**O’reilly**

**生物信息学数据技能**

**使用开源工具进行可重复且可靠的研究**

**文斯布法罗**

**生物信息学数据技能**

学习将大型测序数据集转化为可重复且稳健的生物学发现所需的数据技能。为了这个现实的目标，你应该学习如何使用免费开源的工具从大型且复杂的生物数据集中提取有意义的内容。

**目录**

[绪论 3](#_Toc49046228)

[使用这本书的方法 5](#_Toc1833717195)

[为什么这本书关注测序数据 5](#_Toc747078268)

[观众 5](#_Toc1955049914)

[不同级别的数据技能 5](#_Toc2024105498)

[这本书的假设 5](#_Toc852652759)

[github上的补充资料 5](#_Toc376544082)

[计算资源和设置 5](#_Toc2089562112)

[本书的构成 5](#_Toc1470336993)

[代码约定 6](#_Toc859515322)

[本书的约定 6](#_Toc1899007132)

[代码使用示例 6](#_Toc710905810)

[在线资源 6](#_Toc1742420409)

[如何联系我们 6](#_Toc1772803571)

[确认 6](#_Toc1321499319)

[第一部分 意识形态：强大且可重复的生物信息学的数据技能 6](#_Toc1163177159)

[一、 如何学习生物信息学 6](#_Toc974872672)

[第二部分 先决条件：开始生物信息学项目的基本技能 6](#_Toc1532255341)

[二、 建立和开展生物信息学项目 6](#_Toc2139105010)

[三、 补充Unix Shell知识 6](#_Toc914168643)

[四、 使用远程机器 6](#_Toc1334372263)

[五、 git科学家 6](#_Toc622898620)

[六、 生物信息学数据 6](#_Toc74327215)

[第三部分 实践：生物信息学数据技能 6](#_Toc1529503598)

[七、 Unix 数据工具 6](#_Toc987716996)

[八、 快速介绍r语言 7](#_Toc510960462)

[九、 使用范围数据 7](#_Toc2072864128)

[十、 使用序列数据 7](#_Toc194015)

[十一、 使用对齐数据 7](#_Toc1113326458)

[十二、 生物信息学shell脚本，编写管道和并行化任务脚本 7](#_Toc652763295)

[十三、 补充知识Tabix 和SQLite 7](#_Toc1646230189)

[十四、 总结 7](#_Toc11478575)

[术语表 7](#_Toc1794365442)

[参考书目 7](#_Toc787128873)

[索引 7](#_Toc775702991)

**绪论**

“当我开始学习生物信息学时，我想先读哪本书?”这本书是对我在两年前对我自己提出的这个问题的回答。当我开始在这个领域工作时，我只有python和r方面的编程经验，其他方面的经验很少。我四处寻找关于生物信息学的极好的入门教材，虽然我找到了一些好书，但大多数都不是针对我作为生物信息学家的日常工作所需要的。我找到的一些文本从理论和算法的角度探讨了生物信息学，包括Smith-Waterman比对、系统发育重建、motif查找等主题。尽管它们很吸引人去阅读(并且我确实建议您多阅读这些材料)，但我没有必要在我的日常生物信息学工作中从头开始实现生物信息学算法——这些算法已经存在大量出色的、高度优化的、经过良好测试的实现。其他生物信息学文本采取了更为明确的方法，指导不熟悉计算的读者完成每一步任务，比如运行对齐器或从数据集中下载序列。虽然这些更适用于我的工作，但是，相对而言，他们的书中的大部分内容是过时的。

正如你所猜到的那样，我找不到比这版的《生物信息学数据技能》更好的生物信息学书了。这本书面向那些不确定如何弥合脚本语言和实践生物信息学之间的巨大差距以便以一个健壮的和可复用的方式回答科学问题的读者。要弥合这一差距，必须学习数据技能 - 这种方法使用一套核心工具来操纵和探索您在生物信息学项目中遇到的任何数据。

数据技能是学习生物信息学的最佳方法，因为这些技能使用经过时间检验的开放源码工具，这些工具仍然是操作和探索不断变化的数据的最佳方法。这种方法经受住了时间的考验:高通量测序的出现迅速改变了生物信息学领域，但熟练的生物信息学使用这些相同的工具和技能来适应了这一新数据。毕竟，下一代数据只是数据（差异数据，以及更多数据），而生物信息学大师拥有通过将他们的工具应用于这些新数据来解决问题的基本技能。编写《生物信息学数据技能》是为了帮您学习这些核心工具，帮助您发展相同的技能。

**使用这本书的方法**

许多生物学家从生物信息学起步，往往把“学习生物信息学”等同于“学习如何运行生物信息学软件”。

**为什么这本书关注测序数据**

**观众**

**不同级别的数据技能**

**这本书的假设**

**github上的补充资料**

**计算资源和设置**

**本书的构成**

**代码约定**

**本书的约定**

**代码使用示例**

**在线资源**

**如何联系我们**

**确认**

第一部分 意识形态：强大且可重复的生物信息学的数据技能

1. 如何学习生物信息学

第二部分 先决条件：开始生物信息学项目的基本技能

1. 建立和开展生物信息学项目
2. 补充Unix Shell知识
3. 使用远程机器
4. git科学家
5. 生物信息学数据

第三部分 实践：生物信息学数据技能

1. Unix 数据工具
2. 快速介绍r语言
3. 使用范围数据
4. 使用序列数据
5. 使用对齐数据
6. 生物信息学shell脚本，编写管道和并行化任务脚本
7. 补充知识Tabix 和SQLite
8. 总结

当我打算开始写《生物信息学数据技能》时，起初，我为如何以一种不会在快速发展的生物信息学领域迅速过时的方式，以书的形式呈现中等水平的生物信息学而苦苦挣扎。甚至在我完成这本书的时间里，新的闪亮算法、统计方法和生物信息学软件已经被生物信息学社区发布和采用。有可能(甚至可能)新的测序技术将再次给生物学带来革命性的变化，而生物信息学将需要调整它们的方法和工具。在这个不断变化的环境中，纸质书如何成为一项有价值的学习资产?

我通过每天在工作中使用的工具-Unix，Python和R-找到了这个问题的答案。Unix的时间可以追溯到20世界70年代，它已经有40多年的历史了。Python诞生于1991年，R出现于1993年，他们都有超过20年的历史了。这些工具都经受住了时间的考验，并且是现代数据处理和统计计算的基础。生物信息学和Unix有着几乎不可分割的历史——学习生物信息学技能首先需要学习的是Unix。在基因组学快速发展的同时，生物信息学家继续寻求同样的标准工具来解决新问题和分析新数据集。此外,Unix，Python和R都是可扩展的工具。几乎所有新的生物信息学程序都可以用于Unix命令行。工作期间，使用最新的生物信息学和统计学方法归根结底都是下载和安装新的Python和R包。这本书中庸到的其他工具-GenomicRanges，sqlite3、sam-都是在Unix，Python和R运行的数据工具包。

然后去哪？

术语表

参考书目

索引