第 5 章　基因组功能注释分析

了解：

1. 基因组的组装版本及常见版本间的对应关系。
2. VCF 格式。
3. 纯文本与格式化文本的区别。
4. 常见的文本编辑器。
5. 变异位点注释的内容和常用工具。
6. 基因组功能注释分析平台。

熟悉：

1. GFF 格式。
2. 转换格式的常用工具。
3. 不同操作系统中换行符的差异。
4. 基因组坐标逻辑运算的常用工具。
5. 基因集的富集分析和常用工具。
6. box plot 的含义及其绘制。

掌握：

1. 基因组的两种坐标系统：0-based，1-based。
2. 基因组组装版本间坐标转换的常用工具。
3. FASTA 格式和 BED 格式。
4. 基因组坐标的逻辑运算模式。
5. 序列标识的含义和制作工具。
6. Galaxy 分析平台的使用。
7. 解析图表的策略。
8. 数据处理的“输入 -加工 -输出”三步走策略。

5.1　引言

自从基因组可以被测序以来，从原始的基因组核酸序列中挖掘有用的生物学信息并阐释其生物学含义，即基因组注释（genome annotation），已经成为生物学研究的核心工作之一。基因组注释是调查基因组 DNA 的全貌、描述 DNA 的关键特征的一个过程。基因组注释包括基因组结构注释（structural annotation）和基因组功能注释（functional annotation）两大部分。基因组结构注释是指在基因组序列中寻找基因等功能元件并明确其基本结构；在结构注释的基础上，将进化保守性（evolutionary conservation）和基因本体论（gene ontology）等元数据（meta-data）与功能元件对应起来，找到其生物学功能，这个过程就是基因组功能注释。基因组结构注释的部分内容，在前文有详细的介绍；本章重点介绍基因组功能注释的相关内容。

随着基因芯片（gene chip）和第二代测序技术（NGS，Next Generation Sequencing）等高通量技术的发展与成熟，在基因组水平上进行大规模研究的成本逐渐降低、速度逐渐加快。使用高通量技术（如：外显子组测序，全基因组测序等）对全基因组或基因组上的目标区域进行重测序，把测序

结果与参考基因组进行比较，进而寻找并锁定候选基因（candidate gene）用于后续的实验验证与分析，这已经成为当下研究复杂疾病的主要手段之一。高通量技术会产生海量的数据，实验方法因成本过高无法直接适用于基因组重测序的后续功能分析，因此，使用生物信息学方法对海量数据进行功能注释就成为基因组研究的必须手段。

本章将借鉴复杂疾病研究领域的经验与成果，由浅入深介绍基因组功能注释的相关知识与工具。首先介绍基因组注释的基础知识，包括基因组的组装版本（genome builds）、坐标系统（coordinate system）、注释常用格式以及坐标间的逻辑运算模式（operations on genomic intervals）。因为过于常用，少有书籍介绍这些常识性的基础内容，但如果完全不懂或者理解有误，会直接影响后续功能注释工作的进行及注释结果的解读。之后，在理论知识的基础上，重点通过实例来演示基因组注释中的常见操作，这常常也是进行后续高级注释分析的准备工作，主要包括基因组组装版本间的坐标转换（coordinate transform）、常用格式间的转换以及基因组坐标的逻辑运算。最后，介绍并演示基因组功能注释工作中的几个高级注释内容，包括基因组变异位点的注释、基因集的富集分析（gene set enrichment analysis）和序列标识（sequence logo）的制作，并推荐贯穿全章的生物信息学分析平台——Galaxy。

5.2　基因组组装版本

自从 20 世纪 70 年代 Sanger 测序法诞生以来，DNA 测序技术就在不断发展。虽然越来越多的物种加入被测序的行列，但像人类（*Homo sapiens*）和小鼠（*Mus musculus*）一样，大多数物种的基因组并没有被完全测序。随着测序技术的进步，在全世界科研工作者的努力下，序列不明确的基因组部分会不断被解读出来，测序错误的部分也会被更正。因此，就像操作系统或软件会不时更新、发布新版本一样[1]，基因组数据库也会不定期更新各个物种的现有基因组，或者发布一个新的基因组组装版本。对于同一个物种的基因组来说，不同组装版本间的基因组序列、基因组特征（feature）的坐标等注释信息会有很大不同。因此，在处理基因组数据时，首先要明确使用的基因组组装版本。

对于同一个基因组组装版本，不同数据库中的命名可能不同，但它们本质上完全一样，并存在着一对一的对应关系。在常用的三大综合数据库中，NCBI一般以“Build X”来对基因组组装版本进行命名；Ensembl则以“NCBIX”与之对应；UCSC中的命名比较独特，不同物种使用不同的前缀，如人类基因组的组装版本以“hgX”表示，小鼠的以“mmX”表示[2]。NCBI 和 UCSC 之间的基因组组装版本存在者明确的的对应关系，如：人类基因组的 UCSC hg18 对应 NCBI Build 36.1，发布于 2006 年 3 月；小鼠基因组的 UCSC mm9 对应 NCBI Build 37，发布于 2007 年 7 月。更多物种、更多组装版本间的对应关系可以查阅 List of UCSC genome releases。[3]

5.3　基因组坐标系统

如果把染色体序列看做一条很长的线段，其上的单核苷酸多态性（SNP，Single Nucleotide Polymorphism）[4]等位点就可以看做线段上的点，而基因等特征则好比线段上一条短的子线段。如同数学中使用坐标来精确定位点线一样，在生物学中，描述基因组特征时，为了与基因组序列对应起来，常常使用基因组坐标（coordinate）来将其精确定位。比如，对于人类基因组 hg19 来说，SNP rs1800468 位于 19 号染色体的 41860587 位置，表示为 chr19:41860587；基因 *SAMD11* 位于 1 号染色体上，起始于 861121，终止于 879961，表示为 chr1:861121-879961。

基因组坐标有两种不同的坐标系统，其表示方法略有差异。一种是比较容易理解的全包含的

[1]操作系统或软件版本的例子：XP 与 Win7；迅雷 5 与迅雷 7。

[2]hg、mm、GRCh、GRCm 分别表示什么含义？

[3]对于基因组数据、数据库和软件等，在使用时需要记录下其版本号。

[4]SNP 即单核苷酸多态性，是指等位基因发生了突变，致使不同等位型的出现。SNP 的位点极其丰富，几乎遍及整个基因组。据估计基因组中大约平均每 1000bp 就会出现一个 SNP，这样 SNP 在整个人类基因组的分布就会达到 300 万个。SNP 可以分为两种：纯合 SNP（homozygous SNP ）和杂合 SNP（heterozygous SNP），纯合 SNP 是指两个等位基因都发生了突变，杂合 SNP 是指两个等位基因中的一个发生了突变。

1-based (one-based, fully-closed) 坐标系统，表示为“[start, end]”；另一种是容易引起混淆、但常用且易用的半包含的 0-based (zero-based, half-closed-half-open) 坐标系统[5]，表示为“[start, end)”，此坐标系统有时也被称为以 0-based 开始、以 1-based 结束 (zero-based start, one-based end) 的坐标系统。在处理基因组数据时，如果对两种坐标系统不加区分，很可能会因差之毫厘而谬以千里。

两种坐标系统之间的区别可以用一个例子来进行说明。对于序列“AATTGGCC”来说, 其中的子序列“TG”，如果采用 0-based 的坐标系统，其位置表示为“[3,5)”[6]；如果采用 1-based 的坐标系统，则表示为“[4,5]”。

为了便于理解这两种坐标系统[7]，可以将其和英式英语与美式英语中的 first floor 相类比：在英式英语中，楼房地面与街道相平的楼层叫 ground floor，ground floor 上面的一层叫 first floor；而在美式英语中，与街道相平的楼层叫 first floor，其上面的一层叫 second floor。也可以将这两种坐标系统和编程语言与现实世界的计数方式相比较：在编程语言中，计数一般都是从 0 开始的[8]，如 Perl 语言中数组 @array 的第一个元素表示为 $array[0]；而现实世界中为大家所熟悉的计数方式是从 1 开始的。

在常用的数据格式中，GFF、VCF、SAM 和 Wiggle 等采用的是 1-based 的坐标系统，而 BED、BAM 和 PSL 等采用的则是 0-based 的坐标系统。在常用的数据库和工具中，DAS（Distributed Annotation System）和 UCSC 的 Genome Browser 使用的是 1-based 的坐标系统，NCBI 的 dbSNP 和 UCSC 的 Table Browser 则使用 0-based 的坐标系统。[9]总体来说，主要给研究人员肉眼查看的数据基本上都采用 1-based 的坐标系统，主要用于计算机程序处理的数据则大多采用 0-based 的坐标系统。

5.4 基因组注释常用格式

为了简洁且精确地描述特定的基因组信息，针对序列、特征、变异等多样的基因组数据，科研人员发明了多种格式来存储相应的数据与信息[10]。如：使用 FASTA 等格式存储基因组序列，使用 BED 和 GFF 等格式存储基因组特征，使用 VCF 格式存储基因组序列中的变异信息。针对特定数据设计的多种格式，其存储的信息往往具有一定的相似性，因此这些相关的格式间可以进行转换[11]，如：都是存储基因组特征信息的 BED 和 GFF 格式就可以进行相互转换。

5.4.1 FASTA 格式

FASTA 格式[12]是保存序列最常见的格式之一，其起始标识符为“>”，后面紧跟序列的 ID 以及可有可无的描述信息，下面一行或数行则是具体的序列。严格来讲，每一行最好不要超过 80 个字符，序列中的回车符不会影响序列的连续性及程序对序列的处理。FASTA 格式中的序列使用标准的 IUB/IUPAC 核酸代码和氨基酸代码。此外，还应注意以下情况：

- 允许小写字母的存在，但会转换成大写；
- 单个“-”符号代表不明长度的空位；
- 在氨基酸序列中允许出现“U”和符号“*”；
- 任何数字都应该被去掉或转换成字母（如：不明核酸用“N”表示，不明氨基酸用“X”表示）。

[5]从 0 开始计数的理论依据。（扩展阅读：为什么计数应该从零开始？）

[6]0-based 坐标系统有何优势？

[7]现实生活、学习中存在和这两种坐标系统类似的例子。

[8]宝宝数学很好，2 岁就可以从 1 数到 10 了。后来，我告诉他 0 比 1 还小。今天吃饺子，我说：“宝宝，你数数你想吃几个饺子？”“0，1，2，3。”一边说着一边拿起一个饺子，“这是第 0 个。”老婆怒吼：“下一代还是做程序员的命！”

[9]了解 SAM、BAM、Wiggle、PSL 格式和 DAS 工具。

[10]回顾、复习 GenBank、EMBL、PDB 等数据格式。

[11]txt 与 doc 存储的信息类似，多数情况下可以进行互转；mkv 与 doc 存储的信息完全不同，无法进行转换。

[12]FASTQ 与 FASTA 一个字母之差，它又是什么格式呢？

5.4.2　BED 格式

BED（Browser Extensible Data）格式存储用于展示的特征注释信息，每一行表示一个基因组区域，即基因组特征，又称 BED 记录（record）。BED 格式定义了 12 个项目（即 12 列），包括 3 个必选项目（required BED fields）和 9 个可选的附加项目（additional optional BED fields），换言之，一个 BED 格式文件的列数在 3 ～ 12 之间。在同一个文件中，每个特征（即每一行）的列数必须一致，如果存在可选项目，其出现次序必须遵守 BED 格式的规定。此外，BED 文件中也可以包含对注释信息进行描述或定义的一行或数行内容。

1. **chrom**: Name of the chromosome or scaffold, 染色体名
2. **chromStart**: Starting position of the feature in the chromosome, 特征的起始位置
3. **chromEnd**: Ending position of the feature in the chromosome, 特征的终止位置
4. name: Name of the BED line, 特征名
5. score: A score between 0 and 1000, 分值（0 ～ 1000）
6. strand: Strand, ``+" for forward and ``-" for reverse, 链性（“+”代表正链，“-”代表负链）
7. thickStart: Starting position at which the feature is drawn thickly, 加粗显示的起始位置
8. thickEnd: Ending position at which the feature is drawn thickly, 加粗显示的终止位置
9. itemRgb: RGB value of the form R,G,B (e.g. 255,0,0), RGB 值（如：255,0,0）
10. blockCount: Number of blocks in the BED line, 特征中的区段数
11. blockSizes: A comma-separated list of the block sizes, 区段大小列表（以逗号分隔）
12. blockStarts: A comma-separated list of block starts, 区段起始位置列表（以逗号分隔）

包含全部 12 列的 BED 格式是最完整的 BED，也称为 BED12。此外，一些简化的 BED 格式也比较常见：

BED3　chrom, start, and end；仅含有前三列的 BED
BED4　chrom, start, end, and name；含有前四列的 BED
BED5　chrom, start, end, name, and score；含有前五列的 BED
BED6　chrom, start, end, name, score, and strand；含有前六列的 BED

5.4.3　GFF 格式

GFF（General Feature Format）是用来存储基因组特征的标准数据格式。GFF 格式的文件是用制表符分隔的纯文本文件，通常采用“.GFF”作为其后缀。GFF 格式由注释信息和具体的特征信息两部分组成：注释信息以“##”开头，用来说明格式及其版本号，且必须位于整个文件的第一行；紧随其后的就是基因组特征的相关信息，每行代表一个特征，由 9 列组成。此外，文件中的空行和仅以一个“#”符号开头的行都会被忽略掉[13]。

1. seqid: ID of the landmark, 特征所在的参考序列的 ID
2. source: Program that generated this feature, 产生此特征的程序
3. type: Type of the feature, 特征类别
4. start: Start position of the feature, 特征的起始位置
5. end: End position of the feature, 特征的终止位置
6. score: Score of the feature, 特征的分值
7. strand: Strand of the feature, 特征的链性
8. phase: Reading frame of the first base (only for type ``CDS"), 第一个碱基的阅读相位（仅针对 CDS 类型的特征）
9. attributes: List of feature attributes in the format tag=value, 以“键 = 值”形式表示的特征属性列表

[13]纯文本中常用“#”表示注释，如：加在说明列含义的标题行前以便和数据行区分开来。

5.4.4 VCF 格式

VCF（Variant Call Format）格式是专门用来存储序列变异信息的标准数据格式。VCF 格式由三大部分组成：以“##”起始的数行元信息（meta-information lines）、以“#”起始的一行标题行（header line）和剩余的数据信息行（data lines）。在数据行中，每一行代表基因组中的一个位置，列与列之间用制表符分隔，前 8 列为必选项目，其余为可选的 N+1 列基因型信息（1 个基因型格式说明列和 N 个样本基因型列）。

1. CHROM, Chromosome, 染色体
2. POS, Reference position, 参考基因组中的位置
3. ID, Unique identifier(s), 唯一的名称
4. REF, Reference base(s), 参考基因组上的碱基
5. ALT, Alternate non-reference alleles, 等位基因（非参考基因组上的碱基）
6. QUAL, Phred-scaled quality score, 以 Phred 形式表示的质量值
7. FILTER, PASS if this position has passed all filters, 如果满足所有过滤标准就为 PASS
8. INFO, Additional information, 附加信息

在生物信息学中，以不同格式保存数据的文件多以纯文本形式进行存储，可以使用文本编辑器查看、编辑。

5.5 文本文件与文本编辑器

5.5.1 文本文件

由于结构简单，文本文件被广泛用于记录信息。它能够避免其它文件格式遇到的一些问题。此外，当文本文件中的部分信息出现错误时，往往能够比较容易的从错误中恢复出来，并继续处理其余的内容。文本文件的一个缺点是，它的熵往往较低，也就是说，可以用较小的存储空间记录这些信息。

在 Windows 中，当一个文件的扩展名为“txt”时，系统就认为它是一个文本文件。此外，出于特殊的目的，有些文本文件使用其它的扩展名。例如，计算机的源代码也是文本文件，它们的后缀是用来指明它的程序语言的。.txt 是包含极少格式信息的文字文件的扩展名。.txt 格式并没有明确的定义，它通常是指那些能够被系统终端或者简单的文本编辑器接受的格式。任何能读取文字的程序都能读取带有.txt 扩展名的文件，因此，通常认为这种文件是通用的、跨平台的。在英文文本文件中，ASCII 字符集是最为常见的格式，而且在许多场合，它也是默认的格式。

用文本编辑器打开一个文本文件后，用户可以看到可读的纯文本内容。控制字符有时被编辑器当做文字指令，有时被当做像纯文本那样可编辑的转义字符。尽管文本文件里面有纯文本信息，但是通过特殊方法，文件内的控制字符（尤其是文件结束字符）可以让纯文本不可见。

5.5.2 换行符

ASCII 标准使得只含有 ASCII 字符的文本文件可以在 Unix、Macintosh、Microsoft Windows、DOS 和其它操作系统之间自由交互，而其它格式的文件是很难做到这一点的。但是，在这些操作系统中，换行符并不相同，处理非 ASCII 字符的方式也不一致。

在计算机领域中，换行（newline）或称为 Line break 或 end-of-line（EOL）字符是一种加在文字最后位置的特殊字符，在换行符的下一个字符将会出现在下一行，实际上换行符根据不同的硬体平台或作业系统平台会有不同的编码方式。

换行符可以看作是行的结束符，也可以看作行之间的分隔符，这两种处理方式之间存在一些歧义。如果换行符被当作分隔符，那么文件的最后一行就不需要再有换行符。但是多数系统的做法是在最后一行的后面也加上一个换行符，也就是把换行符看作是行的结束符。这样的程序在处理末行

没有换行字符的文件时，可能会存在问题。相反地，有的程序把换行符看作分隔符，就会把最末尾的换行字符看作是新行的开始，也就是多出了一个空行。

以 ASCII 为基础的或相容的字符集使用 LF（Line feed，换行符，`\n`）或 CR（Carriage Return，回车符，`\r`）或 CR+LF[14]。但不同操作系统对于换行符的表示方式有所区别[15]：

- CR+LF：MS-DOS、微软视窗作业系统（Microsoft Windows）、大部分非 Unix 的系统，文件尾部直接 EOF（文件结束标志）
- LF：在 Unix 或 Unix 相容系统（GNU/Linux, AIX, Xenix, Mac OS X, ...）、BeOS、Amiga、RISC OS，文件最后一行也会增加该字符，然后才是 EOF
- CR：Apple II 家族，Mac OS 至版本 9

微软的 MS-DOS 和 Windows 采用了相同的文本文件格式，它们都使用 CR 和 LF 两个字符作为换行符，这两个字符对应的 ASCII 码分别为 13 和 10。通常，最后一行文本并不以换行符（CR-LF 标志）结尾，包括记事本在内的很多文本编辑器也不在文件的最后添加换行符。

5.5.3　格式化文本

格式化文本（英文：formatted text、styled text、rich text），与纯文本（plain text）相对[16]，具有风格、排版等信息，如颜色、式样（黑体、斜体等）、字体尺寸、特性（如超链接）等。

格式化文本不等同于二进制文件，也不一定就不是 ASCII 文本。因为格式化文本不一定是二进制的，它也可以是一般的文本，如 HTML，RTF，enriched text 等文件，因此可以是 ASCII 文本文件。相反，一个纯文本（plain text）也可以非 ASCII 文件（如 UTF-8 编码的文件）。作为文本文件的格式化文本是用标记语言来写的。Microsoft Word 处理的格式化文本是二进制文件。

5.5.4　文本编辑器

文本编辑器是计算机软件中的一种。主要用于用来编写和查看文本文件。有一些特殊的文本编辑器支持增加自有的格式来丰富文档的表现形式。操作系统或者集成开发环境通常会带有可以查看和编辑纯文本的编辑器，可增加格式的文本编辑器通常是个人用户或者公司在制作需要格式的文件中使用。因为程序和程序的配置文件通常是以纯文本形式储存，所以程序员也需要使用纯文本编辑器来编辑程序和配置文件。

由文本编辑器创建的纯文本文件与由诸如 Microsoft Word、iWork 与 OpenOffice.org 之类的文字处理器创建的文档文件之间，存在着明显的不同：

- 当展现或者编辑纯文本文件时，需要显示文件所含的所有字符。起标记作用的只有那些文本文件所使用的字符集裡的控制字符，即换行符[17]、制表位和换页符，比如简单常用的字符集 ASCII。纯文本文件常用于编写程序和记录配置信息；或者自定义标签的格式如，XML、HTML。
- 文字处理器创建的文档除了包括字符集所定义的字符外，通常还包含由文件格式指定的控制字符。这些控制字符实现了诸如粗体、斜体、字体、列、表等功能。
- 文字处理器通常能够编辑并导出或保存为纯文本文件。但是，这样做必须很小心。当你用文字处理器处理源程序、HTML 代码、配置文件和控制文件的时候，要格外注意。否则，这些文件里就会包含那些文字处理器对应格式裡的特殊字符，这样你的文件就会出现问题。

有些文本编辑器简洁易用，只能编辑纯文本文件，也有些文本编辑器功能丰富复杂，支持大量格式的文本文件。Unix 和类 Unix 操作系统都提供了 vi 编辑器（或者是它的变种），但是也有

[14] 如何检测一个文本文件使用的哪种换行符？
[15] 怎样在不同的换行符间进行转换？
[16] 纯文本与格式化文本有什么区别？
[17] 不同操作系统中的换行符有何不同？

部分操作系统提供了 Emacs 编辑器[18]。这两个文本编辑器以其独特的快捷键操作和可定制的工作环境而在程序员世界中广泛应用[19]；普通用户则喜欢使用 Office 来为文本添加格式。Windows 系统同样提供了可编辑纯文本文件的记事本和一个类似于 Word 的写字板提供了基本的写作功能。也有很多人，尤其是程序员更喜欢 Windows 下功能更丰富，更专业的其它编辑器[20]，如 Notepad++、UltraEdit和Sublime Text 等。OS X 所提供的 TextEdit 可以编辑纯文本也可以编辑多信息文本，类似于 DOS 中的 Wordstar。随着计算机网络的发展，又诞生了可以多人协同编辑的文本编辑器，如 Windows 操作系统下的 Netpad，它可以让多人同时协同编辑或阅读同一个文本文件。

✍ 参考资料

1. 好用、强大的文本编辑器
2. 主流文本编辑器的学习曲线
3. 不同平台间换行符的差异与转换
4. 生物信息学培训资料
5. List of text editors(Wikipedia)
6. Comparison of text editors(Wikipedia)
7. 换行（维基百科）
8. 文本编辑器（维基百科）
9. 文件编辑器比较（维基百科）
10. 编辑器之战（维基百科）

5.6 基因组坐标的逻辑运算

使用基因组坐标进行基因组注释工作时，常常需要对坐标进行比较操作，类似于集合运算（set operations）[21]，主要包括交集（intersect），减法（subtract），合并（merge），串联（concatenate），补集（complement），聚类（cluster），联合（join），等[22]。

基因组坐标的交集是指提取出两组基因组特征坐标中完全重叠的坐标位置或有重叠的基因组特征。比如，在找出含有 SNP 的所有外显子时就需要对外显子数据和 SNP 数据进行交集运算。基因组坐标间的减法与交集相反，是指去除完全重叠的坐标位置，或去除有重叠的基因组特征而只保留完全没有重叠的特征。比如，在找出不含有 SNP 的所有外显子时就需要对外显子数据和 SNP 数据进行减法运算。基因组坐标的合并类似于取并集，是把多个有重叠的坐标位置或基因组特征合并成一个大的坐标或特征。比如，在把有重叠的小的重复元件（repetitive element）合并成大的重复片段时就需要进行合并运算。基因组坐标的串联只是简单的把两组坐标合并起来而已，不进行任何其他操作；如果第一组坐标有 M 条记录（record），第二组坐标有 N 条记录，那么串联后将有 $M+N$ 条记录。比如，在把分别含有外显子和内含子数据的两个文件合成一个大文件时使用的就是串联操作。基因组坐标的补集是指依据基因组坐标全集（如某条染色体的全长，全部基因组等）对当前的这组坐标或特征取补集。比如，以全基因组为全集，对所有基因的坐标取补集，可以得到全基因组上的基因间区域。基因组坐标的聚类[23]是指根据设定的最小坐标间隔以及聚类需要的最小记录数目，将所有符合要求的坐标聚合成一个坐标，或把所有符合要求的特征聚合成一个大的特征。比如，根据某条染色体上的基因数据和设定的标准，将基因进行聚类可以找到染色体上的基因富集区。[24]

[18] Vim 被称为“编辑器之神”；而 Emacs 则被称为“神的编辑器”。

[19] 世界上只有三种/类文本编辑器：一是 Vim，二是 Emacs，第三类就是其他了。

[20] 周围同事不是用 Vim 就是 Emacs，你要是用一 UltraEdit，你都不好意思跟人家打招呼。什么插件呀，语法高亮呀，拼写检查呀，能给它开的都给它开着，就是一个字儿：酷。你说这么牛一东西，怎么着学会也得小半年吧。半年？入门都远着呢，能学会移动光标就不错了，你还别说耗不起，就这还只是左右移动。

[21] 数学中常见的集合运算有哪些？

[22] 基因组坐标的其他逻辑运算模式：coverage、flank、closest、slop、window、……

[23] merge 和 cluster 有什么区别？

[24] 以下常见操作需要使用那种逻辑运算：找到和 LINE 有重叠的基因（intersect），找到和 SINE 没有重叠的基因（intersect -v）；从基因中去除内含子区域，只保留剩余的外显子区域（subtract）；找到人类基因组中没有被重复元件覆

基因组坐标的联合[25]会比较两组坐标或特征，根据坐标的重叠情况，把两组坐标或特征中相应的记录对应起来，第二组坐标或特征中相应的记录会紧跟在第一组坐标或特征的对应记录后面。如：根据外显子数据和 SNP 数据，找到含有 SNP 的所有外显子，并将其上的 SNP 信息附加到对应的外显子记录上。根据参数设定的不同，基因组坐标联合后的输出结果也会有所不同。

可以对基因组坐标进行逻辑运算的常用工具有：

- Galaxy 中的“Operate on Genomic Intervals”
- bedtools: a powerful toolset for genome arithmetic
- BEDOPS: the fast, highly scalable and easily-parallelizable genome analysis toolkit

5.7　总结与答疑

本次课涉及的主要知识点与技能：

- 知识点——基因组注释基础：
  - 基因组组装版本——对应关系。
  - 两种坐标系统——0-based 和 1-based。
  - 四种常用格式——FASTA，BED，GFF，VCF。
  - 坐标逻辑运算——常见模式及其适用范围。
  - 坐标转换、格式转换、逻辑运算的工具。
- 技能——纯文本与文本编辑器：
  - 纯文本与格式化文本。
  - 不同操作系统中的换行符。
  - 文本编辑器——Notepad++，Vim，Emacs。

5.8　回顾与导入

上次课主要介绍了基因组的组装版本、两种常用的坐标系统、四种数据格式（FASTA、BED、GFF、VCF）和基因组坐标的逻辑运算模式，这些都是基因组功能注释的基础知识和准备工作，后续的功能注释还包括变异位点的注释、富集分析等。本次课就简单介绍基因组功能注释工作中的几个高级注释内容，包括基因组变异位点的注释、基因集的富集分析（gene set enrichment analysis）和序列标识（sequence logo）的制作。

5.9　变异位点的注释

随着高通量技术的发展，基因组数据的数据量越来越大。对第二代测序数据等大规模基因组数据进行处理后，往往会得到一大批与参考基因组不同的单核苷酸变异（SNVs，Single Nucleotide Variations）位点[26]。面对成千上万的变异位点，不管是筛选编码区的变异，还是更进一步寻找疾病的候选基因，此时都无从下手，因为最初往往仅有这些变异位点的基因组坐标和 ATGC 四种碱基的变异信息。为了顺利开展后续的工作，首先要做的就是对这些 SNVs 进行注释，即为这些 SNVs 附加相关的基因组注释信息，包括变异位点在 dbSNP 数据库中的 ID、变异位点所在基因的名称（gene name）与索引号（accession number）、变异功能类别（错义突变、同义突变、无义突变等）、导致的氨基酸变化及其在蛋白质产物中的位置等各种相关信息。因为需要注释的 SNVs 数量巨大，所以需要使用专门的工具来对它们进行批量注释[27]，常用的工具有 SeattleSeq Annotation、variant tools 和 SnpEff 等。其中，SnpEff 已经集成到了 Galaxy 中，位于“snpEff”工具集中。

SeattleSeq Annotation 是美国国家心脏、肺和血液研究所（NHLBI, National Heart, Lung and Blood Institute）支持赞助开发的专门用于注释 SNVs 的在线工具。SeattleSeq Annotation 功能强大，可以

盖的区域（complement）；找到 novel SNPs，即不在 dbSNP 和 1000 genome 中的 SNPs（intersect -v）。

[25] intersect 和 join 有什么区别？

[26] SNV 和 SNP 有什么区别？什么是 SNP？

[27] 尝试查找更多用于变异位点注释的工具。

注释已知或未知的 SNVs，此外，还可以对小的插入缺失（indels）进行注释。SeattleSeq Annotation 的注释结果包括 dbSNP 中的 ID、基因名、基因索引号、变异功能类别、氨基酸改变及其在蛋白质序列中的位置、保守性分值、HapMap 数据库中的频率、PolyPhen 对变异危害性的预测、变异与临床疾病的关系等诸多内容。此外，SeattleSeq Annotation 在注释结果中还给出了相关基因组注释数据库或站点的链接，方便用户进一步理解注释结果。

因为 SeattleSeq Annotation 限定了输入文件的格式（GFF，VCF，Maq[28]等），所以需要通过特定的软件直接生成或者通过格式转换工具制作符合要求的输入文件。

在线看到的注释结果，下半部分是以表格形式呈现的具体注释信息；上半部分左侧是对 SNVs 的简单统计，右侧则是各种筛选选项，勾选或点选需要的选项后点击“reset”即可对下半部分表格的呈现方式及内容进行调整。除了在注释结束后立即在线查看注释结果外，也可以通过上传邮箱中接收到的注释结果拷贝，随时在线查看注释结果。在注释结果中，inDBSNPOrNot 列表明 SNVs 是不是在数据库中，可以用来挑选全新（novel）的 SNVs；functionGVS 和 functionDBSNP 两列是对 SNVs 功能类别的说明，可以用来过滤感兴趣的类别；polyPhen 列是对氨基酸改变危害性的预测，scorePhastCons 列是对此位点保守性的评估，结合两者及其他的相关信息可以挑选出理论上比较重要的候选 SNVs；geneList 列给出了变异位点涉及的基因。结合实际工作需要，充分利用 SeattleSeq Annotation 注释结果中的各种信息，可以大大缩小候选范围，减少后续的实验工作量，加快课题进展。

注释完 SNVs 后，根据变异功能类别进行筛选，一般重点关注编码区的非同义多态性（non-synonymous polymorphisms）对蛋白质产物结构和功能的影响，此时，可以使用 SIFT(Sort Intolerant From Tolerant)、PolyPhen-2(**Poly**morphism **Phen**otyping v**2**) 和 SNPs3D 等工具来完成此工作。其中，SIFT 通过序列同源性比较来预测蛋白质中的氨基酸改变对表型的影响，它基于蛋白质进化与蛋白质功能密切相关的前提，即：对蛋白质功能重要的位点在家族聚类中会比较保守，反之亦然；PolyPhen-2 同时分析蛋白质序列的保守性和氨基酸的理化性质，从而推测氨基酸改变对蛋白质结构和功能的影响；SNPs3D 使用基于蛋白质序列（profile model）和结构（stability model）的两个模型来预测氨基酸改变的功能性影响。除了 SNVs，如果还想预测 indels 对蛋白质生物学功能的影响，可以使用 PROVEAN(Protein Variation Effect Analyzer)。SIFT 已经整合到了 Galaxy 中，所以也可以在 Galaxy 中使用 SIFT，它位于“Phenotype Association”工具集中。

SIFT 有着自己的输入文件格式，因此也要先制作符合格式要求的输入文件。

在最终的注释结果中，Transcript ID 和 Protein ID 两列分别表示变异影响到的转录本和蛋白质，Substitution 列包含了氨基酸改变及其位置信息，Region 和 SNP Type 两列表明了变异位点所在的区域及其功能类别，SIFT Score 列是具体的 SIFT 数值，Prediction 列则是根据 SIFT 值给出的文字性描述，据此可以对变异位点进行初步的筛选。如果在 Output Options 中选择了“Gene Name”和“Gene Description”，在注释结果中还可以看到变异影响到的基因及基因的详细描述。

5.10 基因集富集分析

在基因组功能注释工作中，通过对变异位点进行注释和筛选，常常会得到一系列与疾病或表型相关的候选基因，这成百上千的基因构成一个基因列表（gene list），也叫做基因集（gene set）。因为基因功能及其参与的代谢通路的多样性与复杂性，为了对基因集中的所有基因有一个整体上的了解，同时也为了给后续实验提供一个更加明确的方向，一般都需要对得到的基因集进行功能注释，主要是针对 GO(Gene Ontology) 和 KEGG(Kyoto Encyclopedia of Genes and Genomes) 的富集分析（enrichment analysis）[29]。

基因本体（gene ontology，GO）数据库[30]是 GO 组织（GO consortium）在 2000 年构建的一个

[28]Maq 格式是什么样的格式？为什么叫做 Maq 格式？

[29]查阅 GO 和 KEGG 的相关资料。

[30]此部分摘抄自《生物信息学》（李霞，人民卫生出版社），第八章（基因注释与功能分类），第二节（基因注释数

结构化的标准生物学模型，旨在建立基因及其产物知识的标准词汇体系，涵盖了基因的细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）三个方面，目前已经成为应用最广泛的基因注释体系之一。GO 数据库最初收录的基因信息来源于 3 个模式生物数据库：果蝇、酵母和小鼠，随后相继收录了更多数据，其中包括国际上主要的植物、动物和微生物基因组数据库。GO 术语在多个合作数据库中的统一使用，促进了各类数据库对基因描述的一致。

GO 通过控制注释词汇的层次结构使得研究人员能够从不同层面查询和使用基因注释信息。从整体上来看 GO 注释系统是一个有向无环图（directed acyclic graphs），包含三个分支，即：生物学过程（biological process），分子功能（molecular function）和细胞组分（cellular component）。注释系统中每一个结点（node）都是基因或蛋白质的一种描述，结点之间保持严格的关系，即“is a”或“part of”。因此，一个基因或蛋白质可从三个层面得到注释，即基因或蛋白质参与的生物学过程，在细胞内的特定组分，以及分子功能上所扮演的角色。随着生命科学研究的逐步深入，GO 注释数据库正在不断积累和更新。目前 GO 已经成为生物信息领域中一个重要的资源和工具，并正在逐步改变着人们对各种生物学数据的组织和理解方式，它的存在极大地加快了生物数据的整合和利用。

京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）[31]是系统分析基因功能、基因组信息的数据库，它整合了基因组学、生物化学以及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体进行研究。KEGG 提供的整合代谢通路查询十分出色，包括碳水化合物、核苷酸、氨基酸等代谢及有机物的生物讲解，不仅提供了所有可能的代谢通路，还对催化各步反应的酶进行了全面的注解，包含其氨基酸序列、到 PDB 数据库的链接等。此外，KEGG 还提供基于 Java 的图形工具访问基因组图谱、比较基因组图谱和操作表达图谱以及其他序列比较、图形比较和通路计算的工具。因此，KEGG 数据库是进行生物体内代谢分析、代谢网络分析等研究的强有力工具之一。

一个生物过程通常是由一组基因共同参与，而不是单个基因单独完成。富集分析[32]的主要依据是，如果一个生物学过程在已知的研究中发生异常，则共同发挥功能的基因极可能被选择出来作为一个与这一过程相关的基因集合。因此，富集分析方法通常是分析一组基因在某个功能结点上是否过出现（over-presentation）。这个原理可以由单个基因的注释分析发展到大基因集合的成组分析。由于分析的结论是基于一组相关的基因，而不是根据单个基因，所以富集分析方法增加了研究的可靠性，同时也能够识别出与生物现象最相关的生物过程。目前最常用的是基于 GO 或 KEGG 的富集分析。人们通过多种方法获得大量的感兴趣基因，如差异表达基因集、共表达基因模块、蛋白质复合物基因簇等，然后寻找这些感兴趣基因集显著富集的 GO 结点或 KEGG 通路，这有助于知道进一步深入细致的实验研究。

DAVID(Database for Annotation, Visualization and Integrated Discovery) 是一个整合了大量生物学数据和多种分析工具的生物信息数据库，为大规模的基因集或蛋白质集提供系统综合的生物功能注释，帮助用户提取并分析注释信息。它将输入列表中成百上千的基因关联到指定的生物学注释，进而从统计学的层面上，找出最显著富集的注释项目（term）。DAVID 是基因集富集分析中使用最为广泛的工具之一。在富集分析中，除了待分析的基因集外，还需要一个背景集作为对照。一般使用相应物种基因组中的全部基因作为背景集，当然，也可以指定另外一套基因集作为待分析基因集的对照。DAVID 是一个综合工具，不但提供基因富集分析，还提供基因间 ID 的转换、基因功能的分类等。

在最终的富集分析结果中，Term 列是具体的注释项目名称，P-Value 列是通过 Fisher 精确检

据库）。

[31]此部分摘抄自《生物信息学》（李霞，人民卫生出版社），第八章（基因注释与功能分类），第二节（基因注释数据库）。

[32]此部分摘抄自《生物信息学》（李霞，人民卫生出版社），第八章（基因注释与功能分类），第三节（基因集功能富集分析）。

验得到的 P 值，是原始的富集显著性数值，Benjamini 列则是进行多重检验校正（multiple testing correction）（本杰明假阳性率校正）后的显著性数值[33]，一般以此为标准进行筛选。点击打开“Options”可以根据需要调整参数，点击“Download File”可以把富集分析的结果下载保存到本地。如果想对结果中每一列的含义有更加深入的了解，可以点击右上角的“Help and Manual”打开帮助页面，其中对结果和参数都有详细的解释。

DAVID 提供了针对四项分析内容的六个分析工具：

- Gene Name Batch Viewer：把基因 ID 转换成基因名称，从而可以直观地查看基因集，初步判断基因集的质量是否满足要求。
- Gene ID Conversion Tool：在不同数据库的基因 ID 间进行转换，包括 NCBI、Ensembl、UCSC、PIR 和 UniProt 等常用数据库。
- Gene Functional Classification Tool：根据注释信息将功能相关的基因聚成一类，进而在功能注释的层面上分析基因集。
- Functional Annotation Tool：DAVID 最核心的分析内容，包括三个分析工具：
  - Functional Annotation Clustering：基于基因集中基因的功能注释对注释项目进行聚类。
  - Functional Annotation Chart：根据功能注释对基因集中的基因进行富集分析，总共有 80 多个注释项目可供选择。
  - Functional Annotation Table：以表格形式呈现基因集中每个基因在不同数据库中的功能注释。

DAVID 提供的分析工具在分析内容上可能有所重叠，但都有自己的分析侧重点，在实际使用时需要选择最合适的一个或多个工具[34]。

5.11 序列标识

序列标识（sequence logo）是基于 DNA、RNA 和蛋白质的多序列比对信息，把多序列的保守性信息通过图形表示出来。序列标识常用于图形化展示转录因子结合位点（TFBS，Transcription Factor Binding Site）等序列基序（sequence motif）的一致性序列（consensus sequence），但它提供了一个比一致性序列更丰富、更精确的序列相似性描述[35]。每个序列标识由一系列堆叠的核苷酸或氨基酸组成，横轴（x-axis）表示序列的位置（position），纵轴（y-axis）默认是以比特（bits）为计量单位的保守性。在每一个序列位置上用字符堆叠的总高度表示此位置的保守性，堆叠中每个字符的高度表示此位置上核苷酸或氨基酸出现的相对频率。WebLogo 是一个灵活方便的序列标识产生器，最常用的是其网络版本，但也有命令行界面的本地版可供选用。此外，WebLogo 也已经集成到了 Galaxy 中，位于“Motif Tools”工具集中。此外，还有 enoLOGOS、Skylign、Two Sample Logo、ECplot、pLogo、makelogo、CorreLogo、seqLogo、GENIO/logo、CodonLogo、RILogo 等众多类似的工具可供选用。

序列标识图（sequence logo）[36]依次绘出模体中各个位置上出现的碱基，每个位置上所有碱基的累积可反映出该位置上碱基的一致性，每个碱基字母的大小与碱基在该位置上出现的频率成正比。这种表示方法直观地给出模体各个位置上碱基出现的倾向性和整个模体的序列的一致性，应用非常广泛。

为了使用 WebLogo 制作序列标识，需要先进行多序列比对，获得以 CLUSTALW、FASTA 或

[33]什么是多重检验校正？为什么要进行校正？校正的方法有哪些？（文献：Noble, William S. 2009. How does multiple testing correction work? *Nature biotechnology* 27, no. 12 (December): 1135-7. doi:10.1038/nbt1209-1135. http://www.ncbi.nlm.nih.gov/pubmed/20010596.）

[34]查找 DAVID 的使用 protocol。（文献：Huang, D. W., Sherman, B. T. & Lempicki, R. A Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 2009, 4, 44–57.）

[35]序列基序、一致性序列、序列标识之间的关系。

[36]此部分摘抄自《生物信息学》（李霞，人民卫生出版社），第十一章（转录调控的信息学分析），第三节（转录因子结合位点的信息学预测方法）。

MSF 等格式[37]保存的比对结果。

无论何种组织或生物，大多数真核内含子遵照“GT-AG 规则”。此规则的命名来源于这样的事实：几乎所有内含子 DNA 序列 5' 端起始的两个核苷酸总是 5'-GT-3'，而 3' 端的最后两个核苷酸始终是 5'-AG-3'。由于这两个碱基序列的高度保守性和广泛存在性，将其称为 GT-AG 法则，即 5'-GT···AG-3'。

从两个图中可以看出人类基因剪接位点的 GT-AG 规则[38]：内含子的 5' 端是供体位点的 GT，3' 端是受体位点的 AG。对于受体位点 AG 前的 -3 位置来说，其总高度仅有 AG 两个位点高度的一半，说明其保守性远不及 AG 两个位点；此位点中 C 的高度大约是 T 的高度的两倍，说明此位点 C 出现的频率比较高，大约有三分之二，而 T 出现的频率仅为三分之一。

✍ 课外阅读

1. D’haeseleer, P. What are DNA sequence motifs? Nat. Biotechnol. 24, 423–5 (2006).
2. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–100 (1990).
3. Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res. 14, 1188–90 (2004).
4. Workman, C. T. et al. enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. 33, W389–92 (2005).
5. Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics 15, 7 (2014).
6. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22, 1536–7 (2006).
7. A very incomplete list of sequence logo servers and extensions to logos
8. Recommendations for Making Sequence Logos

5.12 box plot

box plot（boxplot、Box-whisker Plot，箱线图、箱须图、盒须图、盒式图、盒状图）于 1977 年由美国着名统计学家约翰 • 图基（John Tukey）发明，是一种用作显示一组数据分散情况资料的统计图，因形状如箱子而得名。它能显示出一组数据的最大值、最小值、中位数、下四分位数（第一四分位数）及上四分位数（第三四分位数）。通过 boxplot 可以粗略地看出数据是否具有有对称性、分布的离散程度等信息，因此特别适合用于几个样本的比较。

box plot 美中不足之处在于：它不能提供关于数据分布偏态和尾重程度的精确度量；对于批量较大的数据集，box plot 反映的形状信息更加模糊；用中位数代表总体平均水平有一定的局限性等等。所以，应用 box plot 时最好结合其它描述统计工具如均值、标准差、偏度、分布函数等来描述数据集的分布形状。

box plot 涉及的主要概念包括：

- 最小值 min，最大值 max
- 中位数 median
- 下四分位数 Q1，上四分位数 Q3
- 四分位数差 IQR（interquartile range），$IQR = Q3 - Q1$
- 内限：$Q3 + 1.5IQR$，$Q1 - 1.5IQR$
- 外限：$Q3 + 3IQR$，$Q1 - 3IQR$
- 异常值（outliers）：处于内限以外的数据

[37]查找 CLUSTALW、MSF 格式的相关资料。

[38]人类基因剪接位点百分之百都符合 GT-AG 规则吗？（分子生物学中的不确定性，总有例外存在。）

- 温和的异常值（mild outliers）：在内限与外限之间的异常值
- 极端的异常值（extreme outliers）：在外限以外的异常值

绘制 box plot 的主要步骤如下所示：

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 $Q3+1.5IQR$ 和 $Q1-1.5IQR$ 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 $Q3+3IQR$ 和 $Q1-3IQR$ 处画两条线段，称为外限。[39]
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心圆表示；极端的异常值用实心点（一说用星号 *）表示。

绘制 box plot 时，可以使用 BoxPlotR、ECplot、Galaxy（“Graph/Display Data”工具集中的 Boxplot）、R 等。

除了传统的 box plot 外，还有 notched box plot、variable width box plot、variable notched box plot、box-percentile plot、histplot、vase plot、violin plot 和 bean plot 等诸多变体。

✍ 课外阅读

1. Streit, M. & Gehlenborg, N. Points of View: Bar charts and box plots. Nat. Methods 11, 117–117 (2014).
2. Kick the bar chart habit. Nat. Methods 11, 113–113 (2014).
3. Krzywinski, M. & Altman, N. Points of Significance: Visualizing samples with box plots. Nat. Methods 11, 119–120 (2014).
4. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. Nat. Methods 11, 121–122 (2014).
5. Kampstra, P. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. Journal of Statistical Software 28, 1–9 (2008).
6. 40 years of boxplots
7. Potter, K. Methods for Presenting Statistical Information: The Box Plot. Visualization of Large and Unstructured Data Sets S-4, 97–106 (2006).
8. Box Plot: Display of Distribution
9. Massart, D. L., Smeyers-Verbeke, J., Capron, X., & Schlesier, Karin. Visual Presentation of Data by Means of Box Plots. LC-GC Europe 18, p215 (2005).

5.13 解析图表

解析图表时，主要从以下几个方面入手：

- 表格
  - 行的含义（每一行代表什么），列的含义（每一列表示什么）
  - 缩写的含义（是什么的缩写）[40]
  - 数值的含义（数值范围，数值大小有何意义）
- 图片

[39] 统计软件绘制的箱线图一般都没有标出内限和外限。

[40] N/A, N.A., NA: not applicable, not available, no answer; NaN: not a number; No., #: numero/number.

  - 生成图片的数据
  - 横轴和纵轴的含义
  - 图片包含的元素（点、线、圆…）
  - 图片元素大小的含义
  - 图片元素颜色的含义

5.14　总结与答疑

本次课涉及的主要知识点与技能：

- 知识点——基因组功能的高级注释：
  - 变异位点的注释——用途，注释工具。
  - 基因集富集分析——功能，分析工具。
  - 序列标识——含义，制作工具。
  - box plot——理解，绘制。
- 技能——解析图表：
  - 表——行列，缩写，数值。
  - 图——数据，横纵轴，图元素，元素大小、颜色。

5.15　回顾与导入

前面介绍了基因组功能注释的基础知识和几个高级注释内容，现在介绍一下贯穿全章的生物信息学分析平台——Galaxy，重点通过实例演示 Galaxy 的使用。

5.16　Galaxy 分析平台

从前文所述可以看出，基因组功能注释的内容丰富多样，涉及的工具更是五花八门。而随着芯片、第二代测序等高通量技术的飞速发展，海量的基因组数据不断积累，分析工作越来越多，注释内容更加复杂。而研究人员则希望数据分析不但要系统化，具有一定的创新性，而且要更能体现个性化。基因组数据与研究人员要求之间的矛盾催生了专业、系统且扩展性高的生物信息学数据分析平台。这些分析平台大多集成了丰富且优秀的生物信息学工具，能够与常用数据库无缝衔接，使研究人员可以轻松快速地进行多样化的基因组功能注释，整个分析工作更加自动化、更具流程性。由 UCSC（加利福尼亚大学圣克鲁兹分校）开发的 Galaxy 便是其中的佼佼者。此外，GenePattern 也是一个不错的选择。

Galaxy 是一个开放的、基于网络的生物信息学分析平台，主要针对基因组的相关分析开发设计，特别适合与第二代测序数据相关的基因组注释工作。Galaxy 整合了各种生物信息学分析工具，并把相关的工具分类汇总到特定的工具集中，包括与公共数据库无缝集成的“Get Data”工具集、处理文本数据的“Text Manipulation”工具集，进行数据格式转换的“Convert Formats”工具集、进行基因组坐标逻辑运算的“Operate on Genomic Intervals”工具集、高级功能注释与分析的“Phenotype Association”工具集、用于统计绘图的“Statistics”和“Graph/Display Data”等工具集、专门用于第二代测序数据分析的“NGS Toolbox”工具集，此外，还集成了 EMBOSS 和 BEDTools 等优秀的开源工具，因此通过该平台可以在不下载和安装任何软件的前提下进行各种各样的生物信息学分析工作，即使没有任何编程经验的研究人员也可以快速简单地运行分析工具或分析流程。通过 Galaxy 还可以方便地构建生物数据分析工作流（workflow），而且在 Galaxy 中进行的工作会以历史记录的形式进行保存，从而使得计算分析过程可以被重复和共享。除了联网使用外，也可以将 Galaxy 部署在服务器中，或者本地安装到基于 UNIX/Linux 平台的个人电脑中离线使用。

Galaxy 界面主要由四部分组成：顶部是刊头（masthead），可以用来切换“分析数据”、“工作流”和“账号”等主界面；左侧栏是工具菜单（tool menu），以工具集的形式罗列着各种各样的分析工具，顶部有一个 search tools 搜索框方便用户查找工具；中间是工作区（work area），点选工

具后此处为参数设置与工具说明界面，查看或修改数据时此处会相应显示数据的内容、信息或其属性（attribute），总之，工作区是最主要的信息输出位置；右侧栏是历史面板（history panel），以历史记录形式存储每一步操作，其中绿色、黄色和灰色的步骤分别表示已经完成、正在运行、尚在排队的工作，而紫色则表示正在上传数据。历史面板除了可以实现下载数据、保存历史等常规操作外，还有一个重要的功能——将已经完成的多步工作提取为工作流，以便日后重复或与他人共享。

Galaxy 平台功能强大，但学习起来并不困难，因为该项目提供了丰富的学习资料。可以先从 Galaxy 101 和 Galaxy Screencasts and Demos 开始，对 Galaxy 有一个直观的认识之后，再学习并练习 Learn Galaxy 上的其他实例。更加深入的学习资料和使用指南可以在 Galaxy Wiki 找到[41]。

5.17 Galaxy 使用演示

5.17.1 人类基因的坐标转换

虽然基因组不同组装版本间的坐标不同，但它们之间是可以相互转换的。liftOver 是由 UCSC 基因组生物信息学组（UCSC Genome Bioinformatics Group）开发的坐标转换工具，它除了可以转换同一物种基因组不同组装版本间的坐标与注释文件外，还可以在不同物种间转化基因组坐标及相应的注释文件。对于 liftOver 来说，有多种形式的版本可供选用，如网页版的 liftOver，单机版的 liftOver，以及集成到 Galaxy 中的 liftOver。

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出
   - 过滤结果：MAPPED VS. UNMAPPED

使用 liftOver 转换坐标后，一般会生成两个文件：转换成功的坐标和转换失败的坐标。一般情况下，只需关注可以成功转换的坐标即可。在 Galaxy 中，标记有“[MAPPED COORDINATES]”即是转换成功的 hg18 坐标；点击右侧的眼睛图标可以直接在 Galaxy 中查看结果，点击软盘图标可以将结果下载保存至本地电脑中。如果有转换失败的坐标，在“[UNMAPPED COORDINATES]”的结果文件中，针对每一条记录都会给出转换失败的原因，常见的有：Partially deleted in new，Split in new，Deleted in new，等。

包括 liftOver，主要有以下一些工具可以用来转换基因组坐标：

- 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。

5.17.2 BED 与 GFF 格式间的互转

受数据库等数据来源的限制，或者为了满足软件与工具对输入文件格式的要求，有时需要把已有数据文件的格式转换成需要的特定格式。一般来说，凡是存储相似基因组信息的数据格式，基本上都可以进行双向转换。但因不同格式包含的信息量不同，有可能在双向转换时会丢失部分信息，甚至只能进行单向转换。BED 格式和 GFF 格式存储的都是基因组特征的信息，此处使用集成到 Galaxy 中的格式转换工具来演示这两种格式间的相互转换。

1. 获取输入

[41] 查找 Galaxy 的相关资料并进行初步学习。

- 输入文件：BED

2. 数据处理
   (a) BED ⇒ GFF
   (b) GFF ⇒ BED
3. 保存输出
   - 查看结果：互相比较

仔细比较最初的 BED 格式文件和最后 GFF 格式转换生成的 BED 格式文件，会发现虽然两者包含的信息基本一样，但文件形式与信息细节上却有一定的差异[42]。因此，在实际工作进行格式转换时，一定要多加留意，仔细检查转换后的格式是否符合要求、有没有丢失重要的信息。

5.17.3　外显子和 SNP 的比较

因实际工作的需要，可能需要对两组基因组坐标进行各种逻辑运算。此时有众多工具可供选用，如集成到 Galaxy 中的“Operate on Genomic Intervals”工具集和单机版的 BEDTools 等。此处使用集成到 Galaxy 中的工具集，通过比较外显子和 SNP 来演示基因组坐标的减法和联合运算。

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join[43]
3. 保存输出
   - 解析结果

从最终的统计结果来看，人类 hg19 基因组 Y 染色体上一共有 4109 个外显子和 27265 个 SNP，其中不含 SNP 的外显子有 3673 个，含有 SNP 的外显子则应该有 436 个。进行联合操作后的结果中有 983 条记录，每一行的前半部分是外显子的信息，后半部分是 SNP 的信息。此文件的记录条数之所以多于理论上的 436 条，是因为有的外显子上有不止一个 SNP。对于处理完的结果，可以直接查看并继续保存在 Galaxy 中用于后续处理，也可以下载保存至本地电脑中。

5.17.4　Galaxy 的综合运用

以人类基因组中的 Y 染色体为例，找到含有 SNP 数目最多的外显子[44]：

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[45]: 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

历史记录的保存，工作流的提取、制作、使用和分享：

1. Save: rename the history as "Exons and SNPs"

[42]尝试一下 GFF ⇒ BED ⇒ GFF，并比较两个 GFF 文件的异同。
[43]尝试一下 intersect，并比较 intersect 和 join 两者的输出结果。
[44]通过更多的实例深入学习 Galaxy。
[45]此处有两种方法：Filter，或者在 Sort 后 Select First。

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)

5.18 “输入 -加工 -输出”三段论

整个数据处理的过程主要涉及最初输入、中间处理、最终输出三个过程：

1. 输入
   - 需要多少输入文件
   - 数据文件的来源
   - 文件格式是否符合要求
   - 是否需要过滤数据
2. 加工
   - 选择合适的工具及其版本
   - 调整设定参数
   - 具体操作步骤
3. 输出
   - 输出文件的格式说明
   - 图表内容的解释
   - 解析、过滤、保存结果

5.19 总结与答疑

本次课涉及的主要知识点与技能：

- 知识点——Galaxy 分析平台：
  - Galaxy——界面，学习，使用。
- 技能——“输入 -加工 -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。

☕ 课后思考

1. hg19 和 mm10 分别代表什么含义？hg19 是和 GRCh37 相对应，还是和 GRCm38 相对应？
2. 常见的基因组坐标系统是哪两种，举例进行说明。
3. 简述 BED 格式前 6 列的含义，能解释实际的 BED 记录。
4. 不同操作系统的换行符有何区别？
5. 基于基因组坐标的常见逻辑运算模式有哪些，画图进行解释。
6. 简述序列标识的含义，能解释实际的序列标识图。
7. 以 SNP 的注释结果为例，论述如何解析一张表。
8. 以 box plot 为例，论述如何解析一张图。
9. 以坐标转换为例，论述“输入 -加工 -输出”的工作流程。