

# 生物信息学课程设计题目（2009 级）

伊现富

2013 年 11 月 19 日

**名称：** 基于 EMBOSS 平台对 \*\*\* 基因进行序列分析

**总学时：** 36

**学分：** 2

**对象：** 生物医学工程系生物技术与生物信息专业本科生

**指导教师：** 伊现富

**课程设计目的：** 在学习分子生物学、生物计算技术、生物网络数据库、生物信息学课程的基础上，培养学生实际进行核酸序列和蛋白质序列分析的能力。此题目旨在帮助学生熟悉强大的 EMBOSS 生物信息学分析平台，并使用此平台进行基因序列与蛋白质序列的常规分析。

## 实验内容

1. 从 NCBI Gene 数据库中获取自己感兴趣的基因的 DNA 序列。
2. 基因的核苷酸序列分析。
  - (a) 统计该基因 DNA 序列的基本信息，如：长度、单核苷酸与二核苷酸的数目及频率、GC 含量，等。
  - (b) 找出该基因 DNA 序列中的 CpG 岛。
  - (c) 提取该基因的编码区序列。
  - (d) 分析编码区序列的密码子使用频率。
  - (e) 寻找该基因序列中的 ORF，与真实的编码区进行比较。
3. 将编码区核苷酸序列翻译成氨基酸序列。
4. 蛋白质的氨基酸序列分析。

- (a) 对翻译得到的蛋白质进行基本的理化性质分析。
  - (b) 寻找蛋白质中的功能与结构基序。
  - (c) 预测蛋白质的二级结构。
5. 利用所有结果，分析该基因的生物学功能与特性。

## 技术指标

1. 下载基因的 **GenBank** 格式的序列文件。
2. 利用 **GenBank** 文件中的注释信息，解读基因的基本信息（如基因全名、长度，所属物种，染色体定位，外显子、内含子等特征的位置，等）。
3. 基因的核苷酸序列分析结果。
  - (a) 基因 DNA 序列的单核苷酸、二核苷酸频率，GC 含量等统计结果。
  - (b) CpG 岛的预测结果及说明。
  - (c) 基因编码区序列的 **GenBank** 格式文件。
  - (d) 编码区的密码子使用频率与偏性分析。
  - (e) 基因 ORF 预测结果及其分析。
4. 基因编码区翻译后的氨基酸序列文件。
5. 蛋白质的氨基酸序列分析结果，即：蛋白质基本理化性质、基序及二级结构的分析结果。
6. 整合上述数据，分析基因的生物学功能与特性。
7. 课程设计报告书：（1）名称；（2）目的和任务；（3）实验步骤；（4）实验结果；（5）实验结果分析和讨论。
8. 课程设计答辩的 PPT 文件（答辩时间 5 分钟）。

## 实验步骤

1. 打开 **NCBI Gene** 数据库，在搜索栏中输入基因名 *G*；在搜索结果中点击某个物种 *S* 的链接，进入 *S* 物种 *G* 基因的信息界面；下拉至 **NCBI Reference Sequences (RefSeq)** 项目，找到紧邻其下的 **Genomic** 子项目，点击其中的 **GenBank**；在新页面的右侧点击 **Send:**，选择 **File** 后，点击 **Create File** 下载 *G* 基因的 **GenBank** 格式的 DNA 序列。
2. 对基因的 DNA 序列进行分析。打开 **EMBOSS explorer** 界面。
  - (a) 组成成分分析。找到 **NUCLEIC COMPOSITION** 分组。

- i. 使用程序 [compseq](#) 对 DNA 序列的基本组成成分进行分析。在 [Input section](#) 项目中, 使用 [upload](#) 上传上一步从 NCBI 下载的 GenBank 格式的 DNA 序列; 在 [Required section](#) 项目中, 把 [word size](#) 分别设成 1、2; 其他参数默认, 或者自行调整; 最后, 点击 [Run compseq](#) 获得组成成分分析结果。
  - ii. 尝试使用程序 [wordcount](#) 进行类似的组成成分分析。尝试使用程序 [chaos](#) 和 [density](#) 将组成成分结果可视化。
- (b) GC 含量分析。找到 [NUCLEIC CPG ISLANDS](#) 分组。使用程序 [geecee](#) 对 DNA 序列的 GC 含量进行分析。与前述类似, 以上传文件的方式提交 DNA 序列, 之后点击 [Run geecee](#) 得到 DNA 序列的 GC 含量。
- (c) CpG 岛分析。找到 [NUCLEIC CPG ISLANDS](#) 分组。
- i. 使用程序 [newcpgreport](#) 寻找 DNA 序列中的 CpG 岛。上传 DNA 序列文件后, 调整相应参数, 点击 [Run newcpgreport](#) 得到 CpG 岛的预测结果。
  - ii. 使用程序 [cpgplot](#) 直观观察 CpG 岛的预测结果。上传 DNA 序列并调整参数, 点击 [Run cpgplot](#) 得到图形化的 CpG 岛预测结果。
  - iii. 尝试使用程序 [cpgreport](#) 和 [newcpgseek](#) 进行类似的 CpG 岛分析。
- (d) 提取 CDS 序列。找到 [FEATURE TABLES](#) 分组。
- i. 使用程序 [coderet](#) 提取基因中的 CDS 序列。上传 DNA 序列文件并调整输出格式为 [Genbank](#), 点击 [Run coderet](#) 得到该基因 Genbank 格式的 CDS 序列。将其保存至本地以备后用。
  - ii. 尝试使用程序 [extractfeat](#) 提取 CDS 序列。尝试使用程序 [showfeat](#) 将整个基因的特征可视化。
- (e) 编码区密码子分析。找到 [NUCLEIC CODON USAGE](#) 分组。
- i. 使用程序 [cusp](#) 计算 CDS 序列中密码子的使用频率。上传上一步保存的 CDS 序列文件, 点击 [Run cusp](#) 得到 CDS 序列的密码子使用频率。
  - ii. 尝试使用程序 [syco](#) 对 CDS 序列中的同义密码子使用频率进行分析。
- (f) ORF 预测与分析。找到 [NUCLEIC GENE FINDING](#) 分组。
- i. 使用程序 [getorf](#) 找到基因中的 ORF。上传基因的 DNA 序列并调整参数, 点击 [Run getorf](#) 获得全部的 ORF 序列。
  - ii. 使用程序 [plotorf](#) 和 [showorf](#) 将基因中的 ORF 以图形方式直观显示出来。
  - iii. 将上述步骤的 ORF 序列结果与图形结果结合起来, 并与基因实际的 CDS 进行比较。

3. 翻译 CDS 序列。找到 [NUCLEIC TRANSLATION](#) 分组。使用程序 [transeq](#) 把 CDS 核苷酸序列翻译成氨基酸序列。上传 CDS 序列，调整参数与输出格式后，点击 [Run transeq](#) 得到翻译后的氨基酸序列。将其保存以备后用。
4. 对蛋白质的氨基酸序列进行分析。
  - (a) 组成成分与理化性质分析。找到 [PROTEIN COMPOSITION](#) 分组。
    - i. 使用程序 [pepstats](#) 分析蛋白质的基本理化性质。上传翻译得到的氨基酸序列，调整参数后点击 [Run pepstats](#) 得到蛋白质的基本理化性质信息。
    - ii. 使用程序 [compseq](#) 查看蛋白质的基本组成成分。
    - iii. 使用程序 [iep](#) 计算蛋白质的等电点。
    - iv. 尝试使用程序 [charge](#) 获得每个氨基酸的带电量数据。尝试使用程序 [octanol](#) 和 [pepwindow](#) 对蛋白质的疏水性进行可视化。尝试使用程序 [pepinfo](#) 对蛋白质氨基酸残基的极性进行分析。尝试使用 [DISPLAY](#) 分组中的程序 [pepnet](#) 和 [pepwheel](#) 对蛋白质氨基酸残基的亲水性和疏水性进行可视化。
  - (b) 基序分析。找到 [PROTEIN MOTIFS](#) 分组。
    - i. 使用程序 [sigcleave](#) 寻找蛋白质中的信号切割位点。上传蛋白质序列后调整参数，点击 [Run sigcleave](#) 找到蛋白质的信号切割为点。
    - ii. 使用程序 [sigcleave](#) 寻找蛋白质中的信号切割位点。上传蛋白质序列后调整参数，点击 [Run sigcleave](#) 找到蛋白质的信号切割为点。
    - iii. 使用程序 [helixturnhelix](#) 寻找蛋白质上的核苷酸结合基序。上传蛋白质序列后调整参数，点击 [Run helixturnhelix](#) 寻找蛋白质上的核苷酸结合基序。
    - iv. 尝试使用此分组中的其他程序（如：[antigenic](#)）寻找更多的功能与结构基序。
  - (c) 结构分析。找到 [PROTEIN 2D STRUCTURE](#) 分组。
    - i. 使用程序 [garnier](#) 预测蛋白质的二级结构。上传蛋白质序列后调整参数，点击 [Run garnier](#) 预测得到蛋白质的二级结构。
    - ii. 使用程序 [tmap](#) 预测蛋白质中的跨膜片段。上传蛋白质序列后，点击 [Run tmap](#) 预测得到蛋白质中的跨膜片段。
    - iii. 尝试使用此分组中的其他程序预测蛋白质的二级结构。
5. 整合上述所有结果，分析基因 *G* 的生物学功能、特性等。

学时分配

设计讲解： 2 学时

实验： 18 学时

总结和课程设计报告： 12 学时

答辩： 4 学时

总共： 36 学时

地点： 教一楼 205 室生物信息学实验室

#### 资源网站

1. EMBOSS 官网: <http://emboss.sourceforge.net/>
2. EMBOSS explorer: <http://emboss.bioinformatics.nl/>; <http://bioinfo.nhri.org.tw/gui/>; <http://genome.csdb.cn/emboss/>
3. EMBOSS GUI: <http://anabench.bcm.umontreal.ca/html/EMBOSS/>; <http://bips.u-strasbg.fr/EMBOSS/>
4. NCBI Gene: <http://www.ncbi.nlm.nih.gov/gene>

**名称：** 基于 Galaxy 平台分析 \*\*\* 物种基因在基因组中的分布

**总学时：** 36

**学分：** 2

**对象：** 生物医学工程系生物技术与生物信息专业本科生

**指导教师：** 伊现富

**课程设计目的：** 在学习分子生物学、生物计算技术、生物网络数据库、生物信息学课程的基础上，培养学生分析大规模基因组数据的能力。此题目旨在帮助学生熟悉易用、强大且日渐流行的 Galaxy 生物信息学分析平台，并使用此平台处理基因组数据，分析基因在每条染色体上的分布情况。

### 实验内容

1. 选择合适的物种  $S$ 。
2. 熟悉 Galaxy 界面。
3. 获取该物种基因组范围的基因信息和每条染色体的长度。
4. 计算每条染色体上的基因数目。
5. 计算每条染色体上的基因密度。
6. 绘制条形图将数据结果可视化。
7. 结合数据表格和条形图分析结果。

### 技术指标

1. 基因组范围的基因数据。
2. 基因组上每条染色体的长度。
3. 每条染色体上的基因数目。
4. 每条染色体上的基因密度。
5. 基因数目与基因密度的条形图。
6. 对所得结果的分析与理解。
7. 课程设计报告书：（1）名称；（2）目的和任务；（3）实验步骤；（4）实验结果；（5）实验结果分析和讨论。
8. 课程设计答辩的 PPT 文件（答辩时间 5 分钟）。

## 实验步骤

1. 选择完成基因组测序且注释比较完善的物种，如：人、小鼠，等。<sup>1</sup>
2. 打开 **Galaxy Test** 主页，熟悉其界面布局。
3. 获取基因与染色体信息数据。
  - (a) 获取基因信息。
    - i. 获取数据。打开 **Galaxy** 的 **Get Data** 分组，点击 **UCSC Main** 进入 **UCSC Table** 界面。在 **UCSC Table** 界面中，**clade** 选择 **Mammal**、**genome** 选择 **Human**、**assembly** 选择 **Feb. 2009 (GRCh37/hg19)**、**group** 选择 **Genes and Gene Prediction Tracks**、**track** 选择 **RefSeq Genes**、**table** 选择 **refGene**、**region** 点选 **genome**、**output format** 选择 **BED – browser extensible data**、**Send output to** 勾选 **Galaxy**、**file type returned** 点选 **plain text**，之后点击 **get output**，新界面中的 **Create one BED record per** 点选 **Whole Gene**，最后点击 **Send query to Galaxy** 即可将基因组中的基因信息载入到 **Galaxy** 平台的工作空间中。
    - ii. 修改属性。为了便于区分工作空间中的不同数据，可以点击 **1: UCSC Main on Human: refGene (genome)** 右侧的铅笔图标，修改数据集的属性，如：修改 **Name** 为 **refGene**，之后点击 **Save** 更新其属性。
  - (b) 获取染色体长度。
    - i. 获取数据。在 **UCSC Table** 界面中，修改 **group** 为 **All Tables**、**database** 为 **hg19**、**table** 为 **chromInfo**，待界面刷新后，直接点击 **get output** 并继续点击 **Send query to Galaxy** 即可将染色体的长度信息载入到工作空间中。
    - ii. 修改属性。为便于区分不同数据，可以点击 **2: UCSC Main on Human: chromInfo (genome)** 右侧的铅笔图标，修改 **Name** 为 **chromInfo**，点击 **Save** 更新数据属性。
4. 计算基因数目。
  - (a) 计算每条染色体上的基因数目。打开 **Galaxy** 的 **Statistics** 分组，点击其中的 **Count** 工具。其中，**from dataset** 选择 **1: refGene**，**Count occurrences of values in column(s)** 点选 **c1**，最后点击 **Execute** 即可计算出每条染色体上的基因数目。同样修改数据的属性，把 **Name** 改为 **geneOnChromAll**。
  - (b) 过滤数据。打开 **Galaxy** 的 **Filter and Sort** 分组，点击其中的 **Select** 工具。其中，**Select lines from** 选择 **3: geneOnChromAll**，**that** 选择 **NOT Matching**，**the pattern** 填写 **\_**，最后点击 **Execute** 过滤数据。修改 **Name** 为 **geneOnChromFilter**。

---

<sup>1</sup>此处以“人”为例详述步骤。

## 5. 计算基因密度。

- (a) 整合染色体上的基因数目数据和染色体的长度信息。打开 Galaxy 的 Join, Subtract and Group 分组, 点击其中的 Join two Datasets 工具。其中, Join 选择 2: chromInfo, using column 选择 c1, with 选择 4: geneOnChromFilter, and column 选择 c2, 其他参数默认即可, 最后点击 Execute 整合两套数据。
- (b) 计算每条染色体上的基因密度。打开 Galaxy 的 Text Manipulation 分组, 点击其中的 Compute 工具。在 Add expression 中填写  $c4/c2*100000000$ , 其他参数默认即可, 最后点击 Execute 计算染色体上每 100 Mb 的基因数目。
- (c) 提取染色体号、长度及其上的基因数目、基因密度等有用信息。打开 Galaxy 的 Text Manipulation 分组, 点击其中的 Cut 工具。在 Cut columns 中填写 c1,c2,c4,c6, 其他参数默认即可, 最后点击 Execute 提取需要的几列。
- (d) 根据染色体长度排序数据。打开 Galaxy 的 Filter and Sort 分组, 点击其中的 Sort 工具。其中, on column 选择 c2, 其他参数默认, 最后点击 Execute 即可根据染色体长度将数据进行排序。修改 Name 属性为 geneNumberDensity。

## 6. 绘制并查看条形图。

- (a) 绘图。打开 Galaxy 的 Graph/Display Data 分组, 点击其中的 Bar chart 工具。其中, 在 Numerical columns 中点选 c3 和 c4, 其他参数适当调整后, 点击 Execute 即可将上一步的结果可视化。
- (b) 查看并保存图片。点击数据右侧的眼睛图表, 可以在 Galaxy 中查看绘制的条形图; 也可点击数据左下方的软盘图表, 将图片保存至本地。
- (c) 尝试使用 Bar chart 工具分开绘制基因数目和基因密度的条形图。

## 7. 分析结果。如: 哪条染色体上的基因数目最多/少? 哪条染色体上的基因密度最高/低? 数目最多和密度最高的染色体是不是同一条? 从中学到了什么<sup>2</sup>? 此外, 还可以把得到的结果和教材或文献中的数据进行比较。

### 学时分配

设计讲解: 2 学时

实验: 18 学时

总结和课程设计报告: 12 学时

答辩: 4 学时

---

<sup>2</sup>提示: 数据要经过标准化后才能互相比较。



总共： 36 学时

地点： 教一楼 205 室生物信息学实验室

#### 资源网站

1. Galaxy Test: <https://test.g2.bx.psu.edu/>
2. Galaxy Main: <https://main.g2.bx.psu.edu/>

**名称：** 基于 Galaxy 平台分析 \*\*\* 物种 SNP 在不同特征区域中的分布

**总学时：** 36

**学分：** 2

**对象：** 生物医学工程系生物技术与生物信息专业本科生

**指导教师：** 伊现富

**课程设计目的：** 在学习分子生物学、生物计算技术、生物网络数据库、生物信息学课程的基础上，培养学生分析大规模基因组数据的能力。此题目旨在帮助学生熟悉易用、强大且日渐流行的 Galaxy 生物信息学分析平台，并使用此平台处理基因组数据，分析 SNP 在不同特征区域上的分布情况。

### 实验内容

1. 选择合适的物种  $S$ 。
2. 熟悉 Galaxy 界面。
3. 获取该物种不同特征区域（如：5' UTR 外显子、编码区外显子和 3' UTR 外显子）的信息。
4. 获取该物种 dbSNP 数据库的 SNP 信息。
5. 计算不同特征区域的 SNP 数目和密度。
6. 将 SNP 密度进行标准化。
7. 比较不同特征区域的 SNP 密度。
8. 对所得结果进行综合分析。

### 技术指标

1. 基因组范围不同特征区域的数据。
2. 基因组范围上 SNP 的数据。
3. 不同特征区域 SNP 的数目。
4. 不同特征区域 SNP 的密度。
5. 标准化之后的 SNP 密度。
6. 对所得结果的分析与理解。

7. 课程设计报告书：（1）名称；（2）目的和任务；（3）实验步骤；（4）实验结果；（5）实验结果分析和讨论。
8. 课程设计答辩的 PPT 文件（答辩时间 5 分钟）。

## 实验步骤

1. 选择完成基因组测序且注释比较完善的物种，如：人、小鼠，等。<sup>3</sup>
2. 打开 **Galaxy Main** 主页，熟悉其界面布局。
3. 获取基因组范围不同特征区域的基本信息。
  - (a) 获取数据。打开 **Galaxy** 的 **Get Data** 分组，点击 **UCSC Main** 进入 **UCSC Table** 界面。在 **UCSC Table** 界面中，**clade** 选择 **Mammal**、**genome** 选择 **Human**、**assembly** 选择 **Feb. 2009 (GRCh37/hg19)**、**group** 选择 **Genes and Gene Prediction Tracks**、**track** 选择 **RefSeq Genes**、**table** 选择 **refGene**、**region** 点选 **genome**、**output format** 选择 **BED – browser extensible data**、**Send output to** 勾选 **Galaxy**、**file type returned** 点选 **plain text**，之后点击 **get output**，新界面中的 **Create one BED record per** 点选 **5' UTR Exons**，最后点击 **Send query to Galaxy** 即可将 5' UTR 外显子的基本信息载入到 **Galaxy** 平台的工作空间中。
  - (b) 修改属性。为了便于区分工作空间中的不同数据，可以点击 **UCSC Main on Human: refGene (genome)** 数据集右侧的铅笔图标，修改数据集的属性，如：修改 **Name** 为 **5UTR**，之后点击 **Save** 更新其属性。
  - (c) 重复操作。重复上述步骤，在 **Create one BED record per** 分别点选 **Coding Exons** 和 **3' UTR Exons**，将编码区外显子和 3' UTR 外显子的基本信息载入到 **Galaxy** 平台的工作空间中。
4. 获取 dbSNP 数据库中 SNP 的信息。
  - (a) 获取数据。在 **UCSC Table** 界面中，修改 **group** 为 **Variation and Repeats**、**track** 为 **Common SNPs(135)**、**table** 为 **snp135Common**，待界面刷新后，点选 **Create one BED record per** 中的 **Whole Gene**，最后点击 **Send query to Galaxy** 即可将 SNP 的基本信息载入到 **Galaxy** 平台的工作空间中。
  - (b) 修改属性。为便于区分不同数据，可以点击 **UCSC Main on Human: snp135Common (genome)** 数据集右侧的铅笔图表，修改 **Name** 为 **SNP135**，点击 **Save** 更新数据属性。
5. 计算不同特征区域中的 SNP 数目及其密度。

---

<sup>3</sup>此处以“人”为例详述步骤。

- (a) 关联特征区域和 SNP 的信息。打开 Galaxy 的 Operate on Genomic Intervals 分组, 点击其中的 Join 工具。其中, Join (First dataset) 选择 5UTR 数据, with (Second dataset) 选择 SNP135 数据, 其余参数默认即可, 最后点击 Execute 即可将特征区域的信息与 SNP 的信息关联起来。<sup>4</sup>
  - (b) 计算特征区域中的 SNP 数目。打开 Galaxy 的 Statistics 分组, 点击其中的 Count 工具。其中, from dataset 选择上一步的结果, Count occurrences of values in column(s) 点选 c4, 最后点击 Execute 即可计算出每个 5' UTR 外显子上的 SNP 数目。
  - (c) 恢复计数结果中的特征区域信息。打开 Galaxy 的 Join, Subtract and Group 分组, 点击其中的 Join two Datasets 工具。其中, Join 选择上一步的结果, using column 选择 c2, with 选择 5UTR 数据, and column 选择 c4, 其他参数默认即可, 最后点击 Execute 整合两套数据。
  - (d) 计算每个特征区域的长度。打开 Galaxy 的 Text Manipulation 分组, 点击其中的 Compute 工具。在 Add expression 中填写 c5-c4, 其他参数默认即可, 最后点击 Execute 计算每个特征区域的长度。
  - (e) 提取数据集中的有用信息。打开 Galaxy 的 Text Manipulation 分组, 点击其中的 Cut 工具。在 Cut columns 中填写 c3,c9,c1, 其他参数默认即可, 最后点击 Execute 提取需要的几列。
  - (f) 按染色体整理数据。打开 Galaxy 的 Join, Subtract and Group 分组, 点击其中的 group 工具。其中, Select data 选择上一步的结果, Group by column 选择 c1; 之后, 点击 Add new Operation, Type 选择 Sum, on column 选择 c2; 之后, 继续点击 Add new Operation, Type 选择 Sum, on column 选择 c3; 其他参数默认, 最后点击 Execute 即可。
  - (g) 过滤数据。打开 Galaxy 的 Filter and Sort 分组, 点击其中的 Select 工具。其中, Select lines from 选择上一步的结果, that 选择 NOT Matching, the pattern 填写 \_, 最后点击 Execute 过滤数据。
  - (h) 计算每个特征区域上 SNP 的密度。打开 Galaxy 的 Text Manipulation 分组, 点击其中的 Compute 工具。在 Add expression 中填写 c3/c2\*1000, 其他参数默认即可, 最后点击 Execute 计算 SNP 的密度。将 Name 属性修改为 5UTRSNP。
6. 标准化 SNP 密度。

- (a) 计算基因组上特征的总长度及 SNP 的总数目。打开 Galaxy 的 Statistics 分

---

<sup>4</sup>注意: 此步舍弃了没有 SNP 的特征区域。

组，点击其中的 **Summary Statistics** 工具。其中，**Summary statistics on** 选择 **5UTRSNP**，**Column or expression** 填写 **c2** 或 **c3**，最后点击 **Execute** 分别计算特征总长  $L$  及 SNP 的总数  $N$ 。

- (b) 将每条染色体上的 SNP 密度进行标准化。首先计算  $M = N/L * 1000$ 。然后，打开 **Galaxy** 的 **Text Manipulation** 分组，点击其中的 **Compute** 工具。在 **Add expression** 中填写 **c4-M**，**as a new column to** 选择 **5UTRSNP**，最后点击 **Execute** 标准化 SNP 密度。将 **Name** 属性修改为 **5UTRSNPNormalized**。

7. 比较不同特征区域的 SNP 密度。

- (a) 重复第 5、6 两步<sup>5</sup>，获得其他特征区域的 SNP 密度。
- (b) 比较不同特征区域的 SNP 密度。打开 **Galaxy** 的 **Join, Subtract and Group** 分组，点击其中的 **Join two Datasets** 工具。其中，**Join** 选择 **5UTRSNPNormalized** 数据，**using column** 选择 **c1**，**with** 选择 **codingSNPNormalized** 数据，**and column** 选择 **c1**，其他参数默认即可，最后点击 **Execute** 整合 5' UTR 和编码区两套数据。以类似的操作，再将 3' UTR 数据整合进去。

8. 综合分析所得结果<sup>6</sup>。如：对于同一条染色体来说，不同特征区域的 SNP 密度有没有差别？对于同一类特征区域来说，不同染色体的 SNP 密度有没有差别？此外，还可以将两者结合起来进行分析，并尝试对分析结果进行解释。

## 学时分配

设计讲解：2 学时

实验：18 学时

总结和课程设计报告：12 学时

答辩：4 学时

总共：36 学时

地点：教一楼 205 室生物信息学实验室

## 资源网站

1. Galaxy Main: <https://main.g2.bx.psu.edu/>
2. Galaxy Test: <https://test.g2.bx.psu.edu/>

---

<sup>5</sup>提示：可以提取工作流程后，修改输入数据，自动化重复前述操作！

<sup>6</sup>敬请注意：因为处理过程中的取舍问题，请不要把此处的计算结果当作常识来使用！