第 5 章　基因组功能注释分析

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（Next Generation Sequencing，NGS）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。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）。 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。

5.1 基因组注释的基础知识

5.1.1 基因组的组装版本

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，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。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。人类和小鼠的基因组组装版本对应关系如表 5.1 所示。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。

此处涉及的数据库与网站如下：

NCBI： http://www.ncbi.nlm.nih.gov

Ensembl：http://www.ensembl.org/index.html

UCSC： http://genome.ucsc.edu

List of UCSC genome releases：http://genome.ucsc.edu/FAQ/FAQreleases.html#release1

表 5.1 人类和小鼠的基因组组装版本

物种 SPECIES	UCSC 版本 UCSC VERSION	发布时间 RELEASE DATE	NCBI 版本 RELEASE NAME
Human	hg19	Feb. 2009	Genome Reference Consortium GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
	hg15	Apr. 2003	NCBI Build 33
Mouse	mm10	Dec. 2011	Genome Reference Consortium GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35
	mm6	Mar. 2005	NCBI Build 34

5.1.2 基因组的坐标系统

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（Single Nucleotide Polymorphism ，SNP）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。

两种坐标系统之间的区别可以用一个例子来进行说明，见图 5.1。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]”。

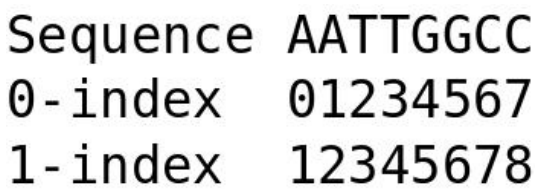


图 5.1 基因组的坐标系统

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。

对于两种坐标系统的详细区别和各自的优缺点，可以参看以下网站：

Database/browser start coordinates differ by 1 base：http://genome.ucsc.edu/FAQ/FAQtracks #tracks1

Coordinate Transforms ： http://genomewiki.ucsc.edu/index.php/Coordinate_Transforms

5.1.3 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。

(1) FASTA 格式

FASTA 格式是保存序列最常见的格式之一，如图 5.2 所示，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，见表 5.2 和表 5.3。此外，还应注意以下情况：

① 允许小写字母的存在，但会转换成大写；

② 单个“-”符号代表不明长度的空位；

③ 在氨基酸序列中允许出现“U”和符号“*”；

④ 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）。

```
>gi|142864|gb|M10040.1|BACDNAE B.subtilis dnaE gene encoding DNA primase, complete cds
GTACGACGGAGTGTTATAAGATGGGAAATCGGATACCAGATGAAATTGTGGATCAGGTGCAAAAGTCGGC
AGATATCGTTGAAGTCATAGGTGATTATGTTCAATTAAAGAAGCAAGGCCGAAACTACTTTGGACTCTGT
CCTTTTCATGGAGAAAGCACACCTTCGTTTTCCGTATCGCCCGACAAACAGATTTTTCATTGCTTTGGCT
GCGGAGCGGGCGGCAATGTTTTCTCTTTTTTAAGGCAGATGGAAGGCTATTCTTTTGCCGAGTCGGTTTC
TCACCTTGCTGACAAATACCAAATTGATTTTCCAGATGATATAACAGTCCATTCCGGAGCCCGGCCAGAG
TCTTCTGGAGAACAAAAAATGGCTGAGGCACATGAGCTCCTGAAGAAATTTTACCATCATTTGTTAATAA
ATACAAAAGAAGGTCAAGAGGCACTGGATTATCTGCTTTCTAGGGGCTTTACGAAAGAGCTGATTAATGA
ATTTCAGATTGGCTATGCTCTTGATTCTTGGGACTTTATCACGAAATTCCTTGTAAAGAGGGGATTTAGT
GAGGCGCAAATGGAAAAAGCGGGTCTCCTGATCAGACGCGAAGACGGAAGCGGATATTTCGACCGCTTCA
GAAACCGTGTCATGTTTCCGATCCATGATCATCACGGGGCTGTTGTTGCTTTCTCAGGCAGGGCTCTTGG
```

图 5.2 FASTA 格式示例

关于 FASTA 格式的更多说明可以参看以下网站：

BLAST help：http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

FASTA format：http://en.wikipedia.org/wiki/FASTA_format

What is FASTA format：http://zhanglab.ccmb.med.umich.edu/FASTA

FASTA format description：http://www.bioinformatics.nl/tools/crab_fasta.html

表 5.2 IUB/IUPAC 核酸代码表

Code 代码	Meaning 含义	Mnemonic 说明
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
R	A or G	puRine
Y	C, T or U	pYrimidines
K	G, T or U	bases which are Ketones
M	A or C	bases with aMino groups
S	C or G	Strong interaction
W	A, T or U	Weak interaction
B	not A (i.e. C, G, T or U)	B comes after A
D	not C (i.e. A, G, T or U)	D comes after C
H	not G (i.e. A, C, T or U)	H comes after G
V	neither T nor U (i.e. A, C or G)	V comes after U
N	A C G T U	aNy
X	masked	
-	gap of indeterminate length	

表 5.3 IUB/IUPAC 氨基酸代码表

1 简写	3 缩写	Meaning 含义	Name 名称
A	Ala	Alanine	丙氨酸
B	Asx	Aspartic acid or Asparagine	天冬氨酸或天冬酰胺
C	Cys	Cysteine	半胱氨酸
D	Asp	Aspartic acid	天冬氨酸
E	Glu	Glutamic acid	谷氨酸
F	Phe	Phenylalanine	苯丙氨酸
G	Gly	Glycine	甘氨酸
H	His	Histidine	组氨酸
I	Ile	Isoleucine	异亮氨酸
K	Lys	Lysine	赖氨酸
L	Leu	Leucine	亮氨酸
M	Met	Methionine	甲硫氨酸

N	Asn	Asparagine	天冬酰胺
O	Pyl	Pyrrolysine	吡咯赖氨酸
P	Pro	Proline	脯氨酸
Q	Gln	Glutamine	谷氨酰胺
R	Arg	Arginine	精氨酸
S	Ser	Serine	丝氨酸
T	Thr	Threonine	苏氨酸
U	Sec	Selenocysteine	硒代半胱氨酸
V	Val	Valine	缬氨酸
W	Trp	Tryptophan	色氨酸
Y	Tyr	Tyrosine	酪氨酸
Z	Glx	Glutamic acid or Glutamine	谷氨酸或谷氨酰胺
X	Xaa	any	不明氨基酸
*		translation stop	翻译终止
-		gap of indeterminate length	不明长度的空位

(2) BED 格式

BED（Browser Extensible Data）格式存储用于展示的特征注释信息，如图 5.3 所示，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），见表 5.4，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。

```
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
```

图 5.3 BED 格式示例

关于 BED 格式的更多说明可以参看以下网站：

BED format：https://genome.ucsc.edu/FAQ/FAQformat.html#format1

BED File Format - Definition and supported options：http://www.ensembl.org/info/website/upload/bed.html

表 5.4 BED 格式定义的 12 个项目

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

表中加粗的 3 个是必选项目，其他则是可选的附加项目。

(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。如图 5.4 所示，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（表 5.5）。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。

```
##gff-version 3
ctg123 . operon          1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  Parent=mrna0001
ctg123 . mRNA           10000 15000  .  +  .  ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon           10000 12000  .  +  .  Parent=mrna0002
ctg123 . exon           14000 15000  .  +  .  Parent=mrna0002
```

图 5.4 GFF 格式示例

表 5.5 GFF 格式定义的 9 个项目

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

关于 GFF 格式的更多说明可以参看以下网站：

Generic Feature Format Version 3 (GFF3)：http://www.sequenceontology.org/

gff3.shtml

GFF format：https://genome.ucsc.edu/FAQ/FAQformat.html#format3

GFF/GTF File Format - Definition and supported options：http://www.ensembl.org/info/website/upload/gff.html

GFF：http://gmod.org/wiki/GFF

General feature format：http://en.wikipedia.org/wiki/General_feature_format

(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。如图 5.5 所示，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，见表 5.6，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）。

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237 .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567 microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

图 5.5 VCF 格式示例

关于 VCF 格式的更多说明可以参看以下网站：

VCF (Variant Call Format) version 4.1：http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

VCF (Variant Call Format) version 4.0：http://www.1000genomes.org/node/101

Encoding Structural Variants in VCF (Variant Call Format)version 4.0：http://www.

1000genomes.org/wiki/Analysis/Variant%20Call%20Format/VCF%20%28Variant%20Call%20Format%29%20version%204.0/encoding-structural-variants

VCF format：https://genome.ucsc.edu/FAQ/FAQformat.html#format10.1

Variant Call Format：http://en.wikipedia.org/wiki/Variant_Call_Format

表 5.6 VCF 格式定义的 8 个必选项目

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

5.1.4 基因组坐标的逻辑运算模式

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。

如图 5.6（A）所示，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，如图 5.6（B）所示，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。如图 5.6（C）所示，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。如图 5.6（D）所示，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。如图 5.6（E）所示，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。如图 5.6（F）所示，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。

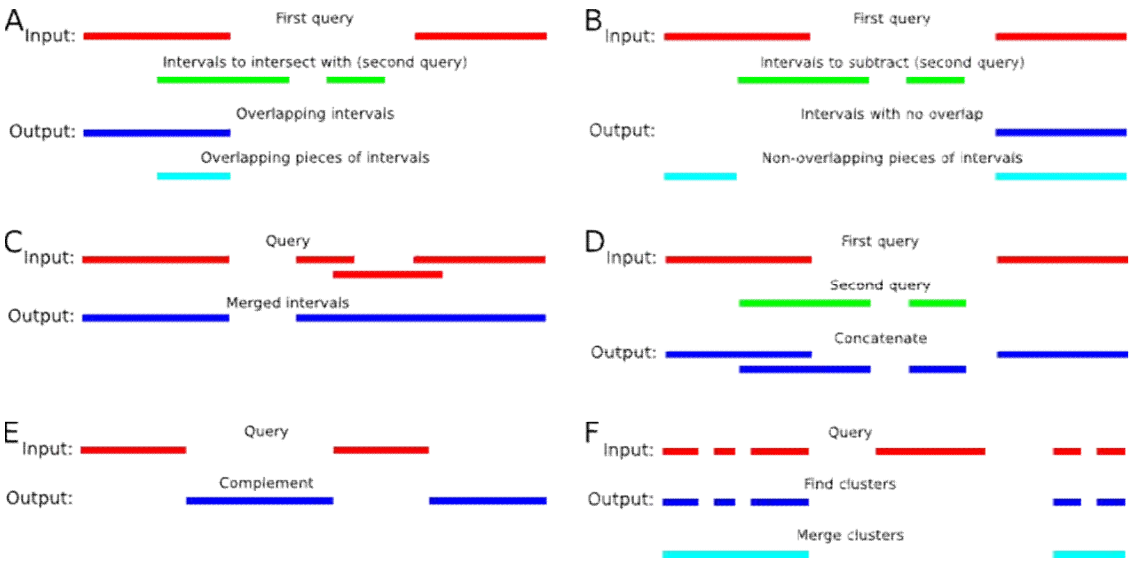


图 5.6 基因组坐标的逻辑运算模式

A. 交集（intersect）；B. 减法（subtract）；C. 合并（merge）；D. 串联（concatenate）；E. 补集（complement）；F. 聚类（cluster）。

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，如图 5.7 所示。

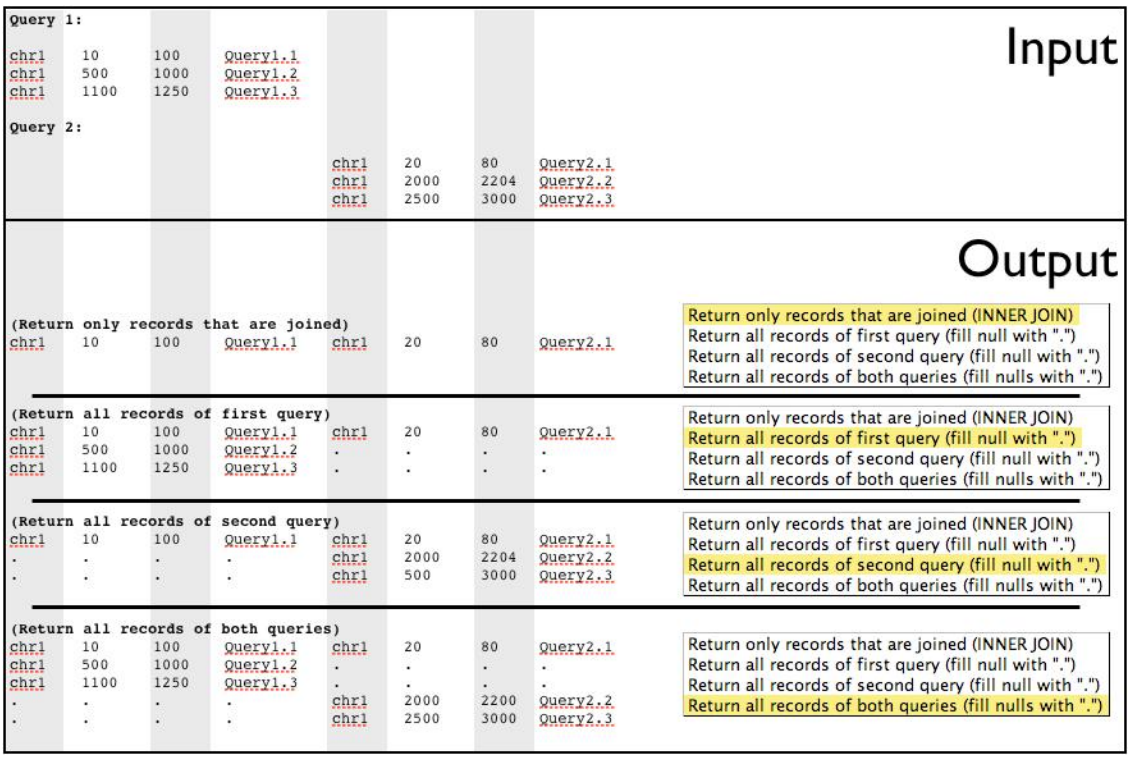


图 5.7 基因组坐标联合（join）操作的示意图

关于基因组坐标逻辑运算模式的更多内容可以参看以下网站：

Interval Operations in Galaxy：http://wiki.galaxyproject.org/Learn/Interval%20Operations

5.2 基因组功能注释的准备工作

5.2.1 基因组组装版本间的坐标转换

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。

例 5.1 人类全基因组基因的坐标转换。

使用集成到 Galaxy 中的 liftOver 工具，把人类全基因组基因的坐标从 hg19 转换至 hg18。

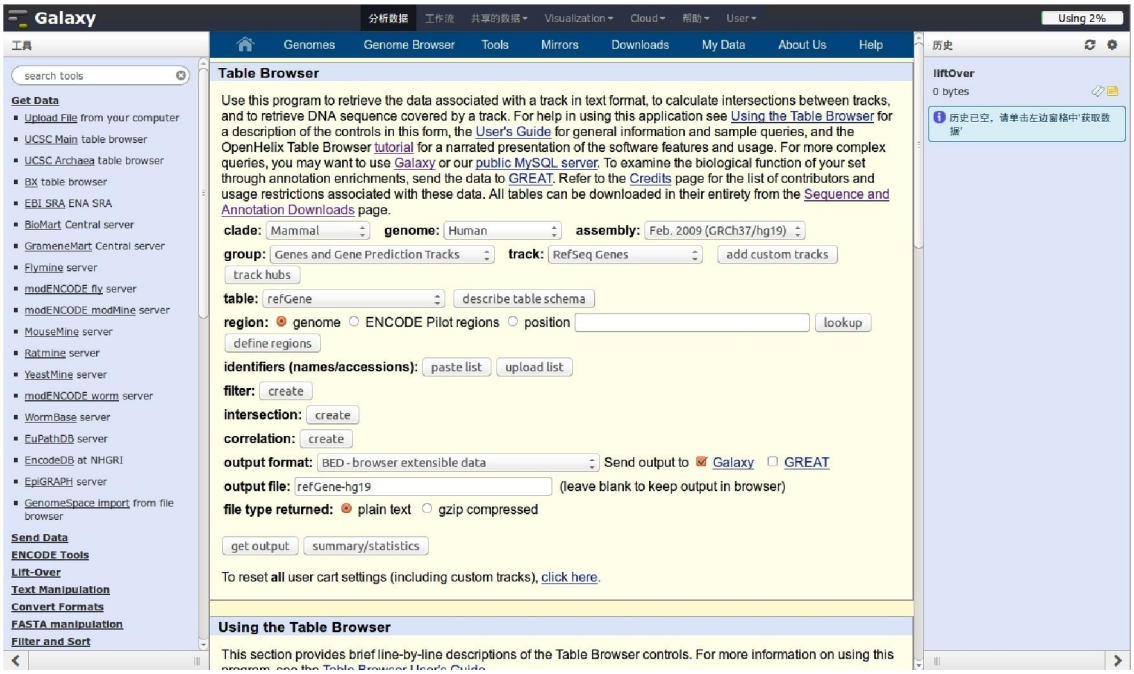


图 5.8 UCSC Table Browser 中获取人类基因信息的参数设置

a）获取输入。使用 Galaxy 中的“UCSC Main”，直接从 UCSC 数据库中提取人类 hg19 基因组的全部基因。打开 Galaxy(https://main.g2.bx.psu.edu/)，在左侧栏的“Get Data”工具集中找到“UCSC Main”，点击可打开 Table Browser 界面，如图 5.8 所示设置参数：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”。点击“get output”按钮后在新的界面中点选“Whole Gene”，最后点击“Send query to Galaxy”即可提取出人类全基因组的基因信息，把数据导入到 Galaxy 中。

b）坐标转换。使用 Galaxy 中的 liftOver 把 hg19 坐标转换成 hg18 坐标。在“Lift-Over”工具中点击“Convert genome coordinates”，在 liftOver 的界面设置参数：Convert coordinates of 选择上一步导入的数据，To 选择需要转换至的目标基因组“hg18”，其他参数默认即可，见图 5.9。点击“Execute”即可完成转换，转换结果同样保存在 Galaxy 中。

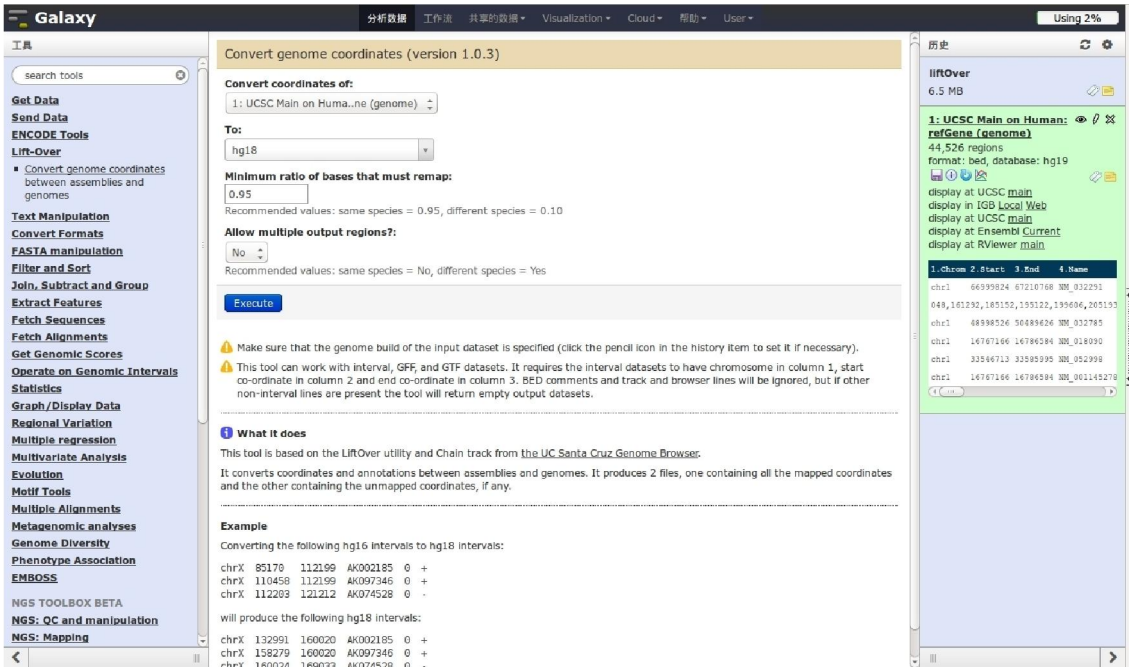


图 5.9 Galaxy 中 liftOver 的参数设置

c）保存结果。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。

Galaxy 和基因组组装版本间坐标转换工具 liftOver 的网址：

Galaxy： https://main.g2.bx.psu.edu

网页版 liftOver：http://genome.ucsc.edu/cgi-bin/hgLiftOver

单机版 liftOver：http://hgdownload.cse.ucsc.edu/downloads.html#source_downloads

5.2.2 常用格式间的转换

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。

例 5.2 BED 格式和 GFF 格式之间的互转。

使用 Galaxy 中的“BED-to-GFF converter”和“GFF-to-BED converter”，实现存储人类 hg19 基因组 Y 染色体上基因信息的 BED 格式和 GFF 格式之间的相互转换。

a）获取输入。使用 Galaxy 中的“UCSC Main”，直接从 UCSC 数据库中提取人类 hg19 基因组 Y 染色体上基因的信息。在 Galaxy 左侧栏的“Get Data”工具集中找到“UCSC Main”，打开 Table Browser 界面，设置参数：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“position”并在其后的输入框内输入“chrY”，output format 选择“BED-browser extensible data”，Send output to 勾选“Galaxy”，file type returned 点选“plain text”。点击“get output”按钮后在新的界面中点选“Whole Gene”，最后点击“Send query to

Galaxy”即可提取人类 Y 染色体的基因信息，把 BED 格式的数据导入到 Galaxy 中。

b）格式转换。先把 BED 格式转换成 GFF 格式，再把转换后的 GFF 格式转换回 BED 格式。

i）把 BED 格式转换成 GFF 格式。在“Convert Formats”工具集中找到“BED-to-GFF”，选择上一步导入的 BED 格式文件，点击“Execute”即可将其转换为 GFF 格式，转换结果直接保存在 Galaxy 中。

ii) 把 GFF 格式转换成 BED 格式。在“Convert Formats”工具集中找到“GFF-to-BED”，选择刚刚转换生成的 GFF 格式的文件，点击“Execute”即可将其转换为 BED 格式，转换结果同样保存在 Galaxy 中。

c）保存结果。对于转换成功后保存在 Galaxy 中的数据，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。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。

Galaxy 的网址：

Galaxy：https://main.g2.bx.psu.edu

5.2.3 基因组坐标的逻辑运算

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。

例 5.3 人类 Y 染色体上外显子和 SNP 的比较。

使用 Galaxy 中的“Operate on Genomic Intervals”工具集，比较人类 Y 染色体上的外显子和 SNP，首先寻找不含 SNP 的外显子，最后把含有 SNP 的外显子和对应的 SNP 关联起来。

a）获取输入。因为要比较人类 Y 染色体上的外显子和 SNP，所以需要两套数据：人类 Y 染色体上的外显子信息和人类 Y 染色体上的 SNP 信息。

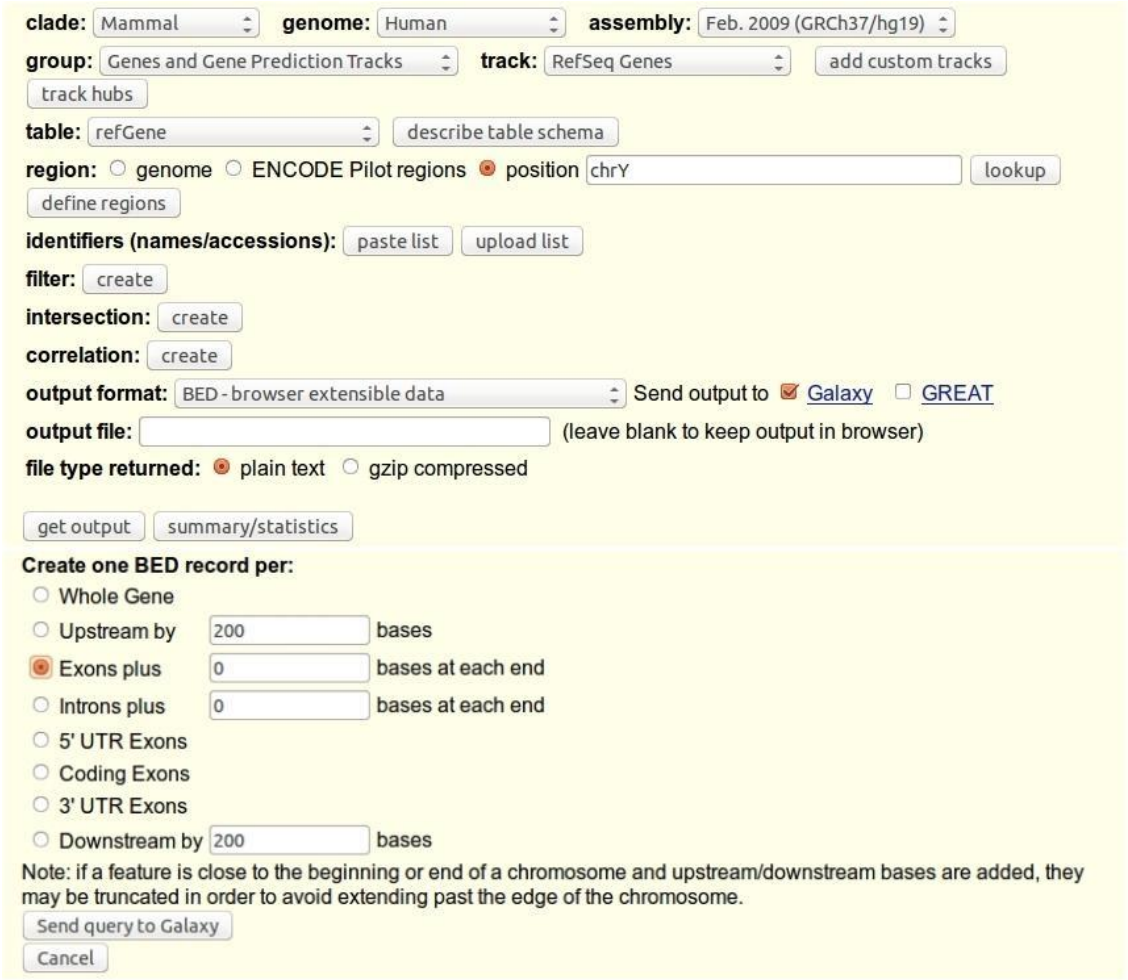


图 5.10 UCSC Table Browser 中获取人类 Y 染色体外显子信息的参数设置

i) 获取人类 Y 染色体上外显子的信息。使用 Galaxy 中的“UCSC Main”，直接从 UCSC 数据库中提取人类 hg19 基因组 Y 染色体上外显子的信息。在 Galaxy 左侧栏的“Get Data”工具集中找到“UCSC Main”，打开 Table Browser 界面，如图 5.10 所示设置参数：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“position”并在其后的输入框内输入“chrY”，output format 选择“BED-browser extensible data”，Send output to 勾选

“Galaxy”，file type returned 点选“plain text”。点击“get output”按钮后在新的界面中点选“Exons”，最后点击“Send query to Galaxy”即可提取人类 Y 染色体的外显子信息，把 BED 格式的数据导入到 Galaxy 中。

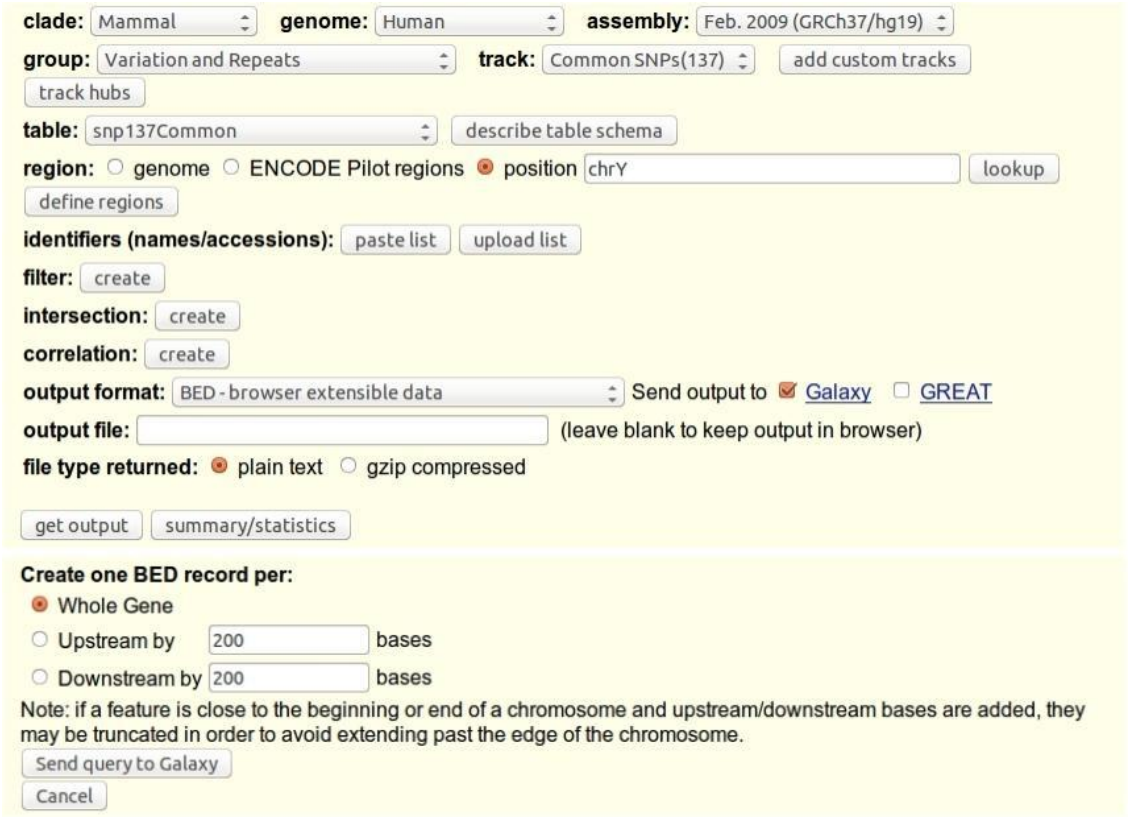


图 5.11 UCSC Table Browser 中获取人类 Y 染色体 SNP 信息的参数设置

ii) 获取人类 Y 染色体上的 SNP 信息。使用 Galaxy 中的“UCSC Main”，直接从 UCSC 数据库中提取人类 hg19 基因组 Y 染色体上 SNP 的信息。在 Galaxy 左侧栏的“Get Data”工具集中找到“UCSC Main”，打开 Table Browser 界面，如图 5.11 所示设置参数：clade 选择“Mammal”，genome 选择“Human”，assembly 选择“Feb. 2009(GRCh37/HG19)”，group 选择“Variation and Repeats”，track 选择“Common SNPs(137)”，table 选择“snp137Common”，region 点选“position”并在其后的输入框内输入“chrY”，output format 选择“BED-browser extensible data”，Send output to 勾选“Galaxy”，file type returned 点选“plain text”。点击“get output”按钮后在新的界面

中点选“Whole Gene”，最后点击“Send query to Galaxy”即可提取人类 Y 染色体的 SNP 信息，把 BED 格式的数据导入到 Galaxy 中。

b）逻辑运算。首先使用基因组坐标的减法提取不含 SNP 的外显子，最后使用基因组坐标的联合操作把外显子和 SNP 的信息关联起来。

图 5.12 Galaxy 中 Subtract 的参数设置

图 5.13 Galaxy 中 Join 的参数设置

i) 提取不含 SNP 的外显子。在“Operate on Genomic Intervals”工具集中找到“Subtract”，参数设置如下：Subtract 即 Second dataset 选择上一步导入的 SNP 数据，from 即 First dataset 选择上一步导入的外显子数据，Return 选择“Intervals with no

overlap”，其他参数默认即可，见图 5.12）。点击“Execute”即可提取出人类 Y 染色体上不含 SNP 的所有外显子。

ii) 将含有 SNP 的外显子和对应的 SNP 关联起来。在“Operate on Genomic Intervals”工具集中找到“Join”，参数设置如下：Join 即 First dataset 选择导入的外显子数据，with 即 Second dataset 选择导入的 SNP 数据，Return 选择“Only records that are joined (INNER JOIN)”，其他参数默认即可，见图 5.13。点击“Execute”即可提取出人类 Y 染色体上含有 SNP 的所有外显子，并将其与对应的 SNP 信息关联起来。

c）保存结果。从最终的统计结果来看，人类 hg19 基因组 Y 染色体上一共有 4096 个外显子和 27265 个 SNP，其中不含 SNP 的外显子有 3660 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。

Galaxy 和 BEDTools 的网址：

Galaxy： https://main.g2.bx.psu.edu

BEDTools：http://bedtools.readthedocs.org/en/latest/

5.3 基因组功能的高级注释

5.3.1 基因组变异位点的注释

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（Single Nucleotide Variations，SNVs）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（National Heart, Lung and Blood Institute，NHLBI）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。

例 5.4 使用 SeattleSeq Annotation 对 SNVs 进行注释。

a）获取输入。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。演示起见，从 SeattleSeq Annotation 网站上下载单个体 Maq 格式的示例文件 example.1Individual.hg19.txt (http://snp.gs.washington.edu/SeattleSeqAnnotation137/HelpDownloadExampleFiles.jsp)。

b）注释 SNVs。在 SeattleSeq Annotation 主页上设置参数，见图 5.14：enter e-mail address 后输入自己的电子邮箱，在文件选择处点击“浏览”上传需要注释的 SNVs 文件（此处即 example.1Individual.hg19.txt），input file format 下点选“Maq”，add more annotation 和 gene locations to use 一般默认即可，也可以根据自己的需要进行选择。选择好输出项目后，最后点击“submit”即可将需要注释的 SNVs 提交到注释服务器，在刷新后的界面中点击“monitor job progress”可以实时监视注释的进度，注释

完成后点击“show table”即可在线查看 SNVs 注释的统计信息与最终的注释表格。同时，在邮箱中将收到一份注释结果的拷贝，可以在本地电脑上进行后续的处理，或用于日后上传以便在线展示注释结果。

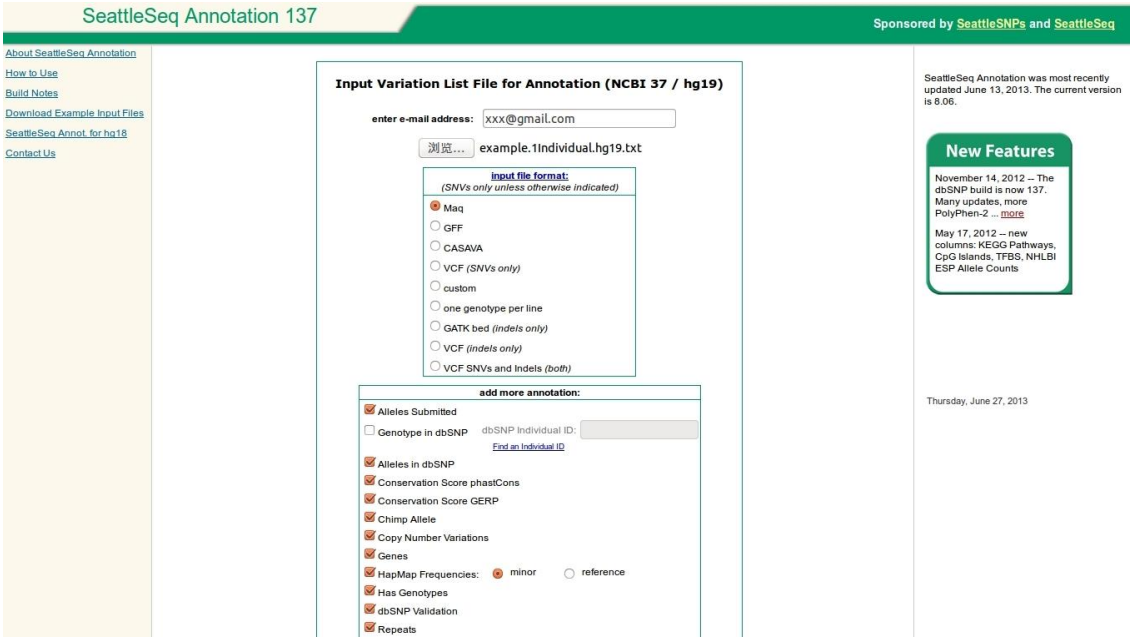


图 5.14 SeattleSeq Annotation 参数设置

c）保存结果。如图 5.15 所示，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 SeattleSeq Annotation 主页上找到 Input Annotation File for Table Display，点击输入框后面的“浏览”上传本地拷贝，点击“submit”后即可在线查看注释结果。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。

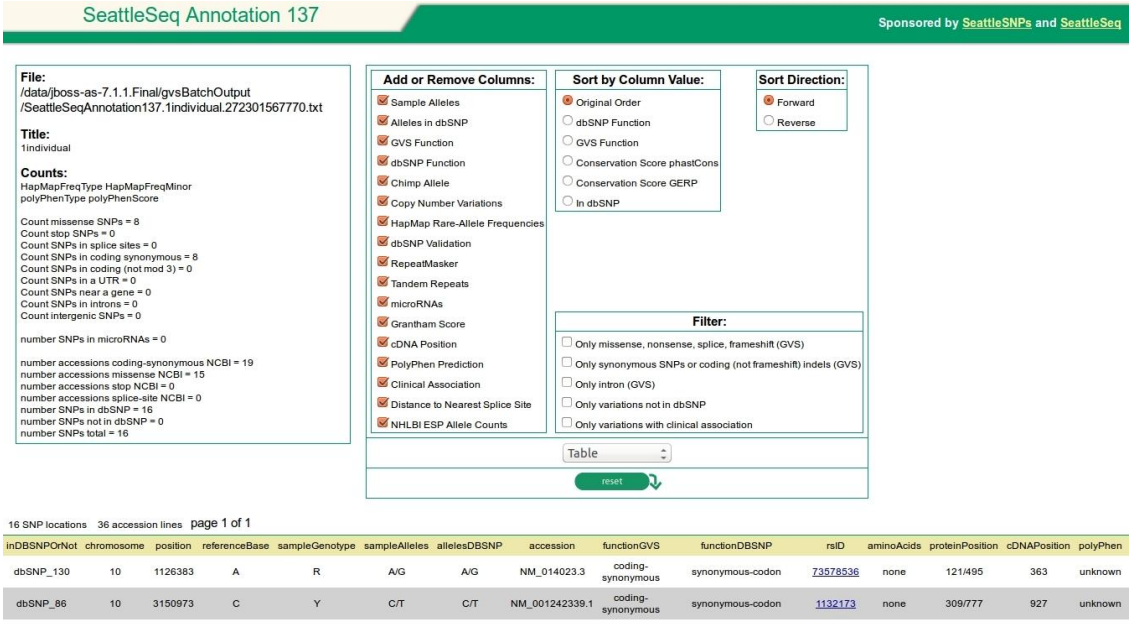


图 5.15 SeattleSeq Annotation 的注释结果

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)。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。

例 5.5 使用 SIFT 对编码区的非同义突变进行注释。

a）获取输入。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。演示起见，使用 SIFT 网站上提供的示例文件（http://sift.jcvi.org/www/chr_coords_example.html）。

b）进行注释。在 SIFT 主页上设置参数，见图 5.16：User Input 区域，Select assembly/annotation version 下选择对应的基因组组装版本，Chromosome Coordinates 输入框中粘贴示例文件，Enter your email address if you want the results through email 下填写电子邮箱以便接收注释结果；Output Options 区域，根据需要选择附加输出项目。最后点击“提交查询”就可将注释工作提交到服务器。

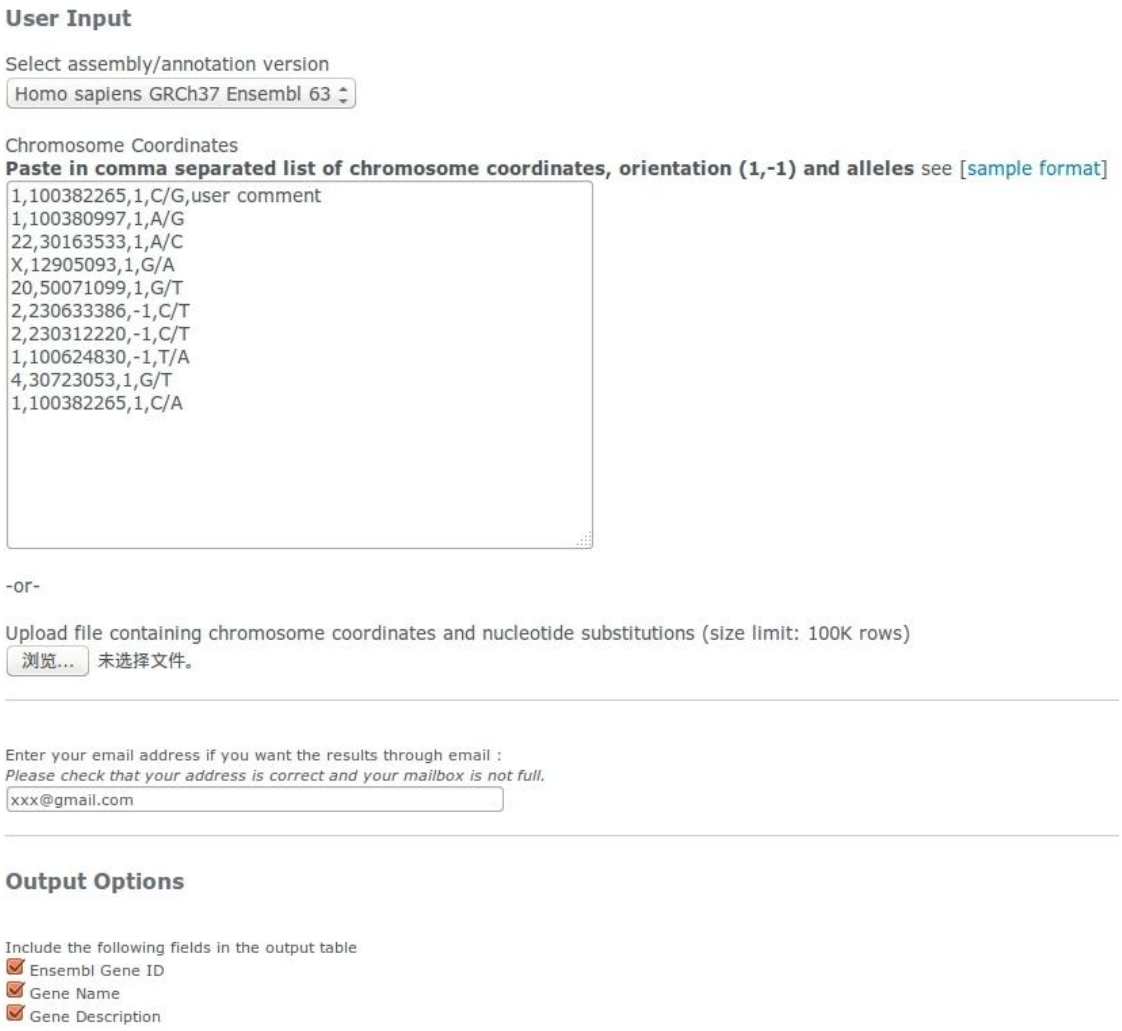


图 5.16 SIFT 参数设置

c）保存结果。如图 5.17 所示，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。

SIFT: PREDICTIONS

Homo sapiens GRCh37 Ensembl 63

User input	Coordinates	Codons	Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction	SIFT Score	Median Information Content	# Seqs at position
1,100382265,1,C/G	1,100382265,1,C/G	CGA-gGA	ENST00000294724	ENSP00000294724	R1487G	EXON CDS	rs12118058:G	Nonsynonymous	TOLERATED	0.46	2.45	74
1,100380997,1,A/G	1,100380997,1,A/G	GAA-GgA	ENST00000294724	ENSP00000294724	E1405G	EXON CDS	rs28730708:G	Nonsynonymous	DAMAGING	0.01	2.45	74
1,100382265,1,C/A	1,100382265,1,C/A	CGA-aGA	ENST00000294724	ENSP00000294724	R1487R	EXON CDS	rs12118058:G	Synonymous	TOLERATED	0.64	2.45	74
22,30163533,1,A/C	22,30163533,1,A/C	GAG-GcG	ENST00000330029	ENSP00000332887	E49A	EXON CDS	novel	Nonsynonymous	DAMAGING	0.02	2.57	97
20,50071099,1,G/T	20,50071099,1,G/T	ACT-AaT	ENST00000371564	ENSP00000360619	T612N	EXON CDS	rs6067785:T	Nonsynonymous	DAMAGING	0	2.81	122
2,230633386,-1,C/T	2,230633386,1,G/A	CAG-tAG	ENST00000283943	ENSP00000283943	Q1910*	EXON CDS	rs1803846:A	Nonsynonymous	N/A	N/A	N/A	N/A
2,230312220,-1,C/T	2,230312220,1,G/A	CCC-CtC	ENST00000341772	ENSP00000345229	P433L	EXON CDS	rs17853365:A	Nonsynonymous	DAMAGING	0.02	2.38	53
4,30723053,1,G/T	4,30723053,1,G/T	AGG-AGt	ENST00000333135	ENSP00000330302	R3S	EXON CDS	rs2631567:T	Nonsynonymous	TOLERATED	0.16	3.12	68
1,100624830,-1,T/A	Reference nucleotide not matched	-					N/A	N/A	Not scored	N/A	N/A	N/A
X,12905093,1,G/A	Reference nucleotide not matched	-					N/A	N/A	Not scored	N/A	N/A	N/A

图 5.17 SIFT 注释结果

基因组变异位点注释工具的网址：

SeattleSeq Annotation： http://snp.gs.washington.edu/SeattleSeqAnnotation137/index.jsp

variant tools： http://varianttools.sourceforge.net

SnpEff： http://snpeff.sourceforge.net

SIFT： http://sift.jcvi.org

PolyPhen-2： http://genetics.bwh.harvard.edu/pph2/index.shtml

SNPs3D： http://www.snps3d.org

PROVEAN：http://provean.jcvi.org/index.php

5.3.2 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）。

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。

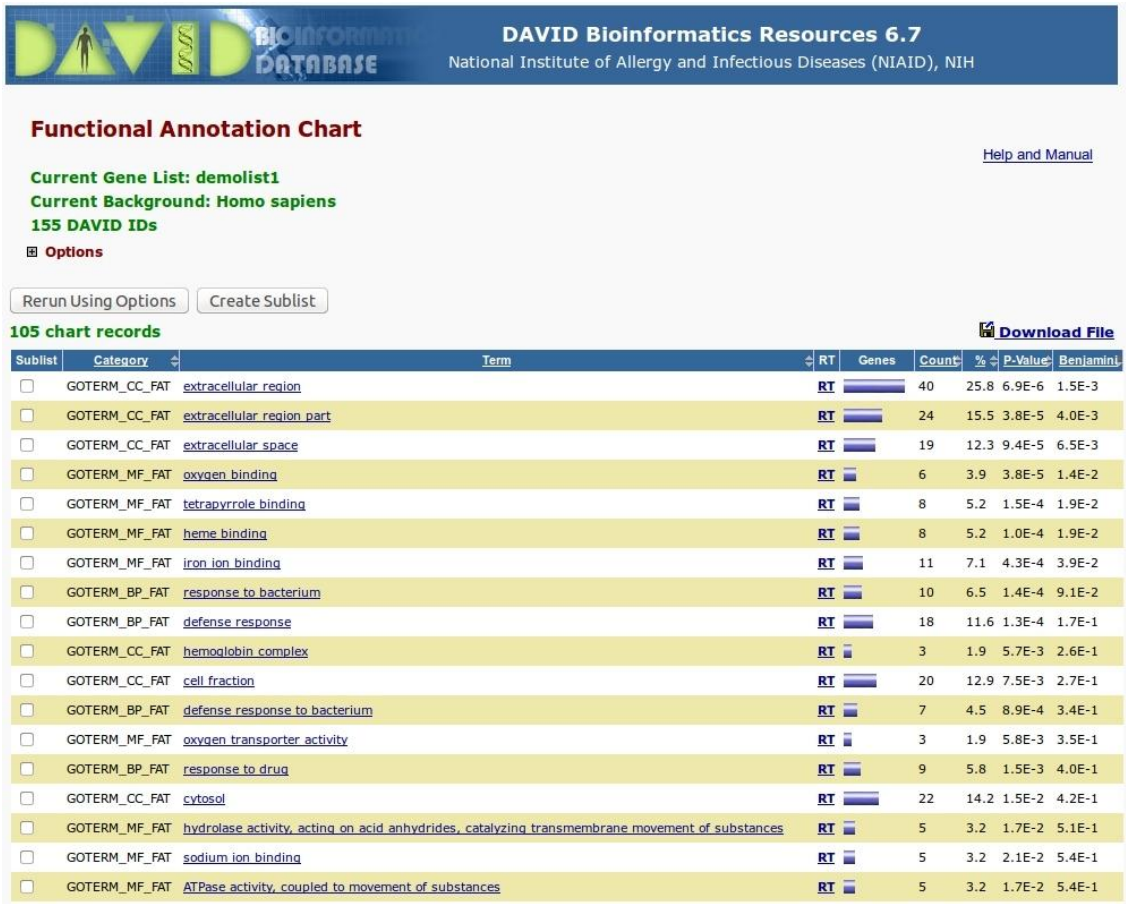
例 5.6 使用 DAVID 对基因集进行 GO 与 KEGG 的富集分析。

a）获取输入。在 DAVID 主页上，点击“Start Analysis”，在“Upload”中上传基因列表，演示起见，点击“Demolist 1”使用 DAVID 自带的数据。之后，在“List”和“Background”中进行相应的设置，此处默认即可。

b）富集分析。点击“Functional Annotation Tool”，进入注释项目选择页面，根据需要添加或重新点选相应的项目即可。此处，仅选择常用的 GO 和 KEGG 项目，其中 GO 包括生物过程（biological process）、细胞组件（cellular component）和分子功能（molecular function）三个子项目。所以最终选择“Gene_Ontology”中的“GOTERM_BP_FAT”、“GOTERM_CC_FAT”、“GOTERM_MF_FAT”和“Pathways”

中的“KEGG_PATHWAY”总共四个注释项目。

c）保存结果。在 Combined View for Selected Annotation 下有“Functional Annotation Clustering”、“Functional Annotation Chart”和“Functional Annotation Table”三种分析工具，此处仅进行富集分析，所以直接点击“Functional Annotation Chart”，在新的页面中可以看到最终的注释结果，见图 5.18。其中，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。



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_FAT	extracellular region	RT		40	25.8	6.9E-6	1.5E-3
	GOTERM_CC_FAT	extracellular region part	RT		24	15.5	3.8E-5	4.0E-3
	GOTERM_CC_FAT	extracellular space	RT		19	12.3	9.4E-5	6.5E-3
	GOTERM_MF_FAT	oxygen binding	RT		6	3.9	3.8E-5	1.4E-2
	GOTERM_MF_FAT	tetrapyrrole binding	RT		8	5.2	1.5E-4	1.9E-2
	GOTERM_MF_FAT	heme binding	RT		8	5.2	1.0E-4	1.9E-2
	GOTERM_MF_FAT	iron ion binding	RT		11	7.1	4.3E-4	3.9E-2
	GOTERM_BP_FAT	response to bacterium	RT		10	6.5	1.4E-4	9.1E-2
	GOTERM_BP_FAT	defense response	RT		18	11.6	1.3E-4	1.7E-1
	GOTERM_CC_FAT	hemoglobin complex	RT		3	1.9	5.7E-3	2.6E-1
	GOTERM_CC_FAT	cell fraction	RT		20	12.9	7.5E-3	2.7E-1
	GOTERM_BP_FAT	defense response to bacterium	RT		7	4.5	8.9E-4	3.4E-1
	GOTERM_MF_FAT	oxygen transporter activity	RT		3	1.9	5.8E-3	3.5E-1
	GOTERM_BP_FAT	response to drug	RT		9	5.8	1.5E-3	4.0E-1
	GOTERM_CC_FAT	cytosol	RT		22	14.2	1.5E-2	4.2E-1
	GOTERM_MF_FAT	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	RT		5	3.2	1.7E-2	5.1E-1
	GOTERM_MF_FAT	sodium ion binding	RT		5	3.2	2.1E-2	5.4E-1
	GOTERM_MF_FAT	ATPase activity, coupled to movement of substances	RT		5	3.2	1.7E-2	5.4E-1

图 5.18 DAVID 的 Functional Annotation Chart 结果

此处仅对 DAVID 中进行富集分析的 Functional Annotation Chart 进行了介绍，但实际上 DAVID 提供了针对四项分析内容的六个分析工具：

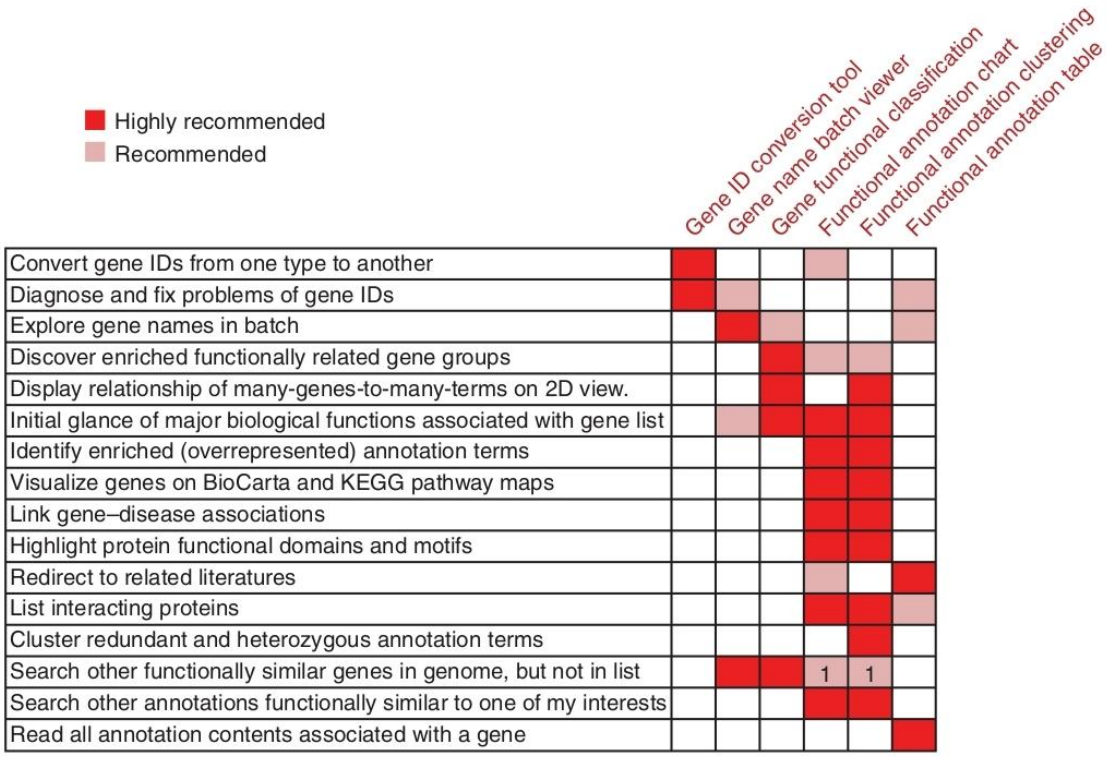


图 5.19 DAVID 分析工具选择示意图（引自参考文献“Huang & Sherman”，2009）

（1）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。

(2) 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。

(3) 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。

(4) Functional Annotation Tool：DAVID 最核心的分析内容，包括三个分析工具：

① 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。

② 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。

③ 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，在实际使用时可以根据图 5.19 选择最合适的一个或多个工具。

基因集富集分析相关数据库与工具的网址：

GO： http://www.geneontology.org

KEGG： http://www.genome.jp/kegg

DAVID：http://david.abcc.ncifcrf.gov

5.3.3 制作序列标识

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（Transcription Factor Binding Site，TFBS）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。

例 5.7 使用 WebLogo 制作人类基因剪接位点的 GT-AG 序列标识。

a）获取输入。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。此处直接使用 WebLogo 官网上 examples(http://weblogo.threeplusone.com/examples.html) 中给出的人类剪

接位点(http://weblogo.threeplusone.com/examples.html#splice) 的数据，包括供体位点（donor sites）和受体位点（acceptor sites）的多序列比对结果。

b）制作序列标识。

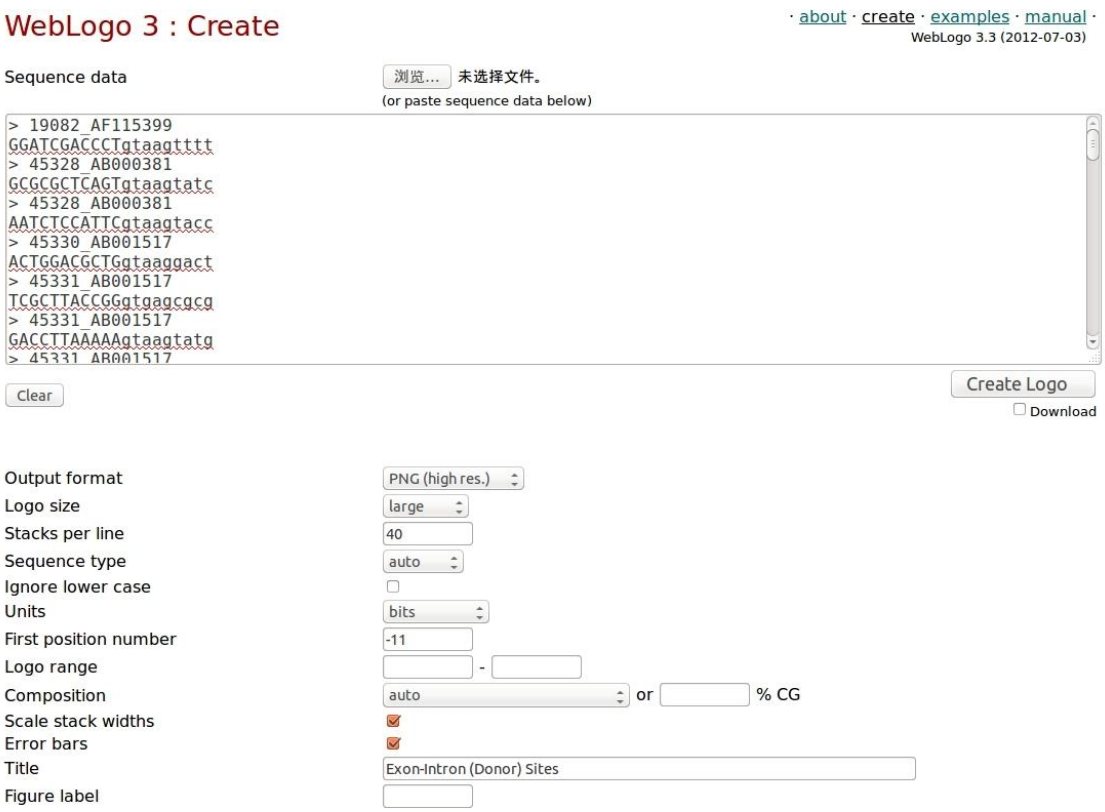


图 5.20 供体位点的 WebLogo 参数设置

i) 制作供体位点的序列标识。点击“Exon-Intron (Donor) Sites”前的“Edit Logo”，打开序列标识制作界面，见图 5.20。供体位点的多序列比对结果自动填充在了“Sequence data”区域。根据需要调整参数，如：Output format 选择“PNG (high res.)”，Logo size 选择“large”，First position number 填写“-11”，清除 Logo range 中的数字，在 Title 后面的输入框中填写“Exon-Intron (Donor) Sites”。最后点击“Create Logo”即可得到供体位点的序列标识。

ii) 制作受体位点的序列标识。点击“Intron-Exon (Acceptor) Sites”前的“Edit Logo”，打开序列标识制作界面，见图 5.21。受体位点的多序列比对结果自动填充

在了“Sequence data”区域。根据需要调整参数，如：Output format 选择“PNG (high res.)”，Logo size 选择“large”，First position number 填写“-21”，在 Title 后面的输入框中填写“Intron-Exon (Acceptor) Sites”。最后点击“Create Logo”即可得到受体位点的序列标识。

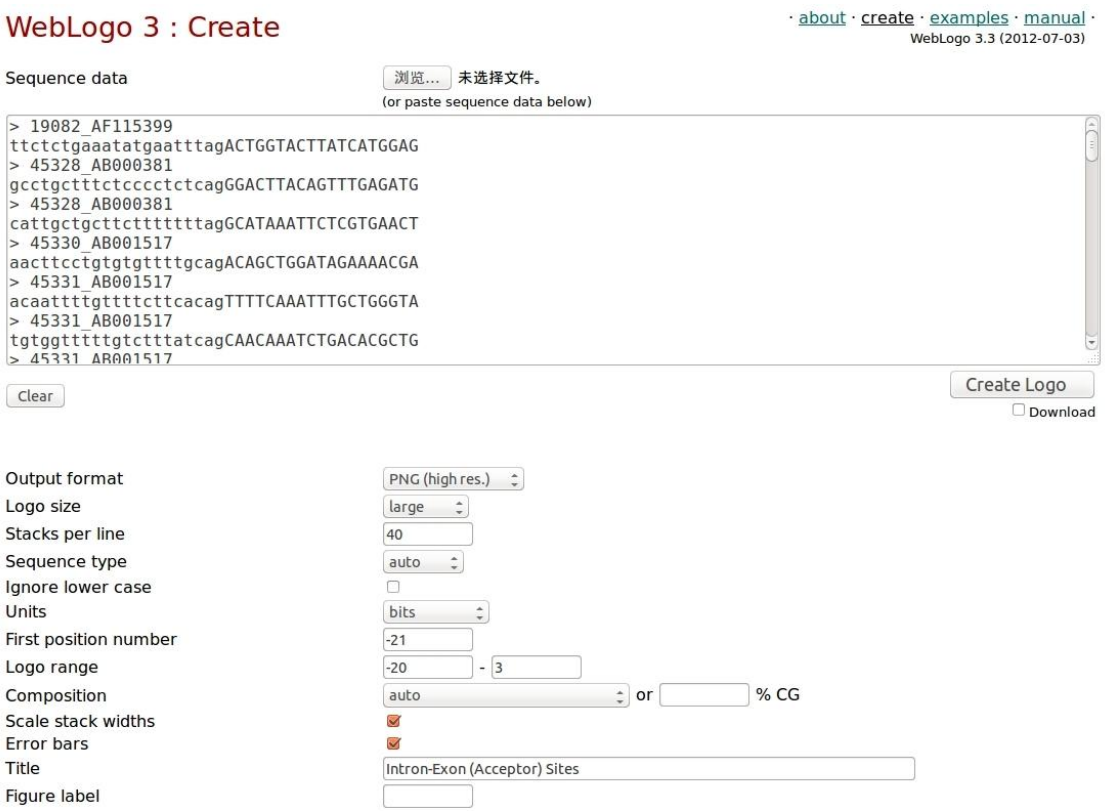


图 5.21 受体位点的 WebLogo 参数设置

c）保存结果。最终制作出如图 5.22 所示的供体位点序列标识和如图 5.23 所示的受体位点序列标识。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。如图 5.23 所示，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。

Galaxy 与 WebLogo 的网址：

Galaxy： https://main.g2.bx.psu.edu

网页版 WebLogo：http://weblogo.threeplusone.com

单机版 WebLogo：http://code.google.com/p/weblogo

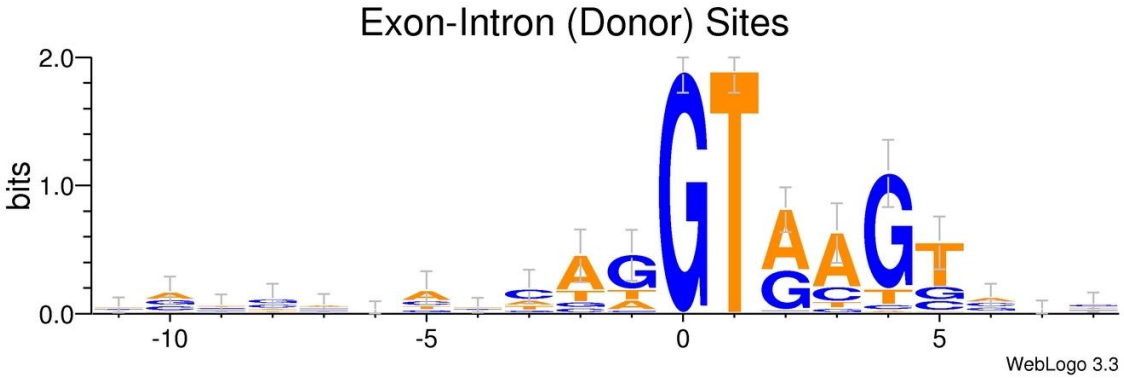


图 5.22 供体位点的序列标识

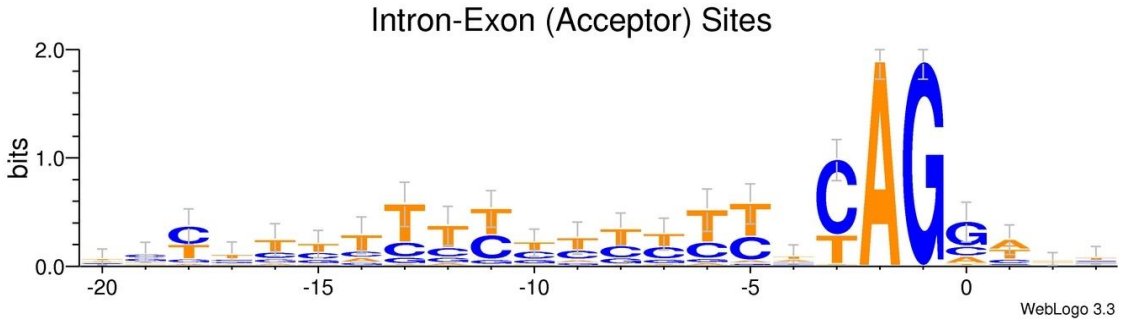


图 5.23 受体位点的序列标识

5.3.4 基因组功能注释分析平台

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。

如图 5.24 所示，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；左侧栏是历史面板（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。

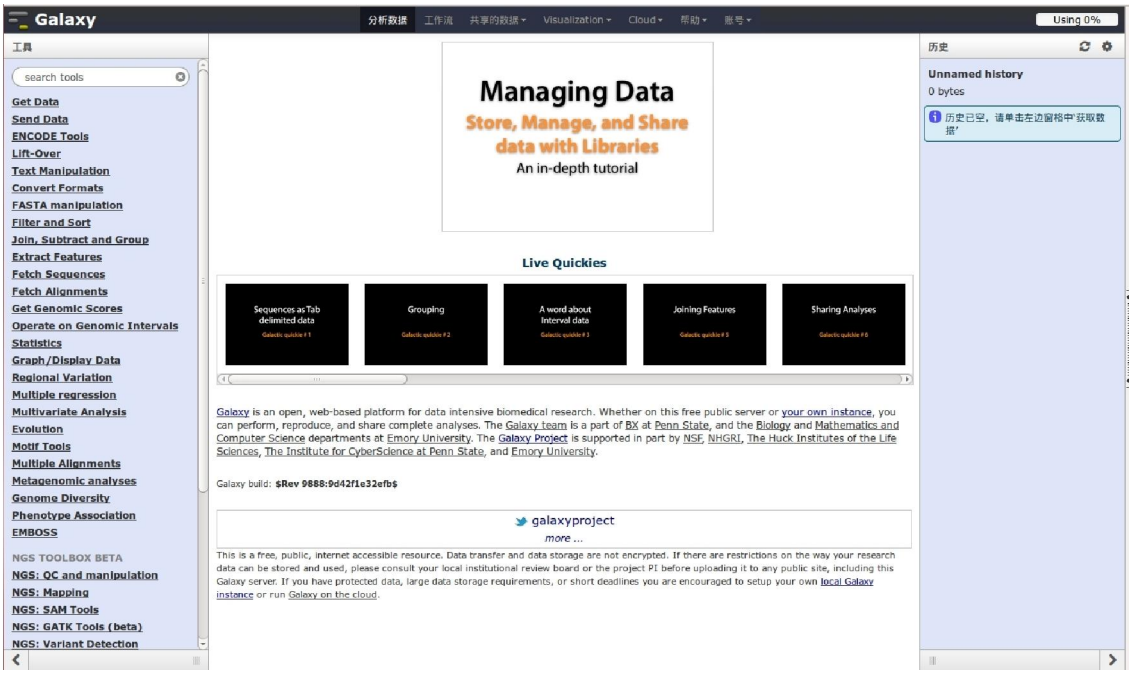


图 5.24 Galaxy 主界面

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。

基因组功能注释分析平台及其学习资料的网址：

Galaxy： https://main.g2.bx.psu.edu

GenePattern：http://www.broadinstitute.org/cancer/software/genepattern

Galaxy 101：http://main.g2.bx.psu.edu/galaxy101

Galaxy Screencasts and Demos：http://wiki.galaxyproject.org/Learn/Screencasts

Learn Galaxy：http://wiki.galaxyproject.org/Learn

Galaxy Wiki： http://wiki.galaxyproject.org/FrontPage

参考文献

[1] Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 2009, 4, 1073–81.

[2] Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nature methods 2010, 7, 248–9.

[3] Yue, P. & Moult, J. Identification and analysis of deleterious human SNPs. Journal of molecular biology 2006, 356, 1263–74.

[4] 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] Huang, D. W., Sherman, B. T. & Lempicki, R. A Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 2009, 37, 1–13.

[6] Schneider T. D., Stephens R. M. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Research 1990, 18, 6097-6100.

[7] Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. WebLogo: a sequence logo generator. Genome research 2004, 14, 1188–90.

[8] Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. Genome research 2005, 15, 1451–5.

[9] Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 2010, 11, R86.

[10] Blankenberg, D. et al. Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology 2010, Chapter 19, Unit 19.10.1–21.

[11] Hillman-Jackson, J., Clements, D., Blankenberg, D., Taylor, J. & Nekrutenko, A. Using Galaxy to perform large-scale interactive data analyses. Current protocols in bioinformatics 2012, Chapter 10, Unit10.5.

[12] Goecks, J., Eberhard, C., Too, T., Nekrutenko, A. & Taylor, J. Web-based visual analysis for high-throughput genomics. BMC Genomics 2013, 14, 397.