生物信息学数据分析丛书

深度测序数据的生物信息学分析及实例

Bioinformatics for Deep Sequencing Data with Examples

沈百荣　主编

科学出版社

生物信息学数据分析丛书

深度测序数据的生物信息学分析及实例

沈百荣　主编

科学出版社

北京

内 容 简 介

本书几乎涵盖了深度测序数据分析及应用的各个方面，适用于从事深度测序数据分析研究的技术人员和学者。在本书中，不仅可以了解到深度测序技术应用的领域，还可以通过具体实例，了解到不同软件的相关算法、原理及使用方法，以帮助选择适合自身研究和应用所需要的深度测序数据分析的解决方案。

本书适合从事生物信息学、系统生物学、医学信息学、转化医学、精准医学、健康管理等研究领域的读者阅读。

图书在版编目（CIP）数据

深度测序数据的生物信息学分析及实例/沈百荣主编. —北京：科学出版社, 2017.9

（生物信息学数据分析丛书）

ISBN 978-7-03-054580-0

Ⅰ. ①深… Ⅱ. ①沈… Ⅲ. ①生物信息论 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字(2017)第 231042 号

责任编辑：李 悦 刘 晶 / 责任校对：郑金红

责任印制：张 伟 / 封面设计：北京图阅盛世设计有限公司

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

北京教图印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2017 年 9 月第 一 版 开本：B5 (720×1000)

2017 年 9 月第一次印刷 印张：13 3/4

字数：271 000

定价：88.00 元

(如有印装质量问题, 我社负责调换)

《深度测序数据的生物信息学分析及实例》
编辑委员会

前　言

近年来，以快速、低成本、高通量为特点的深度测序（又称下一代测序，next generation sequencing，NGS）技术极大地推动了相关科学和产业的进步，是未来精准医疗和健康产业的基石。深度测序产生了海量的数据，需要新的、专业的技术、方法和软件来分析与处理。目前，国内外已有大量优秀的研究人员发表了针对深度测序数据分析的新方法和新软件的论文。但是，国内外全面介绍深度测序数据分析及实例的书籍尚不多见。本书的编写目的就是为不同专业背景的读者提供一本实用的关于深度测序数据分析的书籍。

本书几乎涵盖了深度测序数据分析及应用的各个方面，适用于从事深度测序数据分析研究的技术人员和学者。在本书中，不仅可以了解到深度测序技术应用的领域，还可以通过具体实例，了解到不同软件的相关算法、原理及使用方法，以帮助选择适合自身研究和应用、学习所需要的深度测序数据分析的解决方案。同时，我们构建了本书配套的网站以方便读者进行实例学习，网址为 http://sysbio.suda.edu.cn/NGS_book/index.php。

本书共包括 11 章。第 1 章主要介绍了深度测序技术的常用平台和原理、对现代生物医学研究范式的影响、对生物信息学带来的挑战和机遇，以及深度测序数据分析的常见软件和平台；第 2 章介绍了深度测序相关的数据库和数据格式；第 3 章介绍了碱基识别的方法；第 4 章介绍了基因组序列比对；第 5 章介绍了序列片段的组装；第 6 章介绍了染色质免疫共沉淀测序数据分析；第 7 章介绍了转录组测序数据的分析；第 8 章介绍了 microRNA-Seq 的数据分析；第 9 章介绍了变异检测；第 10 章介绍了单细胞测序数据分析；第 11 章介绍了深度测序数据的可视化软件。本书的编写工作是苏州大学系统生物学研究中心师生多年来共同努力的结果，由于 NGS 领域发展迅速，且我们的时间和学识有限，难免有错误与不当之处，还希望读者反馈指正，我们将在以后再版时进行修改和更正。

本书各章的编写分工如下：前言及第 1 章，沈百荣、钱福良、李庆辉、汤溢飞；第 2 章，吴文涛；第 3 章，王晶；第 4 章，尚婧；第 5 章，张文宇；第 6 章，李庆辉、荆鑫华；第 7 章，严文颖、林宇鑫、汤溢飞；第 8 章，林宇鑫、李粤；第 9 章，崔卫荣、严文颖、蒋峻峰；第 10 章，张文宇；第 11 章，李吟、汤思捷。网站由林宇鑫、刘行云、严文颖开发。

本书内容曾在苏州大学生物信息学本科专业 2007 级、系统生物学硕士专业

2010 级和 2011 级同学中讲授过，感谢当时参加学习和讨论的同学。最要感谢的是中国科学院上海生命科学研究院/上海交通大学医学院健康科学研究所研究员荆清及其课题组的师生，他们在 2009 年就开始与我们共同合作分析 NGS 数据，使我们较早了解该领域，并开始这方面的工作。此外，特别感谢科学出版社的李悦编辑对我们工作的耐心鼓励和支持。

沈百荣

2017 年 7 月

目　录

1 深度测序技术与生物信息学

内容提要：本章主要介绍深度测序的常用平台和原理、深度测序技术对现代生物医学研究范式乃至对社会的影响，讨论生物信息学处理深度测序数据所面临的机遇和挑战，最后对深度测序数据分析的常见软件和平台作简单介绍。

1.1 深度测序的常用平台

近年来，随着多国政府个人基因组计划的启动（例如，2014 年英国启动的“十万人基因组计划”、2015 年美国和中国政府分别启动的“百万人基因组计划”等），基因测序行业逐渐进入了大众视野，深度测序或下一代测序技术已经形成了巨大的产业和市场，是继 Sanger 测序和基因芯片技术之后发展起来的新兴产业。测序的普及使得测序产业的竞争异常激烈，深度测序平台也随时间推移在不断发展更新，市场已从三大主要平台（Illumina、454、ABI）演变到 Illumina 一家独大的局面。2014 年，Illumina 公司被美国权威杂志《麻省理工科技评论》（*MIT Technology Review*）评为“全球创新企业 50 强”第一名，超越了苹果、谷歌等科技巨头。现在 Illumina 占据了基因测序仪市场 70%的份额，其他厂商凭借自身的特点和优势分享余下的份额。在收录的数据方面，以 PUBMED SRA（http://www.ncbi.nlm.nih.gov/sra/）数据库为例，有来自不同测序平台和技术的数据，如 Roche 454 GS System、Illumina Genome Analyzer、Applied Biosystems SOLiD® System、Helicos Heliscope、Complete Genomics、Pacific Biosciences SMRT 等。下面对目前常用的测序平台进行介绍。

1.1.1 Illumina 测序系统

1999 年，Illumina 公司只是一家拥有 25 人的小公司，主要销售传统的微阵列芯片，这种芯片可以检测设计好的特定位点的变化。2007 年，Illumina 以 6 亿美元收购基因测序公司 Solexa。之后 Illumina 逐步从几大测序公司并立的局面中成长为最大的测序公司，Solexa 的基因测序技术比竞争对手快百倍，且价格低廉。Illumina 公司的新一代测序平台非常丰富，有号称测序工厂的 HiSeq X Ten；有 HiSeq X Five、HiSeq 4000、HiSeq 3000、HiSeq 2500；还有小通量灵活型的 MiSeq、针对临床诊断使用的 NextSeq 等。这些测序平台就测序通量和测序成本等方面而言，基本上覆盖了所有的应用及需求层面，这足以说明 Illumina 公司是目前测序

市场产品线最丰富、最强大的公司。

1.1.1.1 HiSeq X Ten

“人类基因组计划”、“曼哈顿计划”、“阿波罗登月计划”是人类自然科学史上的三大计划。其中，“人类基因组计划”耗费了约 30 亿美元，由全球几大主要国家的顶尖科学家参与，耗时 15 年，测定了人类染色体的 30 亿对碱基，完成了第一份人类基因组图谱。随着技术的发展，测序速度和价格发生了惊人的变化。

2014 年初，Illumina 公司在第 32 届摩根大通保健大会上重磅推出了目前最强大的测序仪——HiSeq X Ten。这一套测序系统包括 10 台 HiSeq X 测序仪，适合群体规模的测序项目。HiSeq X Ten 测序平台是当前全球测序能力最强、通量最高的测序平台，也是全球第一款将个人全基因组测序成本降到了 1000 美元以内的平台。

HiSeq X Ten 以 Illumina 成熟的边合成边测序技术为基础，采用了多个先进的设计特点产生超高的通量。HiSeq X Ten 包含有数十亿个纳米孔的流动槽，新的簇生成试剂让数据密度显著增加。利用最先进的光学设备和更高效快速的试剂，HiSeq X Ten 能够比以往更快地测序。每台 HiSeq X 仪器 3 天可产生 1.8Tb 的数据，即每天 600Gb。若同时运行 10 台仪器，人们每年可测序>18 000 个人类基因组（按照每个人的全基因组 30 倍的覆盖深度计算）。

每台 HiSeq X 仪器可运行单流动槽或双流动槽，支持 2×150bp 的读长。在 3 天的时间内，双流动槽的每次运行可产生 1.6~1.8Tb 数据（6 billion SE reads），而单流动槽可产生 800~900Gb 数据（3 billion SE reads）。在 2×150bp 读长时，75%以上的碱基都高于 Q30。在最新版的试剂（V2.5），Q30 最高可达到 90%以上，单个 lane 的数据产量最高可达到 140~150Gb。随着试剂及机器软件的研发升级，单台机器的单位数据产量会越来越高。HiSeq X Ten 在测序市场上的数据产量能力当之无愧排名第一。

目前，HiSeq X Ten 测序系统需 10 台起售，售价 1000 万美元起，国内采购 10 台及相关的配套设备需超过 1 亿人民币的投入。目前全球一共有 10 余家拥有 HiSeq X Ten 的用户，包括科研机构、第三方服务机构、医院等。虽然 HiSeq X Ten 的测序能力非常强大，但是 Illumina 公司为了保护其他产品线的正常运转，对 HiSeq X Ten 的用处做了很大的限制，将其测序的范围限定在只能应用于人类全基因组测序。

1.1.1.2 HiSeq X Five

HiSeq X Five 是由 Illumina 公司在 2015 年的摩根大通保健大会推出的测序系统，是 2014 年推出的顶尖平台 HiSeq X Ten 的缩小版本，包括 5 台 HiSeq X 测序仪。

对于很多测序中心而言，HiSeq X Ten 的价格太高，通量太大。因此，Illumina 公司推出了售价约为 600 万美元的 HiSeq X Five。HiSeq X Five 的推出是为了让那些资金有限的客户也能够享受到 1000 美金的测序福利。HiSeq X Ten 系统在年初上市后反响热烈，大大超乎 Illumina 的预期。起初，Illumina 表示将供应给 5 名客户，但到年底，HiSeq X 测序仪共售出 201 台，客户数达到 18 名。据统计，HiSeq X Five 每年的测序通量超过 9000 例人类全基因组，每个基因组的成本费用大约在 1400 美元。

1.1.1.3 HiSeq 4000

HiSeq 4000 也是 Illumina 公司在 2015 年 1 月份的摩根大通保健大会上带来的新产品。HiSeq 4000 是基于成熟的 HiSeq 2500 系统开发的具有双流动槽的测序仪，但没有了 HiSeq 2500 的“快速运行模式”。HiSeq 4000 的售价为 90 万美元，能够在 3.5 天内测序 12 个基因组、100 个转录组或 180 个外显子组。

与之前的版本相比，HiSeq 3000/4000 的重大改进是流动槽设计。早期的 HiSeq 仪器使用非图案化的流动槽，这样测序簇可在表面的任何地方形成。因此，测序簇的大小、性状和间隔不均匀，而数据分析的步骤之一就是确定它们在哪里。新的仪器（包括 HiSeq X 系列）采用了图案化的流动槽，让测序簇限制在 400nm 的孔中。这种有序的结构使得簇间隔均匀和特征大小均一，便于准确分辨以极高密度成簇的流动槽，使得通量大幅提高。不过，需要注意的是，HiSeq 3000/4000 与 HiSeq 2500 的硬件系统不同，故无法从 HiSeq 2500 直接升级到 HiSeq 3000/4000，两者的试剂也不能混用。

在 HiSeq 4000 系统的每次运行中，最多测序 12 个人类全基因组（30 倍覆盖度）且时间不到 3 天。该系统每次运行最多可测序 180 个外显子组（假定每个外显子组 4 Gb）或 100 个人类全转录组（假定每个样品 5000 万条序列）。

1.1.1.4 HiSeq 3000

和 HiSeq 4000 系统一样，HiSeq 3000 是基于成熟的 HiSeq 2500 系统研发的，且都是在 2015 年 1 月份发布的新测序系统。与 HiSeq 4000 不同的是，HiSeq 3000 只有单个的流动槽。HiSeq 3000 的售价为 74 万美元（比 HiSeq 4000 便宜约 16 万美元），通量为 HiSeq 4000 的一半。

基于成熟的 HiSeq 2500 系统，凭借创新的图案化流动槽技术，HiSeq 3000/4000 系统带来了无可匹敌的速度和性能。双流动槽的 HiSeq 4000 系统提供了最高的通量和每个样品的较低价格，适合广泛的应用。而单流动槽的 HiSeq 3000 系统则享有较低的机器价格和快速运行时间。

在 HiSeq 3000 系统的每次运行中，最多测序 6 个人类全基因组（30 倍覆盖度）

且时间不到 3 天。系统每次运行最多可测序 90 个外显子组（假定每个外显子组 4 Gb）或 50 个人类全转录组（假定每个样品 5000 万条序列）。当需要的时候，HiSeq 3000 可以直接升级到 HiSeq 4000。

1.1.1.5 HiSeq 2500

HiSeq 2500 系统是一台强大而高效的超高通量测序系统，支持最广泛的应用和研究规模。利用 Illumina 成熟的边合成边测序（SBS）原理，无可匹敌的数据质量让 HiSeq 2500 成为全球大型基因组中心和领先机构的首选仪器。新推出的 HiSeq v4 试剂可以在更短时间内获得更多读取和更多数据。HiSeq 2500 适用于生产规模的基因组、外显子组、转录组测序等多种应用。

HiSeq 2500 系统有两种特有的运行模式——快速运行模式和高产量运行模式，能够同时处理一个或两个流动槽。这提供了一个灵活、可扩展的平台，支持最广泛的测序应用和研究规模。在快速运行模式和高产量运行模式中选择，以使得可扩展的产量能满足客户的项目需求。新的试剂能在高产量模式下产生高达 1 Tb 的数据。高产量模式沿用 HiSeq 2000 的运行方式，单次运行可以产出 600G，特别适合样品量较多，或需要最深度覆盖的应用。快速运行模式最多能在一天之内产生 100G 左右的数据量，还能提供 2×150bp 的读长，有助于改善 de novo 应用的组装效果。

1.1.1.6 MiSeq 系列

MiSeq 是行业最准确且最易用的台式测序仪之一，快速简约，适用于小型基因组、扩增子、靶向基因嵌板（panel）、16S rRNA 测序等。新的 MiSeq 试剂以 25 M 测序 reads 和 2×300bp 读长实现了高达 15Gb 的产量，且适用于更多的应用如外显子组、mRNA 测序、靶向基因表达、宏基因组学和 HLA 分型。

MiSeq 是唯一一台在单次运行中产生 2×300bp 双端 reads 和高达 15Gb 数据的台式测序仪。这实现了小型基因组的组装或目标变异的准确检测，特别是在均聚物区域。如今，更多的样品也能在较少时间内处理，同时在每次运行中产生比以往任何版本更多的读段（reads）。所有这些都在“从样品到数据”流程最短的台式测序仪上成为了现实。

通过 MiSeq，可以在单次运行中对多达 96 个样品进行多重分析，获得更高效率；实现准确的双向扩增子测序；产生更完整的 de novo 组装效果。

另外，值得一提的是 MiSeqDx，该仪器是第一台也是唯一一台经过 FDA 批准的体外诊断（IVD）下一代测序（NGS）系统，专为临床实验室的环境而设计。MiSeq Dx 仪器拥有约 0.3m^2 的占地面积、易于上手的流程，以及专为临床实验室的需求而定制的数据产量。此外，整合的软件具有样品追踪、用户可追溯性及结

果解释等功能。利用 Illumina 成熟的边合成边测序（SBS）技术，MiSeqDx 仪器提供了准确而可靠的筛查和诊断检测。

MiSeqDx 仪器上运行的分析采用简单的三步过程，从人类外周血标本中提取出的基因组 DNA（gDNA）开始，通过添加引物、生成带索引的文库、制备测序用的 gDNA 样品，进行同时捕获和扩增；文库可添加到 MiSeqDx 流动槽中，并上样到 MiSeqDx 仪器进行测序。

为了确保系统的正确使用，MiSeqDx 仪器装有 Illumina User Manager Software（用户管理软件）和 MiSeq Operating Software（操作软件）。前者让实验室可控制和追踪系统访问，确保只有经过授权的人员才能运行检测；后者控制 MiSeqDx 仪器，让测序过程自动化，并减少用户的手工操作时间。

1.1.1.7 NextSeq 系列

在2014年和2015年的摩根大通保健大会上，Illumina公司分别推出了NextSeq 500 和 NextSeq 550。

HiSeq X Ten 定位为工厂规模的测序仪，通量超高，价格不菲，而 NextSeq 500 则旨在以 MiSeq 的大小提供 HiSeq 的性能。NextSeq 500 系统集高通量测序的性能和台式测序仪的简约为一体，是目前唯一一款可实现外显子组、转录组和全基因组测序的台式测序仪。它可在两种模式下开展测序实验：高产量（high output）和中等产量（medium output），在单次运行中可获得 20~120Gb 的数据，为用户带来广泛的应用灵活性。

与 MiSeq 一样，NextSeq 500 的整个流程也非常简单。制备好的文库可直接上样到系统。整合的簇生成实现了单分子的自动化克隆扩增。此系统将簇生成、边合成边测序（SBS）和碱基检出整合在单台 NGS 系统中。凭借成熟的 SBS 技术，NextSeq 500 系统带来了行业领先的测序准确性，其中 75%以上的测序碱基都高于 Q30。

NextSeq550 系统整合了高通量测序和芯片扫描功能。在测序方面，NextSeq 550 的测序模块与 NextSeq 500 完全相同；在芯片扫描方法上，它目前支持 Infinium CytoSNP-12、Infinium CytoSNP-850K 和 Infinium Human Karyomap-12 三款芯片，适用于细胞遗传学和生殖健康。NextSeq 550 系统的价格为 27.5 万美元。

1.1.2 Roche 454 测序仪

2005 年底，*Nature* 杂志报道 454 公司推出了革命性的基于焦磷酸测序法的超高通量基因组测序仪 Genome Sequencer 20 System，并分别于 2007 年推出了 Genome Sequencer FLX System、2008 年推出了 GS FLX Titanium 系列试剂和软件，让 GS 测序仪在通量增加的同时，准确性和读长进一步提高。Roche 454 测序仪的特点是

该仪器的最大读长能够达到1000bp；拥有一键式数据处理和分析软件使得包括重测序基因组拼接、参考基因组比对及变异分析等在内的生物信息分析变得非常简单。

GS FLX系统的流程包括以下几个步骤：

（1）样品片段化：GS FLX系统支持各种不同来源的样品，这一步将包括基因组DNA、PCR产物、BAC、cDNA、小分子RNA等在内的样品打断成300~800bp的片段。

（2）文库制备：借助一系列标准的分子生物学技术，将A和B接头（3′端和5′端具有特异性）连接到DNA片段上。接头也将用于后续的纯化、扩增和测序步骤。具有A、B接头的单链DNA片段组成了样品文库。

（3）磁珠链接和纯化：单链DNA文库被固定在特别设计的DNA捕获磁珠上。每一个磁珠携带了一个独特的单链DNA片段。磁珠结合的文库被扩增试剂乳化，形成油包水的混合物，这样就形成了只包含一个磁珠和一个独特片段的微反应器。

（4）乳液PCR扩增：每个独特的片段在自己的微反应器里进行独立扩增，在保证没有其他的竞争性或者污染性序列影响的同时，整个片段文库的扩增平行进行。扩增后原先每条序列产生了数百万个相同的拷贝。

（5）扩增序列的读取：携带DNA的捕获磁珠随后放入PTP板中进行后续的测序。PTP孔的直径（29μm）只能容纳一个磁珠（20μm）。然后将PTP板放置在GS FLX中，测序开始。放置在4个单独的试剂瓶里的4种碱基，依照T、A、C、G的顺序依次循环进入PTP板，每次只进入一个碱基。如果发生碱基配对，就会释放一个焦磷酸。这个焦磷酸在ATP硫酸化酶和萤光素酶的作用下，经过合成反应和化学发光反应，最终将萤光素氧化成氧化萤光素，同时释放出光信号。此反应释放出的光信号被仪器配置的高灵敏度CCD实时捕获。有一个碱基和测序模板进行配对，就会捕获到一分子的光信号；由此一一对应，就可以准确、快速地确定待测模板的碱基序列。

（6）数据分析：GS FLX系统在10h的运行之后产生了数亿个碱基信息。GS FLX系统提供两种不同的生物信息学工具对测序数据进行分析：从头拼接和基因组的重测序。

GS FLX系统的准确率在99%以上。其主要限制来自同聚物，即是相同碱基的连续掺入，如AAA或GGG。由于没有终止元件阻止单个循环的连续掺入，同聚物的长度就需要从信号强度中推断出来。这个过程就可能产生误差。因此，454测序平台的主要错误类型是插入-缺失。

据454官方网站（http://454.com/products/gs-flx-system/index.asp）介绍，该测序仪主要的应用有以下几种：

（1）全基因组测序：Roche 454测序仪的长读长特性给大基因组、复杂基因组的重头测序带来了便利。