

# 实验八 基于Galaxy的基因组数据处理

## 一、实验目的

Galaxy是一个开放的、基于网络的生物信息学分析平台，整合了各种生物信息学分析工具，包括基因组注释的常用工具。完全基于Galaxy平台，利用各种工具及相关操作，寻找人类基因组中22号染色体（chr22）上含有SNP数量最多的外显子。

1. 学习和掌握基因组注释中的常用格式。
2. 学习和掌握基因组坐标的逻辑运算模式。
3. 学习和掌握Galaxy的基本使用方法。

## 二、实验内容

1. 打开Galaxy网站。通过搜索引擎搜索“Galaxy UCSC”，或者直接在浏览器的地址栏中输入网址，打开<https://main.g2.bx.psu.edu/>即可。
2. 获取所需数据。
  1. chr22上的外显子数据。打开Get Data工具集中的UCSC Main工具，调整参数提取人类hg19基因组中22号染色体（chr22）上的所有外显子信息，以BED格式进行存储。
  2. chr22上的SNP数据。同样使用UCSC Main工具提取dbSNP37中chr22上的所有SNP信息，也以BED格式进行存储。
3. 提取含有SNP的外显子。打开Operate on Genomic Intervals工具集中的Join工具，以外显子数据为第一套数据集、SNP数据为第二套数据集，通过坐标比较提取出含有SNP的所有外显子。坐标联合操作的结果中前半部分是外显子的信息，后半部分是SNP的信息。结果显示，不少外显子上含有多个SNP。
4. 对外显子上的SNP进行计数。使用Join, Subtract and Group工具集中的Group工具，针对第4列外显子的ID进行计数，得到的结果中第一列是外显子的ID，第二列则是对应外显子上的SNP数目。
5. 对SNP数目进行排序。使用Filter and Sort工具集中的Sort工具，对第2列中的SNP计数进行排序，采用降序方式，最终含有最多SNP的外显子将排在最前面。
6. 筛选至少含有10个SNP的外显子。使用Filter and Sort工具集中的Filter工具，使用“c2>=10”条件提取第2列SNP计数大于等于10的所有行，即可筛选出至少含有10个SNP的外显子。
7. 附加外显子的原始信息。使用Join, Subtract and Group工具集中的Join two Datasets工具，根据最初外显子数据中的第4列和上一步结果中的第1列（两者均为外显子的ID），把两套数据整合起来，恢复外显子的染色体号、起始终止位置等信息。最终结果中，前面部分是外显子的相关信息，最后一列是相应外显子上的SNP数目。
8. 尝试对其他染色体或全基因组上的外显子进行类似的分析，进一步熟悉在Galaxy中进行基因组注释分析的相关操作。