PacBio

应用说明

Kinnex 全长 RNA 试剂盒 用于异构体测序

引言

在真核生物中，可变剪接 (AS) 通过在同一基因中产生不同的外显子组合表达从而产生功能多样性。准确表征 AS 生成的全长转录本异构体对于生物学和疾病研究至关重要。即便使用最先进的计算工具，也无法通过短读长的 bulk RNA-Seq 全面解析异构体结构，因为 AS 的复杂性阻碍了准确的转录本组装（Stark 等人，2019）。使用 PacBio® 技术（Iso-Seq® 方法）的长读长 RNA-Seq，通过全长 cDNA 测序，避免了转录本组装的需求，并且在众多应用领域中获得了新发现（图 1）。

*Kinnex™ 全长 RNA 试剂盒*以总 RNA 作为起始样本，输出一个可直接用于测序的文库，与典型的 Iso-Seq 文库相比，其通量提高了 8 倍。PacBio 可与 SMRT® Link 软件中 Iso-Seq 分析相结合，实现经济、高效的异构体测序，无需正交测序验证。SMRT Link 软件可生成包含丰度信息的异构体分类报告，可供三级分析工具使用。

102-326-591 REV02 NOV2023

PacBio

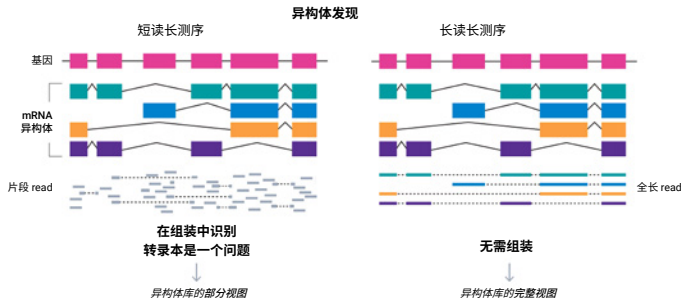

图 1. 长读长 RNA 测序无需进行转录本组装，因为转录本组装无法准确解析异构体结构。采用 PacBio（Iso-Seq 方法）的长读长 RNA-Seq 可以对整个全长 cDNA 进行测序，提供转录组的明确信息。

基于 Iso-Seq 方法的全长 RNA 测序

传统的 RNA-Seq 将 cDNA 片段化用于短读长测序 (100-200 bp)，然后必须通过计算方法来推断原始的转录本异构体。然而，由于可变剪切事件的复杂性，许多异构体具有高度相似的结构，推断出的转录本不准确。PacBio HiFi reads 可对全长 RNA 异构体进行测序，无需进行 cDNA 片段化和转录本组装（图 1），能够准确检测全长异构体。

HiFi 测序对于全长 RNA 测序的优势

- 从 5' 端到 3' 端全长异构体测序。
- 准确表征剪接位点。
- 可发现新基因和新异构体。
- 获得异构体 read 计数信息。

Iso-Seq 方法采用 PacBio HiFi 测序对全长转录本进行测序，该方法已应用于生物学和疾病研究等众多领域。Iso-Seq 方法已应用于人类疾病研究，旨在发现与罕见疾病、表型特征和神经系统疾病相关的异常剪接。在癌症研究中，Iso-Seq 方法已被用于发现引发癌症的突变、融合基因，以及可能用于癌症候选疫苗的新抗原表位（Li 等人，2023）。

Iso-Seq 方法还被应用于植物和动物研究，旨在建立高质量的基因组注释（Zhang 等人，2023），以及识别亲本特异性异构体表达（Wang 等人，2020）。

Kinnex 全长 RNA 试剂盒

*Kinnex 全长 RNA 试剂盒*采用 MAS-Seq 方法来提高 PacBio 长读长测序仪的通量。MAS-Seq 是一种将 cDNA 分子连接成长片段的串联方法（Al'Khafaji 等人，2023）。然后，通过串联分子测序产生的 HiFi reads 经过生物信息学拆分，即可还原原始 cDNA 序列。这种方法能够提高通量，减少测序需求，实现经济高效的异构体测序。

SMRT Link 中的 PacBio Iso-Seq 工作流程能够处理全长 cDNA 序列，然后根据参考注释（例如 GENCODE）对其进行分类，以识别新的基因和异构体。输出结果包括分类的全长异构体，其 read 计数可与三级分析软件兼容。

*Kinnex 全长 RNA 试剂盒*以总 RNA (300 ng) 作为起始样本，生成可用于测序的文库，工作流程需要两天。

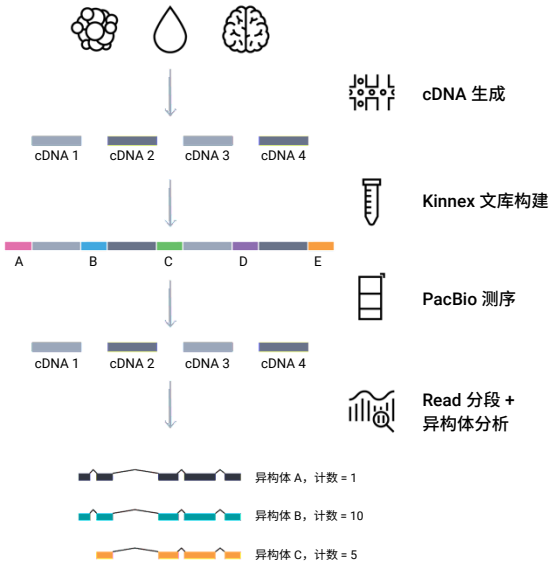

图 2. Kinnex 全长 RNA 测序。将全长 cDNA 分子串联成为长片段插入文库中并进行测序，然后使用 PacBio 软件数据处理。

102-326-591 REV02 NOV2023

Kinnex RNA 文库构建流程

Kinnex 全长 RNA 工作流程（图 3）以总 RNA 为起始样本，最终生成可用于测序的文库。

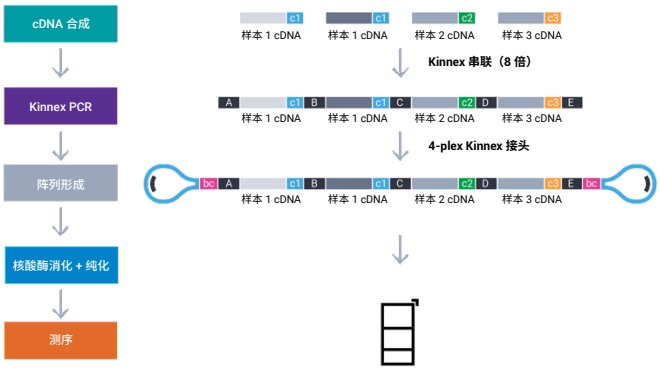

图 3. Kinnex 全长 RNA 文库构建流程。

全长 cDNA 分子合成方法（采用 *Iso-Seq Express 2.0 试剂盒*）与 Kinnex 阵列合成方法兼容。cDNA 条形码将作为 cDNA 扩增的一部分添加至 5' 端。由于部分阵列在核酸酶消化过程中被去除，Kinnex 接头连接可确保在 HiFi 测序之前富集完整阵列。带条形码的 cDNA 最多可支持 12-plex，而 Kinnex 接头在文库水平上支持 4-plex。

通过正确形成完整阵列和充分的测序，Sequel® II/IIe 和 Revio™ 系统上的一个 SMRT® Cell 预计分别能获得约 1500 万和约 4000 万个 cDNA 序列（表 1）。

Kinnex RNA 生物信息学工作流程

SMRT Link *Read segmentation 和 Iso-Seq* 工作流程（图 4）可处理 Kinnex 全长 RNA 文库生成的 HiFi reads，生成异构体分类，其 read 计数与三级分析工具兼容。

异构体聚类

FLNC reads 根据其测序相似性进行聚类，从而生成异构体共有序列。如果没有基因组序列信息，则此步骤是 Iso-Seq 分析的最后一步。

ID	基因	转录本	样本 1 TPM	样本 2 TPM
PB.1.1	ARGN	Novel	100	200
PB.1.2	ARGN	ENST00000379370.7	10	1.4
		...		
PB.10.1	ACTB	ENST00000646664.1	8132	8000

图 4. 使用 *Read segmentation 和 Iso-Seq* 工作流程开展 Kinnex 全长 RNA 分析。

比对

如果提供了基因组，上一步异构体共有序列可进行比对，并通过其外显子结构进一步折叠，以 GFF 文件形式生成异构体，实现可视化。

转录本分类

如果提供了注释（例如，Gencode），则使用 pigeon（PacBio 的 SQANTI3 方法）对异构体进行分类，识别已知和新的基因/异构体。Iso-Seq 工作流程可以联合分析混合的样本 reads，生成统一的异构体注释，其中包含每个样本的 read 计数，既有原始读数，也有归一化读数 (CPM)。

项目	性能
样本制备时间	2 天
预计文库大小	11,000-18,000 bp
靶标 P1 加载	60%–80%
预计 HiFi 产量	150–250 万条 HiFi reads (Sequel II/IIe) 400–600 万条 HiFi reads (Revio)
预计完整阵列百分比 (%)	80%-90%
预计 read 产量	约 1500 万条 reads (Sequel II/IIe) 约 4000 万条 reads (Revio)

表 1. 靶向 Kinnex 全长 RNA 文库性能

102-326-591 REV02 NOV2023

目前，SMRT Link 仅支持人类和小鼠样本的转录本分类。非人类/小鼠样本需要通过命令行运行自定义的注释 GTF 文件。

SMRT Link 注意事项

以下是执行 Iso-Seq 工作流程的一些常见注意事项及建议。

目前，SMRT Link *Read segmentation 和 Iso-Seq* 工作流程支持人类和小鼠参考基因组和注释，可生成具有 read 计数的分类异构体。如果要处理其他生物体，请参阅表 2 中的分析建议。

参考/注释	分析建议
人类或小鼠	采用带有预载人类/小鼠注释的 Iso-Seq 工作流程来获得定位的独特异构体，并且提供分类和 read 计数信息（FASTA、GFF、TXT）。
具备良好注释的模式生物	运行带有上传的参考基因组的 Iso-Seq 工作流程，比对后获得独特异构体（FASTA、GFF）。生成 符合 pigeon 标准的注释，并使用命令行进行带有 read 计数信息的异构体分类 (TXT)。
带有基因组的非模型生物	运行带有上传的参考基因组的 Iso-Seq 工作流程，比对后获得独特异构体（FASTA、GFF）。
无基因组	运行没有参考基因组的 Iso-Seq 工作流程，获得独特异构体 (FASTA)。

表 2. 基于参考基因组和注释信息的 Iso-Seq 数据分析建议。

尽管测序深度因实验目标和样本而异，但表 3 提供了普适性建议。

Kinnex 公共数据集发布

Kinnex RNA 数据集发布资料主要包括 HG002 细胞系和通用人类参考 RNA (UHRR) 组成，另外还包括来自 WTC-11 细胞系、人脑、高粱和小鼠（待发表）的样本作为对比。在经过 *Read segmentation 和 Iso-Seq* 工作流程后，每个样本均获得了超过 20,000 个独特基因（表 4）。转录本长度范围为 100 bp 至 11,000 bp，存在一些细微差异，这些差异似乎与样本和物种相关（图 5）。

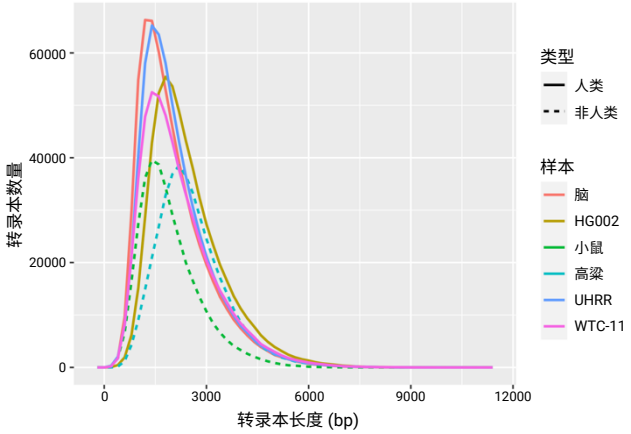

图 5. 不同 Kinnex 文库的转录本长度显示出样本和物种特异性差异，但大部分都在相同的大小范围内。

研究目标	中等至稀有转录本的异构体发现和定量	高表达转录本的异构体发现	物种的综合转录本注释
示例	疾病与正常组织多次重复对比	样本数量超过 20 个的疾病队列	具有多种组织类型的植物或动物
目标深度	每个样本 10M reads	每个样本 5M reads	每个样本 5M reads
文库	4-plex cDNA/1 个 Revio SMRT Cell 或 2-plex cDNA/1 个 SMRT Cell 8M	8-plex cDNA/1 个 Revio SMRT Cell，或 3-plex cDNA/1 个 SMRT Cell 8M	
分析	在 *Read segmentation 和 Iso-Seq* 工作流程中，可选择“pool reads and cluster together”，获得主要异构体分类文件，其中包含每个样本的全长 read 计数		

表 3. 基于不同研究目标的测序示例和分析建议。

102-326-591 REV02 NOV2023

样本	文库	HiFi read	S-read	平均 S-read 长度	已知基因	新基因	已知异构体	新异构体
UHRR	非 Kinnex – Sequel II/IIe	3,194,311	n/a	n/a	12,921	121	27,821	16,323
	Kinnex – Sequel II/IIe	2,720,033	20,453,853	1,918	18,903	1369	53,623	102,059
	Kinnex-Revio	6,546,645	47,250,258	1,914	22,365	4,223	68,087	231,467
HG002	HG002	5,984,046	38,740,671	2,227	18,230	8,448	55,689	281,460
WTC-11	第 0 天 - 重复 1	6,920,750	54,110,504	1,856	19,905	3,636	58,651	212,523
	第 0 天 - 重复 2	8,611,025	67,547,611	1,764	21,170	4,932	63,553	277,189
	第 0 天 - 重复 3	8,124,744	63,251,235	1,864	20,744	4,822	62,349	257,665
	第 1 天 - 重复 1	6,430,958	49,897,067	1,743	20,204	4,629	60,451	213,429
	第 2 天 - 重复 1	7,353,759	58,217,895	1,201	21,169	6,570	64,066	165,496
	第 3 天 - 重复 1	5,483,994	42,173,159	1,844	19,436	2,692	56,533	185,650
	第 3 天 - 重复 2	6,687,580	52,317,384	1,705	21,270	4,482	63,430	241,726
	第 4 天 - 重复 1	7,295,962	57,061,795	1,727	21,751	3,594	63,466	225,636
	第 5 天 - 重复 1	6,645,009	51,741,094	1,751	21,754	3,195	62,217	185,807
	第 5 天 - 重复 2	7,542,604	59,092,202	1,792	21,721	3,613	65,369	228,584
	第 5 天 - 重复 3	6,358,300	49,466,302	1,803	21,389	3,652	59,638	187,394

表 4. Kinnex 全长 RNA 数据集发布资料，包括 UHRR、HG002 和其他合作者的 WTC-11 细胞系。HiFi read 通过 SMRT Link v13.0 中的 Read segmentation 和 Iso-Seq 工作流程进行分析。所有样本均为 Kinnex 文库并在一个 Revio SMRT Cell 上进行测序，非 Kinnex 和 Kinnex 除外，它们在 Sequel II/IIe 系统上进行测序。根据 Gencode v39 注释，使用 pigeon 确定新基因和异构体。

将 Kinnex 与非 Kinnex 文库进行比较后发现，使用 Kinnex 串联或不同的测序平台时，转录本长度没有变化（图 6）。此外，异构体丰度基本保持一致（图 7）。

Kinnex 文库显示出较高的技术可重现性（表 5），与匹配 Illumina 数据中显示的技术可重现性一致（未显示）。

饱和曲线显示，reads 达到 1000 万条时，大多数已知基因和异构体都可以被检测到（图 9）。为了拟合模拟的饱和曲线，我们对 WTC-11 样本分别进行了 500 万条和 1000 万条 reads 的二次抽样，并将它们以不同的富集重数进行汇集，模拟在 Sequel II/IIe 和 Revio 系统上可获得的总产量。如预期的那样，3-plex 5M（Sequel II/IIe 系统上每个 SMRT Cell 8M 总共产出 15M 条 reads）检测到的异构体要少于 8-plex 5M（每个 Revio SMRT Cell 总共产出 40M reads）或者 4-plex 10M 文库（图 10）。

总体而言，饱和曲线和二次采样数据均表明，每个样本 10M 条 reads 时，有约 80% 的已知异构体可以被检测到。增加测序深度同时也会增加检测到的新异构体的数量，尽管大多数新发现的异构体丰度较低。

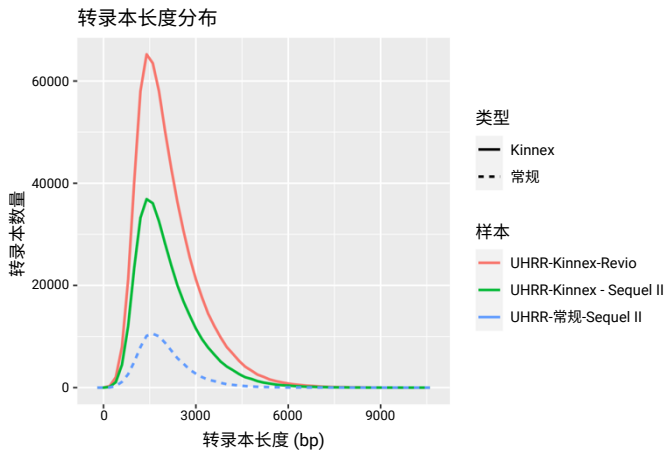

图 6. 对于同一 UHRR 样本，Kinnex（串联）和未串联文库的转录本长度分布并无差异。

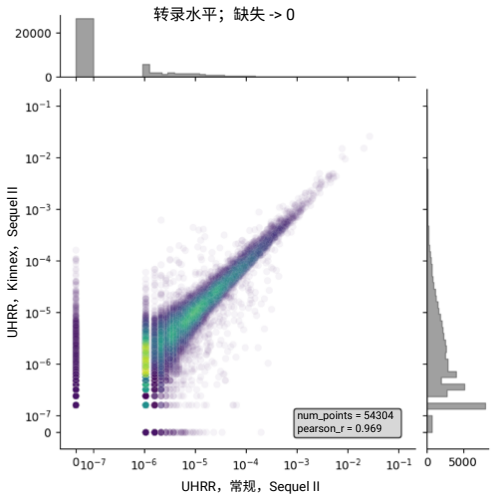

图 7. Kinnex 串联不会使异构体丰度出现偏移。在 Sequel II/e 系统上，Kinnex 和非 Kinnex UHRR 数据之间异构体丰度高度相关。我们仅比较了已知的 Gencode 异构体。与 Kinnex-Revio 数据（未显示）对比也观察到了类似的相关性 (0.964)。

WTC-11 第 0 天

	重复 1	重复 2	重复 3
重复 1	1.00	0.80	0.79
重复 2	0.80	1.00	0.81
重复 3	0.79	0.81	1.00

WTC-11 第 5 天

	重复 1	重复 2	重复 3
重复 1	1.00	0.80	0.80
重复 2	0.80	1.00	0.79
重复 3	0.80	0.79	1.00

表 5. Kinnex 文库具有良好的技术可重现性。WTC-11 样本的三个技术重复实验在第 0 天和第 5 天均显示出较高的异构体丰度相关性，这与匹配 Illumina 技术重复实验观察到的相关值（0.78-0.82，数据未显示）相似。

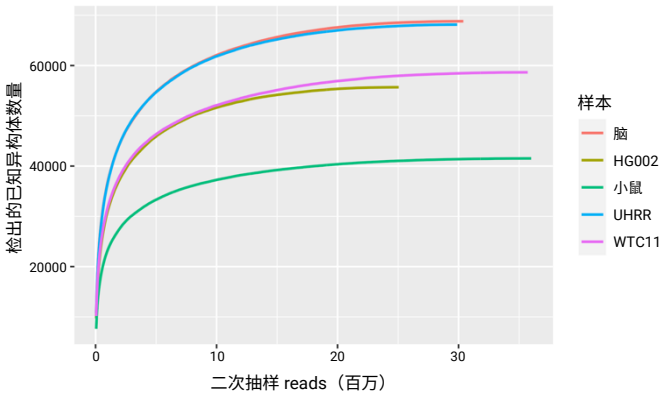

图 9. WTC-11 Kinnex 样本在异构体水平的饱和曲线。在 10M 条 reads 时，检测到了大多数已知的异构体。

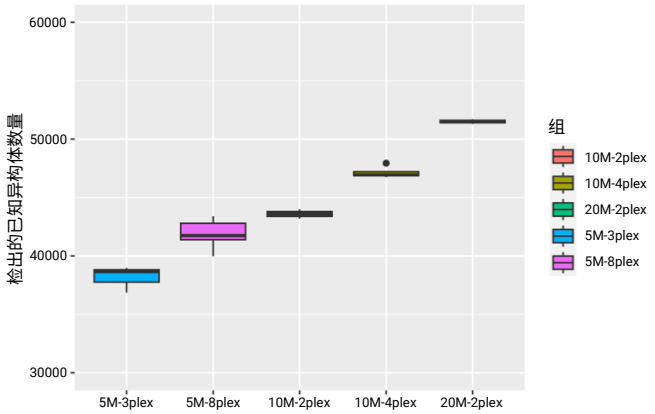

图 10. 在 5M 或 10M 条 reads 的深度模拟下，每个样本检测到的已知异构体的数量。总的汇集测序深度越高，每个样本的 reads 数越多，检测到的已知异构体数量就越多。检测到的新异构体数量范围在 30–70 k（对于 5M-3-plex、10M-2-plex、5M-8-plex）到 90–130 k（对于 10M-4-plex、20M-2-plex）。

102-326-591 REV02 NOV2023

分析 WTC-11 和 UHRR 样本生成的 Kinnex RNA 数据集的结果表明：

- 与常规 Iso-Seq 文库相比，Kinnex 文库不会改变检测到的转录本大小或丰度。
- 检测到的转录本大小因物种和样本类型而异，但在 Kinnex 文库中基本保持不变。
- 该技术制备的文库在 PacBio 长读长测序平台上具有较高的可重现性。
- 在 10M 条 reads 时可检测到大多数已知基因和异构体。检测到的新异构体数量随着测序深度的增加而增加，但会越发稀少。

*Kinnex 全长 RNA 试剂盒*与 SMRT Link 分析相结合，可提供高质量的全长 RNA，从而揭示疾病和生物学研究中的关键信息。

结论

可变剪切产生了复杂转录异构体，短读长 RNA-Seq 不能轻易读取。PacBio Iso-Seq 方法可对全长转录本进行高精度测序，让明确的异构体表征、等位基因特异性异构体表达、差异表达等分析成为可能。

*Kinnex 全长 RNA 试剂盒*采用 MAS-Seq 串联技术，可将通量提高 8 倍，结合灵活的多重策略和 SMRT Link 生物信息学工作流程，用户现在能够以经济高效的方式实现全面的异构体测序。

102-326-591 REV02 NOV2023

资源和参考

相关资源

应用简报 – 使用 Iso-Seq 方法 — 单细胞和批量 RNA 测序获得更完整的癌症转录组。

白皮书 – 用于人类疾病研究的批量和单细胞异构体测序。

应用说明 – 用于全长异构体测序的生物信息学工具。

Kinnex 全长 RNA 数据集：
https://pacb.com/datasets

Iso-Seq 文档：https://isoseq.how/

pigeon 文档：

https://isoseq.how/classification/

参考文献

Al'Khafaji, A. M., et al. (2023). High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, 1-5. https://doi.org/10.1038/s41587-023-01815-7

Li, Z., et al. (2023). An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *bioRxiv*, 2023-04. https://doi.org/10.1101/2023.04.21.536771

Pardo-Palacios, F., et al., (2023). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *bioRxiv*, 2023-05. https://doi.org/10.1101/2023.05.17.541248

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631-656. https://doi.org/10.1038/s41576-019-0150-2

Wang, B., et al. (2020). Variant phasing and haplotypic expression from long-read sequencing in maize. *Communications Biology*, 3(1), 78. https://doi.org/10.1038/s42003-020-0805-8

Zhang, R., et al. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-Seq analysis. *Genome Biology*, 23(1), 149. https://doi.org/10.1186/s13059-022-02711-0

102-326-591 REV02 NOV2023