新生物学丛书

CRC Press
Taylor & Francis Group

RNA-seq Data Analysis
A Practical Approach

RNA-seq数据分析实用方法

〔芬〕E. 科佩莱恩 〔芬〕J. 图梅拉 〔芬〕P. 萨默沃
〔瑞典〕M. 赫斯 〔芬〕G. 旺 编著

陈建国 张海谋 译

编辑推荐
Editor's Choice

科学出版社

新生物学丛书

RNA-seq 数据分析实用方法

〔芬〕E. 科佩莱恩　〔芬〕J. 图梅拉　〔芬〕P. 萨默沃
〔瑞典〕M. 赫斯　〔芬〕G. 旺　编著

陈建国　张海谋　译

科学出版社

北京

图字：01-2017-7081 号

内 容 简 介

本书全面介绍了 RNA-seq 数据分析的基本原理和方法，内容涵盖数据分析的整个工作流程，包括质量控制、作图、组装、统计检验和代谢途径分析等。书中在进行理论讲解的同时，还使用了较多实例，不仅生物信息学家，甚至没有相关分析经验的研究人员也均可参照这些实例进行分析。

本书是一部 RNA-seq 数据分析的实用参考书，可供生物学、医学、遗传学和计算机科学领域的研究人员阅读，也可作为相关专业的高年级本科生课程、研究生课程，以及短期培训班的教材。

图书在版编目（CIP）数据

RNA-seq 数据分析实用方法/（芬）E. 科佩莱恩（Eija Korpelainen）等编著；陈建国，张海谋译. —北京：科学出版社, 2018.3
（新生物学丛书）
书名原文：RNA-seq Data Analysis：A Practical Approach
ISBN 978-7-03-056486-3

Ⅰ. ①R… Ⅱ. ①E… ②陈… ③张… Ⅲ. ①基因组–序列–测试–研究 Ⅳ. ①Q343.1

中国版本图书馆 CIP 数据核字(2018)第 021066 号

责任编辑：王海光 高璐佳 / 责任校对：郑金红
责任印制：张 伟 / 封面设计：刘新新

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京京华虎彩印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2018 年 3 月第 一 版 开本：720×1000 1/16
2018 年 3 月第一次印刷 印张：15 3/4
字数：310 000
定价：120.00 元
(如有印装质量问题，我社负责调换)

《新生物学丛书》专家委员会

丛 书 序

当前，一场新的生物学革命正在展开。为此，美国国家科学院研究理事会于2009年发布了一份战略研究报告，提出一个“新生物学”（New Biology）时代即将来临。这个“新生物学”，一方面是生物学内部各种分支学科的重组与融合，另一方面是化学、物理、信息科学、材料科学等众多非生命学科与生物学的紧密交叉与整合。

在这样一个全球生命科学发展变革的时代，我国的生命科学研究也正在高速发展，并进入了一个充满机遇和挑战的黄金期。在这个时期，将会产生许多具有影响力、推动力的科研成果。因此，有必要通过系统性集成和出版相关主题的国内外优秀图书，为后人留下一笔宝贵的“新生物学”时代精神财富。

科学出版社联合国内一批有志于推进生命科学发展的专家与学者，联合打造了一个21世纪中国生命科学的传播平台——《新生物学丛书》。希望通过这套丛书的出版，记录生命科学的进步，传递对生物技术发展的梦想。

《新生物学丛书》下设三个子系列：科学风向标，着重收集科学发展战略和态势分析报告，为科学管理者和科研人员展示科学的最新动向；科学百家园，重点收录国内外专家与学者的科研专著，为专业工作者提供新思想和新方法；科学新视窗，主要发表高级科普著作，为不同领域的研究人员和科学爱好者普及生命科学的前沿知识。

如果说科学出版社是一个“支点”，这套丛书就像一根“杠杆”，那么读者就能够借助这根“杠杆”成为撬动“地球”的人。编委会相信，不同类型的读者都能够从这套丛书中得到新的知识信息，获得思考与启迪。

《新生物学丛书》专家委员会

主　任：蒲慕明

副主任：吴家睿

2012年3月

前　言

实用性

RNA 测序（RNA-seq）提供了关于转录组的前所未有的信息，但利用生物信息学工具来驾驭这种信息通常是一个瓶颈。本书的目的是使读者能够对 RNA-seq 数据进行分析。本书详细讨论了几个主题，涵盖整个数据分析工作流程，从质量控制、作图和组装到统计检验和代谢途径分析。本书的目的不是最小化与现有教材的重叠之处，而是进行更全面和更实用的介绍。

本书使研究人员能够考察基因、外显子和转录水平上的差异表达，发现新的基因、转录本和整个转录组。为了与非编码小 RNA 的重要调节作用一致，用整整一部分（第 12 章和第 13 章）来介绍非编码小 RNA 的发现和功能分析。

本书是专为学生和高级研究人员写的。实际的例子是以这种原则选择的：不仅生物信息学家，而且没有编程经验的实验室科学家都可以跟着这些例子去做，这使本书适合各种不同背景，包括生物学、医学、遗传学和计算机科学的研究人员。本书可以作为研究生课程、高年级本科生课程，以及暑期培训班的教材。它可以作为主要的 RNA-seq 数据分析方法及如何在实践中使用这些方法的指导手册。

本书在理论与实践之间进行了平衡，每一章以理论背景开始，然后是有关分析工具的描述，最后举例说明它们的用法。我们尽力使本书成为一部能够指导实践的 RNA-seq 数据分析操作指南。重要的是，它也满足对计算机不精通的实验室生物学家的需要，因为除了命令行工具之外，还使用图形化的 Chipster 软件给出了实例。在实例中使用的所有软件都是开放源代码和免费可用的。

内容概要

在第 1 章和第 2 章“引言”部分中，讨论了 RNA-seq 的不同应用，从基因和转录本的发现，到差异表达分析和突变及融合基因的发现。概述了 RNA-seq 数据分析，并讨论了实验设计的重要方面。

第一部分介绍了将读段作图到参考基因组和重新组装的方法。因为这两者都受读段质量的强烈影响，所以还包括关于质量控制和预处理的一章。第 3 章讨论了高通量测序数据特有的若干质量问题，以及用于检测和解决这些问题的工具。第 4 章介绍了将 RNA-seq 读段作图到参考基因组时所面临的挑战，利用实例介绍了一些常用的比对工具；还介绍了用于操作比对文件的工具和在基因组的上下文

中可视化读段的基因组浏览器。第 5 章介绍了转录组组装的一些要素；讨论了数据处理相关的步骤，如过滤、修剪和 RNA-seq 组装中的误差校正；解释了基本的概念，如剪接图、de Bruijn 图和组装图中的路径遍历等；讨论了基因组组装和转录组组装之间的区别。此外，介绍了重构全长转录本的两种方法：基于作图的组装和重新组装，这两种方法都用实例进行了演示。

第二部分主要致力于统计分析，主要利用 R 软件来进行，辅以 Bioconductor 项目所开发的工具。第 6 章讨论了不同的定量化方法和工具，以及基于注释的质量指标。第 7 章介绍了以 R 和 Bioconductor 为基础的 RNA-seq 数据分析的框架，以及如何导入数据；讨论了 R 中的统计学工具和生物信息学工具之间的主要区别。第 8 章和第 9 章讨论了分析基因、转录本及外显子差异表达的不同选项，演示了如何使用 R/Bioconductor 工具和一些单独的工具来进行分析。第 10 章提供了用于注释结果的解决方案，第 11 章介绍了产生信息性可视化效果的不同方式，以显示重要的结果。

本书的最后一部分集中于分析非编码小 RNA，使用基于 web 的或可免费下载的工具。第 12 章描述了非编码小 RNA 的不同类别，刻画了其功能、丰度和序列属性。第 13 章描述了从下一代测序数据集中发现非编码小 RNA 的不同算法，提供了一个实用的方法，带有工作流程和例子，来介绍非编码小 RNA 是如何被发现和注释的；此外，还介绍了可以用来阐明非编码小 RNA 功能的下游工具。

编著者

致　谢

我们感谢 CRC 出版社的工作人员给我们这个机会为 RNA-seq 领域写一本教科书。尤其是 Sunil Nair、Sarah Gelson 和 Stephanie Morkert 在写作过程中引导我们，对我们表现出无限的耐心，并对我们的每次咨询做出快速而及时的回应。

我们还感谢同事和本实验室的成员 Vuokko Aarnio、Liisa Heikkinen 及 Juhani Peltonen 阅读和评论了本书的各个章节，深深感谢他们为此付出的时间和精力。

Tommy Kissa 在本书的最终写作阶段作为助理提供了坚定的和无条件的热心帮助。他把这个工作当作最令人愉快的任务的态度激励了我们，使我们能够获得最终的结果。

最后，我们感谢我们的配偶和家庭成员 Lily、Philippe、Stefan、Sanna 及 Merja，在整个写作过程中，她们除了担任旅馆老板、厨师、女佣和心理治疗师的角色，还充当了研究助理、审稿人、图形艺术家、计算机支持和被征询者。我们亲切地将这部作品献给你们。

编著者

作 者 简 介

E. 科佩莱恩（Eija Korpelainen）

芬兰 CSC-IT 科学中心的生物信息学家，在提供国家级的生物信息学支持方面有十几年的经验。她的团队开发了 Chipster 软件，为芯片和下一代测序数据的分析及可视化工具的全面集合提供了一个用户友好的平台。她还在芬兰和其他国家开设了几门培训课程。

J. 图梅拉（Jarno Tuimala）

芬兰红十字会血液服务中心和 CSC-IT 科学中心的生物学家，致力于生物统计学和生物信息学，也是赫尔辛基大学生物信息学兼职教授。他在使用 R 软件和高通量数据分析方面有十多年的经验。

P. 萨默沃（Panu Somervuo）

2000 年在芬兰赫尔辛基科技大学获得博士学位，从事信号处理、自动语音识别和神经网络的相关研究，后转向生物信息学领域。近 6 年来，他一直参与赫尔辛基大学的微阵列和测序研究项目。

M. 赫斯（Mikael Huss）

在计算生物学方面拥有超过十年的从业经验。2007 年曾在新加坡基因组研究所从事高通量测序的生物信息学博士后研究工作，后来在瑞典国家生命科学实验室（SciLifeLab）的测序机构工作，并设计了 RNA-seq 分析工作流程。目前在 SciLifeLab 的瓦伦堡高级生物信息学机构（Wallenberg Advanced Bioinformatics Infrastructure，WABI）担任“驻站生物信息学家”，WABI 是承担生物信息学分析项目的国家机构。

G. 旺（Garry Wong）

东芬兰大学分子生物信息学教授，澳门大学生物医学教授。在使用、开发 RNA 转录组分析工具方面有十多年的经验，目前的研究重点是利用生物信息学和功能基因组学的工具阐明模式生物中非编码小 RNA 的功能。

目　录

第 1 章　RNA-seq 简介

1.1　引　　言

RNA-seq 指的是用来确定生物样品中 RNA 序列的身份（identity）和丰度（abundance）的实验方法和计算方法的集合。因此，存在于一个单链 RNA 分子中的每个腺嘌呤、胞嘧啶、鸟嘌呤和尿嘧啶核糖核酸残基的顺序被确定。实验方法涉及从细胞、组织或整个动物样品中分离 RNA，制备代表样品中的 RNA 种类（species）的文库，文库的实际化学测序，以及随后的生物信息学数据分析。RNA-seq 与早些时候的方法（如微阵列）最重要的区别是：当前的 RNA-seq 平台的通量非常高，新技术提供了更高的灵敏度，发现新型转录本、基因模型和非编码小 RNA 的能力更强。

RNA-seq 方法由测序技术的代际变化衍生而来。第一代高通量测序通常指 Sanger 双脱氧测序法。由于毛细管电泳被用来解决核酸片段长度的问题，一个标准的毛细管电泳实验可以使用 96 个毛细管，得到的序列长度为 600~1000 个碱基，产生大约 100 000 个碱基的序列。第二代测序，也称为下一代测序（next-generation sequencing，NGS），是指使用类似的测序方法，通过个别核苷酸的合成化学测序，但以大规模并行的方式执行，以便使单次测序实验中的测序反应数目可以达到数百万。一个典型的 NGS 实验可以包含 100 个核苷酸（nt）的 6000 兆测序反应，产生 6000 亿碱基的序列信息。

第三代测序的方法也是大规模并行的，并使用合成化学测序（sequencing by synthesis chemistry），但用单个分子的 DNA 或 RNA 作为模板。第三代测序平台每次实验的测序反应较少，为几百万数量级，但每个反应的序列长度可以更长，可以轻松地对 1500 nt 范围内的序列进行测序。

从 RNA-seq 实验获得的数据可以产生新的知识，从胚胎干细胞中编码蛋白质的新转录本的鉴定到皮肤肿瘤细胞株中过表达的转录本的表征。可以问的问题包括：正常细胞和癌细胞中基因表达水平有什么差异？在缺少一种抑癌基因的细胞株中基因表达水平会发生什么变化？诱变剂处理前后在细胞株中基因表达有什么差异？在大脑发育过程中哪些基因被上调？什么转录本存在于皮肤，但不存在于肌肉中？在氧化应激过程中基因剪接是如何改变的？在人类胚胎干细胞样本中，我们能够发现什么新颖的 miRNA？大家可以看到，可以问的问题