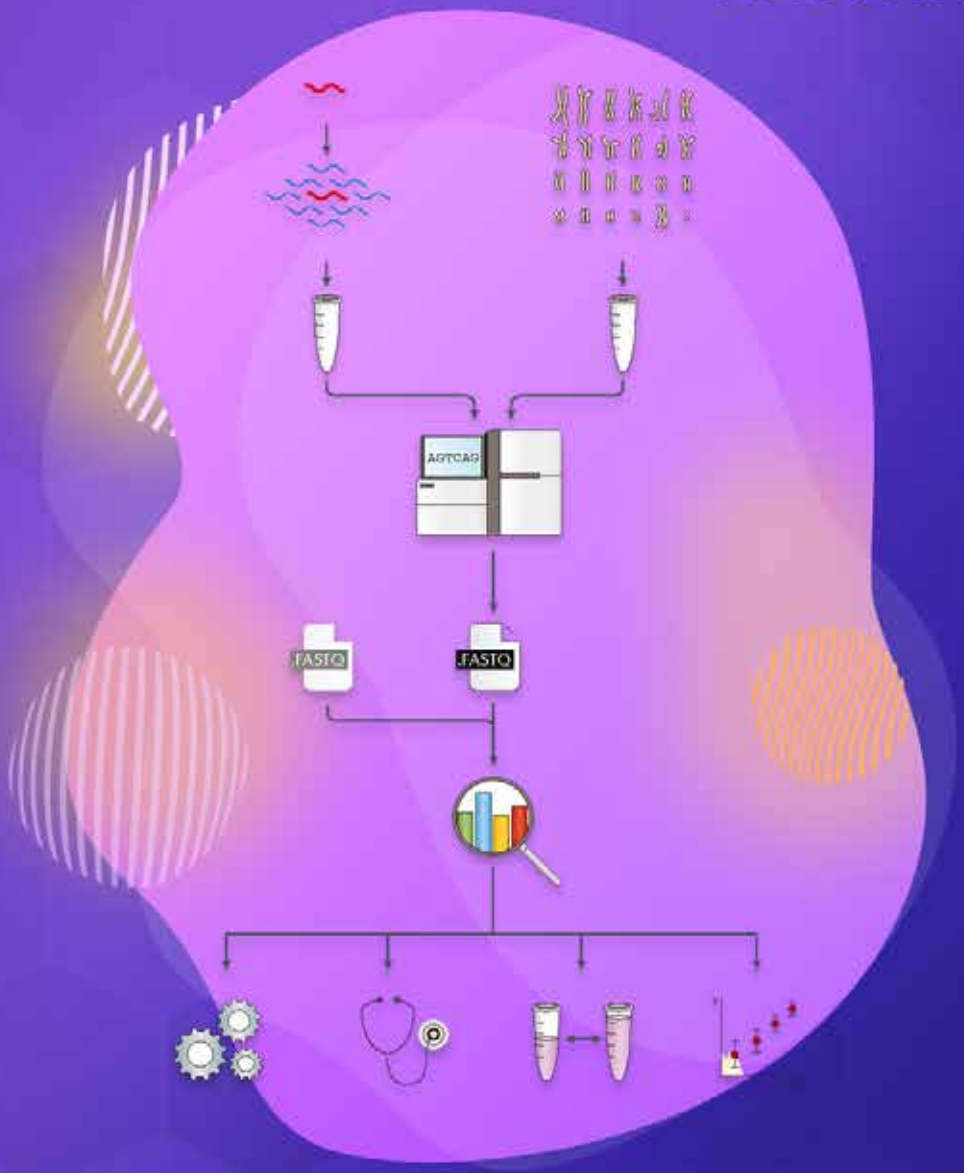
生命奥秘

总126期 2020/6
LIFEOMICS



新一代测序技术的“参照标准”（PART-I）

无奇不有
生命世界
阅读生命
走进科学

目录 CONTENTS

专题:新一代测序技术的参照标准（PART-I）

前言 01

一、新一代测序技术的参照标准 02

1. 参照标准 03
2. 生物参照材料（Biological reference materials） 04
3. Spike-in对照 09
4. 管理问题 11
5. 总结 12

二、Spike-in对照——全基因组分析中被忽视的重要一员 22

1. 何时设置spike-in对照 23
2. 在微球菌核酸酶测序（MNase-seq）中设置spike-in对照 24
3. 在RNA-seq中设置spike-in对照 26
4. 在ChIP-Seq中设置spike-in对照 26
5. 在gDNA-seq中设置spike-in对照 27
6. spike-in对照究竟是什么？ 28
7. 如何使用spike-in对照 29
8. 如何利用spike-in对照进行数据标化 29
9. 数据是否可以进行回顾性标化？ 31

专题

新一代测序技术的参照标准（PART-I）

前言

新一代测序技术（Next-generation sequencing, NGS）让我们有能力对基因组进行更加广泛、深入的研究，而且有助应用于临床诊断，帮助医生发现与疾病有关的那些遗传异常。

一、新一代测序技术的参照标准

目前，如何解读新一代测序技术产生的数据仍是一个大问题，这主要是因为基因组的规模太大，又非常复杂，而且在整个测序工作流程（比如样品的制备、测序、后续数据解读等）中还存在技术误差（technical error）。但是利用参照标准（reference standard），可以发现并减少这些误差带来的影响。所谓参照标准，指的就是由明确的遗传物质，或人工合成的物质作为实验对照物，来校正测序结果。使用参照标准和相应的统计学方法，就能对新一代测序数据进行更好的分析，这也是未来让测序走进临床的必备条件。

新一代测序技术正在被越来越多地应用于临床诊断，这主要是因为这种技术可以发现与疾病有关的遗传变异（genetic variation），发现致癌的融合基因（fusion gene，或者明确致病的病原体（pathogen）。与之前的诊断测序技术不同的是，使用新一代测序技术，用一点点样品，只通过一次实验就可以对DNA或RNA序列进行完全的定量和定性分析，这极大地提高了这种技术的使用范围。

尽管新一代测序技术有上述优点，但是该技术的应用，尤其在数据分析方面目前还存在一定的问题。比如人体的基因组非常庞大，再加之又存在丰富的多样性，这就很难用简单的方法对人类基因组数据进行分析。随着测序广度的拓展，更进一步提高了假阳性和假阴性的风险。此外，现有的用于人类基因组中的重复区域和末端区域的的测序技术还不太成熟，不能很好地完成测序工作，而且在样品的制备、文库构建和测序，以及后续的生物学信息学分析等步骤中也都存在技术上的不稳定性。这些因素都让测序数据分析工作变得更加复杂。

这些技术上的问题都会降低数据分析的准确性，导致诊断错误，很多临床实验室也一直在用Sanger测序法，对复杂或不确定的测序结果进行验证。不过其中很多问题都可以借助参照标准来解决，这也是很多专业机构一直在推荐的做法。用参照标准来评估诊断结果的做法在分析化学（analytical chemistry）等领域都已经得到了很好的开展，这也为我们提供了一个很好的范本，能够移植到其它NGS工作当中。能够作为参照标准的都是同源性足够高，而且某一（几）个特性都非常稳定的对照物，并且已经被证实适合用于这个检测手段。这些特性都是定性的，比如一个DNA分子的序列；也可以是定量的，比如样品里的丰度。由于DNA分子非常稳定，可以以极高的保真度进行复制和准确的检测，所以成为了非常理想的参照标准。虽然RNA分子也有类似的特性，但由于稳定性较弱，所以在处理和储存方面需要更加谨慎。

在NGS领域，参照标准的使用，进展相对还比较缓慢，部分原因也是因为技术的发展太过迅猛。不过，越来越多进入临床实践工作的NGS技术已经开始重视开发配套的参照标准了。接下来，我们将对NGS里参照标准的发展和使用情况作一番综述。综述将重点介绍参照标准设计的原则和统计原理，以及不同方法各自的优缺点。我们认为，如果要开展NGS数据分析，以及NGS在临床诊断上的应用工作，那么都应该好好了解参照标准，以及相关的统计原理。

1. 参照标准

为了尽可能地缩小实验误差，对参照标准也需要用多种不同的方法进行鉴定。对于NGS工作而言，就是用不同的NGS技术，或者Sanger测序法和定量PCR等正交方法（orthogonal method）对其进行鉴定。只有经过了这一系列严格的验证，最后得到的结果，比如基因型，才能够当作参照标准。美国国家标准与技术研究所（National Institute of Standards and Technology, NIST）等度量机构能够提供参照标准认证，这说明它们的成分已经根据现有的技术进行了鉴定，也可以直接在其他检测机构进行追踪。

为了便于使用，参照标准还必须是可以被交换的（commutable），也就是说，在实验中，这些参照标准必须与样品进行比较。比如，可交换性（commutability）要求在对患者的基因组cDNA进行测序操作，包括制备文库、测序和数据分析时，也需对临床DNA参照标准进行同样的操作。如果参照标准的可交换性很低，那么在校正检测时就会出现偏倚，最终无法得到准确的诊断结果。由于可交换性非常重要，因此相关机构也给出了清晰的临床标准指南，这些指南都能够非常容易地应用于NGS参照标准当中（背景知识1）。

一旦具备了可交换性，参照标准就可以对各种不同患者标本的检测结果进行校准。比如，我们可以通过与已知浓度的参照标准进行比对的方法，来判断患者标本里的DNA丰度。当然，这种方法也存在一定的误差。这种校对的方法还可以对多个不同的样本进行标准化检测，还可以用参照标准设定诊断阈值。

1.1 NGS中的参照标准

开发参照标准还需要解决一些NGS特有的问题。目前，NGS技术的测序深度和广度都足以在一次测序中破解基因组中的绝大部分序列，这种技术进步虽然让测序走向了临床诊断和预后判断，但同时也增加了很多假阳性率。为了解决这个问题，我们必须找到一种参照标

准，能够真实地反映NGS实验中基因组或转录组的多样性（diversity）。

由于被测序列的广度有了极大的提升，所以NGS测序的精度就显得尤为重要。比如，在大范围的基因组片段内，其实只有为数不多的罕见突变位点，哪怕很低的假阳性率也会带来大量的突变信息，造成误诊。所以信号过滤就是最常用的提升测序精准度的方法，可是这些操作又会增加假阴性率，容易造成漏诊。为什么很多NGS会得到假阴性诊断（false-negative diagnosis），现在还不是特别清楚，如果没有很好的参照标准，则也很难对其进行检测。

测序技术的覆盖率（sequencing coverage）也是NGS技术中一个非常有意义的因素，这往往与文库的复杂程度和成本考量有关。覆盖率是确保测序灵敏度的关键因素，也是发现变异、了解基因表达情况的必要条件。如果覆盖率较低，那么灵敏度就会降低，不确定性（uncertainty）就会增加。由此可见，参照标准对于覆盖率的检测能力，也是NGS能否得到良好运用的重要保障。

如果DNA样品出现损坏，或者测序出现错误，此时高覆盖率将是降低NGS差错率的手段之一。不过，因为测序设备自身，或者短片段序列拼接错误导致的系统性测序错误（systematic sequencing error）是无法通过提高覆盖率来进行纠正的，此时就需要参照标准的帮助了。

由于参照标准可以提供确认的“真相”，所以预计值和实际测量值之间的差异便是可以根据经验估计出的误差范围。可是这个对于NGS工作却非常困难，因为在NGS的工作流程中往往包含了多个步骤，而每一个步骤都有自己的误差范围。参照标准的作用就是计算整个NGS流程中的这些误差，并且对最终的诊断结果给出一个总的误差范围。

2. 生物参照材料（Biological reference material）

天然的遗传物质也是很好的参照标准，而且相对廉价、易得，同时由于包含了人类基因组和转录组的多样性和完整性，所以应该具备很好的可交换性。不过我们还不清楚这些天然的遗传物质对于各种NGS技术都有哪些限制，这也使其成为了一个很好的参照标准，可以用来对各种不同的NGS技术进行比较（图1）。

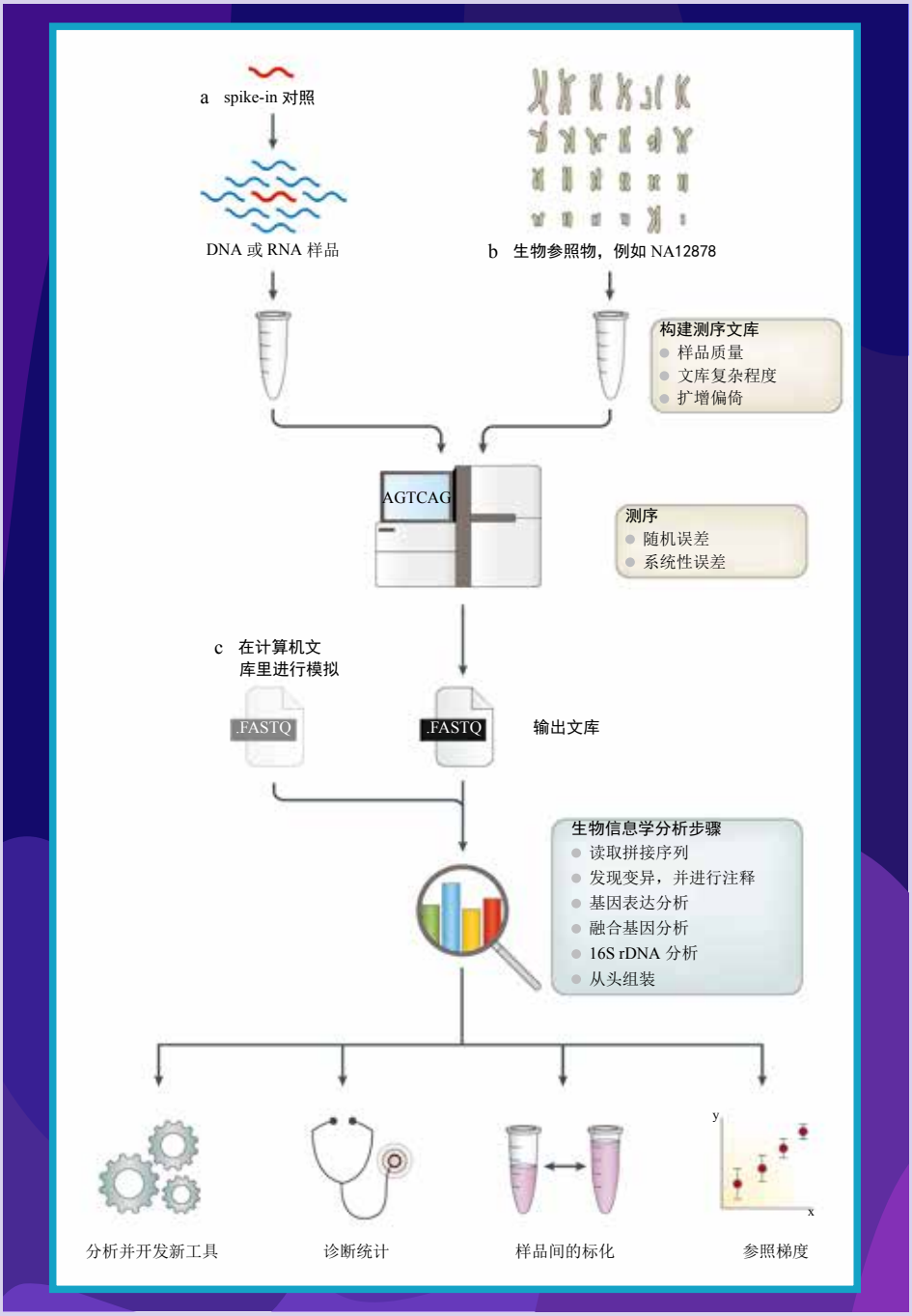


图1 新一代测序工作流程中参照标准的应用。a，在待测DNA或RNA样品中加入对照物，在后续的文库构建、测序和生物信息学分析工作中起到定量和定性的参照物作用。b，明确的生物参照是非常好的测序实验对照品，但是并不能直接添加到待测样品里使用。c，计算机对测序文库进行模拟测序可以快速评估图中蓝色方框内标识的关键生物信息学步骤。参照标准可以对图中米黄色方框内标识的NGS测序步骤的偏倚和误差进行评估。

2.1 人类基因组参照标准

发现与疾病相关的遗传变异，是NGS在临床中最主要的用途之一。可是，使用不同的测序技术和生物信息学分析方法，往往会得到不同的结果。有时在同一个人的基因组内，可能会得到数千个不同的变异结果。这说明我们急需一个可靠的人类基因型图谱，作为人类基因组参照标准。

由于最初的人类基因组参照序列是多个人的基因组序列拼接的结果，所以并不是真正的参照序列。不过，人们后来又绘制了多个人的基因组序列，用来在NGS测序工作中作为参照序列使用。有些人可能会担心这些被测个体的基因组不稳定（genomic instability）和漂移（drift）会带来问题，但是我们可以以这些人的转化细胞系为实验样品，获得稳定的基因组DNA，而且这种实验非常容易，成本也很低。

源自欧洲健康远古女性的基因组NA12878是目前最好的人类基因组参照标准。不同NGS技术在基因组分析工作中的局限性，可以通过将不同测序及数据分析技术得到的结果加以整合的方式得到弥补和改善。通过这种方法就可以得到适用于全人类的、高质量的单核苷酸变异图谱（single nucleotide variant, SNV）和微小插入及删除图谱。比如，长片段测序技术主要用来发现基因组结构变异，也可以用来了解NA12878不同变异的分期信息（phasing information）。虽然我们经过了种种努力，但是在人类基因组中还是有一些区域，由于GC含量太高，或者因为复杂程度太低，以及重复片段太多等原因，目前还不能对其进行测序。这些疑难片段在不同的个体之间往往也有明显差异，而且其中也含有很多具备临床价值的突变。

很多临床实验室都已经常规地在NGS测序工作中将NA12878基因组DNA当作对照物来使用了（背景知识2），可以用高质量的基因型图谱对发现的变异位点进行对照，来判断测序的准确性。对多重重复片段进行测序则可以评估测序每一轮的可重复性（repeatability）和每一次的可重复性（reproducibility）。虽然NA12878使用如此广泛，但是其自身序列却只能用于科学研究，并不能进行商业开发。

人类基因组遗传变异的多样性也是促使我们绘制不同远古祖先参考基因组序列的动力。因此，NIST也正在扩增他们的基因组参照标准文库，涵盖了更多不同种族人群的数据信息。这些基因组数据都是通过一个名为“个体基因组计划（Personal Genome Project）”的项目获得的。而且该项目已经获得信息拥有者的授权，可以将他们的数据应用于很多不同的领域。还有大量区域性的项目也在贡献自己的力量，比如为某些国家绘制参考基因组序列，为某些人群绘制参考基因组序列等。

2.2 用于疾病研究的参考基因组

很多患者的基因组都携带了很多致病的突变，这些信息也是NGS技术指导临床诊断工作必不可少的重要参考数据。不过，在多样性极为丰富的致病突变中，只有极少数的突变是每个患者都会携带的，这就给绘制致病参考基因组图谱带来了很大的麻烦。

携带有致病突变的患者提供的细胞标本可以被转化成细胞系，用作绘制参考基因组图谱的实验材料。“遗传检测用参考材料协作组（Genetic Testing Reference Materials Coordination Program, GeT-RM）”已经对很多携带了遗传性疾病致病突变的细胞系进行了鉴定，同时也对其中的药物遗传学位点的突变情况进行了检测。这些基因组数据都可以作为研究疾病相关突变的参考信息。

基因组编辑技术（genome editing）也可以对细胞系进行改造，引入目标突变，这也是构建细胞系检测材料的一种手段。不过，由于存在脱靶问题，所以更加需要对这类细胞系进行仔细的鉴定。

构建稳定的、经过明确鉴定的、可更新的肿瘤细胞参考基因组细胞材料，是非常困难的。虽然我们通过对肿瘤细胞及其对应的正常细胞进行鉴定获得了非常有价值的参考数据，但同时也发现，肿瘤细胞的基因组非常复杂。不过，鉴于肿瘤标本往往都非常小，而且数量也有限，所以很难有足够的样品用于绘制参考基因组图谱。而且，在一个肿瘤标本中往往含有多种不同的肿瘤亚克隆细胞群体，所以就更难得到稳定的、可靠的、同源的参考肿瘤细胞材料了。培育肿瘤细胞系可以解决这个问题，也可以与其它细胞系混合在一起，模拟真实肿瘤组织里的那种异质性现象。测序质控组织（Sequencing Quality Control, SEQC）等机构也正在培育参照细胞系，用于肿瘤的诊断和研究等工作。

2.3 参考RNA样品

RNA测序的情况就更加复杂了，这是因为细胞转录组的规模和多样性更加庞大，而且RNA样品的质量，以及文库构建方法参差不齐，后续的生物信息学分析步骤也极为复杂。不过无论如何，准确并可靠地获得细胞在各种外部刺激、实验室培养条件等不同情况下的基因表达信息都是非常有价值的。在同一批次实验中，这些不同条件下的细胞转录数据，以及很多参考RNA材料都是不可再生的。

SEQC、GEUVADIS和生物分子学资源协会（Association of Biomolecular Resource Facilities, ABRF）等机构都是用人类参考RNA标本对不同的新一代RNA测序技术和实验流程，以及不同实验室的准确度和可重复性来进行评估。将这些参考实验材料按照已知的比例进行组合，也可以根据差异表达基因的检出情况来对NGS技术的相对准确性进行评估，这种一致性评价手段也已经成为了最好的RNA测序

实验标准。

虽然这些RNA样品都会用多种不同的NGS技术进行测序，但随着测序深度的不断增加，总能发现新的亚型（isoform），这说明转录多样性问题（transcriptional diversity）还需要我们进一步探究。这同时也提醒我们，使用缺乏全面、一致性注释的天然RNA参考标本来评价RNA测序结果的假阳性和假阴性难度非常大。不过即便如此，这些天然的RNA参照标准还是非常重要的，尤其在帮助我们认识疾病相关复杂的转录特征，比如诊断白血病BCR-ABL1融合基因转录等时更有价值。

2.4 微生物组的参照标准

宏基因组测序（metagenomic sequencing）可以为我们提供一个包括样品所在环境微生物信息在内的全面序列，因此可以直接明确是否存在致病微生物，以及包括哪些致病微生物。与以前的技术不同的是，NGS技术可以发现在实验室里无法培养的，或者全新的微生物。不过，微生物丰富的多样性也给NGS分析带来了挑战，因为每一个微生物都有不同的基因组结构。如果缺少参考基因组序列，缺少背景基质DNA信息（比如患者标本里的人体DNA信息），则更难以分析。

NIST已经发布了多个微生物参考基因组，以供大家分析和开发相关工具使用。选择研究哪些基因组，主要是根据这些微生物在食品安全和临床上的重要性来决定的，这些不同的微生物基因组里的GC含量也各不相同，甚至差异很大。FDA建立了一个FDA-ARGOS数据库，里面储存了已经经过鉴定确认的基因组序列信息，其中也包括了多种感染性微生物的相关数据。利用这些数据可以对NGS检测项目进行标化。

人工模拟微生物群落是将多种不同的人工培养微生物，按照不同的浓度混合而成的。这是宏基因组分析里最常用到的参照标准。也可以直接使用提取的基因组DNA进行混合，制成人工模拟微生物群落参照标准。使用有完整参考序列信息和浓度信息的微生物群落，还可以研究基因组定量技术的极限，以及从头组装的能力。这种人工微生物群落也可以用于开展16S rDNA研究时进行多重PCR反应的模板，从而提醒我们是否在分析时漏掉了一些目标微生物。

为了提高不同实验室之间的标化能力，人类微生物学项目联盟（Human Microbiome Project Consortium）收集了大量人工模拟微生物群落的标本，这些微生物涵盖了各种不同大小、不同GC含量、不同重复序列，以及不同种类发生（phylogenetic diversity）的基因组。后续人们又启动了微生物组质控项目（Microbiome Quality Control, MBQC），借助多种不同的参照标准，来评价各种不同测序技术检测人微生物组的能力。最近的工作进一步拓展了参考微生物的应用范围，比如模拟特定环境中的微生物，或者让其更适合NGS测序等。在比较16S数据和鸟枪测序数据时，这些参考微生物组都是非常有用的对照物，既可以评估不同GC含量引起的偏倚，也可以在宏基因组分析时起到一个标尺的作用。

3. Spike-in对照

使用天然的遗传物质作为NGS的参照标准，一大问题是很难在不污染后续分析的前提下，与人体标本很好地结合。而Spike-in对照（Spike-in control）则被设计为可以直接添加到待检样品中和这些样品一起参与文库构建、测序等工作（图1），因此，这些参照物就可以起到很好的对照作用，对待检样品进行定量和定性的分析。

这种参照物的序列往往都是非人类的序列，或者是人造的序列，其中都含有一个特殊的分子标签（molecular barcode），借此就可以分辨出哪些是参照物序列，哪些是待检样品序列。比如，PhiX噬菌体的基因组就是最常用的对照，可以对Illumina测序结果进行最基本的质量控制，了解其质量和出错率。

设计对照序列的灵活性是非常高的，唯一的限制因素就是序列合成技术，因此，这种人工核酸分子可以根据诊断需要快速开发，解决很多在NGS检测时碰到的特殊问题。这些对照分子都是单独制备，可以根据不同的浓度将其混合形成复杂的混合物，这就包含了多种不同特征，也形成了一个“梯度对照体系”，可以更好地进行定量分析。尽管人工对照有这么多优势，但是可交换性却一直是一个大问题，因为人工合成的对照还是无法反映天然DNA和RNA分子复杂的行为。

3.1 基因组spike-in

人工合成的DNA spike-in对照也被用来代表各种人类遗传变异，比如SNV、缺失突变、大规模结构和拷贝数变异等。利用NGS检测发现这些变异，尤其是有临床意义的变异，就是spike-in对照的价值。有很多变异都可以通过简单的DNA spike-in对照被发现，这也进一步拓展了NGS诊断的应用范围。

使用现有的NGS技术，是很难发现有临床意义的变异的。但是人工基因组spike-in就可以代表这些变异，因此可将这种人工DNA spike-in与天然基因组参考材料结合使用。

调整人工DNA spike-in的丰度，也可以模拟基因组生物学的定量特征，比如变异等位基因频率和拷贝数变异等。比如，代表了参考等位基因和变异等位基因的一对DNA spike-in就可以模拟杂合基因型，也可以进一步稀释，模拟更低的体细胞突变等位频率（这种情况在肿瘤细胞中很常见）。这些内部DNA spike-in梯度就可以进行定量统计，给出一个阈值，来区分低等位基因频率的测序误差和真阳性（背景知识3）。

3.2 RNA spike-in

RNA spike-in最开始是由RNA外部质控组织（External RNA Controls Consortium, ERCC）开发的，使其在定量逆转录PCR（quantitative reverse transcriptase PCR）和微芯片实验中作为参照标准，不过后来却被广大的RNA测序工作者广泛采用了。ERCC spike-in由多种不同长度、不同GC含量的多腺苷化的RNA分子组成，其中不含人类基因组同源序列。剪切RNA spike-in可以模拟人类基因复杂的外显子和内含子结构，这有助于我们使用RNA测序技术进一步了解RNA可变剪切和组装的机制。定制的RNA spike-in则有更加特定的功能，比如小RNA测序和检测致癌的融合基因等。

RNA spike-in混合物也可以用于人体基因表达分析。通过与已知浓度的对照进行比对，借助RNA测序的方法对基因表达进行定量分析（背景知识4）。这种方法甚至可以用绝对拷贝数进行定量分析。RNA spike-in的混合方式有很多种，在不同的检测实验中被当作阳性对照或者阴性对照。通过在不同的样品中加入不同的RNA spike-in，还可以在不同的基因表达水平上评价倍数改变的准确性，也可以对伴随样品不同的基因表达的解读工作提供有价值的信息。

RNA spike-in可以提供一整套全面鉴定过的真实信息，能够帮助我们判断假阳性和假阴性结果，这是天然的RNA参考样品无法做到的。这一优势也在SEQC和ABRF等项目中得以很好地体现。这些项目都使用了ERCC参照，因此可以对RNA测序的表现进行全方位的评测，包括不同实验室，以及不同NGS技术之间的敏感性和技术差异的评测。

RNA spike-in还可以用来对不同的样品进行标化（背景知识5）。这在单细胞RNA测序实验中尤为重要，因为这种单细胞实验往往会对数千个细胞的mRNA数据进行比对，而每个细胞mRNA的组成，以及实验的误差给实验结果带来的影响都差异极大。在细胞裂解液中加入RNA spike-in，便可以根据文库中对照的比例估计出每一个细胞内mRNA的含量。如果spike-in比例明显过高，通常说明细胞RNA的含量较低，也可能是实验出了问题。在单细胞层面检测转录体绝对拷贝数的能力让科研人员能够从新的角度去研究转录组动力学问题，这在以前使用传统RNA测序技术时是完全无法想象的。

3.3 生物信息学分析

在整个NGS实验中，生物信息学分析是最为复杂的，也是实验偏倚和误差的一大因素。我们可以快速、方便地构建计算机数据库（*in silico* data set）（图1），这对于开发计算机软件，进行生物信息学分析工作评价也都是非常有价值的。

FASTQ和SAM/BAM等常见的数据库都能够快速地模拟或改变，以形成标定的真实参照物（ground truth），对生物信息学分析结果进行检测。比如，在模拟的数据库里，我们可以按照任何频率，明确罕见的或疑难的变异。而且，在分析流程中，从最开始的基因组定位，到比对，直至最终的分析，每一次模拟过程都可以全程追踪。这就让每一个步骤都可以被评价，能够对整个NGS流程进行快速优化。

科研人员也开发出了不同的计算机软件，根据已知的基因型生成人体基因组，来模拟各种NGS文库。在这些软件中还可以加入各种测序错误，甚至在其他NGS文库中出现过的错误构建模型。对于RNA测序来说，计算机参考数据库也可以当作参照物，对各种已有的分析工具进行评价。最近，也有科研人员针对转录子定量工作开发出了新的软件。

计算机数据库最明显的不足就是只能用于NGS工作中的生物信息学步骤当中，很难用这些模拟数据充分地模拟真实世界里的复杂性和多变性。因此，虽然这些计算机数据对于检验生物信息学工作非常有帮助，但是在临床诊断工作中并不能取代真实数据，因为它无法像真实数据那样充分地反映临床检测工作中的各种变化。

4. 管理问题

根据人基因组序列给出诊断信息（这些有可能会进入每个人的健康档案，伴随一生），必须受到严格、清晰的监管。地方监管机构通常都有责任和义务管理NGS临床检测业务，包括参照标准的使用等。FDA和欧洲医药管理局（European Medicines Agency）是当今世界上最大的两个医药卫生监管机构。出于行文简化，我们只介绍美国的情况，但是这些监管原则同样适用于其他国家。

FDA要求，对于体外检测诊断技术而言，必须明确展示用于检测诊断的目标基因（或突变）的有效数据（validation data）。不过，对于能够在全基因组范围内发现突变的NGS技术而言，这一要求实在太过昂贵，可行性不高。在使用Illumina公司的MiSeqDx测序仪检测CFTR突变等FDA已经批准上市的几个NGS诊断项目中，用参照标准来评价诊断效果已经是非常重要的一个步骤了。

根据美国临床实验室改进修正案（Clinical Laboratory Improvement Amendment, CLIA），NGS检测也可以在有资质的实验室里开展。此时，参照标准同样是证明NGS检测效力的试金石，其中就包括了准确性、精确性、敏感性、特异性、可报告范围和参考区间等全面的分析指标。目前，大多数NGS诊断实验室都在积极地通过这种认证。其中有一部分原因是因为要获得FDA的批准，花费更高，而且时间更长。

获得CLIA认证的NGS检测的实际临床表现，正在常规接受实验室能力比对验证（proficiency testing），也就是将检测样品周期性地寄送给这些实验室，要求他们进行检测，然后返回检测结果，以供评价。很多时候，也可以通过非正式的实验室间交换样品的

方式来进行室间质评。由于NGS检测的变异非常多，所以更适合进行方法学的评价，而非针对某些特定的基因，或突变进行质量评定。使用NGS技术，可以对一组中心参照样品进行质评，对不同实验室，不同NGS检测技术给出独立的、标准化的评价结果。

美国病理学会（College of American Pathologists, CAP）为NGS检测设计了一套最为全面的实验室能力比对验证体系，其中就包括了生殖系突变和体细胞突变检测，以及各种常见的、可干预的融合基因。全球微生物鉴定机构（Global Microbial Identifier）最近也针对微生物基因组的全基因组测序检测，启动了一套实验室能力比对验证系统，要求参与者提交活体微生物和提取的基因组DNA样品，以及NGS检测数据。

由于NGS的数据分析工作非常复杂，在生物信息学分析步骤里也引入了计算机质控（in silico proficiency testing）。这要求参与者用自己的生物信息学分析流程，提交NGS文库数据用于分析。这在评价涉及复杂结构变异的诊断结果，或者评估假阴性率时非常有意义。FDA最近也启动了 precisionFDA项目，不过这还不是正式的质控系统。这是一套在线的系统，参与者可以评估并分享他们的NGS数据，以及他们使用的生物信息学工具，也可以对分析结果进行标化。

5. 总结

高通量的NGS技术让我们可以在一次实验中，对全基因组和转录组进行一番彻底的分析。基于这种优势，NGS技术已经快速地进入了临床疾病诊断领域，尤其是与遗传有关的疾病诊断。不过，临床诊断并非如此简单，尤其是面对规模和多样性都如此庞大的基因组、如此复杂的序列和生物信息学分析工作时，更是如此。因此，参照标准的价值就显得尤为重要，因为借助参照标准，才能让我们发现NGS技术的局限和不足。

前文已经介绍了各种用于NGS的参照标准（表1）。天然的生物样品能够很好地保留基因组的复杂性，也可以呈现很多常见的致病突变（现有的测序技术或生物信息学工具还无法发现这些突变）。然而，人工合成的参照物则可以按照设计者的意愿，进行自由设计，以求解决临床上的特殊问题。而且这些人工参照物还能够对基因组生物学进行定量研究。计算机数据库虽然不适合作为生物参照标准，但是可以对生物信息学分析工作进行优化。每一种参照标准都有各自的优势和不足，最理想的状况是，将不同类型的参照标准组合起来，为NGS的质控和验证工作提供强有力的支持和参照。

目前，在NGS的工作流程中，参照标准已经作为基准被广泛使用。我们估计参照标准的使用还会进一步扩大，这也将催生出一批新的统计学和生物信息学分析工具，其中就包括评价文库统计学和不确定性能力的分析工具，而且有了这些数据，就可以开发出新的生物信息学分析工具。此外，使用参照标准也可以拓展我们对疑难的、复杂的和定量的基因组评定工作，并且对其进行标化，这些新进展也都会进一步推动NGS诊断技术的发展。

技术的持续创新必将会催生出更好的人工合成参照物和被人们认识得更加全面的生物材料。技术创新与参照标准之间的这种关系是相互促进的，因为参照标准反过来也会促进测序新技术的发展和优化。提高临床诊断技术准确率、可靠性和标准化程度，继续发展参照标准是一个相对更容易的方法，这并不需要在NGS技术上有新的突破。因此，随着我们对遗传疾病的认识不断加深，随着NGS不断进入临床诊断工作，参照标准的使用也必将越来越普及。

表1 目前在NGS应用中使用的几种主要参照标准简介

参照标准	实例	优势	劣势
DNA生物材料	●GeT-RM患者细胞系 ● NA12878参考基因 ●GIAB参考基因组 ● 肿瘤参考基因组 ●NIST微生物参考基因组 ● 宏基因组人工模拟群落	●不同患者样品间具备可交换性 ●可以对含有数百万个DNA变异的全基因组的测序结果进行评价（背景知识2） ●不同的NGS技术不可知 ●转化细胞系可以不断地提供基因组DNA材料	● 不代表有临床意义的突变或序列的能力有限 ● 知情同意和隐私方面有隐患（人体样品） ● 很难进行全面的了解 ●要在全基因范围内得到可靠的诊断结果，就会牺牲精确性 ● 细胞系存在基因组不稳定和飘移的问题 ● 不能加入其它样品中，因为可能污染后续的分析工作
人工细胞系	● Horizon Diagnostics通用标准品	● 含有全基因组内的目标突变 ● 在不同患者样品间具备很好的可交换性	● 作为一种外部参照物，不能加入待测样品中 ● 可能存在脱靶问题，以及其它一些意想不到的问题
RNA生物材料	● SEQC和ABRF参考样品 ● *BCR-ABL1*细胞mRNA标准品	● 在不同患者样品间具备很好的可交换性（针对RNA降解问题） ● 各种最新的NGS技术不可知	● 由于批次间的不稳定性，必须使用同一批次的产品 ● 缺乏全转录组注释信息
Spike-in对照	● 人工合成的突变参照 ● DNA突变spike-in ● ERCC参照 ● 剪切RNA spike-in ● 小RNA spike-in	● 可评价文库构建、测序及后续分析质量 ● 能代表任何目标突变序列 ● 能够混合成参照体系，评价NGS定量分析的准确性（背景知识4） ● 可对不同样品起到标化作用（背景知识5）	● 必须首先与生物样品建立可交换性（背景知识1） ● 不能充分代表基因组或转录组的规模和复杂程度 ● 在文库中占据一定比例，通常不超过5%
计算机数据库	● 模拟的或真实的FASTQ数据库	● 数据库能够方便地编辑，以获得任何遗传特征 ● 能够对生物信息学分析工作进行标化	● 只能用于生物信息学评价工作 ● 很难模拟出试验数据里的复杂性和多样性

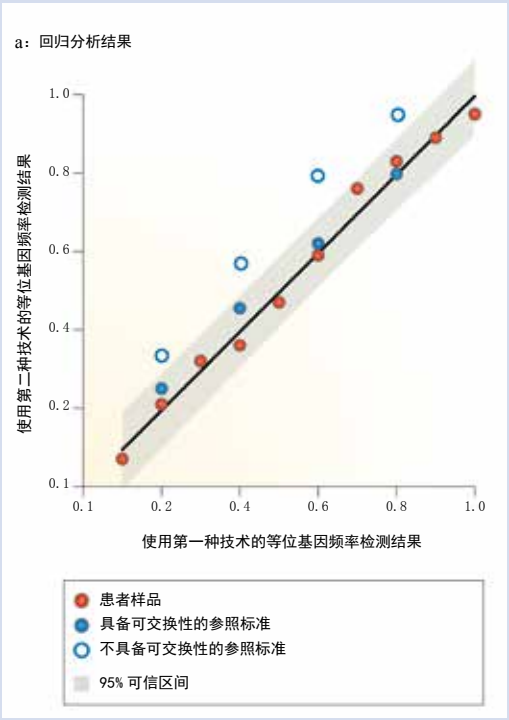
背景知识1：可交换性（commutability）

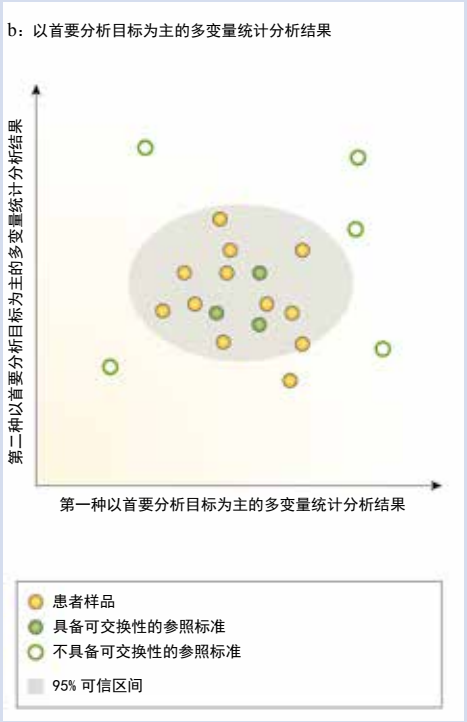
可交换性是指参照标准对不同检测标本之间的比较能力。比如，如果在对不同患者的基因组进行测序和数据分析时表现都比较稳定，那么就可认为人类基因组参考序列具备可交换性。可交换性同时也受到基质效应（matrix effect）的影响，即会受到所有样品组分的影响，而不是只会受到待测目标组分的影响。比如，样品的固定（fixation）就是一种基质效应。如果使用福尔马林固定和石蜡包埋的样品，就会降低参照标准的可交换性。

如果要具备可交换性，参照标准就不能是与被分析物一样，是天然的样品，而应该是和人工合成的spike-in对照一样的样品。对于可变度很高的样品，我们也很难开发出具备可交换性的参照标准。比如，RNA样品，或者已经进行过加帽、加尾等天然转录后修饰的mRNA样品的质量差异都会让检测结果有所不同。如果使用可交换性较低的参照标准对诊断结果进行校正，就会得到不准确的、可变的结果。所以，在使用之前，一定要先确定参照标准的可交换性。

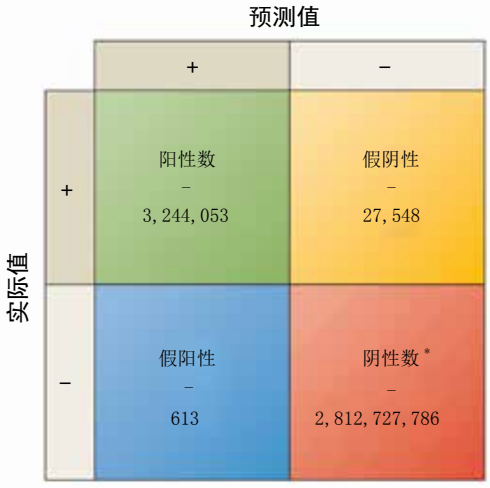
回归分析（regression analysis）可用来帮助我们了解参照标准与被测样品之间是否具有相互依赖的定量关系。如下图a所示，我们用两种不同的测序方法（比如两种NGS方法）对患者样品中变异基因位点的频率（variant allele frequency）进行检测，参照标准就应该沿黑色线条（回归线），分布在图中的阴影区域内。

除此之外，以首要分析目标为主的多变量统计分析也可以对参照标准和患者标本的关系进行分析。患者样品的检测结果通常都会如图b所示，聚集在一个范围内。具备可交换性的参照标准的结果也应该呈现出同样的趋势，而不具备可交换性的参照标准则会离散分布。





背景知识2：诊断统计学



灵敏度 = 阳性数 /（阳性数 + 假阴性）=99. 15%
特异性 = 阴性数 /（阴性数 + 假阳性）=99. 99%
准确性 =（阳性数 + 阴性数）/ 总数 =99. 99%
精确度 = 阳性数 /（阳性数 + 假阳性）=99. 98%
F_1=2 阳性数 /（2 阳性数 + 假阳性 + 假阴性）=99. 56%

* 未平衡组

参照标准能够为使用NGS技术平台的临床诊断测试提供一套真实的样本，来评价这些诊断技术的分类能力（classifier performance）。阳性（true-positive，TP，图中绿色所示）、阴性（true-negative ，TN，图中红色所示）、假阳性（false-positive，FP，图中蓝色所示）、假阴性（false-negative，FN，图中黄色所示）的预测值都可以通过与参照标准比对的方式计算出来。这些结果通常都会被绘制成下图，并计算出一系列的统计学指标，来评价、衡量NGS技术在多个不同方面的检测能力。

为了展现这些检测能力，我们可以参考使用全基因组测序方法（whole-genome sequencing, WGS）来发现人体基因组里的单核苷酸变异（single nucleotide variant, SNV）。通过对NA12878基因组DNA样品进行测序，并与不同已知基因型检测结果进行比较，我们可以评估这些诊断技术的真实检测能力。比如，我们就不太可能在无法检测到的人类基因组区域里发现明确的变异位点。使用目标基因或外显子组测序技术，还可以进一步将报告范围缩小到目标基因区域。报告流程（reporting procedure）也可以限定这个范围。比如，无法明确功能的沉默突变（silent mutation）就不会被报告，以致其也就不会被纳入可靠报告范围（validated reportable range），而且在评价诊断能力时也不会被考虑在内。

还应该指出的是，如果只与明确的基因型进行比对，则可能会高估诊断技术的诊断能力，这是由于很难在可靠报告范围以外的基因组区域里发现突变信息。这种情况也会忽略其它的突变，比如小插入或缺失突变和结构突变等。还需要指出的是，很多突变检测软件都更倾向于在已明确的多态性位点里发现突变。因此，新的突变或者罕见突变，可能都无法通过与HapMap参考图谱里的已知突变位点进行比对的方法来发现。

灵敏度是评价诊断技术最重要的一项指标。对于NGS技术而言，灵敏度往往与测序质量和深度有关。此时，灵敏度则是在NA12878参考基因组序列里发现已知突变的正确比例。假

阳性预测并不会影响这个指标。WGS检测技术可以非常容易地在所有位点发现变异，而且灵敏度也较高。因此，WGS检测也是一种特异性的诊断手段，并且只能够发现突变位点，这同时也提示我们，在基因组里没有哪些突变。由于突变情况在整个基因组中还是非常少的，所以能够排除阴性（true negative）的精确度（precision）指标（又名阳性预测值，positive predictive value），也是评价NGS技术的一项重要参数。在NGS检测中，阳性和阴性并不平衡。进行NGS检测时，在精确度与灵敏度之间，常常需要做一番平衡（背景知识3），此时可以用F_1指标来做判断，所谓F_1指标则是精确度与灵敏度的权重平均值。

背景知识3：测序诊断技术图解

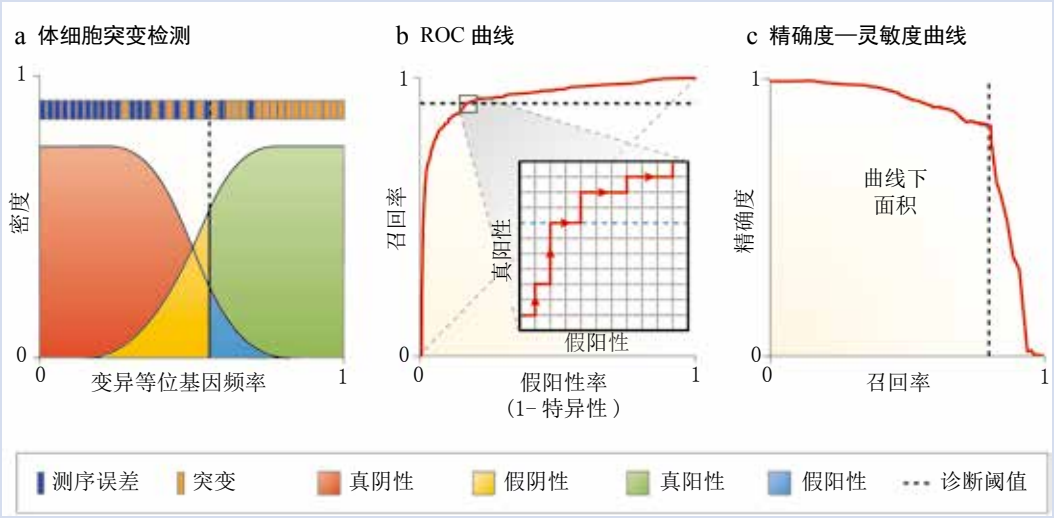
很多检测手段并不会直接给出一个清晰、明确的阳性或阴性结果，只会给出一个连续的定量范围，或者似然值（likelihood values）（图a）。因此，可以使用诊断阈值（图中虚线）区分出阳性范围和阴性范围，为临床诊断提供指导。

为了划定出最佳的诊断阈值，我们常常会用直观的图形来了解真阳性（true-positive）结果和假阳性（false-positive）结果之间的关系。接下来，我们以对癌症患者肿瘤标本中的低频次体细胞突变（low-frequency somatic mutation）进行测序检测为例，来说明这个问题。提高测序的覆盖范围，通常可以提高检测手段的灵敏度，因为这可以发现较低频率的突变等位基因。但是，这也会增加这些较低频率的突变等位基因测序结果的误差，以致很难将其与体细胞突变区分开。因此，应该设定一个等位基因频率阈值，便于尽可能地发现体细胞突变，同时又能够尽量鉴别出测序误差（图a）。

我们可以使用DNA spike-in标准品来帮助设定诊断阈值，这些混合的DNA spike-in就包括了已知的突变，将其与肿瘤标本混合，然后再构建文库，进行测序。这些DNA spike-in可以帮助区别真的体细胞突变，以及测序误差。

在设置恰当的诊断阈值时，绘制受试者工作特征（receiver operating characteristic, ROC）曲线颇有帮助。ROC曲线能够将所有潜在突变根据等位基因频次进行排序之后，展现出真阳性和假阳性的相对习得（relative acquisition）（图b）。曲线下面积（area under the curve, AUC）明确了ROC曲线里真阳性和假阳性率的关系，而且还可用其来比较不同诊断检测的诊断能力。对于不平衡的数据集，在解读ROC曲线时就应该更加谨慎，因为假阳性率任何一点微小的改变都会让预测错误的绝对值发生明显的改变。

在为NGS测试设定诊断阈值时，精确度——灵敏度曲线（precision-recall curve）更加有价值，因为这种检测的真阴性往往要比真阳性多得多（图c）。如图所示，当大量的测序误差没有排除时，精确度——灵敏度曲线明显地展示出等位基因频率。在比较不同的测试，或者不同的筛选策略时，精确度——灵敏度曲线的曲线下面积也是一个有效的定量指标。



背景知识4：准确性定量及回归分析

NGS测序技术能够对一个样品里的DNA或RNA进行序列分析，这已被常规用于基因表达检测、等位基因频率检测和微生物丰度检测等多个领域，而且其中每一种定量检测结果都可以被参照标准校正和评定。

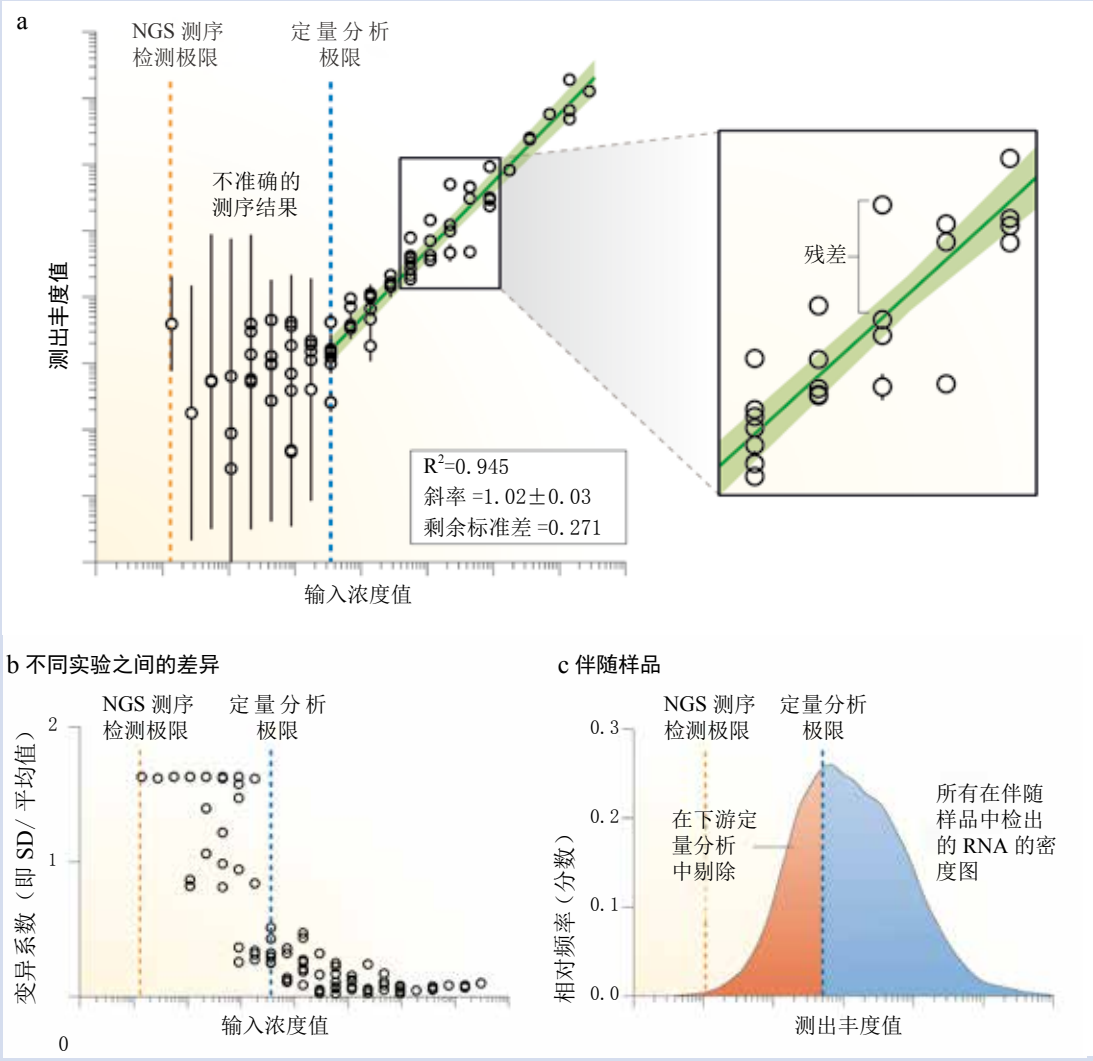
以通过RNA测序对人体RNA标本里的数千种不同基因的表达情况进行检测为例，一式三份将RNA spike-in混合物加入待测样品中，然后构建测序文库，进行测序和数据分析。

将y轴上每一个spike-in的检出丰度，与x轴上已知的spike-in加样浓度相比，就可以对RNA测序结果进行定量分析（图a）。那些无法检出的、丰度很低的spike-in则是该测序实验的检测极限。当然，也可以再设置一条阈值，如图中的蓝色虚线所示，在该阈值以下的基因表达检测结果就被认为不可靠。如果RNA的丰度很低，重复之后就会有一个很高的标准差（standard deviation, SD）和变异系数（coefficient of variation）。

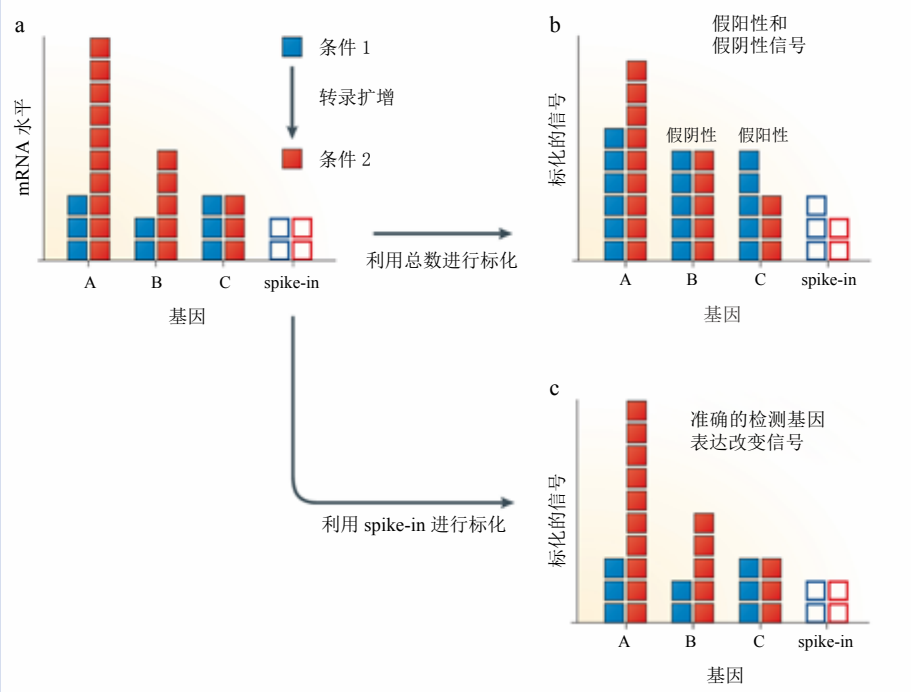
图中绿色曲线表示的是线性回归分析结果，可以代表不同浓度的RNA spike-in的加样浓度与检出丰度结果之间的关系。这也是最小二乘回归法（least-squares regression）中最常用的分析方法。如图a右侧的放大图所示，图中的每一个残差数据点都表示预测值与实际检测值之间的差，这可以表示在整个检测范围内，是否都存在误差。

我们用判定系数（coefficient of determination, R^2）来评价线性关系的强弱。所谓R^2就是独立变量（检出丰度）与根据独立变量（初始加样浓度）预测出的估计值之间的比值。该比值可以评价回归曲线与数据的重合程度，以及RNA测序实验堆spike-in的真实、准确的检出能力。

绘制了回归曲线之后，就可以用来推算每一个待测基因的表达丰度。这时也需要设定一个阈值，比如定量极限等，来排除那些不可靠的数据（图c）。


a
NGS 测序
检测极限
定量分析
极限
不准确的
测序结果
测出丰度值
残差
R^2=0.945
斜率 =1.02±0.03
剩余标准差 =0.271
输入浓度值
b 不同实验之间的差异
NGS 测序
检测极限
定量分析
极限
变异系数（即 SD/ 平均值）
2
1
0
输入浓度值
c 伴随样品
NGS 测序
检测极限
定量分析
极限
0.3
0.2
0.1
0.0
在下游定
量分析
中剔除
所有在伴随
样品中检出
的 RNA 的密
度图
相对频率（分数）
测出丰度值

背景知识5：spike-in对照的标化



NGS数据标化的目的是在进行数据分析时，在保留生物差异的前提下，尽量减少不同样品之间因为技术原因导致的差异，spike-in就是一个很好的工具。

去除非目标变异（remove unwanted variation, RUV）等方法可以通过对一直保留在样品里的spike-in对照进行因子分析（factor analysis），调整技术的影响作用。这些方法假设技术效应对spike-in和样品都有同样的影响，因此，任何样品间spike-in上的差异都是由实验因素导致的，而且也是可以做相应消除的。

在RNA测序工作中，标化的意义就显得尤其重要，这是因为需要对不同细胞、不同条件、不同时间点的基因表达情况进行精确的比较。为了说明这一点，我们以图a中两个不同条件下的基因表达差异分析为例。这两种不同的条件在文库规模上至少应该进行标化，以确保不会因为测序深度的不同导致基因数量上的差异（图b）。

可是，有一些基因在一种条件下会大量表达，但是在另外一种条件下则不会。仅仅对文库规模进行标化，会使非差异表达的基因呈现出表达下调的假象，而高表达基因之间的差异也会缩小（图b）。为了解决这个问题，大多数分析工作都假设大部分基因不是差异表达的，也会根据平均基因表达水平之间的差异，给出一个估计的换算系数（scaling factor）。

可是当整体基因表达都有差异时，这种假设就不成立了，比如肿瘤细胞的基因就会大量表达。在这种情况下，就需要再加入绝对量的RNA spike-in（比如相对于样品中细胞数量足够多的spike-in）使其作为标尺，衡量总体基因表达水平的改变情况（图c）。

虽然在RNA测序工作中用spike-in来标化的情况非常普遍，但这些spike-in也可以在其他不同的NGS应用领域里作为换算系数。比如比较不同基因组拷贝数变异，或者对环境样品里的微生物群落进行标化等。

名词解释：

参照标准（Reference standards）

指使用已知的（比如已知的基因型）材料作为实验对照来开展的测序实验。这样才可以对测序数据进行更好的评估和解读。

可交换性（Commutability）

指在包含多个检测过程的实验中，参照标准能够对患者的样品进行比较的能力。

基质效应（Matrix effect）

指由样品中的任何一种组分，而不是目标分析组分带来的影响作用，这种效应会影响参照标准的可交换性。

变异等位基因频率（Variant allele frequency）

指在肿瘤活检组织等样品中，目标变异等位基因的比例。

精确度（Precision）

又名阳性预测值（positive predictive value），指检测结果中真实阳性的比例。

灵敏度（Sensitivity）

又名召回率（recall），指检测结果中正确预测占所有正确结果的比例。

系统性测序误差（Systematic sequencing error）

指由样品制备、测序流程等因素导致的测序中的非随机误差.

NA12878

指经过详细鉴定的、健康远古女性的基因组序列，常被用来作为基因组分析的参照物。

长片段测序（Long-read sequencing）

指能够对长度达到数kb的核酸片段进行测序的技术，该技术可发现大规模的结构变异。

分型（Phasing）

指明确DNA变异源自哪个染色体的过程。

人工微生物群落（Mock microbial community）

指将多种人工培养的微生物按照已知的浓度进行混合而形成的微生物群落，可对其进行基因组DNA测序，获得参考序列。

Spike-in对照（Spike-in control）

指将已知长度、序列、及丰度的DNA或RNA分子，直接加入待测样品中作为对照，起到定量或定性的作用。

检测极限（Limit of detection）

指检测技术能够检测到的最低浓度。

标化（Normalization）

指对不同样品之间的技术偏倚进行校正，便于对其进行精确比对。

报告范围（Reportable range）

指基因组内，使用NGS测序技术能够获得符合标准的测序数据的一个区域。

参考区间（Reference interval）

指与患者来源相同，但是非病变标本测序得到的变异信息集合。

能力测试（Proficiency testing）

指给待检实验室提供参考样品，待其获得检测结果之后，将结果递交给独立的组织，进行评估。在欧洲，这一流程也被称作外部质量评价（external quality assessment）。

二、Spike-in对照——全基因组分析中被忽视的重要一员

在全基因组范围内对基因表达、转录调控、组蛋白修饰、基因拷贝数变异和核小体定位等进行研究，已经是当代很多实验室的标准研究流程了，这一整套分析也会获得大量与人体健康和疾病有关的数据和信息。不过，其中绝大部分研究都忽视了需要引入spike-in对照（spike-in control）才能开展的标准化问题（normalization）。下文将介绍为什么使用spike-in对照对于全基因组研究如此重要，以及如何设计和使用spike-in对照。我们也会介绍一些方法，帮助大家更好地认识那些因为没有设置spike-in对照，而进行了错误解读的数据。

所有使用芯片（microarray）或NGS开展的研究，都会对两个或多个不同实验条件下的结果进行比较和分析，但是这类研究的前提假设却常常出错。这类假设通常都认为，他们实验所得到的结果，对于样品里的每一个细胞都是一样的。因此，科研人员通常都用同等量的DNA或RNA来进行实验，并且对整体实验结果进行标化。比如，在测序实验中使用每百万reads中来自于某基因每千碱基长度的reads数（reads per million, RPM）来进行校准，而在芯片实验中则使用分位标化法（quantile normalization）进行校准。不过，当不同实验条件下的细胞并没有同等量的DNA或RNA时，这一假设前提就不成立了。而且，在常规的新一代测序数据分析流程里，对于每一个样品都采用同样的数据标化操作，即对所有的测序结果（read）均进行标化。但这只有在基因组的总体增量和总体减量相同的时候才成立，可是这种情况是非常少见的。因此，在基因组研究工作中一直使用的传统标化方法其实都是错误的。为了解决这个问题，就需要在实验里加入spike-in对照，对数据进行更好的标化和更精确的解读，有助于了解不同样品的同一基因组范围内，信号究竟是增加了，还是减少了。

1. 何时设置spike-in对照

只要是使用芯片技术，或测序技术来了解在不同的实验条件下，全基因组范围内信号的绝对变化情况，就应该设置spike-in对照。这些信号包括RNA、DNA、通过微球菌核酸酶（micrococcal nuclease, MNase）降解技术研究的核小体、通过染色质免疫沉淀技术（chromatin immunoprecipitation, ChIP）研究的组蛋白修饰或转录因子等因子与染色体的结合情况等。当总体信号发生改变时，最容易凸显spike-in对照的重要性了（图1a），这种改变在全基因组范围内的所有基因位点上都具有一致性。然而，由于在基因组中部位点发生的局部信号改变也可以导致全局信号改变，所以此时也应该设置spike-in对照。当基因组中某些区域的信号有明显的增加，但是其它区域的信号又没有减少时，对整体信号进行标化则会人为地降低其它区域的信号（图1b）。在研究DNA拷贝数时情况也一样，spike-in对照对于分析基因组中的重复区域也非常重要。在分析DNA甲基化修饰时也需要设置spike-in对照，以帮助我们发现DNA拷贝数有变异区域的甲基化CpG岛的数量（图1c）。综上所述，在进行各种全基因组研究时，都应该设置spike-in对照。

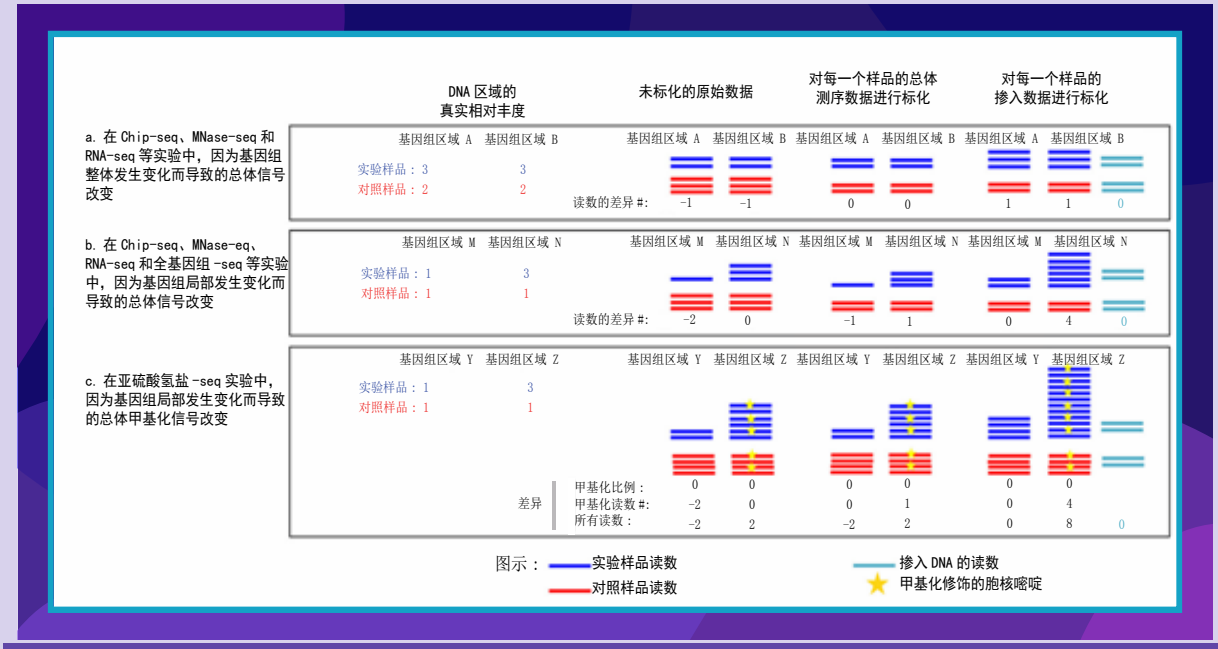


图1 介绍了为何在测序实验中设置spike-in对照会得到更准确的不同样品间的比对结果。图中以基因组中的某一个区域为例进行了介绍。a，当全基因组内都发生了同样程度的改变时，对整体测序数据进行

标化则会隐藏这些变化，然而对spike-in读数进行标化，则可以揭示这些改变。b，当基因组内某个区域里的信号增加时，对不同样品的整体信号进行标化则会引入人为的假象，以致误以为是基因内其它区域的信号降低了，从而得出其它区域里基因表达下调的错误结论。如果使用spike-in对照，就可以纠正这种错误。c，在研究甲基化DNA拷贝数变异问题时，只有设置spike-in对照，才可以得出准确的结果，虽然确定甲基化比例并不需要spike-in对照。

2．在微球菌核酸酶测序（MNase-seq）中设置spike-in对照

我们最近发现，衰老的酵母菌细胞内的正常核心组蛋白（core histone protein）的量会减少一半，而这种核心组蛋白就是构成染色质的关键物质。研究发现，这种核蛋白的减少就是导致细胞衰老的原因。可是，按照标准的操作流程，通过微球菌核酸酶测序技术对细胞总体核小体定位情况进行分析后发现，核小体在基因组的定位并没有发生改变（图2a）。这一奇怪的研究结果让我们对实验方法产生了质疑，这一实验方法就是，对不同实验条件下的样品，在不设置标化对照的情况下，都选取同样的DNA进行分析。下文将详细介绍spike-in对照或标化对照的设计和使用方法。在研究中设置了spike-in对照（即每个细胞都加入同等量的spike-in对照）之后，再进行文库构建、数据标化，结果发现衰老细胞内的核小体的确减少了50%（图2b）。这一发现给了我们很大的提示，让我们意识到在今后的工作中，在研究衰老引起的细胞组蛋白整体减少的情况下的胞转录、组蛋白修饰和基因组稳定性等问题时，一定要设置spike-in对照。将来，我们还需要在设置spike-in对照标化的情况下，对以下关键问题重新研究一番。比如，肿瘤细胞内的高转录和基因组失稳（genomic instability）是否也是因为核小体减少而导致的呢？

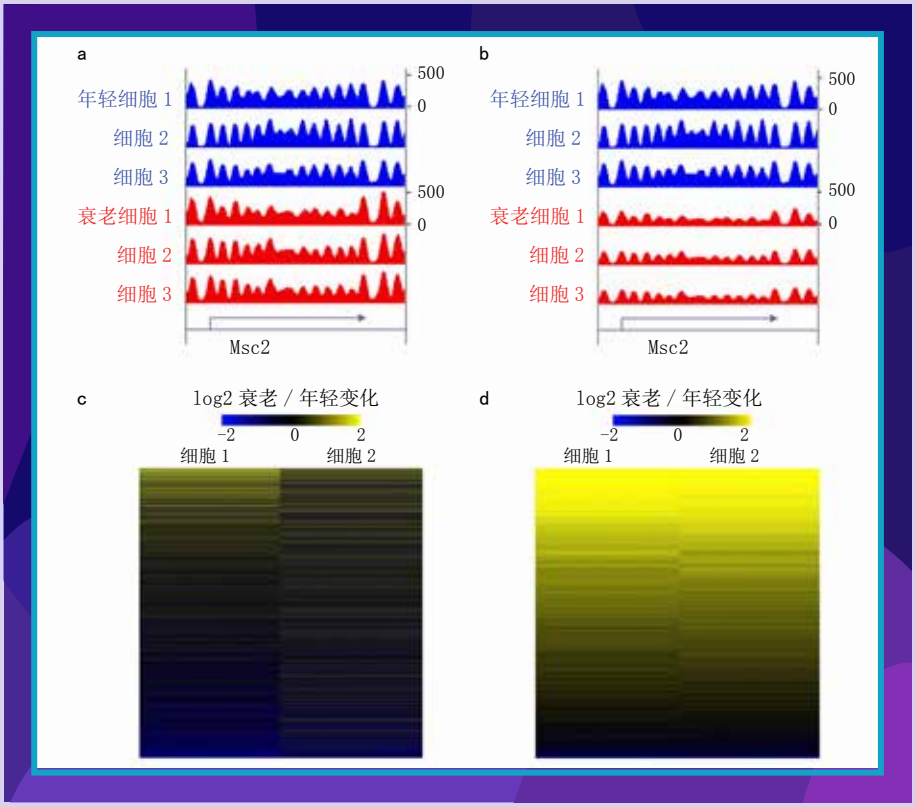


图2 spike-in对照的作用。a，基因组跟踪（genome track）的一个瞬间结果，显示出年轻细胞和衰老细胞内核小体的占位信息。此时使用的是MNase-seq技术，并没有引入spike-in对照。b，在设置了spike-in标化的情况下，基因组里同一区域的实验结果。c，热图（heat map）显示的年轻细胞和衰老细胞内的基因表达情况，此时使用的是RNA-seq技术，并没有引入spike-in对照。d，热图显示的年轻细胞和衰老细胞内的基因表达情况，此时使用的是RNA-seq技术，设置了spike-in对照。此处使用的是全局标化方法，这也是比较不同实验条件下基因组整体变化情况的最理想标化方法。

3. 在RNA-seq中设置spike-in对照

在MNase-seq实验中因为设置spike-in对照而有了重大发现之后，我们又对后续的一系列高通量RNA测序实验进行了改进，也通过设置spike-in对照来研究复制衰老（replicative aging）过程中的转录情况。我们对其他研究者报道过的与细胞衰老有关的部分转录研究成果进行了重复，因为我们认为，细胞总体核小体的减少肯定会影响整体的转录水平。对设置了spike-in对照和未设置spike-in对照的RNA测序结果的解读果然不一样，而且差别还非常大（图2c、d）。通过设置spike-in对照，对RNA测序数据进行了标化之后获得的重复研究结果显示，在衰老的酵母细胞内，超过6,000个基因的转录水平因为核小体缺乏而发生了改变（上调）。这与我们重复的、之前没有设置标化对照的实验结果大相径庭。之前的研究认为，大部分基因的转录水平没有发生改变，只有几百个基因的转录水平有所上调，还有几百个基因的转录是下调的。这也让我们想起了之前对组蛋白缺乏条件下基因表达水平的研究。那些研究认为，染色质不会影响大多数基因的表达，只有几百个基因可能会受到染色质的影响。这些在错误实验方法（未设置合适对照）下得出的错误研究结论会将后续的研究工作引到错误的方向上去。

就在我们开展细胞衰老研究的同时，美国麻省理工学院（Massachusetts Institute of Technology, MIT）的Rick Young课题组也独立发现了spike-in对照对于RNA测序实验的重要性。实际上，他们也因此发现了*cMyc*癌基因其实是一个全基因组延伸因子（elongation factor）。当该基因过表达时，基因组内所有基因的转录水平都会上调。可是过去认为，该基因只是某些特定基因的转录活化因子，只能上调这些基因的表达。这一研究成果也再一次展现出设置spike-in对照的重要性。这提示我们，在不同实验条件下，比如细胞衰老或*cMyc*基因过表达，细胞整体的RNA转录水平是会发生变化的。由此认识到，其它生物研究中肯定也存在同样的问题，所以我们建议所有从事相关研究的科研人员都应该在实验时引入spike-in对照，也应该在设置spike-in对照的情况下对之前做过的实验再做一番验证，及时纠正可能存在的错误。

4. 在ChIP-Seq中设置spike-in对照

此外，我们也强烈推荐在使用ChIP-Seq技术对全基因组范围内的组蛋白修饰，以及各种与基因组结合的因子进行研究时设置spike-in对照。我们在对衰老细胞里DNA损伤反应标志物——H2A组蛋白上第129位丝氨酸的磷酸化（即γH2A）情况进行研究时，就引入了

spike-in对照。结果发现，在衰老的细胞里，这种组蛋白修饰的水平提高了大约3倍，可是在没有设置spike-in标化对照时，得到的结果完全错了。我们在衰老细胞中观察到的这些γH2A的变化非常有意义，观察显示，这种变化并非是全局性的变化，在衰老细胞的基因组中，其实绝大部分区域里的γH2A水平都没有发生改变，但是在核糖体DNA和随着细胞的衰老而转移至胞核里的线粒体DNA里的几个重复区域里，这些γH2A的水平都有明显上升。如果我们对整体测序读数结果进行标化，则会认为在衰老细胞的全基因组范围内，γH2A的水平是下调的，只不过在核糖体DNA和线粒体DNA中的部分区域里，有轻微上调。但是这个结论其实是错误的，也会给其他人带来误导。按照ChIP-seq实验操作规程里的标准化步骤，我们可以纠正蛋白质或翻译后修饰物过多或过少的情况，但这只有在不同实验条件下的染色质样品中特定的翻译后修饰或蛋白质是完全一样时才成立。当不同实验条件下样品里的整体翻译后修饰水平不同时，再按照这种方法进行标化，则会得出组蛋白修饰水平下调这一结论，然而实际上这些组蛋白的修饰水平并没有改变，或者上调。传统的ChIP实验都没有设置spike-in对照，这是因为这些实验都已经根据加入的DNA对每一个DNA序列校正过了。可是，在ChIP-seq实验中，无关的IgG或只有较低测序深度的插入序列并不足以起到标化对照的作用，因为实验结果无法真实地反映ChIP目标的整体变化情况，这点与我们之前介绍过的，对衰老细胞组蛋白水平的研究是一样的。

最近，外源性的表观基因组（epigenome）也被用来当作spike-in对照，来帮助我们了解人体基因组内阻断靶修饰的整体情况。科研人员在实验细胞中添加了黑腹果蝇的染色质，以便对不同ChIP-seq实验的结果进行比对。为了验证这个方法是否可行，他们选择了DOT1L酶的抑制剂EPZ5676进行了实验。DOT1L酶能够使H3组蛋白上第79位赖氨酸（H3 K79me2）发生二甲基化修饰。为了验证EPZ5676的效果，科研人员用这种物质先处理了人体细胞，然后与对照细胞进行比对。在使用传统RPM标化方法时，人们并没有发现实验组细胞与对照组细胞之间有特别大的区别，只有在某些位点上，或者H3 K79me2的metagene图谱里能发现一些细微的不同。但是Western blotting实验已经明确显示出，在经过EPZ5676处理的细胞中，H3 K79me2的整体水平会明显下调。可是引入了黑腹果蝇对照之后则发现，在经过EPZ5676处理的细胞中，H3 K79me2的整体水平出现了明显下调。因为这些实验人员已经知道了黑腹果蝇细胞与实验细胞的比例，所以使用这种方法也可以对ChIP实验中前面几个步骤进行标化。但是这个方法有一个问题就是严重依赖抗体来识别待研究的目标组蛋白修饰。

5. 在gDNA-seq中设置spike-in对照

我们也在年轻细胞和衰老细胞的基因组测序工作中设置了spike-in对照，以发现在线粒体老化过程中，15%的人会出现的细胞内12号染色体扩增1/3的现象。在准确判断染色体倍数（ploidy），以及了解肿瘤细胞基因组中的扩增与缺失区域时，设置spike-in对照非常重要。

6. spike-in对照究竟是什么？

在各种实验中设置的spike-in对照都由好几段不同的DNA序列组成，这些序列不能是目标研究物种里含有的序列，但是其GC含量需和目标研究物种的基因组GC含量一致。在高通量测序实验中，spike-in对照的DNA片段长度应该与测序文库中待测DNA片段长度相差不多。在MNase-seq实验中，spike-in对照的DNA片段长度通常都在150bp左右，这也是一个核小体里包含的DNA长度。这些spike-in对照也都可以应用于ChIP-seq实验和gDNA-seq实验，前提是将研究样品也裂解成150bp大小。我们在实验中使用的spike-in对照都是通过PCR方法，以细菌质粒序列为模板扩增得来的。这些spike-in对照序列和酿酒酵母（*Saccharomyces cerevisiae*）基因组的同源性都很低，GC含量则和酿酒酵母基因组的GC含量相差不多。当然，我们也可以人工合成DNA片段。准确定量之后，就可以按照一定的分子比例，比如1：2：4等比例，将其进行混合。采用哪种比例混合均可，但是一定要确保在数据分析阶段，具有线性扩增。

我们的RNA-seq分析工作使用了Life Technologies公司提供的商业化RNA对照物ERCC（external RNA control consortium）。ERCC的4期测试里集中了96种来自ERCC质粒参考文库不同的转录产物。这些转录产物上还都添加了poly(A)$^+$尾巴，来模拟DNA模板的情况。在转录产物中，79种GC含量在31~51%之间，17种GC含量只有5%。这些转录产物的长度则介于273~2,022bp之间。这些转录产物的浓度则根据拉丁方设计（Latin square design）跨越了数百万倍。因此，这种ERCC参照物体系就可以对不同GC含量、不同片段长度和不同丰度的样品进行标化，适合用于新一代测序平台，对从酵母至人类的绝大多数真核生物进行测序研究。这些ERCC RNA序列和真核生物序列的同源性也非常低，与黑腹果蝇的匹配率只有0.5%，而与人的匹配率更是低至0.01%。这种ERCC最初开发出来原本是用于评估技术平台敏感性和扩增线性等指标的，但是后来也被用来比较不同实验条件下的RNA水平。在文库构建过程中，ERCC spike-in对照的富集效率则取决于RNA纯化的程度。比如，使用poly(A)纯化法富集ERCC的效率就不如RiboZero法。可是，只要使用相同的RNA富集方法，这就不是问题了。在使用了ERCC对照的情况下，RNA测序数据具备了极高的一致性，这一点已经通过对同一RNA样品进行的多次文库制备实验得到了证明，也在对不同poly(A)选择方法的比对中得到了证明。其他实验室也观察到，在文库制备过程中，spike-in对照的扩增存在技术差异，但是借助“非目标差异去除标化法（Remove Unwanted Variation, RUV）”则可以排除这种技术偏倚。

另外一种spike-in就是外源性基因组或表观基因组。如前所述，最近已经有研究证实，使用黑腹果蝇的表观基因组作为标化对照，已经成功地在ChIP-seq实验中对人体细胞的H3K79me2情况进行了分析。

7. 如何使用spike-in对照

在你的实验里，你必须清楚你从每一个实验条件下的样品中分离出了多少细胞。最理想的情况是，从每一个样品中都分离出同等量的细胞。如果无法进行细胞计数（比如样品是人体组织），那么可以在实验开始前用DNA总量对样品进行校正。然而，首先需明确，不同实验条件下的染色体倍数或基因组稳定性不能有差别；然后在等量的样品中加入同样的spike-in对照，开始后续实验。

样品的随意性、文库的复杂性，以及测序深度等因素均可以限制RNA测序的实际应用效果。在一个使用了ERCC spike-in的测序实验中，有5种对照片段没有被检测到，主要原因就是这些片断的浓度太低，而在整个1,000万个读数中，这5种spike-in片段中每一种的理论读数才只有0.6~2.5个。因此，为了检测到所有的spike-in片段，则需要提高spike-in的浓度，让每一种spike-in片段最终的读数至少都能达到2.5。当然，spike-in对照也不能加得太多，因为那样一来，最终得到的就会是大量的spike-in读数，从而掩盖了待测样品的数据。样品与spike-in对照之间最合适的比例应该是1,000：1~50：1，即最终的读数中，spike-in片段的读数大约占到0.1~0.2%。

8. 如何利用spike-in对照进行数据标化

在6个数量级范围内，测序读数的数量与ERCC spike-in浓度之间具有很好的线性关系（Pearson’s r>0.96），这说明可以用读数来评估RNA的丰度。不同文库spike-in读数也具有同样的线性关系（Pearson’s r>0.98），这说明spike-in的读数不会受到内源性RNA复杂程度的影响。在ChIP-seq等实验中使用定制的spike-in对照时，对spike-in对照的读数进行分析，可以确保实验是按照各spike-in片段之间的比例等比例扩增的。比如，如果spike-in中各片段的比例是1：2：4，那么最终这些片段读数的比例也应该是100：200：400或300：600：1200。

为了在这些spike-in对照的基础上分析数据，我们首先需要根据spike-in对照的浓度，对spike-in对照的读数进行标化（图3，步骤1）。比如在一个简化的模型里，当读数与样品浓度之间存在线性关系时，就可以根据总体的线性曲线来进行标化。通过将spike-in标化的读数与原始读数进行对比，就可以获得每一个实验的标化功能，例如实验条件1的n=1.5 × r，实验条件2的n=0.5 × r。此处的n代表标化的读数，而r代表原始读数。然后，你可以使用这些公式将标化的读数意义对应到整个基因组里的每一个核酸上（图3，步骤3）。

在更加复杂的情况下，人们可能会需要

好几十种不同的spike-in片段，以及更加精密的标准化公式，来纠正测序中产生的误差。比如，高浓度的片段可能就比低浓度的片段更容易扩增，读数与样品浓度之间的关系可能也是非线性的。有一种解决方案就是进行分位数标化（quantile normalization），然后再利用非线性回归模型（nonlinear regression model）来模拟每一种样品的标准化功能。基因组中的每一个基因的读数，或每一个碱基对的读数都可以根据回归模型来进行分析。另外一个潜在的问题就是spike-in对文库构建的影响。我们已经观察到，在不同实验室构建的文库中，ERCC读数的比例是不一样的，可是同一文库不同测序循环中的读数的比例却是恒定不变的。利用RUV标化方法可以有效地去除这种技术偏倚。这种标化方法使用了一组负向调控基因，假定这些基因的表达和spike-in对照一样，是不会变化的，但是它也估计了所有基因的非目标因子。 就其本身而言，RUV方法使用的假设与回归分析或全局化定位等标化方法的假设（此时的假设认为，所有技术上的影响对spike-in对照和待测目标都是一样的）均不同，但是这一假设却更常见。在这种情况下，RUV就可以纠正spike-in扩增过程中的技术误差，这对于研究特定基因的表达变异非常有价值。可是，RUV方法并不适合用于研究全基因组都可能发生改变的情况，因为这些改变会被RUV标化过程消除掉。在这种情况下，使用spike-in标化对整体信号进行标化就更有价值。因此，我们建议使用spike-in对照进行标化时应与RUV等其它标化方法相结合。

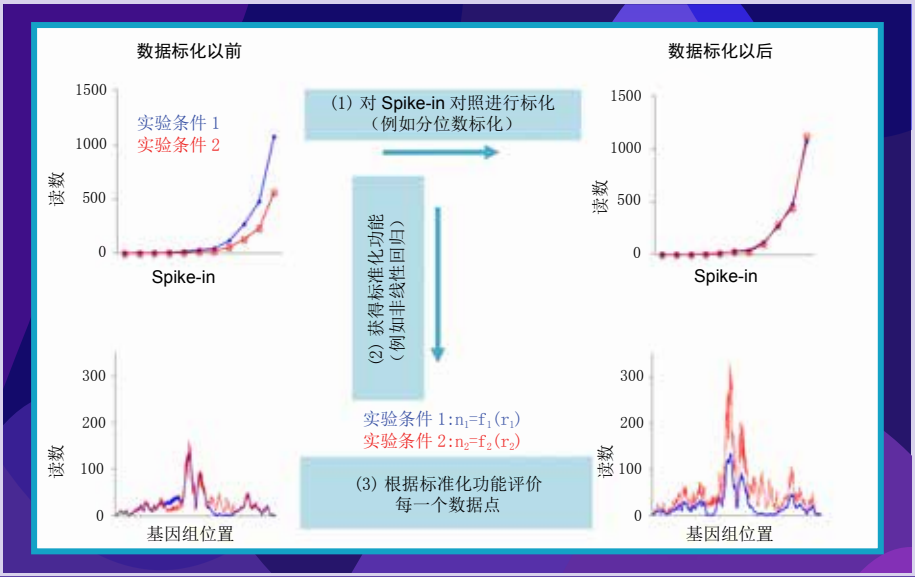


图3 使用spike-in对照对测序数据进行标准化处理。步骤1，对不同实验条件下的每一个spike-in（左上图及右上图x轴所示）的原始读数（左上图y轴所示）都需要进行标化。步骤2，对标化之后的读数（右上图y轴所示）与原始的读数进行比对，就可以得到一个每一种实验条件下的标化功能（见中图所示）。在步骤3中使用这些功能，对每一个基因组位置上的标化读数进行标化（见下图所示）。这就是在spike-in对照的帮助下，对全基因组进行标化的操作流程，也是比较不同实验条件下全基因组变化的最理想标化流程。

9. 数据是否可以进行回顾性标化？

近几十年来，我们已经对数百万个不同的样品进行了全基因组分析，获得了大量的数据，但是这些实验都没有设置合适的对照。那么这些数据是否还有价值呢？还是应该彻底抛弃它们？我们建议，对这些数据进行抢救性再分析。比如，对于基因表达分析，首先需要确定在所有细胞中都稳定表达的标志基因（marker gene），然后将其作为spike-in对照，对其它基因的表达数据进行标化。另外，也可以对相关细胞系的标志基因的转录产物进行定量逆转录PCR（quantitative reverse transcription-PCR, RT-PCR）分析，还可以判断基因表达水平的差异，然后再根据这些数据对其它基因表达水平进行标化。如果基因组里还存在其它类似于标志基因的、可以定量的因子或组蛋白修饰情况，也可以将它们作为spike-in对照，对ChIP-seq实验的数据进行重新标化。

原文检索：

1. Simon A. Hardwick Ira W. Deveson & Tim R. Mercer. (2017) Reference standards for next-generation sequencing.NATURE REVIEWS | *GENETICS,* 18:473-484.

2. Kaifu Chen, Zheng Hu, Zheng Xia, Dongyu Zhao,Wei Li, Jessica K. Tyler. (2016) The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Molecular and Cellular Biology,* 36(5): 662-667.

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：

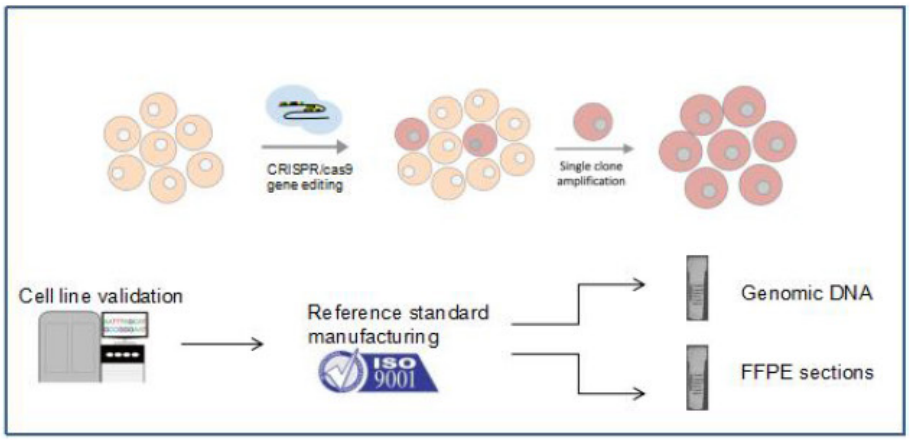
1.具备基因组学、蛋白组学、生物信息学、细胞生物学等生命科学学科背景；
2.具备良好的生命科学前沿触觉；
3.具备较高的外文文献翻译、编译水平；
4.具备较强的选题策划、资料搜集、组织能力，以及专业稿件撰写能力；
5.具有高级职称；或者拥有（正在攻读）该领域的最高学位。

有意者请将个人简历发送至 editor@lifeomics.com

GeneCopoeia™
Expressway to Discovery

OncoSpot™ NGS标准品

GeneCopoeia 可以提供高特征、生物学相关的纯合子 OncoSpot™ NGS 标准品，与多种基因组平台兼容，例如 NGS、Sanger 测序和 qPCR，含有基因组 DNA 和福尔马林固定石蜡包埋（FFPE）两种产品形式。



产品应用

可作为确定工作流程标准的理想方法，包括：

- DNA 提取
- NGS 文库构建
- 样品制备
- 测序

产品优势

- 在检测病人样本前优化 DNA 提取方案
- 建立检测限
- 在质量控制过程中准确地评估批间可变

GeneCopoeia™
Expressway to Discovery

gDNA 产品列表

货号	规格	产品
RM001	5 μg	野生型 HCT116 标准品
RM002	1 μg	EGFR ΔE746-A750 标准品
RM003	1 μg	EGFR L858R 标准品
RM004	1 μg	EGFR T790M 标准品
RM005	1 μg	EGFR T790M, C797S 标准品
RM007	1 μg	BRAF V600E 标准品
RM008	1 μg	cKIT D816V 标准品
RM009	1 μg	AKT E17K 标准品
RM010	5 μg	野生型 RKO 标准品
RM012	5 μg	野生型 NCI-H3122 标准品
RM013	1 μg	EGFR C797S 标准品
RM014	1 μg	KRAS G13D 标准品
RM015	1 μg	KRAS G12D 标准品
RM016	1 μg	EGFR L858R 标准品
RM017	1 μg	KRAS G12C 标准品
RM018	1 μg	NRAS G12D 标准品
RM019	1 μg	EGFR G719S 标准品
RM020	1 μg	EGFR S768I 标准品
RM021	1 μg	EGFR L861Q 标准品
RM022	1 μg	EGFR V769-D770insASV 标准品
RM023	1 μg	KRAS G13C 标准品
RM024	1 μg	PIK3CA E545K 标准品
RM025	1 μg	NRAS Q61K 标准品

FFPE 产品列表

货号	规格	产品
RM201	1 片	野生型 HCT116 标准品（FFPE）
RM202	1 片	EGFR ΔE746-A750 标准品（FFPE）
RM203	1 片	EGFR L858R 标准品（FFPE）
RM204	1 片	EGFR T790M 标准品（FFPE）
RM205	1 片	EGFR T790M, C797S 标准品（FFPE）
RM206	1 片	KRAS G13C 标准品（FFPE）
RM207	1 片	BRAF V600E 标准品（FFPE）
RM208	1 片	cKIT D816V 标准品（FFPE）
RM209	1 片	AKT E17K 标准品（FFPE）
RM210	1 片	EGFR C797S 标准品（FFPE）
RM211	1 片	KRAS G12D 标准品（FFPE）
RM212	1 片	EGFR G719S 标准品（FFPE）
RM213	1 片	KRAS G12C 标准品（FFPE）
RM214	1 片	EGFR L858R 标准品（FFPE）
RM215	1 片	KRAS G13D 标准品（FFPE）
RM216	1 片	KRAS G12D 标准品（FFPE）
RM217	1 片	KRAS S768I 标准品（FFPE）
RM218	1 片	KRAS L861Q 标准品（FFPE）
RM219	1 片	EGFR V769_D770insASV 标准品(FFPE)

*备注：产品列表逐渐更新中，如果以上产品没有您感兴趣的基因或突变，请与我们客服联系。


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