李靖宇,程景民,贺培凤,等. 基于文献挖掘视角的组学研究脉络梳理[J]. 中华医学图书情报杂志,2018,27(3):44-49.

DOI:10.3969/j.issn.1671-3982.2018.03.009

·情报研究与方法·

基于文献挖掘视角的组学研究脉络梳理

李靖宇,程景民,贺培凤,田 玥,于 琦

[摘要]目的:梳理组学研究脉络,分析组学研究热点趋势,帮助科研工作者把握研究方向,为科研管理部门在组学研究方面的资助提供决策支持。方法:利用 E-utilities 下载 1986-2016 年 PubMed 数据库收录的含有"-Omics/Omic/Omes/Ome"的题录信息 27 040 819 条,在题录信息的题目和摘要中对以"-Omics/Omic/Omes/Ome"结尾的词进行标注、筛选,并利用词频分析法、共现分析法、社会网络分析法对筛选结果进行分析。结果:获得组学相关记录共 346 977 条,文献发表量逐年递增,自 2000 年后呈井喷式增长。各类单一组学的文献发表量均呈逐年递增的趋势,其中基因组学文献发表量最多。2014 年之后,转录组学的文献发表量赶超了蛋白质组学,成为继基因组学之后的第二大组学类型。转录组学、线粒体基因组学与基因组学共现的文献量在某一时间节点后呈井喷式增长,其余组学类型与基因组学共现的文献量都稳定增长。结论:后基因组时代把组学研究推向了高潮,无论是数量还是种类都出现了井喷式的增长。多类型组学的融合研究越来越受科研人员的关注。

[关键词]组学;文献挖掘;共现分析;可视化

[中图分类号]Q819;TP311.13 [文献标志码]A [文章编号]1671-3982(2018)03-0044-06

Literature mining-based summarization of studies on omics

LI Jing-yu, CHENG Jing-min, HE Pei-feng, TIAN Yue, YU Qi

(Shanxi Medical University Management School, Taiyuan 030001, Shanxi Province, China)

Corresponding author: YU Qi

[**Abstract**] **Objective** To analyze the hotspots and trend in studies on omics in order to help scientific researchers to keep their research direction and scientific management departments in deciding their fund support for the studies on omics. **Methods** A total of 27 040 819 papers with their titles containing -omics/omic/omes/ome covered in PubMed from 1986 to 2016 were downloaded using E-utilities. The words with a suffix of -omics/omic/omes/ome in the titles and abstracts of these papers were tagged. The picked out words were analyzed by word frequency analysis, co-occurrence analysis and social network analysis, respectively. **Results** A total of 346 977 words containing -omics/omic/omes/ome were picked out from the 27 040 819 papers. The number of published papers increased progressively year after year and has been increasing in a blowout way since 2000. The number of published papers on each omics tended to increase year by year with genomics ranked first followed by transcriptome because the number of published papers on it has been more than that on proteomics since 2014. The co-occurrence of words with a suffix of -omics/omic/omes/ome with a suffix of -omics/omic/omes/ome in the published papers on transcriptome, mitochondrial genomics and genomics increased in a blowout way at a certain time point while a stable increase was found in that of the other omics and genomics. **Conclusion** The post genome era has pushed the studies on genomics to a new high tide with a blowout increase both in

[基金项目]国家自然科学基金面上项目"基于多元分析的科研文献微观实体评价理论与实证研究——以生物医学为例"(71573162)

[作者单位]山西医科大学管理学院,山西 太原 030001

[作者简介]李靖宇(1992-),男,山西太原人,在读硕士研究生。

[通讯作者]于 琦(1982-),男,山西临汾人,博士,副教授,主持国家自然科学基金 2 项,发表 SCI、SSCI 论文 13 篇,研究方向为数据驱动的生物医学知识发现。E-mail:yuqi351@gmail.com

their number and in their types. Collaborative studies on multiomics have attracted the attention of more and more scientific researchers.

[**Key words**] Omics; Literature mining; Co-occurrence analysis; Visualization

随着科学研究的进展,人们发现单纯研究某一方向无法解释全部生物医学问题,科学家便提出从整体出发研究人类组织细胞结构、基因、蛋白及其分子间相互的作用,通过整体分析反映人体组织器官功能和代谢的状态,因此便产生了“组学”的概念。从分子生物学角度,组学主要涵盖基因组学、蛋白组学、代谢组学、转录组学、脂类组学、免疫组学、糖组学和 RNA 组学等。Omics 是组学的英文称谓,其词根“-ome”在英文中是指一些种类个体的系统集合。Genomics(基因组学)是最早提出的组学类型,由美国科学家 Thomas Roderick 于 1986 年提出[1],之后其他类型的组学相继出现。笔者通过查阅分析国内外大量组学相关综述后发现,现阶段的组学研究综述都是关注某一种组学的最新进展,缺少从宏观角度分析多种组学的融合研究。就目前组学研究的态势而言,多种组学技术融合已成为必然趋势。因此,全面研究组学的整体发展趋势和各类组学之间的脉络关系,显得十分重要。文本挖掘技术和信息计量学方法的发展为从海量的科研文献中梳理组学研究脉络提供了可能[2]。

文献是科研成果的主要产出和表达形式,是由科研工作者对其创造性研究成果进行理论分析和科学总结并公开发表的文体,也是医学事业不断发展的重要科技信息源,是记录医学科技进步、重大发明和改革的历史性文件[3]。文献挖掘[4]是数据挖掘领域中一个重要研究方向,其处理对象是文本类型的文献数据。一般通过统计方法获取所关注的文献,再使用自然语言处理方法从中抽取出特定的事实信息,并对内容进行分析,从非结构化的数据中分析出隐藏的一些规律。文献挖掘方法已在多个领域中得到了广泛的应用,如生物学、医药学、生物医药学以及科学计量学等。文献挖掘技术[5]主要包括信息检索、实体识别和信息抽取。实体识别[6]旨在发现文献中重要的实体,该技术中常见的方法为基于特征、基于词典或者基于规则进行实体识别。而信息抽取技术主要把文献中含有的重要信息或者事实抽取出来,并用形式化的结构表示,依据共现关系[7]和自然语言处理技术[8]进行文本内容关系的抽取。

文献计量分析[9]有助于全面了解某一研究领域的国内外文献发表情况,目前以所有组学为对象的文献计量分析少之又少。通过分析国内外文献发表情况,方便该领域研究人员了解组学的研究现状及发展方向,有助于科研管理机构在项目评审、资助中合理分配资源,有助于其科研选题、成果发表及选择研究合作方并调整研究方向[10]。

本文拟利用文献计量学方法,借助 PubMed 数据库及相关文献挖掘、分析方法对“组学(Omics)”相关英文文献进行统计和分析,探寻组学的研究轨迹,为研究人员更加深入系统地开展组学研究提供参考。

1　资料来源与方法

本文使用的数据集来自 PubMed 数据库[11]。美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)提供的 The Entrez Programming Utilities(E-utilities)编程工具,是访问 NCBI Entrez 查询和 PubMed 数据库的稳定接口,可以实现 PubMed 数据库记录的批量下载。本文使用 E-utilities 中的 Esearch 和 Efetch 2 种工具获取 PubMed 记录,时间跨度为 1896-2016 年,共获得 27 040 819条记录,包含所有的出版类型。本文关注的主题为“组学(Omics)”。组学是研究一些种类个体的系统集合的学科,如基因组是构成生物体所有基因的组合,基因组学这门学科是研究这些基因以及这些基因间的关系,因此我们将组(Omes)与组学(Omics)同等对待。具有组学含义的单词均有一个共同的特征,即以“-ome”“-omes”“-omic”或“-omics”结尾,故本文选取文献的“title”或“abstract”为统计窗口,从中识别具备上述特征的单词,在数据集中共识别出 19 268 个具备上述特征的单词。通过删除噪音单词(如“some”“home”等),最后得到 77 个出现频次不低于 10 次的“-Omics”单词(表

1),以供下一步分析。将上述 77 个“-Omics”单词重新在原始数据的题目和摘要中用 Python 语言编写的程序进行匹配,27 040 819 条原始数据中含有 77 个“-Omics”单词中的任何一个的记为有效数据,共得到 346 977 条记录作为本文的数据集。

可视化分析采用 VOSviewer,它是一款用来构建和查看文献计量图谱的免费文献计量分析软件,基于文献的共引和共被引原理,可用于绘制各个知识领域的科学图谱。将所有类型组学的共现数据经过处理后导入 VOSviewer 进行可视化,得到网络可视化图。图中圆圈和标签代表关键词,圆圈及标签大小代表其重要性的高低,拥有相同颜色的圆圈属于同一个聚类[12]。

主题河是一种被证明为可有效反映文本之间的时间属性的方法。在这种可视化方法中,时间被表示为从左往右的一条水平轴,然后用不同的颜色条带代表不同的主题,条带的宽度代表该主题在该时间的一个度量。这样人们可以跟踪任何一个主题在量上随时间的变化,也能比较不同的主题在同一个时刻相对规模的大小[13]。

表 1 出现频次不低于 10 次的“-Omics”单词

组学类型	频次	组学类型	频次	组学类型	频次
genomics	328827	immunoproteomics	246	antivenomics	65
proteomics	83232	chemogenomics	358	oncoproteomics	63
transcriptomics	40858	proteogenomics	316	sphingolipidomics	29
metabolomics	18565	exposomics	407	ecogenomics	60
microbiomics	14914	immunomics	268	paleogenomics	44
metagenomics	6414	physiomics	249	ecotoxicogenomics	55
pharmacogenomics	5500	radiomics	343	nanoproteomics	41
epigenomics	4655	n-glycomics	514	ribonomics	22
interactomics	4250	venomics	187	infectomics	18
secretomics	3997	fluxomics	110	phospholipidomics	20
mitogenomics	4120	neuroproteomics	174	o-glycomics	15
phosphoproteomics	3076	ionomics	241	sociogenomics	19
lipidomics	1872	radiogenomics	130	pharmacometabonomics	36
connectomics	2250	n-glycoproteomics	24	salivaomics	24
metabonomics	1839	oncogenomics	85	palaeogenomics	16
glycomics	1668	rnomics	87	nutriproteomics	19
kinomics	1284	pathogenomics	93	pharmacoepigenomics	17
peptidomics	1437	metalloproteomics	72	ecotoxicoproteomics	20
toxicogenomics	1240	cellomics	73	altitudeomics	14
phenomics	921	pharmacoproteomics	80	allergenomics	13
phylogenomics	693	immunogenomics	69	peptidogenomics	13
glycoproteomics	475	culturomics	95	redox-proteomics	12
degradomics	114	vaccinomics	97	neurolipidomics	11
metatranscriptomics	534	neurogenomics	60	physiogenomics	10
cytomics	284	toxicoproteomics	20	circomics	10
metaproteomics	312	foodomics	85		

2 结果与分析

从上述数据集中筛选出关于组学的相关文献共计 346 977 篇,包括期刊论文 345 549 篇(占 99.59%)和综述 1 428 篇(占 0.41%)。

2.1　文献年度变化趋势

文献的年度分布情况可以从一定程度上反映该领域的发展情况。分析文献量与时间变化的关系可以反映研究主题的发展情况，可以大体揭示该主题的发展阶段与规律。本文将组（-ome/-omes）与组学（-omic/-omics）同等对待，美国科学家 Thomas Roderick 于 1986 年最先提出的是 Genomics（基因组学），而第一篇提到“基因组（genome）”的文献则出现在 1943 年。1943-2016 年全世界组学相关文献发表情况如图 1 所示。从 1943 年之后组学相关研究的发表量整体呈逐年递增趋势，从 1999 年的 4 331篇迅速增长到 2000 年的 5 288 篇，到 2016 年文献发表量已达 40 590 篇。

人类基因组计划（Human Genome Project，HGP）由美国于 1987 年启动，2000 年 6 月 26 日参加人类基因组工程项目的美国、英国、法国、德国、日本和中国等 6 国科学家共同宣布，人类基因组草图的绘制工作已经完成，后基因组时代来临。组学领域的研究文献呈现了井喷式的增长，已有越来越多的国内外科学工作者投入到组学研究中，并获得了大量的研究成果，组学已逐渐成为生物医学研究领域的热点之一。

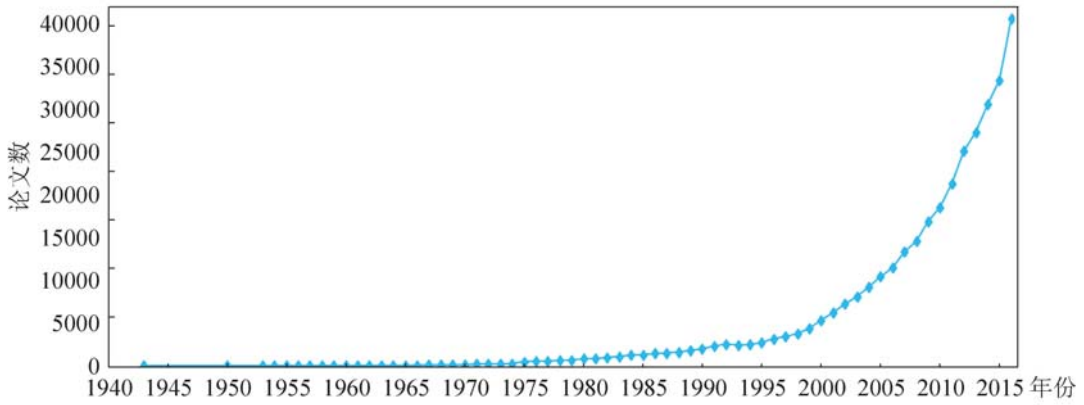

图 1　组学研究论文发表情况

2.2　各类组学文献情况

2000 年之前组学研究类型较单一，之后各类组学研究相继涌现，并呈现出不同的变化。我们选取数据中论文总数排名前 10 的组学类型进行比较，发现各类组学都呈现了逐年递增的现象。其中基因组学的文献发表量遥遥领先，蛋白质组学和转录组学的文献发表量紧随其后。2000-2014 年，蛋白质组学的文献发表量一直高于转录组学，2014 年之后转录组学的文献发表量赶超了蛋白质组学。原因在于，从 2008 年开始，第二代测序技术利用一系列高通量测序技术（high throughput sequencing）进行大规模的基因组 DNA 或 RNA 测序，能够快速准确地获得基因组编码序列，满足极短时间内对基因组进行高分辨率检测的要求。随着第二代测序技术高通量、高准确率、低成本等优点的实现，转录组学测序技术也随之得到了更广泛的应用[14]。因此，转录组学的关注度逐渐升高并且超过了蛋白质组学的关注度。

2.3　各类组学共现情况

统计不同类型的“组学”之间在同一篇文献的题目和摘要中出现的情况，便可形成多组学研究的相关关系。将多组学共现类型细分为在同一篇文献中分别出现 2 种类型、3 种类型、4 种及 4 种以上类型，并进行分类计量。1995 年首次出现多组学共现的文献。2 种类型组学共现的文献量一直处于遥遥领先的状态，3 种类型组学共现和 4 种及 4 种以上组学类型共现的文献量较 2 种类型组学共现的文献量还有一些差距，但总体来说各种共现情况都随着时间的增长呈现出逐年递增的趋势。

将多组学共现数据导入 VOSviewer 中，其结果以可视化图谱的形式展示出来。如图 2 所示，可以看出基因组学、转录组学、蛋白质组学、代谢组学是组学共现研究的热点，文献数量居于前列。通过连线可以看出，基因组学与转录组学的共现文献量最多，基因组学与蛋白质组学的共现文献量次之。各

类组学之间都存在着错综复杂的关系。

研究结果表明，多组学的结合研究已成为组学研究领域的趋势，整合多组学数据用于药物重定位和个性化医疗越来越受到重视[15]。因此，相关领域科研人员未来要注意多组学类型的结合研究，从而促进组学研究的进一步发展。

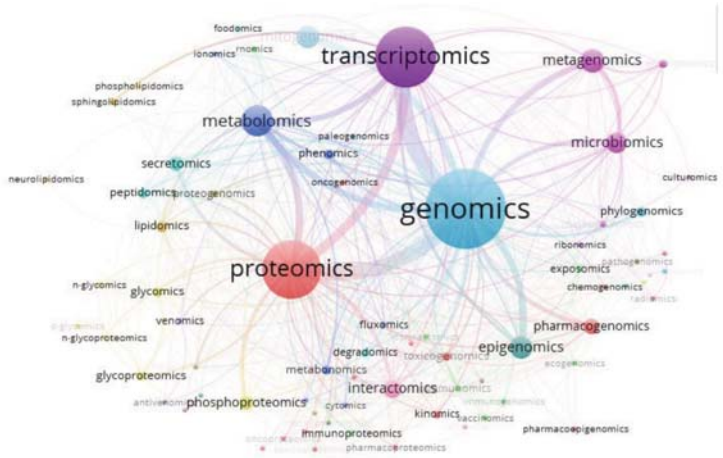

图 2 所有类型组学共现情况

2.4 基因组学脉络研究

“基因组学”为最早出现的组学类型，且与各类组学都有共现的情况，因此以“基因组学”为主脉络，展示其余各类组学与“基因组学”共现研究的相关情况，通过主题河图进行呈现。选取与“基因组学”共现文献总量排名前 15 的组学类型，年份从出现多组学共现的第一篇文献的 1995 年到 2016 年进行研究（图 3）。图 3 中河流的宽窄代表各类组学与基因组学共现的文献数的比例，横坐标为年份的变化。

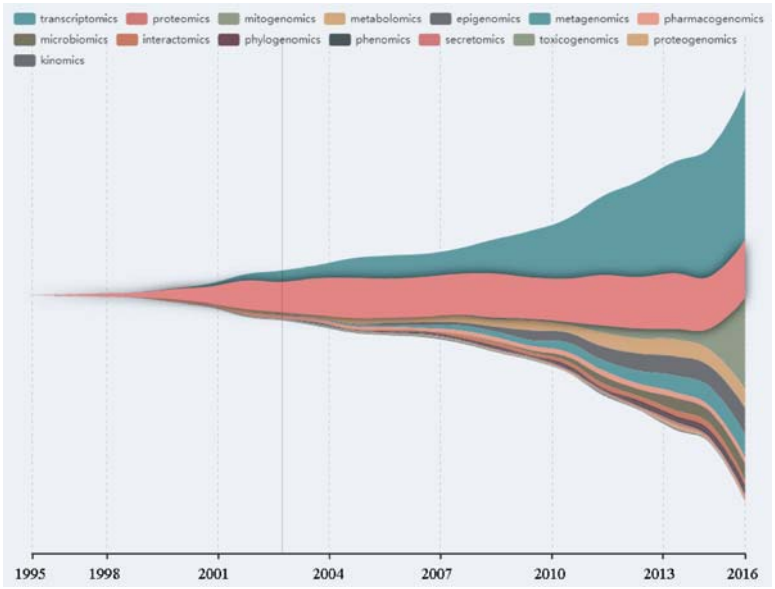

图 3 “基因组学”与其他各类组学共现论文数变化情况

2.4.1 稳定型增长类型

最早与“基因组学”共现的是“蛋白质组学”。“蛋白质组学”这个概念由 Marc Wikins 1994 年首次提出[16]。在 1995 年“基因组学”与其他类型组学共现的 5 篇论文中，4 篇是“基因组学”与“蛋白质组学”的共现，“蛋白质组学”与“基因组学”的共现论文数一直处于稳定增长的趋势。究其原因，一方面，从分层递阶结构来说，蛋白质系统的粒度较基因组

系统粒度粗,蛋白质系统数据处理的复杂度不会超过基因组系统数据处理的复杂度;另一方面,蛋白质的功能性研究距离我们所期望的在细胞水平上研究分子生物学更近,或者说距离在实际应用中所需要的功能研究更近,如在药物基因组学中的关键蛋白质组的寻找[17]。

2.4.2　井喷型增长类型

“转录组学”与“基因组学”的共现文献量随时间的变化呈井喷式增长,到 2016 年已成为与“基因组学”共现占比最大的基因类型。究其原因,一是转录组学是功能基因组学研究的重要组成部分,是一门在整体水平上研究细胞中所有基因转录及转录调控规律的学科[18-19];二是随着新一代高通量基因测序技术运用到转录组学研究之中,转录组学研究中提供的数据量呈现爆炸式的扩增,拓宽了转录组学研究解决科学问题的范围[14]。

“线粒体基因组学”与“基因组学”的共现论文数量从 1995 年到 2014 年一直处于缓慢增长态势,然而到 2015 年共现文献量呈现井喷式增长,成为 2016 年当年排在“转录组学”之后的第二大共现组学类型。究其原因,是由于“线粒体基因组学”在 2008 年后随着中国科研人员的加入,半翅目昆虫线粒体基因组测序进入了迸发阶段,在 2008-2015 年共获得了 89 种昆虫的线粒体基因组,其中 81 种在中国完成测序。截至 2015 年 5 月,美国国立生物技术信息中心共收录 100 种半翅目昆虫的线粒体基因组,其中 83 个为全线粒体基因组,17 个近似完整的线粒体基因组[20]。线粒体基因组的获取完成在极大程度上推进了线粒体基因组学与基因组学的共同研究。

3　结束语

文献资料中涵盖了大量重要信息,能够从海量的文献资料中快速挖掘出人们所需求的信息知识,是文献挖掘技术日益受重视的主要原因。我国“文献挖掘”多采取在数据库中检索所研究的主题对结果进行分析的方式。本文采用获取 PubMed 数据库 1896-2016 年的全数据的方法,通过对所研究主题的词根进行识别挖掘,运用社会网络分析的方法和可视化技术,从组学相关文献的年度变化趋势和共现情况方面进行分析,为传统的文献挖掘提供了一种新的思路,为学者和研究人员创造了一个知识共享平台。同时通过分析研究数据,发现后基因组时代的到来把组学研究推向了高潮,无论是数量还是种类都出现了井喷式的增长。多类型组学的融合研究越来越受科研人员的关注,已成为未来组学研究的热点趋势。本文的不足在于只从英文文献着手,研究方法还不够完备,对多种类型的数据的处理与挖掘还不完善。

【参考文献】

[1] Kuska B. Beer, Bethesda, and biology: how “genomics” came into being[J]. Journal of the National Cancer Institute, 1998, 90(2): 93.

[2] 李　伟,王士泉. 基于专病队列的重大疾病临床样本生命组学数据库建设[J]. 中华医学图书情报杂志,2017,26(6):11-16.

[3] 梅升辉,陈瑞玲,朱乐亭,等. 基于 Web of Science 的蛋白质组学文献计量分析[J]. 中南药学,2016,14(4):358-362.

[4] Lok C. Literature mining: speed reading[J]. Nature, 2010, 463(7280): 416.

[5] Yandell MD, Majoros WH. Genomics and natural language processing[J]. Nature Reviews Genetics, 2002, 3(8): 601.

[6] Konkol M, Brychcín T, Konopík M. Latent semantics in Named Entity Recognition[J]. Expert Systems with Applications, 2015, 42(7): 3470-3479.

[7] Cohen AM, Hersh WR, Dubay C, *et al*. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts[J]. Bmc Bioinformatics, 2005. 6(1): 1-15.

[8] Thessen AE, Cui H, Mozzherin D. Applications of natural language processing in biodiversity science[J]. Advances in Bioinformatics, 2012, 2012(2): 391574.

[9] 李　旋,郝继英. 学者的学术影响力评价方法[J]. 中华医学图书情报杂志,2016,25(8):48-52.

[10] 陈冠初. 文献计量学与非文献计量学在期刊评价中的应用[J]. 编辑学报,2006,18(6):472-474.

[11] 吕　霖. PubMed 及其衍生数据库在医药领域的应用[J]. 中国发明与专利,2014(8):107-113.

[12] 宋秀芳,迟培娟. Vosviewer 与 Citespace 应用比较研究[J]. 情报科学,2016,34(7):108-112,146.

[13] 詹　建,高民权. 基于主题河的网络舆情可视化关联分析方法[J]. 情报资料工作,2014,35(6):17-22.

[14] 王　跃,毛开云,王恒哲,等. 转录组学测序技术应用与市场分析[J]. 生物产业技术, 2017(5):11-17.

[15] 刘　阳,白　卉,陶　欢,等. 面向药物发现和精准医疗的基因表达谱分析[J]. 生物化学与生物物理进展,2016,43(10):923-935.

[16] 尹　稳,伏　旭,李　平. 蛋白质组学的应用研究进展[J]. 生物技术通报,2014(1):32-38.

[17] 唐旭清,朱　平. 后基因组时代生物信息学的发展趋势[J]. 生物信息学,2008,6(3):142-14.

[18] 祁云霞,刘永斌,荣威恒. 转录组研究新技术:RNA-Seq 及其应用[J]. 遗传,2011,33(11):1191-1202.

[19] 李小白,向　林,罗　洁,等. 转录组测序(RNA-seq)策略及其数据在分子标记开发上的应用[J]. 中国细胞生物学学报,2013(5):720-726,740.

[20] 郭仲龙,袁明龙. 半翅目昆虫线粒体基因组学研究进展[J]. 中国科学:生命科学,2016,46(2):151-166.

[收稿日期:2018-02-20]

[本文编辑:黄思敏]