第 5 章　基因组功能注释分析

标题页

5.1　课前甜点

5, -00:05--00:00 (^1.1-)

播放第二代测序的动画视频。

5.2　引言

5, 00:00--00:05 (-1.1-)

结构注释与功能注释

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。

本章提纲

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。本章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。

5.3　基因组组装版本与坐标系统

15, 00:05--00:20 (-1.1-)

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，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。

XP 与 Win7；迅雷 5 与迅雷 7

提问：举个操作系统或软件版本的例子。

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。NCBI 和 UCSC 之间的基因组组装版本存在着明确的的对应关系，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。

人类和小鼠的基因组组装版本对应关系

提问：hg、mm、GRCh、GRCm 分别表示什么含义？

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。

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）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-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，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。

基因组坐标系统示意图

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)”；如果采用 1-based 的坐标系统，则表示为“[4,5]”。

提问：0-based 坐标系统的优势何在？

提问：[课后思考] 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？）

提问：举一个和这两种坐标系统类似的例子。

英式和美式英语中的 first floor

为了便于理解这两种坐标系统，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，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。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。

提问：[课后思考]SAM、BAM、Wiggle、PSL 格式；DAS 工具。

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。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，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。

FASTA 格式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）。

提问：[课后思考]FASTQ 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

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！”

数据格式与坐标系统

20, 00:20--00:40 (-1.1-)

txt 与 doc；mkv 与 doc

FASTA 格式示例

核酸与氨基酸代码

BED 格式示例

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）

GFF 格式示例

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。

1. seqid: ID of the feature, 特征的 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

VCF 格式示例

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

提问：[课后思考] 学习 GenBank、EMBL、PDB 等其他数据格式。

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打开查看。

提问：[课后思考] 纯文本（plain text）与格式化文本（formatted text，styled text，rich text）的区别。

提问：[课后思考] 文本编辑器与换行符。（扩展阅读：好用、强大的文本编辑器；主流文本编辑器的学习曲线；不同平台间换行符的差异与转换；生物信息学培训资料；List of text editors(Wikipedia)；Comparison of text editors(Wikipedia)；文本编辑器（维基百科）；文件编辑器比较（维基百科）；编辑器之战（维基百科））

5.5 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），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。

提问：数学中常见的集合的运算有哪些？

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。

基因组坐标的联合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。

提问：intersect 和 join 的区别；merge 和 cluster 的区别。

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

提问：[课后思考] 基因组坐标的其他逻辑运算模式：coverage、flank、closest、clop、……

文本编辑器学习曲线

Vim 被称为“编辑器之神”；而 Emacs 则被称为“神的编辑器”。

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。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。

25, 00:40--00:15 (-1.1-1.2-)

集合运算示意图

集合运算 ⇒ 逻辑运算

逻辑运算模式

join 示意图

25, 00:15--00:40 (-1.2-)

5.6　操作演示

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。

提问：[演示] 人类基因的坐标转换。

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。

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。

提问：[演示]BED 与 GFF 格式间的互转。

1. 获取输入
   - 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。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。

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。

提问：[演示] 外显子和 SNP 的比较。

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
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。

5.7 总结与答疑

10, 00:40--00:50 (-1.2$)

本次课需要掌握的知识点与技能：

- 知识点：
  - 基因组组装版本的对应关系。
  - 两种坐标系统——0-based 和 1-based。
  - 四种常用格式——FASTA，BED，GFF，VCF。
  - 逻辑运算模式。
  - 坐标转换、格式转换、逻辑运算的工具。
- 技能：
  - “输入 -加工 -输出”三段论。
  - 获取输入。
  - 数据处理。
  - 解析输出。

10, -00:10--00:00 (^2.1-)

5.8　课前甜点

播放 Galaxy 的动画视频。

5, 00:00--00:05 (-2.1-)

5.9　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，并推荐贯穿全章的生物信息学分析平台——Galaxy，最后通过一个实例重点演示 Galaxy 的使用。

10, 00:05--00:15 (-2.1-)

5.10　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。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，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 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。

提问：[课后思考]Maq 格式是什么样的？为什么叫做 Maq 格式？

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（nonsynonymous 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, http://provean.jcvi.org/index.php)。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。

提问：[课后思考] 尝试查找更多的用于变异位点注释的工具。

5.11 基因集富集分析

15, 00:15--00:30 (-2.1-)

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）。

提问：[课后思考] 查阅 GO 和 KEGG 的相关资料。

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。

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）后的显著性数值，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。

提问：[课后思考] 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.）

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 工具选择图

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。

提问：如何查找一个工具的使用方法与步骤？（期刊：*Nature protocols*，*Current Protocols*（*in Bioinformatics*），*SpringerProtocols*，*Methods in Molecular Biology* ）

提问：[课后思考] 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.）

5.12　序列标识

20, 00:30--00:50 (-2.1$)

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。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 可供选用。

提问：[课后思考] 序列基序、一致性序列、序列标识之间的关系。

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 MSF 等格式保存的比对结果。

提问：[课后思考] 查找 CLUSTALW、MSF 格式的相关资料。

提问：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。）

从两个图中可以看出人类基因剪接位点的 GT-AG 规则：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。

供体位点和受体位点的序列标识图

5.13 Galaxy 分析平台

10, 00:00--00:10 (^2.2-)

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。

提问：[课后思考] 查找 Galaxy 的相关资料并进行初步学习。

30, 00:10--00:40 (-2.2-)

5.14　Galaxy 使用演示

找到含有 SNP 数目最多的外显子（以 Y 染色体为例）：

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

两种方法：Filter，或者先 Sort 后 Select First。

从历史记录中提取工作流并进行分享：

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy
4. Modify: modify a parameter in the workflow in galaxy
5. Rerun: run workflows with galaxy
6. Share: share workflows with galaxy
7. Create: generate workflows from scratch with galaxy Find the 50 longest exons

提问：[课后思考] 通过实例深入学习 Galaxy。

5.15　总结与答疑

5, 00:45-00:50 (-2.2$)

本次课需要掌握的知识点与技能：

- 知识点：
  - 变异位点注释的用途及注释工具。
  - 基因集富集分析的功能及分析工具。
  - 序列标识的含义与制作工具。
  - Galaxy 分析平台的使用方法。
- 技能：
  - 查找工具使用的 protocol。
  - 学习新工具的方法与步骤。
  - 数据处理流程的保存与共享。