天津医科大学理论课教案首页

课程名称：生物信息学　　**课程内容/章节：**第五章（5.3）Galaxy 分析平台

教师姓名：伊现富　　**职称：**讲师　　**教学日期：**2020 年 5 月 29 日 8:00-9:40

授课对象：2017 级基础班（本）　　**听课人数：**18

授课方式：理论讲授　　**学时数：**2　　**教材版本：**生物信息学：基础及应用

教学目的与要求（分掌握、熟悉、了解、自学四个层次）：

- 掌握 Galaxy 分析平台的基本使用方法。
- 熟悉 Galaxy 分析平台；数据处理的基本策略。
- 了解基因组功能注释分析平台。
- 自学 Galaxy 分析平台的高级使用技巧。

授课内容及学时分配：

- (5’) 回顾与导入：回顾基因组注释的基础知识和高级注释，介绍基因组功能注释分析平台。
- (10’) Galaxy 分析平台：介绍 Galaxy 分析平台、Galaxy 中的常用工具集和主界面、Galaxy 的相关资料。
- (30’) Galaxy 的基本使用：通过坐标转换、格式转换、坐标逻辑运算等实例演示、讲解 Galaxy 的基本使用方法。
- (40’) Galaxy 的综合运用：通过实例演示 Galaxy 在生物信息学工作中的综合运用，介绍 Galaxy 的高级使用技巧。
- (5’) 总结与答疑：总结授课内容中的知识点与技能，解答学生疑问。

教学重点、难点及解决策略：

- 重点：Galaxy 分析平台的使用。
- 难点：Galaxy 分析平台的使用。
- 解决策略：通过实例的逐步演示，详细讲解 Galaxy 的使用方法与技巧。

专业外语词汇或术语：

工作区（work area）　　属性（attribute）
历史面板（history panel）　　工作流（workflow）

辅助教学情况：

- 多媒体：Galaxy 分析平台的界面。
- 板书：数据处理的主要步骤。
- 操作演示：Galaxy 分析平台的使用。

复习思考题：

- Galaxy 分析平台的基本使用方法。
- 以坐标转换为例，论述“输入-加工-输出”的工作流程。

参考资料：

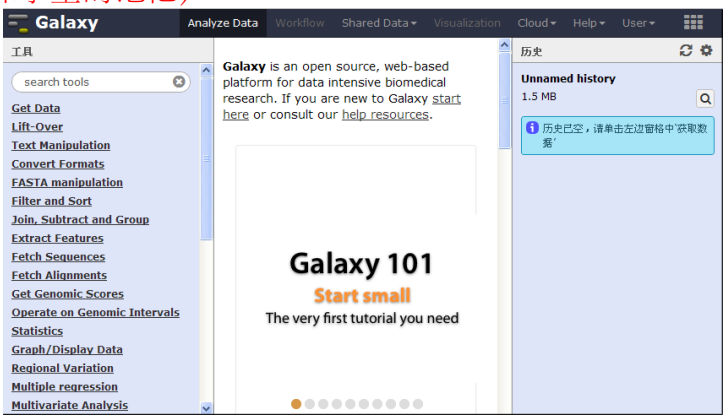
- Galaxy
- 维基百科

一、回顾与导入（5 分钟）

1. 基因组注释
   - 基础知识：基因组组装版本、坐标系统、常用格式、坐标的逻辑运算
   - 高级注释：变异位点的注释、基因集的富集分析、序列标识
2. 生物信息学分析平台：Galaxy，GenePattern，…

二、Galaxy 分析平台（10 分钟）

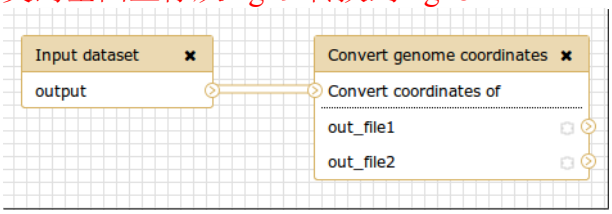
1. 主界面（通过讲解每部分的具体功能加深学生的理解）
   - 顶部是刊头：切换“分析数据”、“工作流”和“帐号”等主界面
   - 左侧栏是工具菜单：以工具集的形式组织罗列着各种工具
   - 中间是工作区：工具参数设置、使用说明和数据内容、属性等信息的输出位置
   - 右侧栏是历史面板：以历史记录的形式记录存储着每一步操作
2. 工具集（展示工具集中的具体工具，加深学生的记忆）
   - Get Data：从公共数据库提取数据
   - Text Manipulation：处理文本数据
   - Convert Formats：数据格式转换
   - Operate on Genomic Intervals：坐标的逻辑运算
   - Statistics 和 Graph/Display Data：统计绘图
   - NGS Toolbox：分析第二代测序数据
   - ……



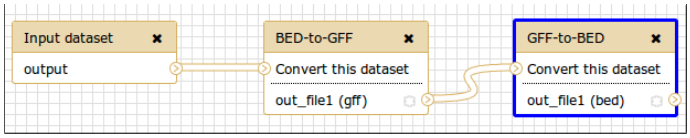
3. 学习资料（先易后难，由浅入深）
   - Galaxy 101
   - Galaxy Screencasts and Demos
   - Shared Pages, Histories & Workflows
   - Learn Galaxy
   - Galaxy Wiki

三、 **【重点、难点】**Galaxy 的基本使用（30 分钟）

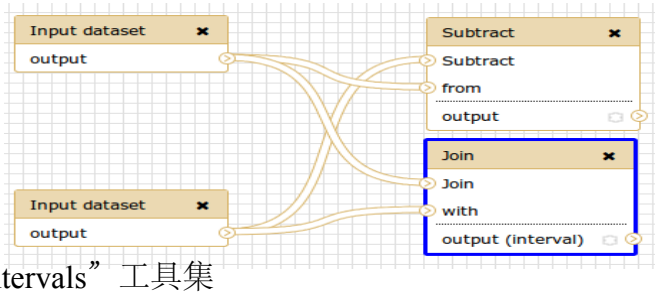
1. 坐标转换：使用集成到 Galaxy 中的 liftOver 把人类的基因坐标从 hg19 转换到 hg18
   - 获取输入。输入文件：hg19 的基因坐标
   - 数据处理。设置参数：hg19 ⇒ hg18
   - 保存输出。过滤结果：MAPPED vs. UN-MAPPED
   - 坐标转换的常用工具
     - liftOver：支持 BED 和“chrN:start-end”格式的输入
     - Galaxy 中的 liftOver：支持 BED、GFF 和 GTF 格式的输入
     - NCBI Remap：支持 BED、GFF、GTF 和 VCF 等格式的输入
     - Ensembl assembly converter：支持 BED、GFF、GFT 和 PSL 格式的输入，但输出都是 GFF 格式的
     - pyliftover：仅支持点坐标（point coordinates）的转换，无法对区段（ranges）坐标进行转换



2. 格式转换：使用 Galaxy 实现 BED 与 GFF 格式的互转
   - 获取输入。输入文件：BED
   - 数据处理。格式互转：BED ⇒ GFF；GFF ⇒ BED
   - 保存输出。查看结果：互相比较



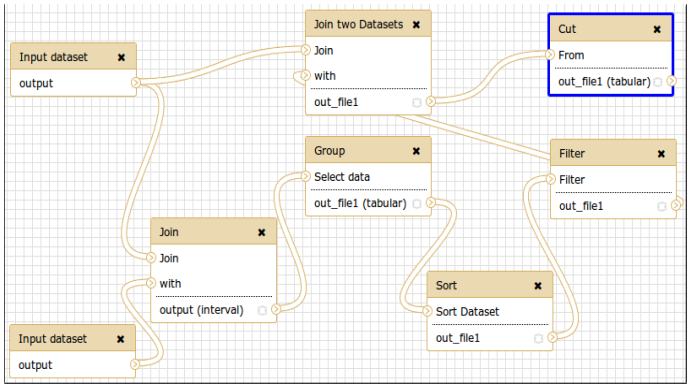
3. 逻辑运算：使用 Galaxy 进行外显子与 SNP 的比较
   - 获取输入。输入文件：exon，SNP
   - 数据处理。不含 SNP 的 exon：subtract；含有 SNP 的 exon：join
   - 保存输出。解析结果（一个 exon 上可能含有多个 SNP）
   - 逻辑运算的常用工具
     - Galaxy 中的“Operate on Genomic Intervals”工具集
     - bedtools: a powerful toolset for genome arithmetic
     - BEDOPS: the fast, highly scalable and easily-parallelizable genome analysis toolkit



四、 **【重点、难点】**Galaxy 的综合运用（40 分钟）

以寻找 Y 染色体上含有 SNP 数目最多的外显子为例进行操作演示：

1. Input: Getting exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Joining exons with SNPs
3. Group: Counting the number of SNPs per exon
4. Filter: Filtering exons that have ten or more SNPs
5. Compare two Datasets: Recovering exon info
6. Visualize: Display data in genome browser
4. Sort: Sorting exons by SNPs count
4. Select first: Selecting top ten
5. Join[Join two Datasets]: Recovering exon info



扩展介绍“工作流”的思想及其优势，以及 Galaxy 中工作流的提取、制作、使用和分享：

1. Save: rename the history as ”Exons and SNPs”
2. Workflow: extract workflow from history
3. Modify: open workflow editor and modify the parameter
4. Rerun: run workflow on whole genome data
5. Share: share or publish workflow
6. Create: create workflows from scratch (e.g. Find the 50 longest exons)

五、总结与答疑（5 分钟）

1. 知识点
   - Galaxy 分析平台：界面，学习，使用
2. 技能（数据处理的“输入-处理-输出”三段论）
   - 获取输入：格式、来源、过滤
   - 数据处理：工具、版本、参数
   - 解析输出：格式、注释、解析