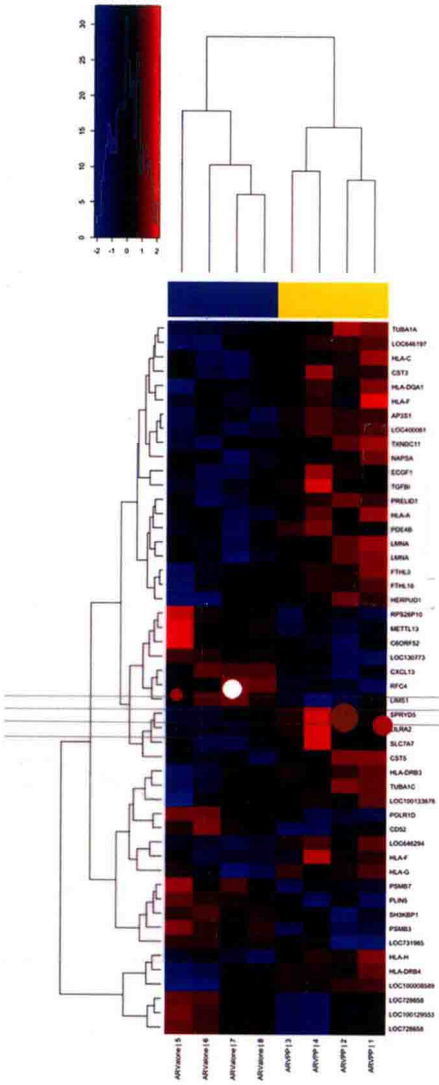
国家自然科学基金项目
“生物胁迫下拟南芥环形RNA的生物信息学分析和功能研究”
（项目编号：31571306）

高通量测序技术在选择性剪接研究中的应用



孙晓勇 著

中国农业出版社

国家自然科学基金项目“生物胁迫下拟南芥环形 RNA 的生物信息学分析和功能研究”(项目编号:31571306)

高通量测序技术在选择性剪接研究中的应用

孙晓勇 著

中国农业出版社

图书在版编目（CIP）数据

高通量测序技术在选择性剪接研究中的应用/孙晓勇著．—北京：中国农业出版社，2018.1
ISBN 978-7-109-23518-2

Ⅰ.①高… Ⅱ.①孙… Ⅲ.①基因组—序列—测试—研究 Ⅳ.①Q343.1

中国版本图书馆 CIP 数据核字（2017）第 278175 号

中国农业出版社出版
（北京市朝阳区麦子店街 18 号楼）
（邮政编码 100125）
策划编辑 黄 宇 李 蕊 杜 然
文字编辑 李兴旺

三河市君旺印务有限公司印刷 新华书店北京发行所发行
2018 年 1 月第 1 版 2018 年 1 月河北第 1 次印刷

开本：700mm×1000mm 1/16 印张：10.75 插页：2
字数：210 千字
定价：45.00 元
（凡本版图书出现印刷、装订错误，请向出版社发行部调换）

内容简介

本书结合生物信息学课题组多年从事选择性剪接的研究成果，系统阐述了高通量技术在选择性剪接领域的应用与最新进展，重点讲述了分析高通量数据的软件，包括通用软件 BioIDMapper、选择性剪接分析识别软件 SplicingTypesAnno 和 nagnag、可视化软件 prettycloud 和 pairheatmap。在此基础上，进一步探讨了通过基因芯片和高通量测序分析挖掘选择性剪接的方法和流程，探索性地研究了与编码区相关的 NAGNAG 选择性剪接、非编码区相关的未知转录本和长链非编码 RNA 及相关的调控网络。最后介绍了反向剪接产生环形 RNA 的最新研究成果。

本书以 R 语言作为主要分析工具，内容新颖，叙述深入浅出，适用于生物学和生物信息学及其相关专业的研究者阅读参考。

前言 FOREWORD

选择性剪接是中心法则中一个关键的环节，不仅产生蛋白质多样性，而且最后产生表现型的多样性。研究者已经发现：92%～94%的人类基因和61%的拟南芥基因有选择性剪接的现象。最复杂的一些人类基因，如MOG等，具有70多个转录本。最近研究显示：95%的人类基因能够产生100 000个完全不同的选择性剪接现象。细胞通过选择性剪接去除内含子和保留外显子，可以产生大量来自同一基因的转录本，从而产生蛋白质复杂性，并影响整个发育过程。选择性剪接是许多人类疾病，如癌症、帕金森病、心脏病、心血管疾病、血液凝固、胆固醇体内平衡等的主要驱动因素。

高通量测序技术是对传统测序一次革命性的改变，提供了前所未有的机遇。此技术可以一次对几十万，甚至几百万条DNA分子进行序列测定，因此在有些文献中称其为下一代测序技术（next generation sequencing）。2005年，454 Life Sciences公司首先推出二代测序仪。2006年，Solexa推出了Genome Analyzer。到今天为止，高通量测序的出现超过了10年，应用到生物、医学、农业等各领域，并且逐渐从实验室进入临床检验。目前针对上百万人类基因组的研究屡见不鲜，展现了蓬勃的生机及巨大的发展空间。

传统的生物学研究集中于单个基因中的单个选择性剪接事件，而且侧重的是单个基因中蛋白的表达和功能。随着高通量技术的不断发展和成熟，研究已经转向系统生物学角度，从全转录组中分析选择性剪接，从而系统地研究整个基因调控网络中其相关的特征和功能。最新研究将不同组学数据集中在一起，分析挖掘选择性剪接，不仅从RNA水平，而且结合DNA水平和蛋白水平，更系统更全面地分析研究这一现象。目前有证据显示，表观基因组学与选择性剪接密切相关。将其与转录组学共同分析，能够揭示DNA甲基化与选

择性剪接的关系与作用机理。

以前人们一直认为RNA是线性的，环形RNA是来自于实验误差或者选择性剪接异常引起的。近几年随着高通量测序技术的发展，科学界才逐渐意识到环形RNA不仅大量存在，且具有特殊功能，并且广泛存在于不同的真核生物中。传统RNA检测和分析方法都要求有自由的5’端和3’端，而这种经典方法完全忽略并低估了环形RNA的表达量和生物重要性。自2011年以来的多篇文章，证明了这种RNA的大量表达，因而近年来成为国际研究的热点。环形RNA的研究具有重要研究价值。首先，环形RNA是由非编码RNA介导表达调控网络的重要组成元件：环形RNA竞争性与microRNA结合，阻断microRNA对其靶标的负抑制，功能特殊，在mRNA表达中起重要的调控作用，是阐明表达调控网络的重要一环。其次，环形RNA具有潜在的巨大应用价值：由于环形RNA组织特异性强并且结构特殊，在细胞质中稳定性强，可以起到长期细胞调节功能，因而在农业和医学上具有广泛的应用前景。在农业上，环形RNA可以通过与microRNA的互作，不仅应用于防治病虫害，而且通过microRNA与产量的关系，影响农艺性状；在医学上，环形RNA可以在医疗诊断中作为潜在的生物标志物，并且对治疗疾病具有巨大的潜在应用价值。

为了分析挖掘高通量测序数据中的选择性剪接事件，笔者以课题组开发的相关高通量数据分析软件为工具，针对特殊的选择性剪接事件，包括NAGNAG选择性剪接、非编码区的剪接、基因调控网络以及环形RNA，开展了具体的剪接识别、数据分析、特征挖掘等研究。

本书涉及的内容主要反映了作者10年来从事高通量测序分析技术和选择性剪接研究的工作。其中，环形RNA的相关工作是承担国家自然科学基金项目（31571306）取得的部分成果。本书在整理过程中得到了笔者学生的大力支持和帮助，包括王凯、宋坤、石传宏、王冲、姜鑫、宋永康、张庆雷、李瑞、叶家震、李树章、张钦然、唐文浩的辛勤工作和成果。其中，王凯完成了本书所有R图的制作，

宋坤、石传宏、王冲、姜鑫、宋永康负责文章统稿，李瑞参与材料整理工作，张庆雷、叶家震、李树章、张钦然、唐文浩负责部分数据分析，在此一并表示最诚挚的谢意。

鉴于多方面的原因，特别是作者水平有限，疏漏在所难免，诚恳广大读者不吝指正。

著　者

2017年9月

CONTENTS 目录

第1章 BioIDMapper：生物大分子编码转换系统

1.1 生物大分子编码转换

随着越来越多的物种测序完成，许多针对基因和蛋白质的新数据库不断开发出来以适应研究的需要（Sun 等，2009）。如何在不同数据库、不同平台进行数据的转换，并且分析生物重要的功能成为一个巨大的挑战。DNA、RNA、蛋白质、小分子不能独立起作用。最近的研究清楚地表明，单一来源的数据远远无法解释基因调控、蛋白质修饰、信号网络等复杂的生物学功能。因此目前有一个新兴的趋势：整合来自于不同来源的数据，包括 miroarray、高通量测序技术、SAGE 技术、GC/LS 等。通过比较积累的数据，从不同的生物学层面揭示复杂且系统的生物体系。然而，各种数据库不同的标识符以及不同的标准造成了合并数据的障碍，因此研究者迫切需要能够综合不同平台和数据库的软件工具。

目前，流行的数据库包括 Entrez Gene、UniProt、Gene Ontology、EMBL、OMIM、PubMed、KEGG 等。BioIDMapper 可以对 NCBI、UniProt、KEGG 等不同数据库之间的 DNA、RNA、蛋白质等大分子进行相互转换，通过集成各种 ID 系统，提供了一个完整的生物学视野（Bussey 等，2003；Alibes 等，2007）。BioIDMapper 软件包是基于 NCBI 和 UniProt（Apweiler 等，2004）网站，并在另外两个 R 软件包（XML 和 RCurl）的基础上开发完成。

1.2 BioIDMapper 下载资源

BioIDMapper 自 2008 年开发完成后，2008—2010 年首先存储在 CRAN 上（https：//cran. r-project. org/），2010 年以后，软件包存储在 sourceforge 上。下载地址是 https：//sourceforge. net/projects/bioidmapper/。

1.3 系统要求

BioIDMapper 需要安装以下程序和 R 软件包。

1. 安装 RCurl 软件包

(1) 安装 curl (7.14.0 或更高版本) https：//curl.haxx.se。

(2) 安装 R 软件包：RCurl。

```
install.packages("Rcurl")
```

2. 安装 XML (具体问题请参阅 XML 安装说明)

(1) 安装 libxml2 (≥2.6.3)。

(2) 安装 R 软件包：XML。

```
install.packages("XML")
```

1.4 软件构架

BioIDMapper 是一个集成两个世界最大数据库 (NCBI 和 UniProt) 的 R 软件包 (图 1-1)。NCBI 是世界最大的存储基因相关信息的数据库，UniProt 则提供最大的蛋白质信息存储。BioIDMapper 可以用于转换不同数据库中 59

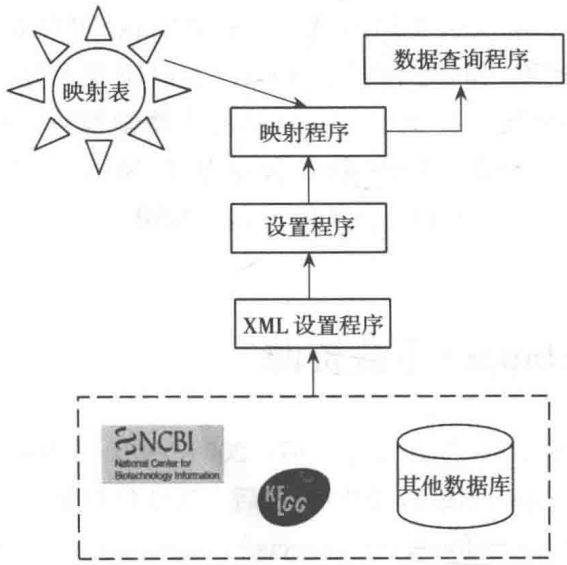


图 1-1 软件构架

种 ID，包括 DNA、RNA、蛋白质和代谢等生物大分子。此软件包需要互联网连接，并已在 Windows 和 Linux 上测试完成。由于 NCBI 和 UniProt 网站的查询流量限制，每次只能查询有限的 ID，因此 BioIDMapper 会自动进行分组，并根据要求分批查询用户要求的所有数据。详细信息可以查看安装指南。

BioIDMapper 采用两个类：Configure 和 MAP。Configure 类包含 parseXML、setupMap 和 setupBridge 3 个函数。MAP 类则包括 Convert 和 Validate 两个函数。具体信息见图 1-2。

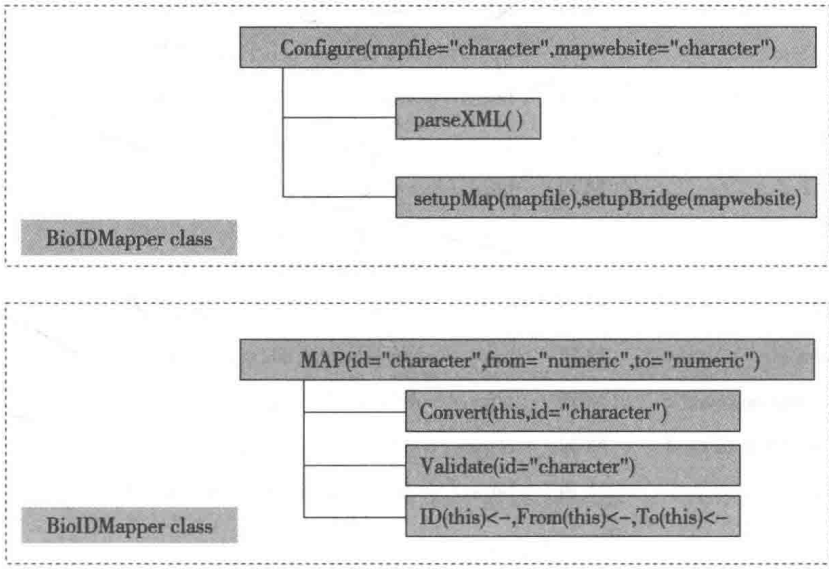


图 1-2　BioIDMapper 的两个类：Configure 和 MAP

GI 编号是沟通 NCBI 和 UniProt 的桥梁（图 1-3）。BioIDMapper 可以通过 GI 编号查询 NCBI 和 UniProt 数据库，并转换 59 种生物 ID（表 1-1）。

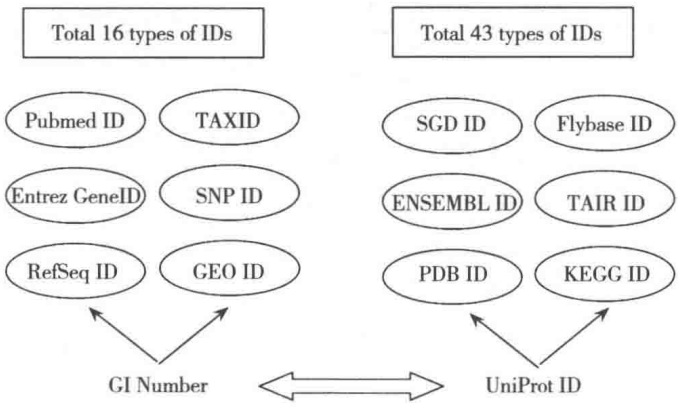


图 1-3　ID 转换示意

表 1-1　59 种生物大分子 ID 信息

编号	生物编号	来源数据库	网　址
1	GI number	NCBI	https://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html
2	Pubmed id	NCBI	https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/#converter
3	GEO id	NCBI	https://www.ncbi.nlm.nih.gov/geo/
4	OMIM id	NCBI	https://www.ncbi.nlm.nih.gov/omim
5	SNP id	NCBI	https://www.ncbi.nlm.nih.gov/snp
6	UniGene cluster id	NCBI	https://www.ncbi.nlm.nih.gov/unigene
7	UniSTS id	NCBI	ftp://ftp.ncbi.nih.gov/pub/ProbeDB/legacy_unists/
8	Popset id	NCBI	https://www.ncbi.nlm.nih.gov/popset
9	MMDB id	NCBI	https://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
10	3D SDI id	NCBI	https://www.ncbi.nlm.nih.gov/cdd
11	PSSM id	NCBI	https://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi
12	TAXID	NCBI	https://www.ncbi.nlm.nih.gov/taxonomy
13	Genome id	NCBI	https://www.ncbi.nlm.nih.gov/genome/
14	PubChem Compound id	NCBI	https://pubchem.ncbi.nlm.nih.gov/
15	PubChem Substance id	NCBI	https://pubchem.ncbi.nlm.nih.gov/
16	PubChem BioAssay id	NCBI	https://preview.ncbi.nlm.nih.gov/pcassay
17	NNNNNN	Boundary	
18	GI number	UniProt	https://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html
19	UniProtKB Accession	UniProt	http://www.uniprot.org/
20	UniProtKB id	UniProt	http://www.uniprot.org/
21	PIR Accession	UniProt	https://www.ncbi.nlm.nih.gov/protein
22	Enzyme Commission	UniProt	https://en.wikipedia.org/wiki/Enzyme_Commission_number
23	GO id	UniProt	http://geneontology.org/
24	Entrez Gene id	UniProt	https://www.ncbi.nlm.nih.gov/gene
25	EMBL id	UniProt	http://www.ebi.ac.uk/ena
26	ENSEMBL id	UniProt	http://ensemblgenomes.org/
27	UniGene id	UniProt	https://www.ncbi.nlm.nih.gov/unigene
28	TAIR id	UniProt	http://www.arabidopsis.org/
29	TIGR id	UniProt	http://blast.jcvi.org/euk-blast/index.cgi?project=osa1
30	KEGG id	UniProt	http://www.genome.jp/kegg/
31	NCBI Taxon id	UniProt	https://www.ncbi.nlm.nih.gov/taxonomy

（续）

编号	生物编号	来源数据库	网　址
32	OMIM id	UniProt	https://www.ncbi.nlm.nih.gov/omim
33	Ecogene id	UniProt	http://www.ecogene.org
34	Flybase id	UniProt	http://beta.flybase.org/
35	GENEDB SPOMBE id	UniProt	http://www.uniprot.org/
36	GERMONLINE id	UniProt	http://www.uniprot.org/
37	GRAMENE id	UniProt	http://gramene.org/
38	HIV id	UniProt	http://www.uniprot.org
39	IPI	UniProt	http://www.uniprot.org
40	PDB id	UniProt	http://www.rcsb.org/pdb/home/home.do
41	REBASE id	UniProt	http://www.uniprot.org
42	Refseq Accession	UniProt	https://www.ncbi.nlm.nih.gov/refseq
43	SGD id	UniProt	https://www.yeastgenome.org/
44	TRANSFAC id	UniProt	http://gene-regulation.com/pub/databases.html
45	WORMPEP id	UniProt	http://www.uniprot.org
46	UniRef100 id	UniProt	http://www.uniprot.org
47	UniRef90 id	UniProt	http://www.uniprot.org
48	UniRef50 id	UniProt	http://www.uniprot.org
49	InterPro id	UniProt	http://www.ebi.ac.uk/interpro/
50	Medline id	UniProt	https://www.nlm.nih.gov/bsd/pmresources.html
51	PFAM id	UniProt	http://pfam.xfam.org/
52	PIRSF id	UniProt	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml
53	PRINTS id	UniProt	http://www.uniprot.org
54	PRODOM id	UniProt	http://www.uniprot.org
55	PROSITE id	UniProt	http://www.uniprot.org
56	PMID	UniProt	http://www.uniprot.org
57	SMART id	UniProt	http://www.uniprot.org
58	TAXGRPID	UniProt	http://www.uniprot.org
59	TIGRFAMs id	UniProt	http://www.uniprot.org
60	TRANSFAC id	UniProt	http://www.uniprot.org

1.5　功能简介

软件系统核心是生物 ID 转换表，由 bio.type() 函数调用。基于转换表可

以提供3种功能：ID转换、连接和数据分析。转换功能是由bio.convert()函数来实现的；连接功能可以使用bio.link()函数，而数据分析则需要由两个函数完成：bio.sum()和bio.select()。

1.5.1 生物ID转换表（bio.type）

```
bio.type <- function(type2id)
```

此函数显示所有系统涉及的生物大分子ID信息。目前系统支持59种生物ID。执行此函数只需要一个参数：type2id。

type2id是生物ID转换表中的固定编码。如果此参数为空，则显示系统支持的所有生物ID(59种)，也就是展示整个生物ID转换表；如果此参数是生物ID转换表中的编码，则返回相应的生物大分子名称；如果参数是生物大分子名称，则会返回相应的生物ID转换表中的编码。

```
> bio.type()
```

Biokey number	BioIDs	Sources
[1,] "1"	"GI number"	"NCBI"
[2,] "2"	"Pubmed id"	"NCBI"
[3,] "3"	"GEO id"	"NCBI"
[4,] "4"	"OMIM id"	"NCBI"
[5,] "5"	"SNP id"	"NCBI"
[6,] "6"	"UniGene cluster id"	"NCBI"
[7,] "7"	"UniSTS id"	"NCBI"
[8,] "8"	"Popset id"	"NCBI"
[9,] "9"	"MMDB id"	"NCBI"
[10,] "10"	"3D SDI id"	"NCBI"

……

```
> library(BioIDMapper)
> bio.type(5)
```

[1] "SNP id"

使用此函数可以将“SNP id”转换成SNP的生物编码号“5”。

```
> bio.type("SNP id")
```

[1] 5