# 课程大纲
# COURSE SYLLABUS

| 1. | 课程代码/名称<br>Course Code/Title | **生物医学组学数据分析 / Biomedical Omics Data Analysis** |
|---|---|---|
| 2. | 课程性质<br>Compulsory/Elective | 专业选修课 / Elective |
| 3. | 课程学分/学时<br>Course Credit/Hours | 3/48 |
| 4. | 授课语言<br>Teaching Language | 英文或中英 / English or Bilingual Teaching |
| 5. | 授课教师<br>Instructor(s) | 帅世民（医学院）/ Shimin Shuai (School of Medicine) |
| 6. | 是否面向本科生开放<br>Open to undergraduates<br>or not | 否 / No |
| 7. | 先修要求<br>Pre-requisites | 本科水平的分子生物学、遗传学、细胞生物学或近似生物学课程。推荐有一定的概率论与数理统计知识、任意一门编程语言的入门经验，但不必须。<br>College–level molecular biology, genetics, cell biology or similar. Knowing basic probability theory, statistics, and any entry–level programming is recommended but not mandatory. |
| 8. | 教学目标<br>Course Objectives | |

从人类基因组 2001 年问世以来，相关组学技术快速发展。面对测序获得的庞大数据，如何有效分析是不少科研人员在实际研究中经常遇到的问题。《生物医学组学数据分析》专注于组学技术在生物医学研究中的应用。本课程强调实际操作，旨在教会学生基本的组学数据分析流程。在学完本课程后，学生能掌握基本的 Unix 环境和 R 语言编程技巧，并能使用 Shell 和 Bioconductor 来初步处理课题中碰到的常见组学数据（全基因组/外显组测序、RNA-Seq、单细胞 RNA-Seq、ChIP-Seq、ATAC-Seq 等）；或者能更好地知晓自己的分析需求，从而与生物信息学家能更高效地沟通。除此之外，本课程还希望教会学生常见的数据分析方法背后的大致原理（不涉及算法细节），让学生了解相关数据分析步骤背后的理由和逻辑。最后，本课程还希望教会学生如何更好地设计组学实验从而让之后的分析变得轻松、让结论更可靠。

Since the advent of the human genome in 2001, related omics technologies have developed rapidly. Facing the huge data obtained by sequencing, how to effectively analyze them are a problem that many researchers often encounter in daily research. *Biomedical Omics Data Analysis* focuses on the applications of omics technologies in biomedical research. This course also emphasizes hands-on experience with the aim to teach students the basic omics data analysis workflow. After completing this course, students are expected to master the basic Unix environment and R language programming skills and can use Shell and Bioconductor to process common omics data encountered in their thesis research (whole genome/exome sequencing, RNA-Seq, Single-cell RNA-Seq, ChIP-Seq, ATAC-Seq, long-read sequencing etc.); or can better understand their analysis needs, so that they can communicate with bioinformaticians more efficiently. In addition, this course also aims to teach students the general principles behind common data analysis methods without getting into too many algorithmic details, so that students can understand the reasoning and logics behind each analysis step. Finally, this course also hopes to teach students how to better design omics experiments to make subsequent analysis easier and make conclusions more reliable.

| | | |
|---|---|---|
| | | |
| **9.** | 教学方法<br>**Teaching Methods** | |
| | 课堂讲授、随堂演示、课后数据分析作业、期末项目（包含项目报告和 PPT 演示）。<br>Lectures, in-class demo, programming assignments, and final projects with written report and presentation. | |
| **10.** | 教学内容<br>**Course Contents** | |
| | **Section 1** | Overview of omics technologies / 组学技术概述<br><br>    This section includes a general overview of the history, classification, and application of modern omics technologies in biomedical research. |
| | **Section 2** | Introduction to common omics data format / 常见组学数据格式简介<br><br>    This section teaches data formats commonly used in omics data analysis, such as FASTA, FASTQ, BAM/SAM/CRAM, GTF, BED, Wig/Bigwig, VCF and TSV/CSV etc. |
| | **Section 3** | Basics of the Unix environment / Unix 环境基础<br><br>    This section introduces the Unix environment, such as how to log into computing clusters, manipulate files, execute scripts, manage permissions, and install softwares etc. |
| | **Section 4** | Basics of R and Bioconductor / R 语言和 Bioconductor 基础<br><br>    This section introduces the base R programming and how to use Bioconductor to install and manage R packages related to omics data analysis. |
| | **Section 5** | RNA-Seq data analysis (I) / RNA 测序数据分析（I）<br><br>    This section starts with introduction of the RNA sequencing technology and then we will demonstrate how to analyze RNA-Seq data from FASTQ files to aligned data. |
| | **Section 6** | RNA-Seq data analysis (II) / RNA 测序数据分析（II）<br><br>    This section will continue the RNA-Seq analysis workflow. We will introduce commonly used downstream analyses such as differential expression analysis, gene set over representation and enrichment analysis etc. |
| | **Section 7** | Single-cell RNA-Seq data analysis / 单细胞 RNA 测序数据分析 |

| | This section will introduce the general scRNA-Seq data analysis workflow (data import, quality control, PCA/UMAP/tSNE, clustering, marker gene discovery etc.) based on R package Seurat. |
|---|---|
| **Section 8** | WGS&WES data analysis / 全基因组和全外显组测序数据分析<br><br>This section will introduce DNA sequencing data analysis, focusing on sequence alignments, variant calling, variant annotation, and association analysis etc. |
| **Section 9** | ChIP-Seq data analysis / ChIP-Seq 数据分析<br><br>This section will introduce ChIP-Seq data analysis for transcription factors and histone modifications. We will focus on how to align and clean data, call and annotate peaks, find enriched sequence motifs etc. |
| **Section 10** | ATAC-Seq data analysis / ATAC-Seq 数据分析<br><br>This section covers the general data analysis workflow for ATAC-Seq, including alignment, QC, peak finding, peak annotation, motif finding etc. We will also compare ATAC-Seq to ChIP-Seq data analysis. |
| **Section 11** | Final project proposal presentation / 期末项目计划展示<br><br>Each student will be asked to prepare a short (~5 mins) presentation showing the dataset and research question they will use in their final projects. |
| **Section 12** | Long-read sequencing data analysis / 长读长测序数据分析<br><br>This section focuses on the emerging long-read technologies (Nanopore and PacBio). We will introduce the benefits of long reads and challenges of their analysis. We will also cover common tools used in long-read data analysis. |
| **Section 13** | Multi-omics data integration / 多组学数据整合<br><br>This section focuses on how to integrate different types of omics data computationally and demonstrates examples of such integrations in literatures. |
| **Section 14** | Applications of omics in medical research / 组学在医学研究中的应用<br><br>This section shows active research areas that use different omics technologies. Examples of recent studies will be discussed to demonstrate the extensive usage of omics in biomedical research. |

| Section 15 | Final project presentation (I) / 期末项目展示（I）<br><br>Students will be asked to give a presentation (10–15 mins per person) about their final projects, including background, methods, results and discussions. |
|---|---|
| Section 16 | Final project presentation (II) / 期末项目展示（II）<br><br>Students will be asked to give a presentation (10–15 mins per person) about their final projects, including background, methods, results and discussions. |

| 11. | 课程考核<br>**Course Assessment** |
|---|---|

1. 考核形式 Form of examination：本课采取期末项目报告+期末项目 PPT 展示的方式考核。The form of examination will be final project report and presentation.
2. 分数构成 Grading policy
   - 出勤 / Attendance 10%
   - 课堂参与 / Participation 10%
   - 课后作业 / Assignments 40%
   - 期末项目报告+PPT 展示 / Final project report and presentation 40%

| 12. | 教材及其它参考资料<br>**Textbook and Supplementary Readings** |
|---|---|

No mandatory textbook for this course. We will recommend reading materials for each section during the course.

Recommended resources for Unix and R:
- Basic introduction to the Unix environment: www.ee.surrey.ac.uk/Teaching/Unix
- Basic R concept tutorials: www.r-tutor.com/r-introduction

Recommended book for statistics:
- Modern Statistics for Modern Biology: free at https://www.huber.embl.de/msmb/

Recommended book for bioinformatics:
- The Biostar Handbook: not free at https://www.biostarhandbook.com/