

公共选修课课程介绍

课程名称	生命大数据与 R 语言：数据驱动的生命科学探索		总学分：2 总学时：34		其中	理论	34	
						实验	0	
开课院部	生命科学与技术学院	教研室	微生物与合成生物学	教学对象	不限			
教材名称	R for Data Science	主编	Wickham H	年/版	2023 第 2 版	出版社名称/自编	O'Reilly Media	

课程简介（课程的背景、教学目的、主要内容等，不少于 600 字，）：

课程背景

随着生命科学、医学和药学的迅速发展，生物数据的规模和复杂性日益增加，高通量测序、基因组学、蛋白组学等领域的大数据分析需求急剧上升。R 语言作为一种功能强大、灵活且广泛应用的数据分析工具，在生物信息学、统计分析、数据可视化及机器学习等领域发挥着关键作用。然而，许多生命科学相关专业的学生由于缺乏编程基础，难以高效地利用 R 语言进行数据分析。

主要内容

本课程针对生命科学、医学、药学及相关交叉学科的本科生，特别面向具有实验学科背景但缺乏编程基础的学生，旨在系统介绍 R 语言在生物信息学中的应用，培养学生的数据分析和科研思维。课程采用理论讲授与实践操作相结合的教学方式，涵盖 R 语言基础、数据处理与统计分析、高通量测序数据分析、生物数据可视化、机器学习与多组学整合、科研数据管理及生物信息学伦理 等核心内容，帮助学生掌握生物数据分析的基本方法并具备独立开展数据分析工作的能力。

教学目的

总目标：

本课程面向生命科学、医学、药学及相关交叉学科的学生，旨在系统讲解生物信息学的基本理论、核心方法和数据分析技能。通过理论讲授与实践操作结合，培养学生掌握计算工具、生物数据处理和可视化分析的能力，并能独立完成转录组数据分析等核心任务。同时，课程融入国家生物信息战略，增强学生的科研诚信意识、数据安全意识和科技自立自强精神，为未来科研与行业应用奠定基础。

五个课程子目标：

1. 掌握生物信息学基础知识与技能
 - 了解生物信息学的基本概念、学科定位及其在生命科学中的应用价值。
 - 熟悉基因组学、转录组学、蛋白质组学等核心研究领域，掌握基础的数据分析方法。
 - 了解计算机基础知识，包括计算机系统原理、Linux 操作、R 语言及数据可视化工具的基本使用。
2. 提升生物数据分析能力（方法与实践）
 - 学习生物数据的基本处理方法，如数据库检索、序列数据格式解析、基于 R 语言的云计算环境的简单应用。
 - 具备基础的数据分析能力，能够完成 RNA-seq 数据的基本处理，包括数据预处理、简单的差异表达分析和功能分析。
 - 掌握数据可视化的基础方法，能够使用 R 语言或其他工具呈现和解读生物数据。
3. 培养科研思维与解决实际问题的能力（综合应用）
 - 通过案例学习和实践操作，培养学生运用数据解决生物学问题的基本能力。
 - 具备使用 R 语言进行简单数据分析的能力，并理解其在生命科学研究中的作用。
 - 能够对常见的生物数据分析结果进行合理解释，具备基础的科研思维。
4. 增强国家科技自立自强意识与数据安全伦理意识（思政目标）
 - 结合我国生物信息学发展历程，引导学生认识生物数据主权的重要性，增强科技自立自强意识。
 - 通过国家生物信息战略、数据安全法规案例，让学生理解数据合规、伦理道德和科研诚信的核心价值。
 - 引导学生关注精准医学、新药研发等领域的社会价值，培养其责任感和科研使命感，促进科技报国精神的形成。

讲授提纲（每一章节的名称）

第 1 章：绪论

1. R 语言的历史与发展
2. R 语言在数据分析、生物信息学、大数据科学中的应用
3. R 语言与其他编程语言的对比
4. 课程目标、学习路径及评估方式

第 2 章：R 语言环境配置

1. Windows/macOS 环境下安装 R 和 RStudio
2. RStudio 界面介绍与个性化设置
3. 线上 R 分析环境的获取与使用
4. 练习数据集的获取与管理
5. 课堂上机操作

第 3 章：R 语言基础

1. R 语言的三种运行模式（交互模式、脚本模式、项目管理）
2. R 语言的基本语法与变量类型
3. R 语言函数：定义、调用、参数传递与返回值
4. 获取帮助文档及错误、警告信息解析

第 4 章：R 语言扩展功能（R 包管理）

1. R 包的概念、分类及应用场景
2. R 包的安装、更新与管理（CRAN、Bioconductor、GitHub）
3. R 包文档的获取与阅读
4. 常见 R 包介绍
5. 课堂上机操作

第 5 章：数据处理与导入导出

1、数据格式与存储方式

- 结构化与非结构化数据的区别
- R 语言中的数据类型（数据框、矩阵、列表）
- 数据存储方式（文本文件、Excel、数据库、RData）

2、数据导入方法

- 读取文本文件（`read.table()`、`read.csv()`）
- 读取 Excel 文件（`readxl`、`openxlsx`）
- 读取 RData 格式数据（`load()`、`save()`）
- 读取 SQL 数据库（DBI、RSQLite）

3、数据清理与预处理

- 处理缺失值（`na.omit()`、`impute()`）
- 处理异常值（箱线图分析、标准化）
- 数据转换与标准化（`scale()`、`mutate()`）

4、数据转换与合并

- 数据筛选与子集提取 (`subset()`、`dplyr::filter()`)
- 数据合并 (`merge()`、`dplyr::left_join()`)
- 长数据与宽数据转换 (`tidyr::pivot_longer()`、`pivot_wider()`)

5、数据导出

- 导出文本与 CSV 文件 (`write.table()`、`write.csv()`)
- 导出 Excel 文件 (`writexl::write_xlsx()`)
- 保存 RData 格式数据 (`save()`、`saveRDS()`)

第 6 章：数据可视化

1、数据可视化基础

- 数据可视化的意义与作用
- 数据可视化在生物医药领域的应用（精准医学、公共健康监测、基因组数据可视化等）
- R 语言中的数据可视化工具概述（base R、`ggplot2`、`ComplexHeatmap` 等）

2、`ggplot2` 数据可视化

- `ggplot2` 的基本概念与语法（`aes` 映射、美学属性）
- `ggplot2` 数据层、几何对象（`geom_point`、`geom_bar`、`geom_line` 等）
- `ggplot2` 的主题、坐标轴、图例调整

3、高级数据可视化

- 单变量数据可视化（直方图、密度图、箱线图、小提琴图）
- 双变量及多变量数据可视化（散点图、回归分析、分组条形图、点图、热图）
- 复杂生物数据可视化（PCA、相关性分析可视化、维恩图、火山图、复杂热图）

第 7 章：高级数据分析（3 学时）

- 相关性分析
- 维恩图绘制与解读
- 数据聚类方法（层次聚类、K-means、t-SNE）
- 变量间关系探索

第 8 章：R 语言在 Bulk RNA-seq 数据分析中的应用（7 学时）

1、RNA-seq 数据解析与预处理

- RNA-seq 数据格式解析（FASTQ、Counts、TPM）
- RNA-seq 数据标准化与质量控制

2、差异表达分析（DEA）

- DESeq2、edgeR、limma 及其应用
- 差异基因筛选、表达矩阵处理

3、功能富集分析

- PCA（主成分分析）在转录组数据中的应用
- GSEA（基因集富集分析）
- GO/KEGG（基因本体及信号通路富集分析）

4、结果可视化

- 火山图（Volcano plot）展示差异表达基因
- 热图（Heatmap）展示基因表达模式

GSEA、GSVA、GO、KEGG 等富集分析图

考核方式或评分标准（笔试、论文、实际操作考察等）：

序号	成绩类别	考核方式	考核要求	评价权重（%）
1	过程性考核（平时成绩）	考勤 10%、课堂表现 10%、小组任务 10%和作业 10%	课堂提问、讨论、小组案例分析作业、编程作业	40%
2	期中作业	项目报告	代码编写、数据分析报告撰写	10%
3	期末成绩	客观题、项目报告	对重点知识点进行客观题考察、完成期末项目分析报告	50%

任课教师简介（不少于 50 字）：

夏云，工学博士，中国药科大学生命科学与技术学院副教授，独立 PI，研究生导师。研究领域涵盖生物医学大数据整合、新药发现、肿瘤免疫及生物标志物识别。团队开发诊断与预后模型，构建肿瘤与免疫分析数据库，并申请多项国内外专利。2024 年在 Nature Immunology（封面文章）、Briefings in Bioinformatics、Thyroid、iMeta 等期刊发表论文。曾任职于华中科技大学科学技术发展院，具专利代理师资格，擅长生物医药领域的专利挖掘与成果转化。