遗传 HEREDITAS (Beijing) 2013 年 4 月, 35(4): 545－546
ISSN 0253-9772 www.chinagene.cn

DOI: 10.3724/SP.J.1005.2013.00545

斑马鱼核心数据库简介

肖安, 张博

北京大学生命科学学院, 细胞增殖与分化教育部重点实验室, 北京 100871

本文主要介绍斑马鱼(*Danio rerio*)研究工作中经常用到的生物信息学数据库, 主要包括物种专用综合数据库 ZFIN, 以及通用基因组注释数据库 Ensembl、Vega 和 UCSC Genome Browser。

1 斑马鱼模式物种综合数据库 ZFIN

斑马鱼模式物种数据库 ZFIN(Zebrafish Model Organism Database, 原名 Zebrafish Information Networks)是斑马鱼的核心数据库。ZFIN 由美国 University of Oregon 维护, 收集的信息主要可分为以下几个方面：

(1) 斑马鱼基因信息。以斑马鱼基因为核心, 包括各种分子标记等。每个基因的页面中, 包含了跟这个基因相关的表达信息(包含大量图片, 特别是原位杂交图片和相应的描述)、已知突变或基因敲低(Knock down)的研究报道、表型, 以及到其他核酸或蛋白质相关数据库、Gene Ontology 数据库、蛋白结构域数据库、基因组注释数据库、文献数据库等的链接等。

(2) 斑马鱼相关资源。包括各种质粒克隆、已报道的 morpholino 和抗体、转基因载体等; 这些均和相应的基因存在相互链接。同时也收集各类斑马鱼野生型品系、突变品系、转基因品系资源, 分别链接到相应的资源中心数据库。

(3) 斑马鱼基因组图谱和作图(Mapping)相关的信息。

(4) 斑马鱼研究工作信息。ZFIN 收集了大量斑马鱼相关的工作、会议、在网站注册的斑马鱼研究机构、公司和实验室/个人的信息。同时也提供斑马鱼研究相关的新闻。其他交流资源还包括各种实验操作技术、*The Zebrafish Book* 的在线版本等。

(5) ZFIN 同时也具体负责斑马鱼组织结构的解剖学定义、斑马鱼品系和基因命名规范等工作。

2 基因组注释数据库 Ensembl

Ensembl 是由 EBI(欧洲生物信息研究所)和英国 Sanger 研究所共同维护的一个脊椎动物基因组自动注释平台。斑马鱼是其 3 大核心物种之一(另外两个是人和小鼠)。在每个新的斑马鱼基因组拼装版本公布之后, 或每隔 2~3 个月的更新中, Ensembl 会利用一套自动注释程序, 将其他各数据库的最新信息在基因组序列上进行注释整合。在 Ensembl 中, 处于核心地位的两组页面是基因组页面和基因页面。

2.1 基因组页面

基因组页面通过图形化显示染色体的一个区段, 可以移动或缩放所显示的范围。通过缩略图和细节信息图标注该区段所涵盖的基因的转录本、外显子和内含子、编码区域等信息, 同时, 通过使用页面左下方的“Configure this page”可设置显示或隐藏类别非常多的其他条带(track), 通过“Manage your data”来添加其他来源的数据, 包括用户自己上传的基因组注释信息。视图中的每一个条带基本上都可以点击并链接到相关的细节页面。

需要注意的是, 基因组页面主视图中间的条带代表染色体本身以及其基因组拼装结构, 位于此染色体条带上方的信息表明其对应转录本转录方向和染色体人为规定的方向一致, 位于下方则表明该转录方向跟染色体方向相反。而在下面将要介绍的基因页面中, 大部分信息则是按照基因的转录方向展

网络出版时间: 2013-1-30 9:54:58
URL: http://www.cnki.net/kcms/detail/11.1913.R.20130130.0954.002.html

示。因此，两者条带的方向有可能正好相反。

2.2 基因页面

基因页面中包含了跟基因相关的所有信息，主要包含以下 3 个部分：

(1) 基本信息：用文字或图形表示各种基因的位置、剪接变体、外显子和内含子、编码区和 UTR、具体序列、来源于外部数据库的信息整合及相应链接等(例如：在“Splice variants”中可以看到各蛋白结构域和外显子的对应关系)。

(2) 比较基因组学信息：包括物种内或物种间的同源基因、系统发生树、蛋白家族等。

(3) 遗传变异信息：包括 SNP 等。

此外，基因页面下还分出各转录本和蛋白产物的子页面，其中包括外部数据库的支持证据、具体的外显子、cDNA 和蛋白质的序列、探针和 Gene Ontology 信息、结构域等各种蛋白注释信息等。

Ensembl 数据库给各个基因、转录本、蛋白质、外显子、蛋白家族等均建立了一套 ID 系统，在注释版本升级过程中这些 ID 是稳定不变的，可用于外部引用和追踪变化。对于斑马鱼基因，ID 的格式为“ENSDARG”后面加 11 位数字，其中“ENS”代表由 Ensembl 注释，“DAR”为斑马鱼拉丁名缩写、“G”代表基因(相应的，“T”、“P”、“E”则分别代表转录本、蛋白质、外显子)；而“ENSFM”后面加 14 位数字则表示某个基因家族的 ID(并不针对某个特定的物种)。Ensembl 显示的斑马鱼基因符号，优先选用 ZFIN 确定的斑马鱼基因命名，若无则采用人类直系同源基因的命名，若均未知则直接使用 Ensembl ID。区分前两者的方法是 ZFIN 确定的斑马鱼基因符号一定全部由小写字母组成，而来源于人类的基因符号则全部为大写字母，有时还会在后面有“(*1 OF 2*)”等补充，表明这一人类基因符号在斑马鱼中被用于多个基因的命名。

2.3 Ensembl 提供的搜索和比对工具

Ensembl 除了直接显示相关注释信息之外，还提供了两大类信息搜索和导出的工具：BLAST/BLAT 和 BioMart。BLAST/BLAT 是两个常用的序列比对工具，其中 BLAST 适用于核酸或蛋白质的比对，其参数设置较为丰富；BLAT 则在比对速度上远高于 BLAST，但是仅适用于核酸比对，且对较短的序列(<30 bp)的比对效果可能不如 BLAST。

BioMart 是一套数据库批量检索系统，相对于逐个搜索可以大大地提高数据检索的效率。Ensembl 中的 BioMart 服务操作使用顺序是：

(1) 选择数据库和子库(例如依次选择最新的 Ensembl 注释版本、最新的斑马鱼基因组数据库)。

(2) 确定一些搜索条件，又称为过滤器(Filters)；例如可以指定目标必须属于染色体的某一区段范围、必须是一些指定基因(可通过填写或额外上传文件的方式提供一批 Ensembl 或外部数据库的 ID 列表)、必须具有某些表达或蛋白结构域信息、必须具有某些特征(如转录本数量、注释状态)等等，并可将这些条件进行组合限制。

(3) 选择需要获取的目标或需求(Attributes)，即指定从满足上述条件的搜索结果中(需要)给出哪些内容，包括基因特性、同源基因、基因结构分布、序列和变异等非常大量的选择。在确定了搜索条件和搜索需求后，点击“Count”可以对结果进行计数，点击“Result”可以预览部分结果，并以文本文档、PDF 文档、电子表格等格式下载全部结果或去除完全重复的记录后的结果。

3 其他基因组注释数据库和其他常用数据库

除了 Ensembl 之外，Vega(http://vega.sanger.ac.uk/)和 UCSC Genome Browser(http://genome.ucsc.edu/)也是常用的基因组注释数据库。其中 Vega 采用和 Ensembl 类似的网页界面系统，但其注释数据为人工注释而非自动注释。当前斑马鱼基因组的人工注释覆盖范围还比较有限。UCSC Genome Browser 则是一个老牌的基因组注释系统，界面样式和 Ensembl 有所差别，注释更新频率可能也略低；但不少希望在基因组上呈现(用户提供的)外部数据信息的研究者习惯于使用这一数据库。

其他斑马鱼常用的物种通用数据库包括核酸数据库 GenBank/EMBL/DDBJ、蛋白数据库 UniProt、蛋白结构域和二级结构整合数据库 InterPro 以及 InterPro 所参考的各个原始数据库、基因功能知识库 GO 等等。可以参考 *Nucleic Acids Res* 杂志年度数据库专刊和其在线数据库收藏(http://www.oxfordjournals.org/our_journals/nar/special_collections.html 中的 Molecular Biology Database Collection，亦为年度更新)。