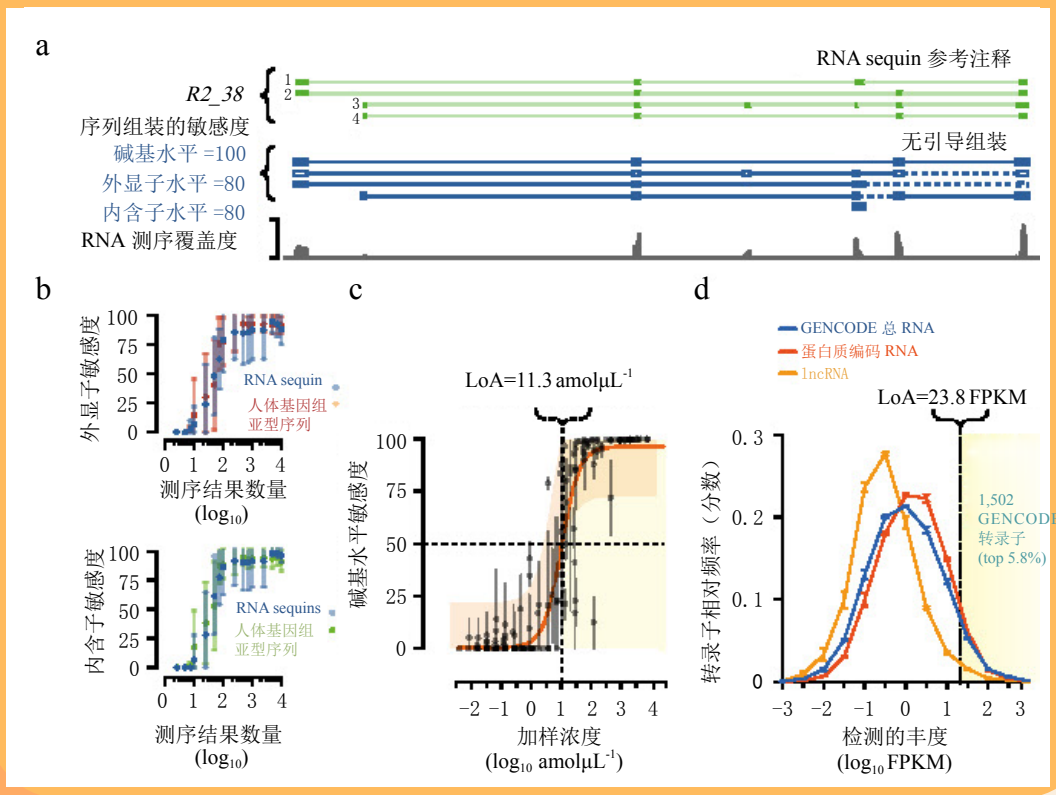


图2 使用RNA sequin来评估转录子组装工作。a，没有借助参照注释，使用Tophat2和StringTie软件进行转录子组装。虚线表示缺失的特征。b，用散点图表示在各种测序覆盖水平下，用试验的测序结果组装RNA sequin及相匹配的人类同系物情况。根据正确组装的外显子（顶部）或内含子（底部）的百分比来评价组装情况。c，用散点图表示碱基水平组装的敏感度和RNA sequin加样浓度之间的关系。试验结果为S状曲线（R^2 = 0.858），根据该曲线可以估算出某一RNA分子完成50%可靠组装时所需要的加样浓度。橙色的阴影部分表示90%预测区间，黄色阴影部分表示表达量能够超过组装极限（limit of assembly, LoA）的RNA的比例。d，该直方图表示不同表达水平的内源性人体转录子与相应K562样品之间的相对比例，黄色阴影部分表示表达水平超过了LoA阈值以上的部分。红色表示外显子同系物的频率，橙色表示lncRNA的频率，蓝色表示总GENCODE的频率。试验重复三次，误差条表示标准差。

1.3 使用RNA sequin分子检测同型物组装

为了更进一步展现RNA sequin分子的作用，我们将RNA sequin分子混合物与复制了3次的K562细胞的RNA总提取物进行混合，RNA sequin分子的终浓度为10%。不过我们建议选择1~3%的终浓度，因为这样可以减少测序深度。我们对这种混合样品进行了RNA测序，然后将测序结果与Hg38及人工染色体进行比对。能够与人工染色体匹配的测序结果被二次采样作为spike-in，用于对不同的样品或不同复制次数的样品进行校正。

对RNA sequin分子的测序结果与K562细胞RNA样品的测序结果进行比较，结果发现，不论是每个核酸（per-nucleotide），还是每条测序结果（per-read），我们都无法分辨两者之间的误差，而且两者在测序结果长度方面的Phred质量评分（Phred quality score）的分布也是一样的。同时，这两者在注释转录子长度覆盖的标准化方面也都是相似的。

足够的序列覆盖是保证转录子顺利组装的前提，因此，lncRNA这类表达量很低的RNA，往往就很难组装成功。RNA sequin可以用来帮助判断在组装可变剪接同型RNA时，足够进行转录子组装所需要的最低表达量。用同型RNA分子组装的比例与各自的加样浓度进行作图，就可以得到一个S状的剂量-反应曲线（sigmoidal dose–response curve），该曲线的R^2 = 0.893（图2C）。这条曲线就可以帮助我们确定，如果要进行成功组装，所需要的最低加样浓度。比如，最低浓度为11.3 amol μL^{-1}，或23.8个FPKM（fragments per kilobase of exon model per million mapped reads，表示每百万测序碱基中每千个转录子测序碱基中所包含的测序片断数）时，大约可以组装50%的参考注释基因转录子，即碱基水平的灵敏度为50。相比之下，在K562转录组中，仅仅发现有1502个转录子的表达水平超过了这个阈值，这在整个GENCODE数据库里已经被注释过的25,975个转录子中，只占到了5.8%，而lncRNA由于其表达量太低，组装成功率就更低了（图2d）。这说明，如果使用标准的RNA测序文库，那么在K562转录组里，大部分新发现的转录子其实都是难以组装成功的。

1.4 使用RNA sequin了解基因表达及可变剪接情况

RNA测序已经成为了研究基因表达情况的第一选择。为了评价RNA测序技术的定量准确度，我们采用StringTie软件对sequin及人体内源基因及其同系物的表达情况进行了检测和分析。对sequin的加样浓度和检测到的丰度进行作图，得到了一个强线性曲线，基因样品的R^2 = 0.922，同系物样品的R^2 = 0.899（图3a、b）。使用分段线性回归分析发现，基因表达和同系物表达的低位拐点（lower inflection point）分别为0.472 amol μL^{-1}和2.52 amol μL^{-1}。这个低位拐点就是定量检测的极限，如果低于这个值，sequin的表达量与加样浓度

之间就不再具有线性关系，此时基因样品的R^2 = 0.247，同系物样品的R^2 = 0.0323（图3a、b）。

这条线性曲线的斜率分别为0.955（基因）和1.05（同系物），据此也可以在加样浓度与检测丰度之间进行换算，也可以根据sequin来判断在K562转录组中相应基因的表达情况（图3）。基因和同系物的LoQ分别为3.13 FPKM和5.17 FPKM。由于在K562转录组中，有9,175个基因和5,457个同系物超过了这个阈值，因此，足以进行准确的定量分析。比例更小的lncRNA同样再次没有达到这个LoQ阈值（在K562样品中，所有lncRNA和基因转录子中只有6.5%和4.73%被检出），这就是对K562样品进行基因表达分析的极限范围。

我们也对可变剪接同系物进行了半定量分析，此时检测的指标是每一个sequin基因的最低同系物丰度与最高同系物丰度之间的比值（图3g）。不过总体准确度不够高，R^2 = 0.509。这种方法在检测高表达基因时的准确度有明显提升，当丰度超过30.2 amol μL^{-1}时R^2 = 0.887，曲线斜率为0.807；而当丰度不高时，R^2 = 0.256，斜率为0.434。在丰度较低时，同系物的检测分辨率也达不到，当基因的丰度低于0.472amol μL^{-1}时，就很难对其进行准确的检测和分析了。其它的剪接复杂情况则更进一步影响了定量检测的准确度，因此，当sequin基因的转录同系物达到4个时（n = 3），就不能被正确检测，而当只有一个sequin基因的转录同系物为3个（n = 12）时可以被正确检测。为了更进一步研究可变剪接问题，我们将sequin的注释打散成883个组份外显子计数单元（constituent exon counting bin），然后对预计的和实测的每一个外显子计数单元拼接之后的百分比进行作图分析。结果发现，总体预计值和实测值之间的关联并不明显，R^2 = 0.603，但是在高表达基因中，这种关系就变得非常明显了，R^2 = 0.909，只不过在低表达基因中关联性会减弱，R^2 = 0.403。

最后，我们又观察了外显子组装和定量分析中的系统性偏倚和内在偏倚。我们用每一个外显子的丰度对应其加样浓度进行了作图分析，结果发现，当LoQ低于1.89 amol μL^{-1}（或每个外显子的测序结果少于26.3个）时，我们就很难对该外显子的表达量进行准确的检测（图3c）。当内部外显子（internal exon）的浓度为6.72 amol μL^{-1}（碱基水平敏感度为50）时，R^2 = 0.905，而两端外显子（terminal exon）的浓度为15.1 amol μL^{-1}（碱基水平敏感度为50）时，R^2 = 0.825，可见内部外显子的RNA测序不论在组装，还是在准确性方面都要优于两端外显子。这可能是因为在测序时存在边缘效应（edge effect），所以转录子末端的测序覆盖率与其内部不太一致。这也提示我们，在使用RNA测序技术分析完整的RNA亚型结构，了解可变转录起始及终止位点时，还存在一定的困难。

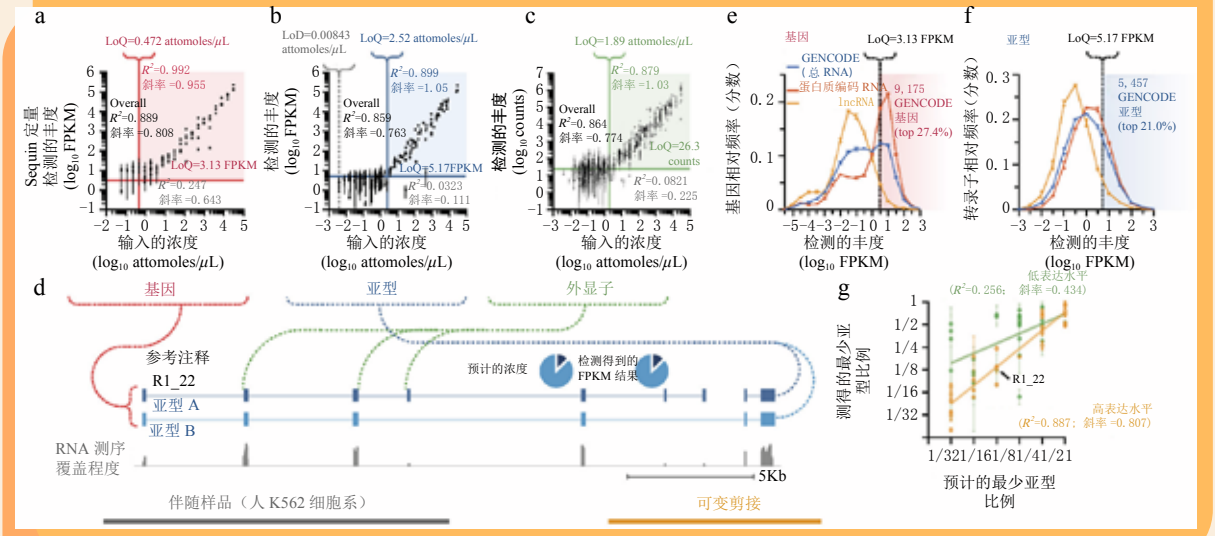


图3 用RNA sequin检测基因表达及可变剪接情况。（a-c），散点图表示RNA sequin输入浓度与基因表达水平之间的相关性。（a-c），分别表示基因、亚型和外显子的情况。灰色虚线表示检测极限（limit of detection, LoD），彩色实线表示定量极限（limit of quantification, LoQ,）。（d），用基因组浏览器显示了*R1_22* sequin参考注释，图下部分的灰色直方图表示RNA试验测序操作的覆盖情况，浅蓝色和深蓝色分别表示检测到的和预估的亚型表达量。（e）和（f）分别用直方图表示人类基因和亚型的频率与K562细胞转录组的表达水平。LoQ阈值表示达到了足以被准确检测表达丰度的基因及亚型的比例。红色表示外显子，橙色表示lncRNA。（g），散点图表示每一个sequin基因的主要亚型里，最少亚型占比的实测结果与预估结果之间的相关性。黄色标识高表达基因（R^2 = 0.887），绿色标识低表达基因（R^2 = 0.256），可见不同表达水平下的检测分辨率差异较大。试验重复三次，误差条表示标准差。

1.5 使用RNA sequin评估不同样品之间的基因表达情况

不同成份配比的RNA sequin混合物为我们提供了一套“标尺”，可以对不同的样品进行标化处理，可检测不同样品的基因表达差异，以及可变剪接等情况（图4a）。我们制备了一种混合物B，其中56种基因和126种亚型的浓度都与混合物A有所不同，以此来模拟两种不同的基因表达及可变剪接情况。这两种混合物里RNA sequin的差异倍数在基因水平可以分为9个亚群，在亚型水平则可以分为19个亚群。

为了展示RNA sequin在基因谱分析方面的作用，我们在1 μg GM12878细胞总RNA提取物中加入了混合物B。为了与qRT-PCR进行定量比较，我们选择了一些能够代表所有混合物差异倍数预测值的RNA sequin，结果发现实测差异倍数值与预测差异倍数值在RNA sequin与qRT-PCR之间存在密切的相关性，R^2分别为0.988和0.929。

我们使用DESeq2方法对不同混合物里RNA sequin的差异倍数情况也进行了分析，结果发现实测差异倍数值与预测差异倍数之间的线性关系也很紧密（图4b），R^2 = 0.830，斜率为0.756。DESeq2方法也正确预测出所有的阴性对照基因（n = 22，校正后的p值 > 0.1），只有两个假阳性结果例外。这两个假阳性结果都属于极端的亚型转换，可变亚型之间在长度上差异极大，但其总体基因表达水平却一样。我们用模拟的数据库也得到了类似的结果。在A、B两种混合物中总共有56个基因的表达水平有差异，DESeq2分析漏掉了其中的20个基因。这20个基因的加样浓度都非常低，说明在检测低丰度转录子时还存在一定的困难。

我们还对5个亚群的sequin基因绘制了比例检测极限曲线（limit of detection of ratio curve），了解其差异倍数方面的情况，以及每一种基因能够准确检测其表达水平的平均表达极限值（图4c、d）。在用该曲线检测K562细胞及GM12878细胞的基因表达水平差异时，则表示超出了这个阈值的所有内源基因的比例（图4e）。与预计的一样，随着差异倍数的增加，诊断能力也相应的提高了（图4f）。

为了评价两种sequin混合物检测可变剪接事件的能力，我们也为这两种sequin混合物在亚型水平和外显子水平绘制了差异基因曲线。同时，使用DEXSeq来评价不同外显子的用途。据估计，在所有883个组份外显子计数单元中，212个的表达量可能是恒定的（即真阴性），其它671个的表达量可能是变化的（即真阳性），变化范围可达512倍之多。DEXSeq正确的发现了164个真阴性结果，正确率达到了77.4%，但是只发现了305个真阳性结果，正确率仅为45.5%。这可能是因为低丰度的亚型RNA分子的测序结果数量不够（这种情况对于末端外显子尤为突出），这也揭示了在分析不同样品间可变剪接事件时还存在一定的困难。

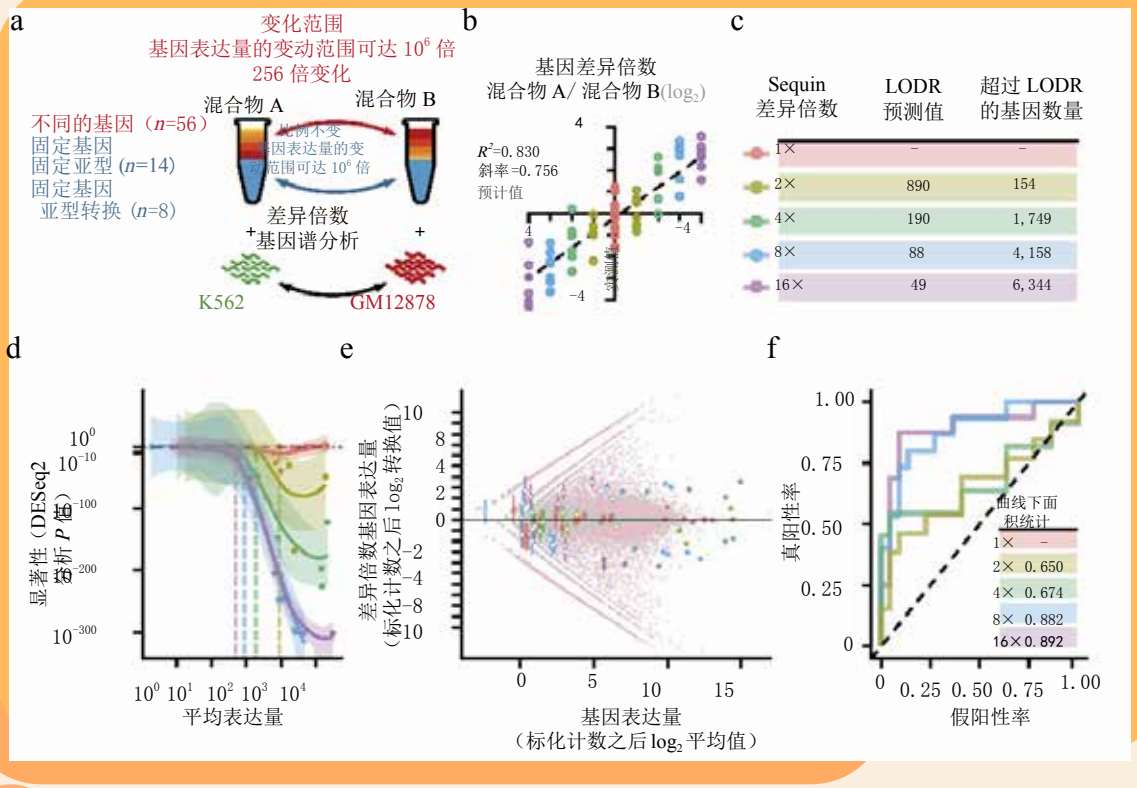


图4 使用RNA sequin混合物来比较不同检测样品之间的基因表达情况。（a） 每一种RNA sequin混合物里都含有同样的成份（蓝色）和不同的成份（黄色、橙色和红色）。（b） 散点图表示A、B两种RNA sequin混合物里RNA sequin位点log2差异倍数（log2-fold change, LFC）预期值与实测值之间的关系。（c） 表中不同的颜色表示每一种sequin基因的差异倍数情况、相应的比例检测极限（limit of detection of ratio, LODR）预测值，以及在K562和GM12878转录组中超过了该极限的，在GENCODE数据库中有注释的基因的数量。d，LODR曲线表示平均表达量、差异倍数，以及检测不同基因表达水平的能力，还给出了每一种sequin基因平均表达量的DESeq2分析P值。伪发现率阈值（false-discovery threshold）用黑色虚线表示，每一种差异倍数的LODR预测值用垂直的彩色虚线表示。e，表示了实测LFC与平均表达量之间的关系。彩色点表示的sequin与在K562及GM12878细胞中有差异表达的人体内源基因有很好的重叠，红色点表示有明显的意义。空心及实心圆圈分别表示在相应的LODR预测值之下和之上的sequin。误差条表示标准差。f，sequin基因受试者工作特征曲线 (receiver operating characteristic, ROC)表示在检测差异倍数时的诊断能力。

1.6 使用sequin来分析融合基因

染色体重排（Chromosomal rearrangement）可以将染色体上之前相互独立的两个位点融合到一起，这是导致肿瘤发生的主要原因。RNA测序技术就可以发现融合基因，也可以发现新型染色体重排，或复杂的染色体重排。可是，准确的诊断染色体重排需要在染色体断裂位点处，对于跨越该位点的内含子区域进行足够的序列比对。如果存在同源的，或者重复的DNA序列，就会进一步增加分析的难度。

我们重新设计了RNA sequin，使其能够代表人工染色体*chrIS*_R 中12个平衡倒位（balanced inversion）形成的24种融合基因（图5）。这些融合的RNA sequin能够代表各种不同的基因，涉及大小、外显子数目和剪接复杂程度等指标。同时还有一套相对应的、未修改过的原始基因数据。为了验证这套融合基因RNA sequin的功效，我们进行了试验，结果发现不论在何种测序深度上，这些RNA sequin在检测的敏感度方面与人体内源性的致癌融合基因不相上下。

体外转录出融合RNA sequin之后，使用上述方法对等摩尔比例的混合物进行测序。在充分的测序覆盖下，使用STAR-Fusion软件发现了所有的融合基因。使用各种浓度梯度的融合RNA sequin混合物就可以对融合基因进行定量检测，我们也对这种方法的准确性进行了评估，结果发现跨越了融合断点的分隔测序结果与初始融合RNA sequin加样浓度具有明显的相关性（R^2 = 0.929），而与两个不同基因匹配的跨越配对测序结果（spanning pair）的R^2仅为0.479。

融合基因常常通过可预测的亚克隆频率（subclonal frequency）表达。因此，我们根据正常基因的表达情况，相应地设计了一套丰度范围跨越64倍的融合基因sequin亚型分子，将正常基因与相应的融合基因混合在一起，形成了表达动态范围达到128倍的混合体系。这种混合物再与普通的RNA sequin混合，最后以较高的终浓度与K562细胞的RNA总提取物混合，完成文库构建及后续的RNA测序工作。实验发现，这些融合基因sequin分子的预测值与实测值之间存在密切的线性关系（R^2 = 0.907），灵敏度极限为7.56 amol μL^{-1}。融合基因的表达倍数与普通基因的表达倍数之间也存在很好的线性关系（R^2 = 0.853）。我们在人工染色体里设计了24个融合基因，最终发现了其中的20个，检测灵敏度达到了83.3。使用STAR-Fusion软件发现了21个假阳性融合基因，准确度为48.8。我们要求所有的融合基因都必须满足以下条件，即至少>1条分隔测序结果，>1对跨越测序结果，两个融合断点都能够与一个参考外显子的剪接位点匹配上。我们筛查了942个K562细胞基因组中可能的融合基因位点，最后锁定了12个可疑目标，其中有7个基因与其它基因之间存在同源序列或重复序列，2个是之前已经被确认过的融合基因，分别是*BCR–ABL1*融合基因和*NUP214–XKR3*融合基因。

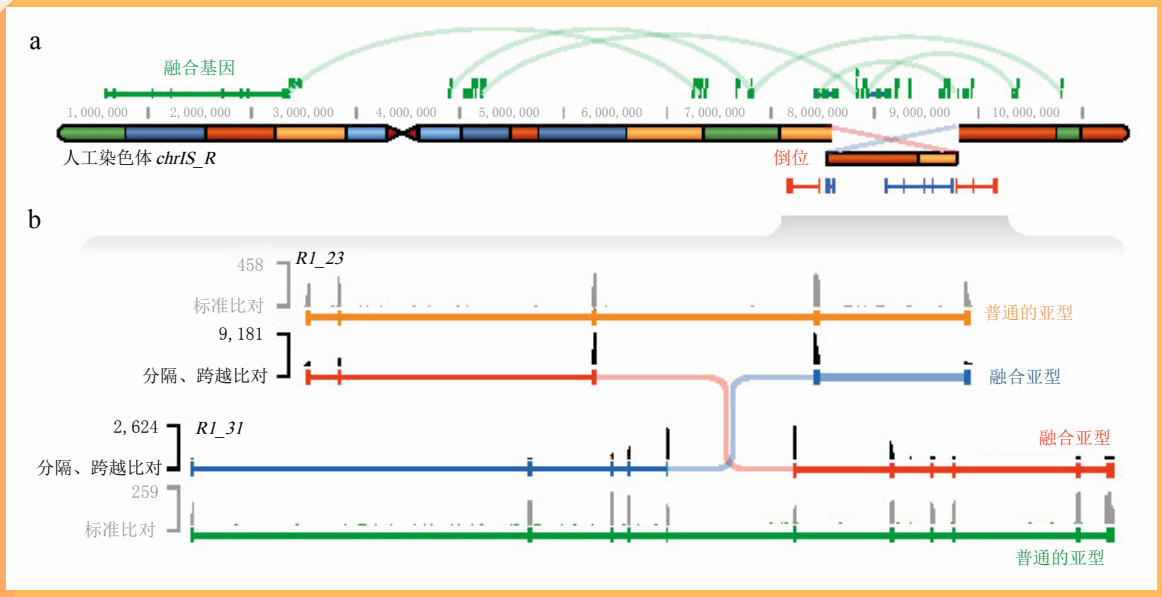


图5 融合RNA sequin简介。a，人工染色体*chrIS*_R经人工重排之后就可以得到人工融合基因RNA sequin。b，杂合型重排可以在*R1_23*和*R1_31*位点得到两种相反的融合基因。RNA测序覆盖显示了横跨融合连接处的正常的比对（灰色直方图）和分隔、跨越比对（黑色直方图）。

2. 讨论

RNA sequin是一种新型的RNA测序分析工具，可以帮助了解RNA测序技术的极限和偏倚，提高对转录组的认识，同时也可以充当标准化的参照系，用于测序文库构建、测序和生物信息学软件等多个方面的校准和比较工作。我们也可以提供RNA sequin混合物（详见http://www.sequin.xyz），并同时提供相应的人工染色体序列以及基因注释和支持软件等信息。

认识RNA测序技术的检测极限和偏倚情况，是将该技术应用于临床诊断工作的必要前提。在RNA测序诊断工作中常规使用内部对照不仅能够提高诊断的准确性和可靠性，也可以对不同患者的诊断检测结果进行标化，帮助制定诊断标准。

在研究工作中，我们设计了一系列RNA sequin分子来模拟人体的转录组。这些RNA sequin可以帮助研究任何基因组生物学问题。同时，Deveson等人也设计了一套DNA sequin系统，用于全基因组测序及全外显子组测序工作。这些DNA及RNA sequin系统可以帮助科研人员在测序试验中设置好内部对照，提高NGS的准确性，进一步拓展其应用范围。

原文检索：

Ira W Deveson, Wendy Y Chen, Ted Wong, Simon A Hardwick, Stacey B Andersen, Lars K Nielsen, John S Mattick & Tim R Mercer. (2016) Representing genetic variation with synthetic DNA standards. *Nature methods,* 13(9): 784-791.

Simon A Hardwick, Wendy Y Chen, Ted Wong, Ira W Deveson, James Blackburn, Stacey B Andersen, Lars K Nielsen, John S Mattick & Tim R Mercer. (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *NATURE METHODS*, 13(9): 792-798.

Eason/编译

特约编辑招聘启事

为了及时收集生命科学最新资讯、提高《生命奥秘》办刊质量，现面向从事生命科学或对这学科有浓厚兴趣的科研人员、学生诚聘特约编辑（兼职）。

职位职责：

独立完成《生命奥秘》专题的策划：对基因组学、蛋白组学、生物信息学和细胞生物学等学科的发展以及生物医学领域相关技术（例如基因诊断技术、干细胞和克隆技术、生物芯片技术等）的应用进行翻译及深入评述。

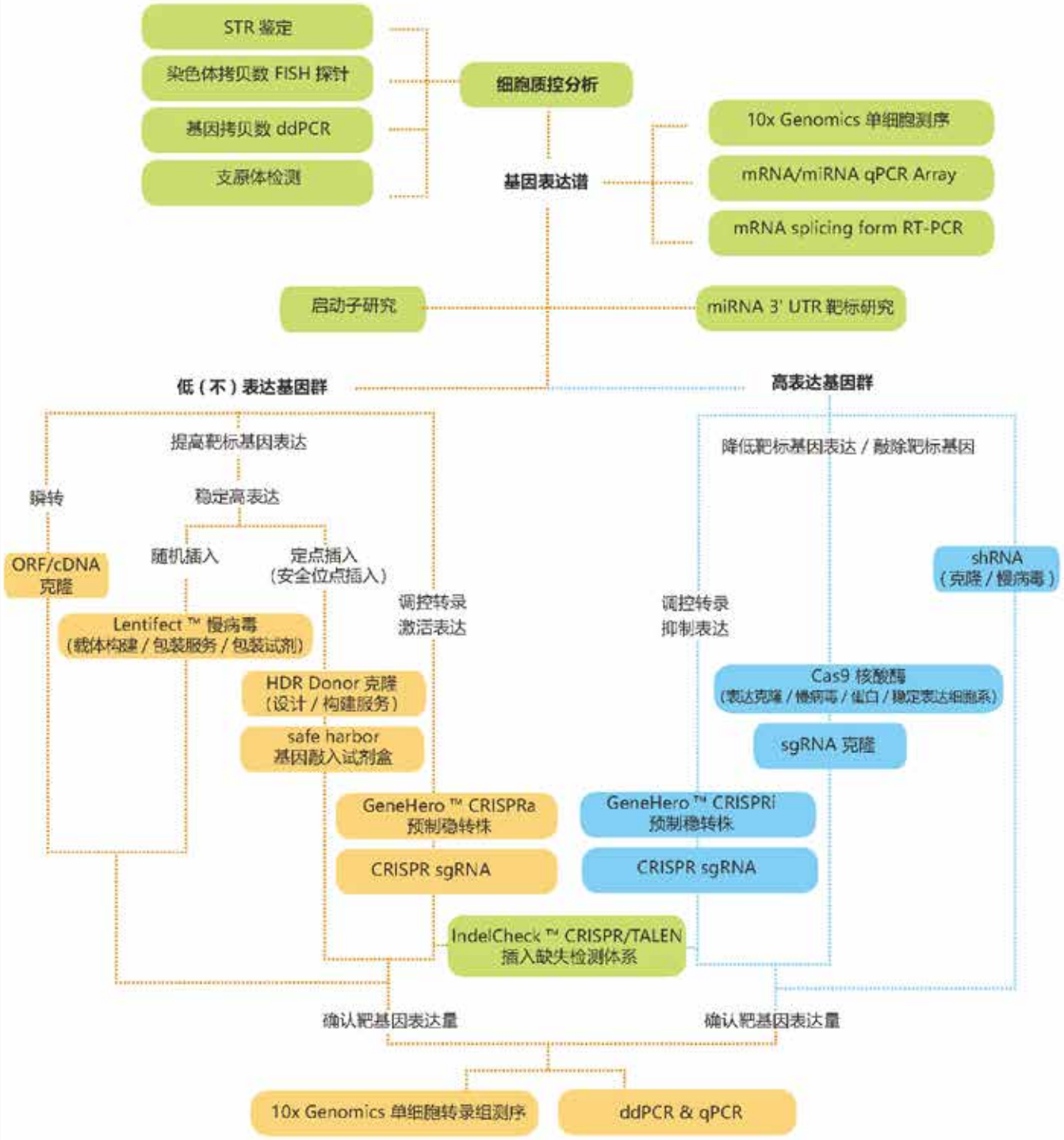
选题要求内容新颖、评述精辟、注重时效和深入浅出。尤其欢迎以自身系统研究为基础的高水平译述与评论，结合所从事的科研工作提出自己的见解、今后设想或前瞻性展望。

要求：

1.具备基因组学、蛋白组学、生物信息学、细胞生物学等生命科学学科背景；
2.具备良好的生命科学前沿触觉；
3.具备较高的外文文献翻译、编译水平；
4.具备较强的选题策划、资料搜集、组织能力，以及专业稿件撰写能力；
5.具有高级职称；或者拥有（正在攻读）该领域的最高学位。

有意者请将个人简历发送至 editor@lifeomics.com

基因功能研究产品和服务指南



- 相关产品及服务：【转染试剂】【AAV 包装服务】【荧光素酶检测试剂盒】【细胞系构建服务】
- 客户发表文章超过 100,000 篇

扫码进入网页版

合办专题专刊
网站广告合作
邮件群发推广
请致电(020)32051255

www.LifeOmics.com