PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024,51(4):797~808
www.pibb.ac.cn

嵌合RNA检测和验证技术*

王广富[1,2,3,4)] 丁咏伟[6)] 唐 悦[1)**] 秦付军[1,2,3,4,5)**]

([1)] 郑州大学基础医学院微生物学与免疫学系，郑州 450001；[2)] 郑州大学医学科学院医学表观遗传学中心，郑州 450052；
[3)] 郑州大学医学科学院分子病理中心，郑州 450052；[4)] 郑州大学医学科学院转化医学平台，郑州 450052；
[5)] 郑州大学省部共建食管癌防治国家重点实验室，郑州 450052；[6)] 郑州大学基础医学院病理与病理生理学系，郑州 450001）

摘要 嵌合RNA（chimeric RNA）是由来自不同基因的外显子片段组成的融合转录本。传统的嵌合RNA检测方法有染色体核型分析、荧光原位杂交（FISH）等，但这些技术的特异性、灵敏性和准确性较差。随着测序技术的发展，二代测序技术展现出强大的数据处理能力，可以通过高通量序列分析来检测嵌合RNA，目前基于高通量测序的检测方法有FusionCatcher、SOAPfuse、EricScript等。目前较为常用的对检测到的嵌合RNA的验证方法有聚合酶链反应（PCR）、核糖核酸酶保护实验（RPA）、琼脂糖凝胶电泳、Sanger测序等。多种检测技术的开发使得越来越多的嵌合RNA被发现，但现有的检测技术各有优劣，主要体现于检测成本、假阳性率、检测时间等方面的差异。本文对嵌合RNA的检测方法、验证方法及各方法的优劣性进行阐述。

关键词 嵌合RNA，检测技术，验证技术，RNA-Seq，生物信息学
中图分类号 R34，R-1 **DOI**：10.16476/j.pibb.2023.0234

20世纪60年代，费城染色体（Ph）首次在慢性骨髓性白血病（chronic myelocytic leukemia，CML）中被发现[1]，根据其融合基因*BCR-ABL*而设计的药物格列卫（imatinib），是第一个用于治疗融合基因所引起癌症的靶向药物，同时也开创了融合基因靶向治疗癌症的先河[2-3]。基因组的不稳定性和重排是癌症的重要标志之一[4]。缺失、插入、倒置或易位均可产生染色体内或染色体间的重排[5]，其中由两个基因连接在一起形成的新基因被称之为融合基因（fusion gene）[6]。融合基因可产生新的蛋白质产物，进而导致癌基因的激活、抑癌基因的失活等，影响癌症的进展[7-9]。

长期以来，嵌合RNA通常被认为是由于在DNA水平上染色体发生易位、缺失或染色体倒置所形成的产物[10]。但随着新一代测序技术以及生物信息技术的发展，新的研究发现，嵌合RNA的产生既可以来源于DNA水平的改变也可以来源于仅RNA转录水平的改变，即单独转录的基因mRNA转录本的反式剪接[11-12]与相邻基因间忽略的基因边界的顺式剪接[13-14]（图1）。那么嵌合RNA如何检测，检测后如何验证，又如何保证检测的嵌合RNA不是假阳性结果，是一个值得深思的问题。本文主要从嵌合RNA的检测、验证技术进行论述。

* 国家自然科学基金（81972421），国家自然科学基金（NSFC）-河南联合项目（U2004135），郑州大学高层次人才启动项目（32340177），河南省高等学校大学生创新创业训练计划（202210459126，202210459161）和郑州大学教育教学改革研究与实验项目（2022ZZUJGXMLXS-017）资助。
** 通讯联系人。
唐悦 Tel：0371-67781950，E-mail：tangyue@zzu.edu.cn
秦付军 Tel：0371-66658776，E-mail：fujun_qin@zzu.edu.cn
收稿日期：2023-06-15，接受日期：2023-07-20

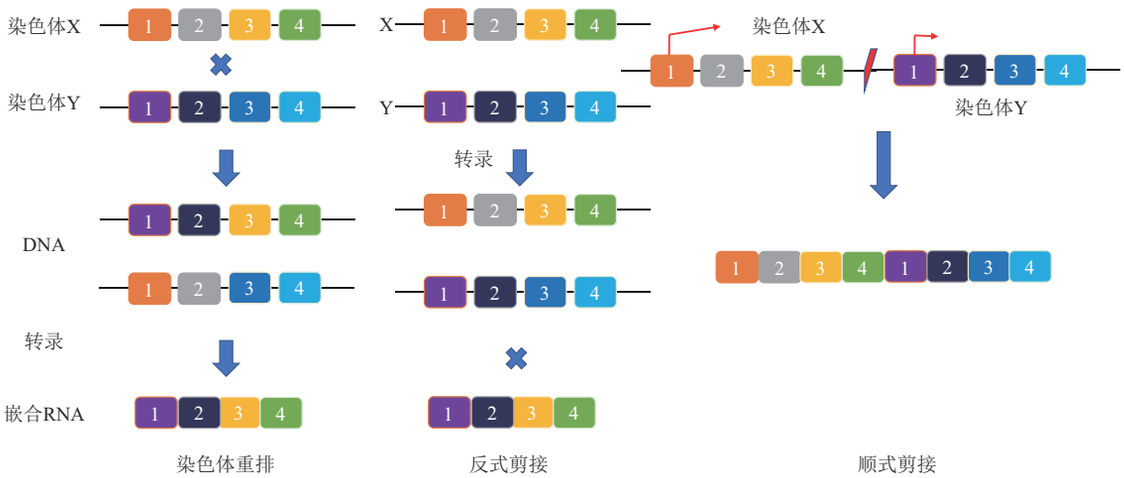


Fig. 1　Chimeric RNA formation mechanism

图1　嵌合RNA形成机制

1　染色体核型分析

核型分析是一种广泛的分析，提供了关于染色体数量和结构的额外信息。以分裂中期染色体为研究对象，根据染色体的长度、着丝点位置、长短臂比例、随体的有无等特征，并借助显带技术对染色体进行分析、比较、排序和编号，根据染色体结构和数目的变异情况来进行嵌合RNA的验证和疾病的诊断，如嵌合RNA *ETV6-NTRK3*，此方法适用于染色体异常形成的嵌合RNA[15]（表1）。

恶性血液病经常携带获得性染色体变化，产生具有致病和诊断意义的融合基因，如*NRIP1-MIR99AHG*嵌合RNA（表1）。但此方法需从新鲜组织样本中成功培养细胞，时间周期长，分辨率仅限于染色体臂/带的微观水平，且需要借助荧光原位杂交（fluorescence *in situ* hybridization，FISH）或者反转录-聚合酶链反应（RT-PCR）等方法进一步验证[16-17]（图2）。

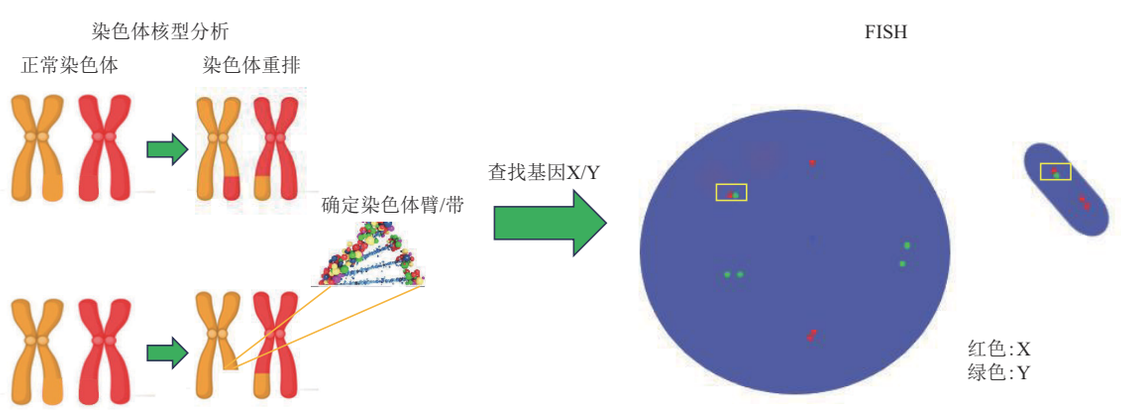


Fig. 2　Chromosome karyotype analysis and FISH

图2　染色体核型分析与FISH

2　荧光原位杂交（FISH）

FISH技术是探针利用碱基配对互补的原理，将荧光染料标记的DNA与特定靶DNA序列进行互补杂交，通过荧光显微镜观测荧光信号的位置、大小及数量来判断DNA序列的缺失、扩增、融合、

断裂及易位等，此方法适用于发生DNA重组或RNA反式剪接形成的嵌合RNA。

目前，FISH检测用的探针大部分为双色标记，主要分为染色体计数探针（缺失和扩增）和位点特异性识别探针（融合、重排、断裂等）。融合基因的检测是用含有两种不同颜色荧光素标记的探针，分别对应各自独立的靶基因位点，针对其是否发生融合而产生不同的颜色进行区分。如Ruffle等[18]利用橙色或绿色探针去验证*PAN3-NONE*融合基因（表1）。

FISH检测嵌合RNA只是通过宏观观察颜色的变化判断[19]，且只能显示RNA层面发生的改变，只能作为嵌合RNA的验证手段，而不具备作为发现和筛选嵌合RNA的优势。但对于临床实验室来说，其是目前检测最快速的方法之一，可在24~48 h内得出结果，但检测范围有限。Jeck等[16]针对此缺陷设计了一种利用Oxford Nanopore Minion测序系统的方法，可以提供长读数的快速测序，缩短检测融合癌基因的周转时间，对需要快速结果的临床检验可能是一种有效的方法。此外，Pacbio测序技术也使得直接分离和筛选嵌合RNA成为可能。

3　高通量嵌合RNA检测技术

3.1　基于RNA−Seq的单端数据和双末端数据

新一代测序技术（next-generation sequencing，NGS）以Roche公司的454技术、Illumina公司的Solexa、Hiseq技术和ABI公司的Solid技术为代表，相比FISH和RT-PCR等方法[20-21]，NGS大大降低了测序成本，它可以分析更多的目标，在可接受的周转时间内节省时间和材料，而且拥有高度敏感性和高准确性。同时，它不会受限于无法鉴定距离很近或者序列未知的基因，也没有实验过程中杂交和荧光信号使用产生的严重杂音信号。虽然NGS允许在临床实践中不断合并新发现的生物标志物，但是缺乏标准化仍然是一个比较致命的问题[22]。

与传统的方法相比，RNA-Seq可以准确量化转录本的表达，发现新的转录本，识别可变剪切事件，检测基因融合，从而揭示不同条件下转录组的动态变化。多年来，已经开发了一些软件工具来从RNA测序（RNA-seq）数据中检测基因融合[23]。在嵌合RNA的研究中，RNA-seq可以检测发生在RNA水平的基因间剪接融合，检测多种可选的剪接变异体造成的融合。低成本和快速周转时间是其最大优点。

RNA-seq实验提供了一组短读数，可以有两种形式：单端序列或双端序列。单端测序就是将DNA样本随机打断后，在DNA片段的一端连接上引物序列，然后在末端加上接头，对每个片段进行测序。双端测序是在DNA片段两端的接头都加上测序引物结合位点，引导互补链在原位置再生和扩增，然后合成互补链并测序。以Illumina测序平台的序列结果为例，其测序结果输出格式是fastq[24]，如果是双端数据，数据存储到*_1.fastq，*_2.fastq文件，代表来自两个不同末端的测序数据。

嵌合RNA可以从不同长度的单端测序（single-end reads）[25]或双端测序（paired-end/mate-paired reads）[26]中检测到。由于分析软件的限制，利用双端测序数据分析嵌合RNA更为方便和快捷，原因是双端测序方法不仅能提供更大的动态范围，而且可以更准确地预测基因融合位置及两侧序列信息，便于进行下一步的实验验证。

根据RNA-seq数据，一系列利用短读长检测融合的技术已经开发出来[27]，如FusionPlex、AmpliSeq、QIAseq、RNAscan、Oncomine Focus Assay、TruSight Tumor 170、SureSelect XT等[28-31]。常用的检测效率较高的FusionCatcher（https://github.com/ndaniel/fusioncatcher）检测技术可以用于检测来自脊椎动物疾病样本的双端RNA测序数据中新的和已知的体细胞融合基因[32]。FusionCatcher的主要目标是：良好的精确度和灵敏度。首先对测序数据预处理和过滤，再利用FusionCatcher在RNA水平进行Ensembl基因组注释和Bowtie aligner对测序read比对，使用4种不同的方法和4种不同的比对装置组成的集成方法来识别融合连接（图3）。

基于双端测序所形成的嵌合RNA中包括跨链嵌合RNA（cscRNA），即两条相反DNA链编码的转录本融合产生的RNA嵌合体，Wang等[33]构建生物信息学比对方法cscMap检测cscRNA。cscMap的设计专门用于从头鉴定带注释或未注释的RNA转录本之间的非规范跨链连接事件。简单原理是将配对末端RNA-seq读数与参考基因组和转录组比对，然后未比对的读数进行第二轮比对，该比对将配对末端RNA-seq读数的一端分解为两个片段，分别寻求这些片段与两条相反DNA链的比对。

使用RNA-seq检测融合要求检测的灵敏度和特异性，其取决于测序深度、长度和质量，以及所

使用的生物信息学方法和参数。RNA-seq数据利用短读数即可进行表达谱和基因表达的定量，然而短读数在识别复杂的基因组重排、重复区域或全长转录本方面效率较低，利用生物学算法进行融合识别时重要的变异容易遗漏[34]。目前针对短读数出现的问题常用以下两种方法来进行规避。一为优先比对：首先识别基因组重排后不一致的读数；二为优先组合：将读数组装成更长的转录本序列以识别融合转录本[35]。因此，克服短读读长限制的一种方法是合成长读长，将短读长编译在一起，构建更长的读长。此方法已成功识别良性结肠黏膜、原发性结肠癌和转移性结肠癌中此前未识别的融合转录本[36]。

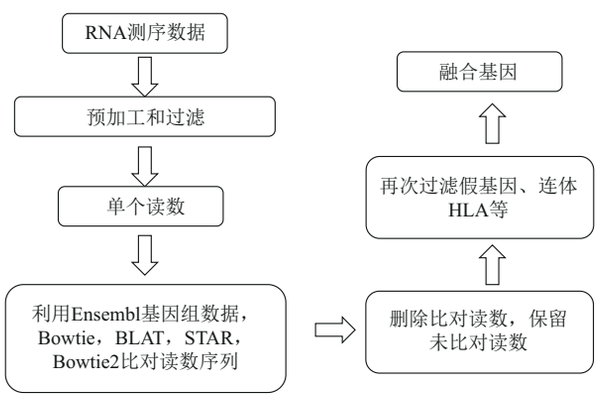


Fig. 3 FusionCatcher analysis process

图3 FusionCatcher分析流程

3.2 基于RNA-Seq的嵌合RNA检测技术

目前最流行的嵌合RNA检测算法都是将DNA或cDNA序列比对到参考转录组数据库。通过序列比对的方法找出形成嵌合RNA的两个来源母基因序列及信息，以便对其产生机制和功能与母基因的异同进行更为深入的研究。Winters等[37]利用RNA-Seq检测出包括*LPP-FOXP1*、*KIF5B-RET*在内的多个癌症中嵌合RNA，并使用RT-PCR等方法进行了验证。鉴于目前嵌合RNA发现工具在特异性、敏感性、效率以及计算机内存使用方面表现不尽相同，目前常用的嵌合RNA检测分析软件包有：deFuse、EricScript、JAFFAL、FusionCatcher、FusionMap、PRADA、SOAPfuse、STAR-FUSION、Arriba和Chimeraviz等。以上所述软件，其工作原理和机制基本类同，但各自算法进行了独立的优化，现在以常用的软件分析包SOAPfuse、EricScript及近年文献提到的方法来发现和挖掘新的嵌合RNA。

3.2.1 SOAPfuse

SOAPfuse (http://soap.genomics.org.cn/soapfuse.html) 是华大基因公司基于perl语言开发的RNA-Seq数据全基因组范围内检测嵌合RNA的开放性工具。SOAPfuse通过将RNA-Seq双端序列比对结果与人类基因组参考序列和注释基因做比对，根据RNA序列是否比对到两个不同的基因，进而发现可能的嵌合RNA（图4）。它通过两种序列形式来支持一个融合事件，一组不一致的连接候选嵌合RNA的双端序列（span-read）以及一组来确认准确连接位置的连接序列（junc-read）。SOAPfuse采用一种改进的算法来有效地构建一个融合位置文库，并采用一系列参数条件和数据质量控制措施来从序列和假阳性结果中判别出可能的阳性结果。该软件可以生成融合转录本的高可信度结果列表，其中包含连接位点在单核苷酸分辨率上的精确位置。此外，SOAPfuse还可以预测并提供融合转录本交接处侧翼各200 bp的序列，该序列可用于设计嵌合RNA特异引物来为RT-PCR验证做准备。此外，SOAPfuse还可以创建示意图，可以显示连接序列上所匹配的序列（span-reads和junc-reads）的排列情况，以及每个基因对的外显子的表达水平。SOAPfuse可以区分RNA-Seq数据的具体参数，如插入大小和读取长度。因此，即使单个样本包含不同类型的双端RNA-Seq数据，它仍然可以很好地工作[38]。此种方法适用于大多数嵌合RNA类型的检测，但主要用于研究顺式剪接形成的嵌合RNA，如嵌合RNA *D2HGDH-GAL3ST2*[39]（表1）。

3.2.2 EricScript

Benelli等[40]提出了一种新的嵌合转录检测算法——EricScript (https://sites.google.com/site/bioericscript/home)，该算法检测嵌合RNA是基于双端RNA-Seq数据。其新颖之处在于对外显子结合位点有一个有效的重新校准过程，能够增加灵敏度和特异性，并减少运行时间。EricScript分析主要流程：a. 根据转录组对序列进行映射；b. 不一致地排列识别和生成外显子结合位点参考序列；c. 外显子结合位点参考序列的重新校准；d. 对候选的嵌合RNA进行评分和筛选（图5）。EricScript引入了3个新的参数，分别是真正的结合位点参数（genuine junction score，GJS）、边缘参数（edge score，ES）、一致性评价（uniformity score，US）。

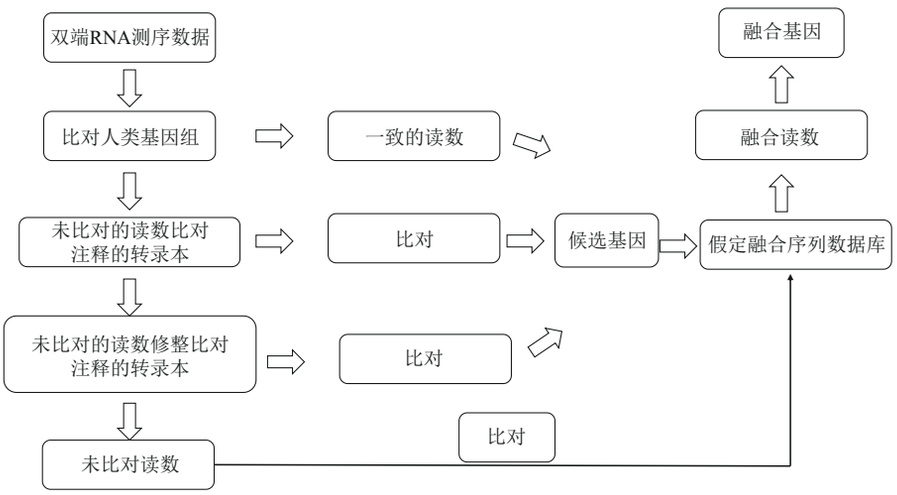


Fig. 4　SOAPfuse analysis process

图4　SOAPfuse分析流程

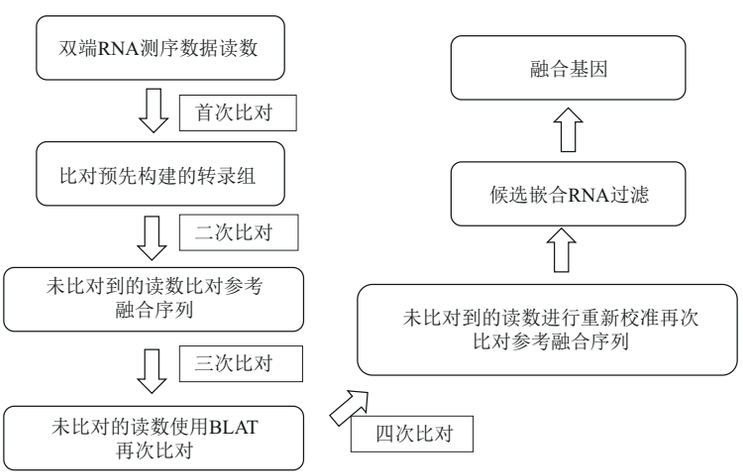


Fig. 5　EricScript analysis process

图5　EricScript分析流程

它能够高精度地区分真正的嵌合RNA和假阳性事件，并减少数据分析产生的大量调用。

由于EricScript测序和分析过程中会产生出大量的假阳性事件，因此加入了一系列筛选条件：a. 丢弃所有完全比对到相同位置的双端序列；b. 仅保留评分良好的候选嵌合RNA并对其进行分类；c. 使用BLAT对Ensembl转录组比对到野生型结合位点周围的100 bp序列的区域，若比对结果同源性和相似度比例均很高，则排除该候选嵌合RNA。

EricScript生成了模拟不同覆盖水平的不同读取长度的合成数据集，合成数据集可用于训练自适应boosting（AdaBoost）从而对分析结果进行排序。同时，在EricScript包中还有用于模拟基因融合的程序。研究表明，EricScript能够实现对嵌合RNA检测分析的高灵敏度和特异性，且运行效率高时间短，此种方法适用于大多数嵌合RNA类型的检测，如嵌合RNA *FOXP1*－*RYBP*[40]（表1）。

3.3　优化的嵌合RNA检测技术

近年来也有其他学者提到一些更有优势的检测方法。Heyer等[41]提出了靶向RNA-seq的方法。此方法使用生物素化的寡核苷酸探针来富集目标RNA转录本。首先设计探针靶向融合基因的外显

子，同时制备基因特异性文库，接下来将探针与基因杂交，进行RNA测序，从而鉴定新型嵌合RNA。此方法相比于STARFusion和FusionCatcher方法，拥有更高的准确度、灵敏度且能够检测罕见的融合基因。但其假阳性风险增大，且需对结果信息进行多次验证，并解释说明。随着统计学方法的进步，增大的假阳性率可以使用，如Depest[42]的算法降低。另外，Engvall等[43]同样也提出了靶向RNA-seq的方法，并在白血病患者中进行了应用诊断。该方法使用一种锚定多重PCR，将基因特异的引物与包含通用引物结合位点的接头结合，在事先不知道配对序列或特定断裂点的情况下扩增目的序列，再使用嵌套的基因特异性引物进行第二次PCR。该方法流程简单、时间短、特异性高。但该方法容易产生第一次链式反应的产物污染。Iyer等[44]建立了ChimeraScan方法，该方法使用Bowtie将双端读段比对至参考基因组，其中无法比对的序列重新筛选，通过一定的筛选标准得到推测的嵌合体，再过滤掉读取较低的嵌合体从而得到嵌合基因。该方法拥有处理长（>75 bp）双端读段、处理模糊映射读段、检测跨越融合连接的读段、支持标准化的SAM格式和HTML报告等优点。Li等[45]建立了ChimeRScope方法，是一种新的基于双端RNA-Seq数据的无比对融合转录预测算法，它是根据RNA-Seq双端序列的基因指纹（如k-mers）图谱从而准确地预测融合转录本。此方法比较适合于大规模的融合转录本数据分析。可在精准医学中用于有效诊断和预后的生物标志物。Friedrich等[46]提出了STfusion方法，在顺式剪接机制或染色体重排引起的融合转录本的情况下，5'基因中没有Poly（A）尾，可以定位到3'基因的Poly（A）尾数量，以及5'基因上没有Poly（A）尾巴的数量，被用来指示融合转录本，并使用前列腺癌中的*SLC45A3-ELK4*（表1）嵌合RNA进行验证。Liu等[47]开发了LongGF来高效地从长RNA-seq数据中检测基因融合，运用C++实现，运行速度快，只需几分钟和<3 GB的内存便可从转录组中读取50 000个长片段进行基因融合检测，但此方法只能检测已知的基因融合，且容易造成假阳性。Chiu等[48]提出了利用RNA-Bloom和PAVFinder融合基因检测方法Fusion-Bloom，该方法自动化运行3个分析阶段：组装、比对和分析，并以BEDPE格式报告结果，具有良好的敏感性和特异性，但此方法运行需求时间较长。其他的也有如：基于Read-mapping和从头融合转录本组装方法[35]；长链转录组测序检测融合基因的JAFFAL方法[49]；基于功能基因组学方法[50]；基于筛选条件的方法：TSscan筛除由于RT期间的错误导致的假阳性结果，但也会忽略一些真正的嵌合RNA[51]；Map Splice不依赖于连接位点的特征或内含子的长度，不局限于经典的拼接位点来鉴定候选嵌合RNA序列，但可能产生假阳性结果[52]；Segemehl算法考虑错配和插入/缺失等因素[53]。NCLscan软件有助于区分反式剪接，环形或基因重排衍生融合[54]。Jin等[55]提出的scFusion方法支持序列和深度学习模型筛除单细胞测序中的假阳性结果。因此，选择合适的软件来分析特定的RNA-Seq数据是至关重要的。

3.4 RNA−Seq数据的局限性

RNA-Seq仅对转录组中的一小部分（约2%）进行测序[56]。除传统的基因融合外，RNA-seq还检测仅发生在RNA水平的基因间剪接融合。更低的时间成本和经济成本使得RNA-Seq在融合转录本研究中非常受欢迎。然而，RNA-Seq还存在许多局限性：a. 不能检测到涉及非转录事件中的融合基因[57]；b. 只能检测未发生在DNA水平的融合事件，不能排除DNA本身发生了融合导致转录组的序列基因结构变化；c. 在人类转录组研究中，组织的特异性和表达的广阔动态范围使RNA-Seq数据分析复杂化[58]。有研究报道，有几个重要的基因融合是利用全基因组测序（whole genome sequencing，WGS）被检测到的[59-60]。因此，利用生物信息学分析软件将RNA-seq和WGS数据结合起来是嵌合RNA检测的最佳数据来源。

4 嵌合RNA验证方法

4.1 聚合酶链反应（PCR）

PCR是一种高效评估转录水平的方法，由于其敏感性和特异性，特别适合检测验证嵌合RNA。Ruffle等[18]使用PCR方法检测TM值，与其阴性对照结果不同的样品认为是潜在的阳性嵌合RNA，并对其进行测序，如嵌合RNA *ESRRA-C11orf20*[61]（表1）。PCR方法具有高效直观的优点，但利用逆转录构建cDNA文库的方法都容易由于模板转换而产生假阳性，且只对与特定一对相关基因的融合敏感，甚至可能对某些不常见的变种不敏感[16]。

4.2　核糖核酸酶保护实验（RPA）

RPA的基本原理是探针和样品液杂交，而后进行酶切消化，聚丙烯酰胺凝胶电泳鉴定。基于这一原理，RPA用作评估RNA水平以及定位基因的转录起始位点[62]。同时也作为一种敏感的方法来识别选择性拼接的转录本和不同的转录起始位点[63]。然而，在cDNA文库构建中使用逆转录的方法容易产生因模板转换而出现的假阳性，因为逆转录酶会产生在原始生物体中不存在的嵌合RNA[64-65]。但本方法不采用逆转录程序，从而避免了在逆转录过程中可能出现的假阳性结果。

4.3　琼脂糖凝胶电泳和Sanger法测序

琼脂糖凝胶电泳以琼脂凝胶作为支持物，利用DNA分子在泳动时的电荷效应和分子筛效应，达到分离混合物的目的。通过生信分析工具获得嵌合RNA，进行RT-PCR，验证嵌合RNA的转录表达水平，筛选出表达差异明显的嵌合RNA将其扩增后的cDNA进行琼脂糖凝胶电泳，分离纯化cDNA，以便后续对其进行Sanger法测序[66]。

Sanger测序是DNA测序技术的金标准，曾在人类基因组计划中发挥了关键推动作用，并且现在仍被用来获得高度准确且可信赖的测序数据。其基本原理是碱基互补配对原则及缺乏3'-OH基团的随机终止。Sanger法测序的主要目的是验证SOAPfuse/EricScript等方法中提供的连接位点参考序列，从而保证筛选出来的嵌合RNA的真实性。

4.4　NanoString nCounter

NanoString nCounter分析系统是最新的多重基因定量检测技术，该技术是基于核酸分子与探针杂交后，对探针上的颜色分子条形码标记直接探测、计数而实现多重定量的检测技术。其核心技术原理包括分子条形码技术和单分子成像数字计数技术。此技术可直接检测条形码探针标记的单个mRNA转录子，并通过数字计数进行定量。

NanoString nCounter检测技术能够检测mRNA、miRNA及DNA拷贝数变异，敏感性高，且因无需逆转录扩增，避免了假阳性。但其不能检测未知mRNA。Lira等[67]应用NanoString技术对肺非小细胞癌标本多种断裂点的ALK融合基因进行检测，可确定各种断裂点亚型的融合基因及其表达水平，可用于肺非小细胞癌的分子诊断。临床中也使用此技术对患者的活检组织进行融合检测[68]。

Table1　Chimeric RNA illustration

表1　嵌合RNA例证

检测验证工具	嵌合RNA	参考文献
染色体核型分析	*ETV6-NTRK3*	［15］
	NRIP1-MIR99AHG	［17］
FISH	*PAN3-NONE*	［18］
SOAPfuse	*D2HGDH-GAL3ST2*	［39］
EricScript	*FOXP1－RYBP*	［40］
STfusion	*SLC45A3-ELK4*	［46］
PCR	*ESRRA-C11orf20*	［61］

5　总结与展望

随着新一代测序技术的研发，嵌合RNA的发现与研究成为新兴的研究热点，其检测及验证方法更是关键性技术。现如今科研人员从利用染色体核型分析、FISH到SOAPfuse等高通量测序方法在非小细胞肺癌、前列腺癌、宫颈癌等疾病中发现众多新的嵌合RNA[69-71]，并利用PCR等方法成功验证其真实性，利用各种功能实验发现了其抑癌功能。

高通量测序技术的快速发展使得嵌合RNA能够以低成本、快速、高准确的方式被发现，但检测方法本身仍然存在假阳性、特异性不好、灵敏度不足等问题。染色体核型分析的周期长，分辨率低；FISH方法只能用于RNA层面；高通量测序方法SOAPfuse假阳性率较高。另外，其验证方法如PCR、琼脂糖凝胶电泳和Sanger法测序由于存在序列大量扩增，有一定概率会引入错配碱基致使假阳性率升高，错误地验证嵌合RNA。即使通过DNA-seq数据的补充验证如NanoFG[72]，嵌合RNA的检测仍然存在许多问题。另外，环状融合RNA（circular fusion RNAs）的发现对合适的检测工具需求更加迫切[73]。因此，二代测序技术带来的问题如读取序列的长度已不能满足当今科学的发展。

三代及四代测序技术的出现一定程度上改善了这些问题，PacBio公司的SMRT作为三代测序技术的代表和Oxford Nanopore Technologies纳米孔单分子测序技术作为四代测序技术的代表[74]。与前两代测序技术相比，其最大的特点就是单分子测序，测序过程无需进行PCR扩增，并且理论上可以测定无限长度的核酸序列，即可以获得长读长。长读可以解析复杂的多外显子并识别长转录本和更多的剪接变体[75]，识别天然的RNA修饰[76]。Mitsuhashi等[77]利用此方法获得*LTR-RBM26*融合转录本的全长及其不同的剪接形式。长读长转录组

测序的另一个优点是能够识别双跳和桥接融合[78]。但长读也存在读长数量少、读长深度低、碱基准确性低的问题[79-80]。这种问题可以通过如isONcorrect计算工具和PacBio公司的High Fidelity方法获得更高的读数准确性[81-82]。另一种解决方法是，可以通过混合短读和长读进行融合检测[83]。目前Rautiainen等[84]开发了AERON，即长读比对融合检测工具，称为GraphAligner；Zhang等[85]开发的CIRI-long和Xin等[86]开发的isoCirc利用长读检测circRNA。未来基于三、四代测序结果的分析，将有更多嵌合RNA被挖掘发现，阳性率也将提高。

本文揭示了嵌合RNA的检测方法，从传统的检测技术到高通量测序方法，从染色体核型分析到RNA-Seq数据的处理，以及生物信息学筛选，实验验证其真实性等。虽然目前嵌合RNA研究领域尚处于初级阶段，但是其检测技术已经初具规模。随着新一代测序技术和生物信息学分析软件的进步，对嵌合RNA的研究将会不断地深入，形成更全面的认识。

参 考 文 献

[1] Melo J, Gordon D, Cross N, *et al*. The ABL-BCR fusion gene is expressed in chronic myeloid leukemia. Blood, 1993, **81**(1): 158-165

[2] Sun Y, Li H. Chimeric RNAs discovered by RNA sequencing and their roles in cancer and rare genetic diseases. Genes (Basel), 2022, **13**(5):741

[3] Taniue K, Akimitsu N. Fusion genes and RNAs in cancer development. Noncoding RNA, 2021, **7**(1): 10

[4] Hanahan D, Weinberg R A. Hallmarks of cancer: the next generation. Cell, 2011, **144**(5): 646-674

[5] Gupta S K, Jea J D, Yen L. RNA-driven JAZF1-SUZ12 gene fusion in human endometrial stromal cells. PLoS Genet, 2021, **17**(12): e1009985

[6] Mertens F, Johansson B, Fioretos T, *et al*. The emerging complexity of gene fusions in cancer. Nat Rev Cancer, 2015, **15**(6): 371-381

[7] White N M, Feng F Y, Maher C A. Recurrent rearrangements in prostate cancer: causes and therapeutic potential. Curr Drug Targets, 2013, **14**(4): 450-459

[8] Berthold R, Isfort I, Erkut C, *et al*. Fusion protein-driven IGF-IR/PI3K/AKT signals deregulate hippo pathway promoting oncogenic cooperation of YAP1 and FUS-DDIT3 in myxoid liposarcoma. Oncogenesis, 2022, **11**(1): 20

[9] Kuravi S, Baker R W, Mushtaq M U, *et al*. Functional characterization of NPM1-TYK2 fusion oncogene. NPJ Precis Oncol, 2022, **6**(1): 3

[10] Edwards P A. Fusion genes and chromosome translocations in the common epithelial cancers. J Pathol, 2010, **220**(2): 244-254

[11] Li H, Wang J, Mor G, *et al*. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science, 2008, **321**(5894): 1357-1361

[12] Li H, Wang J, Ma X, *et al*. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle, 2009, **8**(2): 218-222

[13] Kannan K, Wang L, Wang J, *et al*. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci USA, 2011, **108**(22): 9172-9177

[14] Nacu S, Yuan W, Kan Z, *et al*. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics, 2011, **4**: 11

[15] Tognon C, Knezevich S R, Huntsman D, *et al*. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell, 2002, **2**(5): 367-376

[16] Jeck W R, Lee J, Robinson H, *et al*. A nanopore sequencing-based assay for rapid detection of gene fusions. J Mol Diagn, 2019, **21**(1): 58-69

[17] Kerbs P, Vosberg S, Krebs S, *et al*. Fusion gene detection by RNA-sequencing complements diagnostics of acute myeloid leukemia and identifies recurring NRIP1-MIR99AHG rearrangements. Haematologica, 2022, **107**(1): 100-111

[18] Ruffle F, Audoux J, Boureux A, *et al*. New chimeric RNAs in acute myeloid leukemia. F1000Res, 2017, **6**: ISCB Comm J-1302

[19] Panagopoulos I, Andersen K, Gorunova L, *et al*. Fusion of the HMGA2 and BNC2 genes in uterine leiomyoma with t(9;12)(p22;q14). *In Vivo*, 2022, **36**(6): 2654-2661

[20] Schoch C, Schnittger S, Bursch S, *et al*. Comparison of chromosome banding analysis, interphase- and hypermetaphase-FISH, qualitative and quantitative PCR for diagnosis and for follow-up in chronic myeloid leukemia: a study on 350 cases. Leukemia, 2002, **16**(1): 53-59

[21] Bustin S A, Nolan T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. J Biomol Tech, 2004, **15**(3): 155-166

[22] Bruno R, Fontanini G. Next generation sequencing for gene fusion analysis in lung cancer: a literature review. Diagnostics (Basel), 2020, **10**(8): 521

[23] Kumar S, Vo A D, Qin F, *et al*. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep, 2016, **6**: 21597

[24] Cock P J, Fields C J, Goto N, *et al*. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res, 2010, **38**(6): 1767-1771

[25] Maher C A, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. Nature, 2009, **458**(7234): 97-101

[26] Maher C A, Palanisamy N, Brenner J C, *et al*. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci USA, 2009, **106**(30): 12353-12358

[27] Singh S, Li H. Comparative study of bioinformatic tools for the identification of chimeric RNAs from RNA Sequencing. RNA Biol, 2021, **18**(sup1): 254-267

[28] Bergeron D, Chandok H, Nie Q, *et al*. RNA-Seq for the detection of gene fusions in solid tumors: development and validation of the JAX fusionSeq 2.0 assay. J Mol Med (Berl), 2022, **100**(2): 323-335

[29] Heydt C, Wolwer C B, Velazquez Camacho O, *et al*. Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation. BMC Med Genomics, 2021, **14**(1): 62

[30] Peng H, Huang R, Wang K, *et al*. Development and validation of an RNA sequencing assay for gene fusion detection in formalin-fixed, paraffin-embedded tumors. J Mol Diagn, 2021, **23**(2): 223-233

[31] Qu X, Yeung C, Coleman I, *et al*. Comparison of four next generation sequencing platforms for fusion detection: oncomine by ThermoFisher, AmpliSeq by illumina, FusionPlex by ArcherDX, and QIAseq by QIAGEN. Cancer Genet, 2020, **243**: 11-18

[32] Nicorici D, Şatalan M, Edgren H, *et al*. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv, 2014. doi: https://doi.org/10.1101/011650

[33] Wang Y, Zou Q, Li F, *et al*. Identification of the cross-strand chimeric RNAs generated by fusions of bi-directional transcripts. Nat Commun, 2021, **12**(1): 4645

[34] Dorney R, Dhungel B P, Rasko J E J, *et al*. Recent advances in cancer fusion transcript detection. Brief Bioinform, 2023, **24**(1): bbac519

[35] Haas B J, Dobin A, Li B, *et al*. Accuracy assessment of fusion transcript detection *via* read-mapping and *de novo* fusion transcript assembly-based methods. Genome Biol, 2019, **20**(1): 213

[36] Liu S, Wu I, Yu Y P, *et al*. Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer. Commun Biol, 2021, **4**(1): 506

[37] Winters J L, Davila J I, Mcdonald A M, *et al*. Development and verification of an RNA Sequencing (RNA-Seq) assay for the detection of gene fusions in tumors. J Mol Diagn, 2018, **20**(4): 495-511

[38] Jia W, Qiu K, He M, *et al*. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol, 2013, **14**(2): R12

[39] Qin F, Song Z, Chang M, *et al*. Recurrent cis-SAGe chimeric RNA, D2HGDH-GAL3ST2, in prostate cancer. Cancer Lett, 2016, **380**(1): 39-46

[40] Benelli M, Pescucci C, Marseglia G, *et al*. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics, 2012, **28**(24): 3232-3239

[41] Heyer E E, Deveson I W, Wooi D, *et al*. Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun, 2019, **10**(1): 1388

[42] Dehghannasiri R, Freeman D E, Jordanski M, *et al*. Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. Proc Natl Acad Sci USA, 2019, **116**(31): 15524-15533

[43] Engvall M, Cahill N, Jonsson B I, *et al*. Detection of leukemia gene fusions by targeted RNA-sequencing in routine diagnostics. BMC Med Genomics, 2020, **13**(1): 106

[44] Iyer M K, Chinnaiyan A M, Maher C A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics, 2011, **27**(20): 2903-2904

[45] Li Y, Heavican T B, Vellichirammal N N, *et al*. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. Nucleic Acids Res, 2017, **45**(13): e120

[46] Friedrich S, Sonnhammer E L L. Fusion transcript detection using spatial transcriptomics. BMC Med Genomics, 2020, **13**(1): 110

[47] Liu Q, Hu Y, Stucky A, *et al*. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. BMC Genomics, 2020, **21**(Suppl 11): 793

[48] Chiu R, Nip K M, Birol I. Fusion-Bloom: fusion detection in assembled transcriptomes. Bioinformatics, 2020, **36**(7): 2256-2257

[49] Davidson N M, Chen Y, Sadras T, *et al*. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. Genome Biol, 2022, **23**(1): 10

[50] Li J, Lu H, Ng P K, *et al*. A functional genomic approach to actionable gene fusions for precision oncology. Sci Adv, 2022, **8**(6): eabm2382

[51] Wu C S, Yu C Y, Chuang C Y, *et al*. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. Genome Res, 2014, **24**(1): 25-36

[52] Wang K, Singh D, Zeng Z, *et al*. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res, 2010, **38**(18): e178

[53] Hoffmann S, Otto C, Kurtz S, *et al*. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol, 2009, **5**(9): e1000502

[54] Chuang T J, Wu C S, Chen C Y, *et al*. NCLscan: accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision. Nucleic Acids Res, 2016, **44**(3): e29

[55] Jin Z, Huang W, Shen N, *et al*. Single-cell gene fusion detection by scFusion. Nat Commun, 2022, **13**(1): 1084

[56] Sboner A, Habegger L, Pflueger D, *et al*. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol, 2010, **11**(10): R104

[57] Kim D, Salzberg S L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol, 2011, **12**(8): R72

[58] Taylor B S, Ladanyi M. Clinical cancer genomics: how soon is now?. J Pathol, 2011, **223**(2): 318-326

[59] Pleasance E D, Stephens P J, O'meara S, *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature, 2010, **463**(7278): 184-190

[60] Link D C, Schuettpelz L G, Shen D, *et al*. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA, 2011, **305**(15): 1568-1576

[61] Salzman J, Marinelli R J, Wang P L, *et al*. ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. PLoS Biol, 2011, **9**(9): e1001156

[62] Goodall G J, Wiebauer K, Filipowicz W. Analysis of pre-mRNA processing in transfected plant protoplasts. Methods Enzymol, 1990, **181**: 148-161

[63] Zhang Y, Gong M, Yuan H, *et al*. Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. Cancer Discov, 2012, **2**(7): 598-607

[64] Adams M D, Kelley J M, Gocayne J D, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 1991, **252**(5013): 1651-1656

[65] Lu Z, Matera A G. Vicinal: a method for the determination of ncRNA ends using chimeric reads from RNA-seq experiments. Nucleic Acids Res, 2014, **42**(9): e79

[66] 江梅, 周裕儒, 詹媛, 等. 转录组测序分析融合基因在染色体核型正常髓系白血病诊断中的应用. 中华医学杂志, 2021, **101**(13): 939-944
Jiang M, Zhou Y R, Zhan Y, *et al*. National Medical Journal of China, 2021, **101**(13): 939-944

[67] Lira M E, Kim T M, Huang D, *et al*. Multiplexed gene expression and fusion transcript analysis to detect ALK fusions in lung cancer. J Mol Diagn, 2013, **15**(1): 51-61

[68] Gimenez-Capitan A, Sanchez-Herrero E, Robado De Lope L, *et al*. Detecting ALK, ROS1, and RET fusions and the METΔex14 splicing variant in liquid biopsies of non-small-cell lung cancer patients using RNA-based techniques. Mol Oncol, 2023, **17**(9): 1884-1897

[69] Wang Q, Chen J, Singh S, *et al*. Profile of chimeric RNAs and TMPRSS2-ERG e2e4 isoform in neuroendocrine prostate cancer. Cell Biosci, 2022, **12**(1): 153

[70] Dang X, Xiang T, Zhao C, *et al*. EML4-NTRK3 fusion cervical sarcoma: a case report and literature review. Front Med (Lausanne), 2022, **9**: 832376

[71] Lei Y, Lei Y, Shi X, *et al*. EML4-ALK fusion gene in non-small cell lung cancer. Oncol Lett, 2022, **24**(2): 277

[72] Stangl C, De Blank S, Renkens I, *et al*. Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. Nat Commun, 2020, **11**(1): 2861

[73] Huang Y, Zhang C, Xiong J, *et al*. Emerging important roles of circRNAs in human cancer and other diseases. Genes Dis, 2021, **8**(4): 412-423

[74] Athanasopoulou K, Boti M A, Adamopoulos P G, *et al*. Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. Life (Basel), 2021, **12**(1): 30

[75] Guan J, Yin S, Yue Y, *et al*. Single-molecule long-read sequencing analysis improves genome annotation and sheds new light on the transcripts and splice isoforms of *Zoysia japonica*. BMC Plant Biol, 2022, **22**(1): 263

[76] Liu H, Begik O, Lucas M C, *et al*. Accurate detection of m(6)A RNA modifications in native RNA sequences. Nat Commun, 2019, **10**(1): 4079

[77] Mitsuhashi S, Nakagawa S, Sasaki-Honda M, *et al*. Nanopore direct RNA sequencing detects DUX4-activated repeats and isoforms in human muscle cells. Hum Mol Genet, 2021, **30**(7): 552-563

[78] Namba S, Ueno T, Kojima S, *et al*. Transcript-targeted analysis reveals isoform alterations and double-hop fusions in breast cancer. Commun Biol, 2021, **4**(1): 1320

[79] Karaoglanoglu F, Chauve C, Hach F. Genion, an accurate tool to detect gene fusion from long transcriptomics reads. BMC Genomics, 2022, **23**(1): 129

[80] Creason A, Haan D, Dang K, *et al*. A community challenge to evaluate RNA-seq, fusion detection, and isoform quantification methods for cancer discovery. Cell Syst, 2021, **12**(8): 827-838

[81] Sahlin K, Medvedev P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nat Commun, 2021, **12**(1): 2

[82] Miller A R, Wijeratne S, Mcgrath S D, *et al*. Pacific biosciences fusion and long isoform pipeline for cancer transcriptome-based resolution of isoform complexity. J Mol Diagn, 2022, **24**(12): 1292-1306

[83] Weirather J L, Afshar P T, Clark T A, *et al*. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res, 2015, **43**(18): e116

[84] Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol, 2020, **21**(1): 253

[85] Zhang J, Hou L, Zuo Z, *et al*. Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. Nat Biotechnol, 2021, **39**(7): 836-845

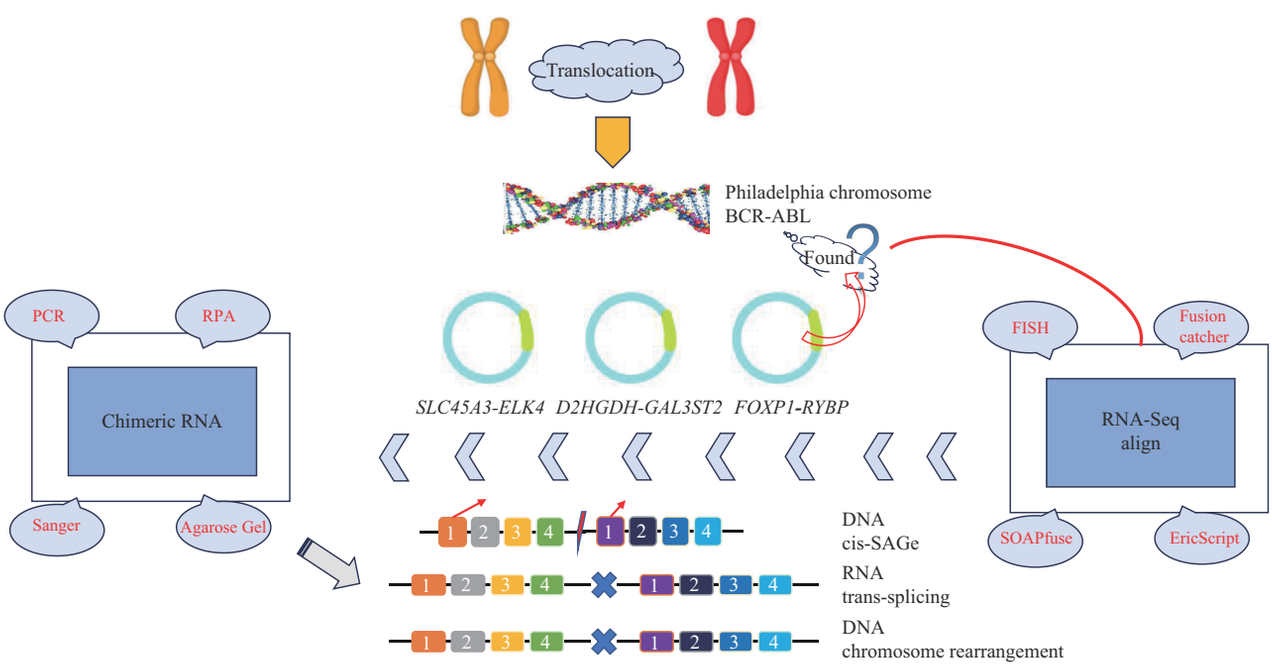
[86] Xin R, Gao Y, Gao Y, *et al*. isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. Nat Commun, 2021, **12**(1): 266

Detection and Validation of Chimeric RNA*

WANG Guang-Fu[1,2,3,4], DING Yong-Wei[6], TANG Yue[1]**, QIN Fu-Jun[1,2,3,4,5]**

([1]*Department of Microbiology and Immunology, School of Basic Science, Zhengzhou University, Zhengzhou* 450001, China;
[2]*Center for Medical Epigenetics, Academy of Medical Sciences, Zhengzhou University, Zhengzhou* 450052, China;
[3]*Molecular Pathology Center, Academy of Medical Sciences, Zhengzhou University, Zhengzhou* 450052, China;
[4]*Translational Medicine Platform, Academy of Medical Sciences, Zhengzhou University, Zhengzhou* 450052, China;
[5]*State Key Laboratory of Esophageal Cancer Prevention & Treatment, Zhengzhou University, Zhengzhou* 450052, China;
[6]*Department of Pathology and Pathophysiology, School of Basic Medicine, Zhengzhou University, Zhengzhou* 450001, China)

Graphical abstract



Abstract　Chimeric RNA is a fusion transcript comprising of exon fragments from different genes. There are three splicing types: chromosome rearrangements, trans-splicing, cis-splicing, and the recently mentioned circular chimeric RNA. The traditional methods for the detection of chimeric RNA includes chromosome karyotype

* This work was supported by grants from The National Natural Science Foundation of China (81972421), Joint Program NSFC-Henan (U2004135), Zhengzhou University High-level Talent Start-up Project (32340177), Innovation and Entrepreneurship Training Program for College Students of Henan Province (202210459126, 202210459161), and Education and Teaching Reform Research and Practice Project of Zhengzhou University (2022ZZUJGXMLXS-017).

** Corresponding author.
TANG Yue. Tel：86-371-67781950，E-mail：tangyue@zzu.edu.cn
QIN Fu-Jun. Tel：86-371-66658776，E-mail：fujun_qin@zzu.edu.cn

Received：June 15，2023　Accepted：July 20，2023

analysis, FISH, DNA microarray, *etc*., but their specificity, sensitivity and accuracy for the detection of chimeric RNA are poorly understood. With the development of sequencing technology, second-generation sequencing technology has shown strong data processing capabilities and can detect chimeric RNA through high-throughput sequence analysis. Currently, detection methods making use of high-throughput sequencing datasets includes FusionCatcher, SOAPfuse, EricScript, *etc.* For validation of the detected chimeric RNA, the commonly used methods include PCR, RPA, agarose gel electrophoresis, sanger sequencing, *etc*. The development of newly introduced techniques has led to the discovery of different novel chimeric RNA, the third and fourth generation sequencing has also been developed and nearly mature, and the sequencing technology taking PacBio as an example has also brought a new dawn to the discovery of chimeric RNA, but each of them has its advantages and disadvantages, mainly focusing on its cost, false positive rate, detection time, *etc*. This paper basically describes various different techniques that can be utilized for the detection and validation of chimeric RNA.

Key words chimeric RNA, detection techniques, validation techniques, RNA-Seq, bioinformatics
DOI：10.16476/j.pibb.2023.0234