

天津医科大学实验课教案首页

(共 3 页、第 1 页)

| | | |
|----------------------------|-------------------------|---------------------------------|
| 课程名称：生物信息学 | 实验名称：基于 Galaxy 的基因组数据处理 | |
| 教师姓名：伊现富 | 职称：讲师 | 教学日期：2014 年 4 月 30 日 14 时 -17 时 |
| 授课对象：生物医学工程学院 2011 级生信班（本） | 实验人数：27 | |
| 实验类型（验证型、综合型、设计型、创新型）：验证型 | 实验分组：一人一机 | |
| 学时数：3 | 教材版本：生物信息学实验讲义（自编教材） | |

实验目的与要求：

- 掌握基因组注释中常用的 BED 格式。
- 掌握基因组坐标的逻辑运算模式。
- 掌握 Galaxy 的基本使用方法。

实验内容及学时分配：

- (10') BED 格式：回顾 BED 格式使用的坐标系统及其每一列的含义。
- (10') 逻辑运算：回顾交集、减法、联合等逻辑运算模式。
- (10') Galaxy 简介：简单介绍 Galaxy 分析平台的主界面、工具集及学习资料。
- (120') 实验操作：寻找人类基因组中 22 号染色体上含有 SNP 数目最多的外显子。

主要仪器和实验材料：

- 实验材料：人类基因组（hg19）中 22 号染色体（chr22）上的外显子和 SNP。
- 主要仪器：联网的计算机。
- 分析工具：Galaxy 分析平台。

实验重点、难点及解决策略：

- 难点：基因组坐标的联合运算；解决策略：通过实例进行讲解。
- 重点：Galaxy 的使用；解决策略：根据资料进行学习，通过练习熟练掌握。

思考题：

- BED 格式使用的哪一类坐标系统？
- BED 格式每一列的含义是什么？
- 如何进行基因组坐标的联合操作？

参考资料：

- Galaxy

主任签字：

年 月 日

教务处制

一、BED 格式 (10 分钟)

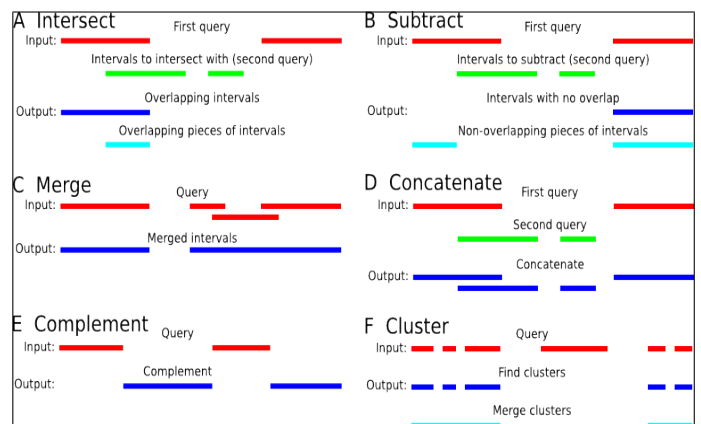
BED 格式: 3+9=12 列 (BED12) , 0-based (与 1-based 的区别)。

- BED3: chrom, start, end
chr1 11873 14409
- BED4: chrom, start, end, name
chr1 11873 14409 uc001aaa.3
- BED5: chrom, start, end, name, score
chr1 11873 14409 uc001aaa.3 0
- BED6: chrom, start, end, name, score, strand
chr1 11873 14409 uc001aaa.3 0 +

| | | | | | | | | |
|------|-----------|-----------|------|---|---|-----------|-----------|---------|
| chr7 | 127471196 | 127472363 | Pos1 | 0 | + | 127471196 | 127472363 | 255,0,0 |
| chr7 | 127472363 | 127473530 | Pos2 | 0 | + | 127472363 | 127473530 | 255,0,0 |
| chr7 | 127473530 | 127474697 | Pos3 | 0 | + | 127473530 | 127474697 | 255,0,0 |
| chr7 | 127474697 | 127475864 | Pos4 | 0 | + | 127474697 | 127475864 | 255,0,0 |
| chr7 | 127475864 | 127477031 | Neg1 | 0 | - | 127475864 | 127477031 | 0,0,255 |
| chr7 | 127477031 | 127478198 | Neg2 | 0 | - | 127477031 | 127478198 | 0,0,255 |
| chr7 | 127478198 | 127479365 | Neg3 | 0 | - | 127478198 | 127479365 | 0,0,255 |
| chr7 | 127479365 | 127480532 | Pos5 | 0 | + | 127479365 | 127480532 | 255,0,0 |
| chr7 | 127480532 | 127481699 | Neg4 | 0 | - | 127480532 | 127481699 | 0,0,255 |

二、逻辑运算 (10 分钟)

- intersect, 交集: 保留重叠的坐标
- subtract, 减法: 去除重叠的坐标
- merge, 合并: 合并重叠的坐标
- concatenate, 串联: 合并多组坐标
- complement, 补集: 取坐标的补集
- cluster, 聚类: 聚合符合要求的坐标
- join, 联合: 根据坐标重叠把两组记录对应起来 (与交集的区别)



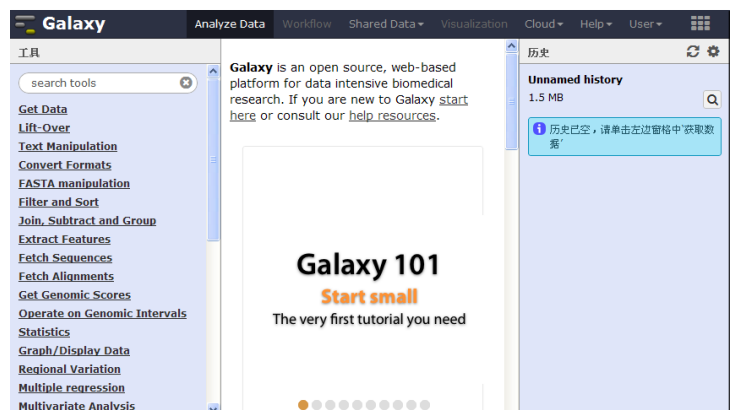
三、Galaxy 简介 (10 分钟)

1. 主界面

- 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

- 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



四、实验操作 (120 分钟)

寻找人类基因组 (hg19) 中 22 号染色体 (chr22) 上含有 SNP 数目最多的外显子。

- 获取数据 (选择正确的格式, 把结果导出到 Galaxy 中)
 - 外显子数据: Get Data, UCSC Main, hg19, chr22, RefSeq Genes, BED 格式
 - SNP 数据: Get Data, UCSC Main, hg19, chr22, dbSNP137, BED 格式
- 提取含有 SNP 的外显子: Operate on Genomic Intervals, Join (注意数据集的顺序)
- 对外显子上的 SNP 进行计数: Join, Subtract and Group, Group (注意选择正确的列)
- 筛选至少含有 10 个 SNP 的外显子: Filter and Sort, Filter (学习编写筛选表达式)
- 附加外显子的原始信息: Join, Subtract and Group, Compare two Datasets (注意数据集的选择, 同时根据每一列的含义选择正确的列)

