基于家蚕中肠 RNA-seq 数据的新基因发掘及初步分析

周凯月[1], 唐 健[1], 李玉霞[1], 郝长富[1], 徐安英[1,2*]

(1.江苏科技大学生物技术学院,江苏镇江 212018;2.中国农业科学院蚕业研究所,江苏镇江 212018)

摘要 对前期获得的家蚕中肠组织转录组数据进行进一步生物信息学分析,利用 Cufflinks 软件对 Mapped Reads 进行组装,并与参考基因组注释信息进行比对,共发掘新基因 788 个。利用 Blast 软件将发掘的新基因与 NR,Swiss-Prot 数据库比对,其中 746 个新基因得到功能注释。这些新发掘基因在家蚕 28 个连锁群上都有分布,其中在第 15 连锁群上最多。有 198 个新基因注释到 KEGG 数据库中,分布于 85 条已知的通路中,将新基因与 COG 数据进行比对,并进行功能注释与分类,总共有 258 个新基因得到了注释,被分为 22 个 COG 类别,部分新基因的克隆分析,与预期结果一致,这些发现为以后进一步研究新基因的功能提供了基础信息。

关键词 转录组测序;新基因;序列比对;功能分析

中图分类号 S881.2+6 **文献标识码** A **文章编号** 0517-6611(2018)24-0060-05

New Gene Discovery and Analysis Based on the Silkworm Midgut RNA-seq Data

ZHOU Kai-yue, TANG Jian, LI Yu-xia et al (1.School of Bio-technology, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212018; 2.Institute of Sericulture, Chinese Academy of Agricultural Sciences, Zhenjiang, Jiangsu 212018)

Abstract In this research, the previous silkworm midgut RNA-seq data were used for further bioinformatics analysis. We used cufflinks software to assemble mapped reads, and finally found 788 new transcripts. After blasted to NR, Swiss-Prot database, 746 genes were annotated. These new genes were distributed in 28 chromosomes, and the most of which were on the 15th chromosome. 198 new genes were annotated into the KEGG database, and distributed in 85 known pathways. A total of 258 new genes were annotated of 22 COG categories after functional annotation and classification with the COG database. Cloning and analysis of some new genes, consistent with expected results. These new findings provide basic information for further study of the function of new genes.

Key words RNA-sequencing; New genes; Sequence alignment; Functional analysis

我国是茧丝绸生产和出口大国,家蚕作为重要的吐丝昆虫,具有很高的经济价值,在生物反应器和模式生物中也显现出广阔的应用前景。随着家蚕研究的深入,越来越多的新基因被发掘出来,一些新基因往往发挥着极为重要的作用,甚至是某些疑难问题的突破点,人们对新基因功能的研究也越来越重视。

转录组研究是一个发掘功能基因的重要途径,与基因组学相比,转录组学只研究被转录的基因,研究范围缩小,针对性更强[1],越来越多地被运用到基因功能的相关研究中。转录组是特定组织或细胞在某一发育阶段或功能状态下转录出来的所有 RNA 的总和[2],主要包括 mRNA 和非编码 RNA(ncRNA)。转录组测序即通过第二代高通量测序技术对特定组织或细胞的转录产物(主要是全部 mRNA)反转录后测序并对其进行生物信息学分析的技术[3],是当前在全基因组水平上研究基因表达模式的主要技术[4],目前该技术已被广泛应用于生物信息学研究的多个领域[5]。笔者拟采用 RNA-seq技术对所构建的家蚕中肠的转录组进行测定,并在基因组水平上进行转录组分析,对新基因进行初步发掘,以期为新基因的功能鉴定基础。

1 材料与方法

1.1 材料 试验所用家蚕品种为抗 BmNPV 家蚕品种 QFN 和常规品种 QF,来源于中国农业科学院蚕业研究所蚕资源中心课题组。

基金项目 现代农业产业技术体系建设专项(CARS-18);镇江市现代农业重点研发计划项目(NY2017017)。

作者简介 周凯月(1992—),女,河南周口人,硕士研究生,研究方向:家蚕分子遗传学。*通讯作者,研究员,硕士生导师,从事蚕资源保存与利用研究。

收稿日期 2018-04-19;**修回日期** 2018-04-25

1.2 方法

1.2.1 转录组测序。首先进行样品检测、构建 RNA 文库,文库质控合格后,采用 HiSeq2500 进行高通量测序, Illumina HiSeq2500 高通量测序获得 Reads 或碱基信息,筛选除去冗余后得到 Clean Reads,通过 solexa QA 软件对其进行质量检测可得到高质量的 Clean Reads[6]。

1.2.2 转录组数据比对。对于 Clean Reads 需要用高效的序列比对软件 TopHat2 将其与参考基因组进行序列比对,得到 Mapped Reads。比对效率可以直接反映出转录组数据的利用率[6]。

1.2.3 新基因分析。通过 Cufflinks 软件对 Mapped Reads 进行组装,将得到的序列与参考基因组注释信息进行比对,寻找未知的新基因。再利用 Blast 软件对新基因进行功能注释,然后利用各个数据库分别对新基因的 NR 注释信息、COG 功能注释及其分类、KEGG 注释通路进行分析,获得新基因的相关注释信息。

1.2.4 部分新基因的克隆分析。以家蚕抗性品种 QFN 的中肠组织 cDNA 为模板,随机挑选 3 个新基因(Silkworm New Gene 1、Silkworm New Gene 5、Silkworm New Gene 20)设计引物,进行 RT-PCR 扩增。PCR 反应体系为(25 μL):1 μL cDNA 模板,2.5 μL 10×PCR Buffer,2 μL 10 mmol/mL dNTP,1 μL 20 pmol/mL 上下游引物,0.3 μL 5 U/μL Ex*Taq*DNA 聚合酶,加双蒸水补充至 25 μL。PCR 反应后检测扩增片段的大小,与目的片段大小是否相符。所需引物序列见表 1。

采用 SanPrep 柱式 DNA 胶回收试剂盒进行目的片段的回收和纯化。试验前确认 Wash Solution 中是否加入乙醇,将胶回收纯化的 PCR 产物连接到 pMD™18-T Vector 上,用冷

热刺激的方法转化感受态细胞,过夜培养后挑取独立菌落进行培养,再进行菌液 PCR 扩增判断是否为目的条带,挑选阳性菌液送生工生物工程(上海)股份有限公司测序鉴定。

2　结果与分析

2.1　测序数据　为研究家蚕感染 BmNPV 后蚕体内基因表达调控情况,以家蚕抗性品种 QFN 和常规品种 QF 中肠组织为材料进行转录组测序分析(表 2)。参考基因组组装能否满足信息分析的需求,可以通过转录组数据与参考基因组序列的比对结果评估(表 3)。

经筛选,2 个文库共获得 12.8 Gb Clean Data,QF、QFN 的≥Q30 的碱基百分比分别为 86.03%、87.01%。

表 1　部分新基因 RT-PCR 引物序列

Table 1　Partial new gene RT-PCR primer sequence

序号 No.	引物 Primer	引物序列 Primer sequence
1	1F	AAACCAAAACTAAAAGCA
2	1R	ACAGACACAGTCTCACCA
3	2F	TGGGTGATGGTGAGGTCC
4	2R	ATAAGAGTCAGCGGGGTT
5	3F	GAGAGAAGAAATAGGGGG
6	3R	TGGAAATCGTTTTGAAAG

表 2　样品测序数据统计

Table 2　Statistics of sample sequencing data

样品 Sample	Read 数目 Read number	总碱基数 Base number	GC 含量 GC content//%	≥Q30 的碱基百分比 Percentage of ≥Q30 base//%
QF	25 420 394	6 402 063 943	49.64	86.03
QFN	25 500 762	6 422 482 754	49.85	87.01

表 3　Clean Data 与参考基因组比对结果统计

Table 3　Statistics of comparison between clean date and reference genome

样品 Sample	总数目 Total reads	比对序列 Mapped read	比对比例 Mapped ratio//%	唯一比对序列 Uniq mapped reads	唯一比对序列比例 Uniq mapped ratio//%
QF	50 840 788	35 265 473	69.36	29 824 772	58.66
QFN	51 001 524	35 824 800	70.24	30 984 105	60.75

文库中比对到参考基因组上的 Reads 在 Clean Reads 的效率达 69.36%和 70.24%,其中比对到参考基因组唯一位置的 Reads 在 Clean Reads 中所占比例分别为 58.66%、60.75%。

2.2　新基因分析

2.2.1　新基因发掘及基因结构的分析。对测序得到的序列进行拼接和组装,与原有的一些基因组注释信息进行比对,寻找未被注释的新基因。该研究过滤掉编码的肽链过短(少于 50 个氨基酸残基)或只包含单个外显子的序列,得到了 788 个新基因。如 Silkworm New Gene1007 位于 nscaf2794 基因序列 172~4 311 的正链上,包含 5 个外显子;Silkworm New Gene1008 位于 nscaf2795 基因序列 2 402 828~2 404 107 的正链上,包含 3 个外显子,部分新基因的文件见表 4。

表 4　部分新基因的文件

Table 4　Some new gene files

染色体号 Seq ID	类型 Type	起始端 Start	终止端 End	特征序列所在的正负链 Strand	属性 Attributes
nscaf2794	gene	172	4 311	+	ID=SNG_1007
nscaf2794	mRNA	172	4 311	+	ID=SNG_1007.1;Parent=SNG_1007
nscaf2794	CDS	172	306	+	Parent=SNG_1007.1
nscaf2794	CDS	1 001	1 093	+	Parent=SNG_1007.1
nscaf2794	CDS	1 746	1 862	+	Parent=SNG_1007.1
nscaf2794	CDS	2 632	2 742	+	Parent=SNG_1007.1
nscaf2794	CDS	4 055	4 311	+	Parent=SNG_1007.1
nscaf2795	gene	2 402 828	2 404 107	+	ID=SNG_1008
nscaf2795	mRNA	2 402 828	2 404 107	+	ID=SNG_1008.1;Parent=SNG_1008
nscaf2795	CDS	2 402 828	2 402 987	+	Parent=SNG_1008.1
nscaf2795	CDS	2 403 062	2 403 222	+	Parent=SNG_1008.1
nscaf2795	CDS	2 403 302	2 404 107	+	Parent=SNG_1008.1
nscaf2795	gene	700 616	702 250	−	ID=SNG_1010
nscaf2795	mRNA	700 616	702 250	−	ID=SNG_1010.1;Parent=SNG_1010
nscaf2795	CDS	700 616	700 780	−	Parent=SNG_1010.1
nscaf2795	CDS	701 651	701 749	−	Parent=SNG_1010.1
nscaf2795	CDS	702 075	702 250	−	Parent=SNG_1010.1

注:SNG 为 Silkworm New Gene 简写

Note: SNG is short for silkworm new gene

2.2.2 新基因功能注释。使用 Blast 软件将发掘的新基因分别同 NR[7]、Swiss-Prot[8]、GO[9]、COG[10]、KEGG[11]数据库进行序列比对[12]，结果发现 788 个新基因中 746 个得到了注释，各数据库得到注释的基因数分别为 746、418、443、187、224 个。新基因功能注释结果统计详见表 5。

表 5 新基因功能注释结果统计

Table 5 Statistics of new annotation of gene function results

序号 No.	注释数据库 Annotated databases	新基因数 New gene number
1	COG	187
2	GO	443
3	KEGG	224
4	Swiss-Prot	418
5	NR	746

同时发现，其中 746(94.67%)个新基因与数据库中匹配到的序列具有显著的相似性(E<lE-5)，其余 42(5.33%)个新基因与匹配到的序列相似性较低，匹配 E 值的分布见图 1。

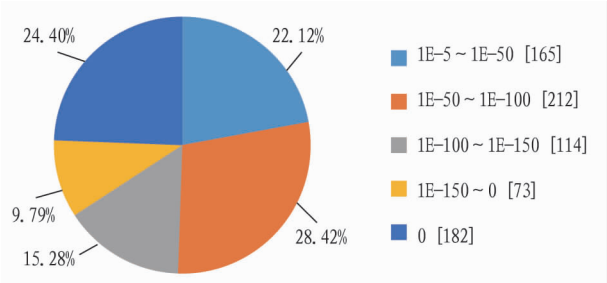


图 1 注释基因 E 值分布

Fig.1 Distribution of E-value for annotated genes

2.2.3 与注释基因匹配的物种分布。利用 BlastX 将组装出来的 unigene 序列与 NR 数据库进行比对后，共找到 746 个 unigene 与其他近缘生物的已知基因具有不同程度的同源性，746 个注释的基因中，有 611 条(81.90%)基因与家蚕(*Bombyx mori*)序列同源，99 条(13.27%)与黑脉金斑蝶(*Danaus plexippus*)序列同源、4 条(0.54%)与玉带凤蝶(*Papilio polytes*)序列同源、4 条(0.54%)与赤拟谷盗(*Tribolium castaneum*)序列同源、4 条(0.54%)与致倦库蚊(*Culex quinquefasciatus*)序列同源、3 条(0.40%)与柑橘凤蝶(*Papilio xuthus*)序列同源、2 条(0.27%)与印度跳蚁(*Harpegnathos saltator*)序列同源、2 条(0.27%)与佛罗里达弓背蚁(*Camponotus floridanus*)序列同源、2 条(0.27%)与毕氏粗角猛蚁(*Cerapachys biroi*)序列同源，仅有 15 条(2.01%)与其他物种序列相匹配(图 2)。

2.2.4 新基因 GO 富集分析。利用 Blast2Go 软件对筛选到的基因进行 GO 富集分析，结果显示，基因主要注释到细胞组分、分子功能和生物学过程 3 个分支中，分别有 947、551 和 1 777 个(图 3)。在细胞组分模块中(图 3A)，注释到细胞(cell)、细胞部分(cell part)的基因数目较多，分别占 19.6% 和 20.0%；在分子功能模块中(图 3B)，注释到黏合(binding)和催化活动(catalytic activity)的基因数目较多，分别占41.0% 和 37.7%。在生物学过程模块中(图 3C)，注释到细胞过程(cellular process)和单组织过程(single organismprocess)及代谢过程(metabolic process)的基因数目较多，分别占13.8%、13.6%和 13.9%。

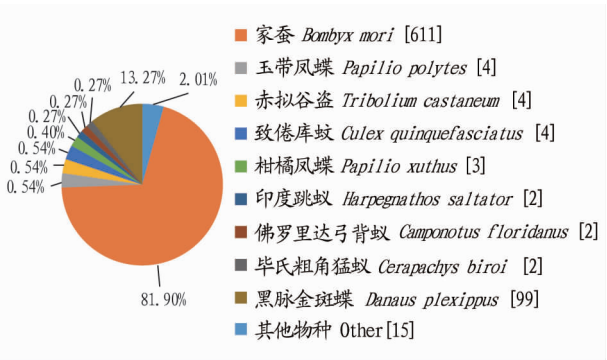


图 2 与注释基因匹配的物种分布

Fig.2 Distribution of species match to the annotated genes

2.2.5 新基因在家蚕基因连锁群上的分布。将新基因的 Locus 与家蚕基因连锁群进行比对发现，有 545 个新基因分布在不同的染色体上，且在 28 条染色体上都有分布(图 4)，在 18 号染色体上分布数量最多，有 58 个，其中在 nscaf2902 上分布有 54 个；在 15 号染色体上分布有 40 个，其中在 nscaf2888 上分布有 36 个；在 26 号染色体上分布最少，仅有 7 个。

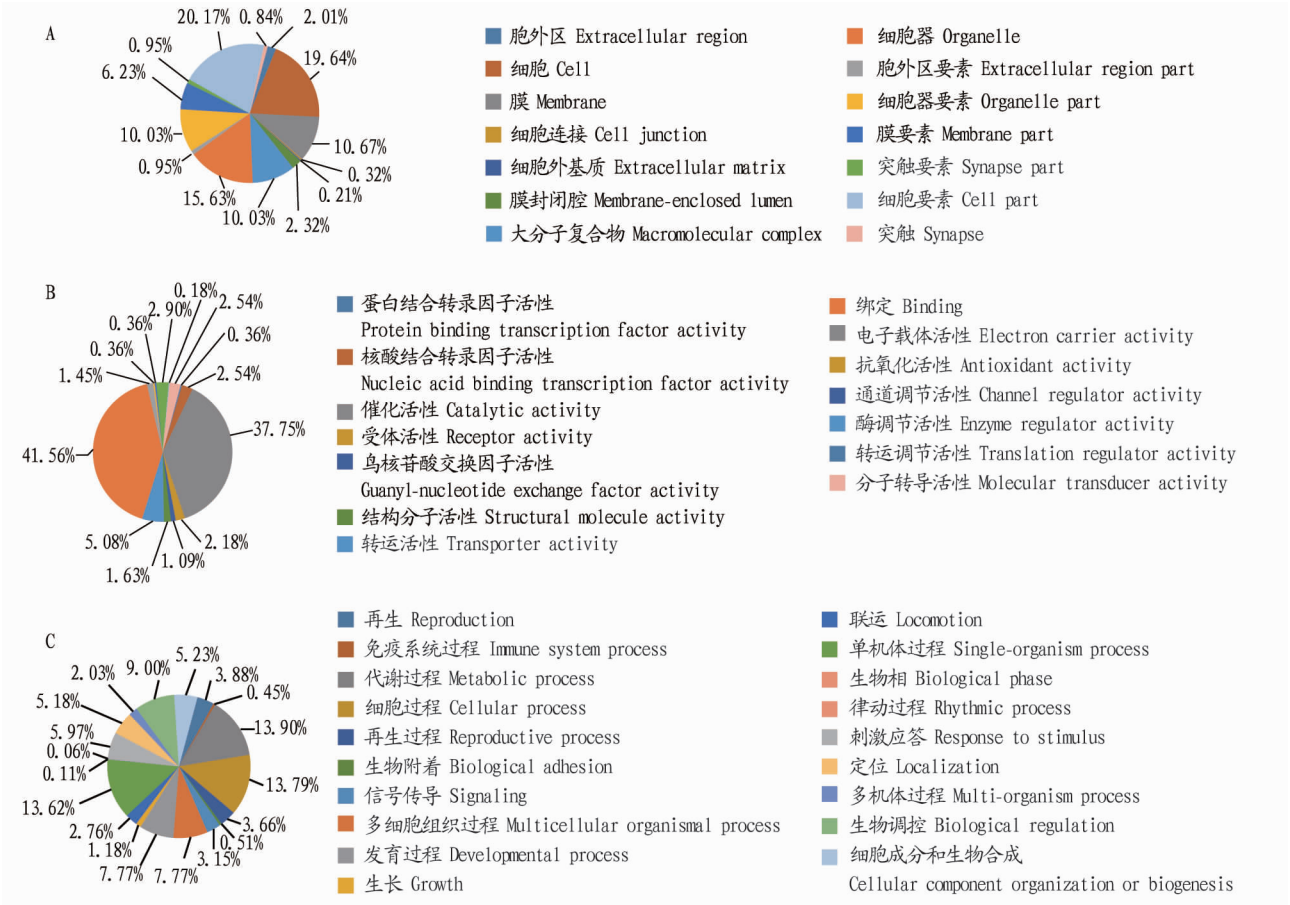
2.2.6 新基因通路富集分析。通过 KEGG 分析，对新基因进行通路富集分析发现，共有 198 个新基因注释到 KEGG 数据库中，分布于 85 条已知的通路中。映射基因最多的 5 个通路分别为剪接体(splicesome)(ko03040、12 条)、RNA 转运(RNA transport)(ko03013、9 条)、真核细胞核糖体合成(ribosome biogenesis in eukaryotes)(ko03008、6 条)、过氧物酶体(peroxisome)(ko04146、5 条)、内吞(endocytosis)(ko04144、5 条)。映射到的信号通路见图 5。

2.2.7 新基因 COG 数据库功能注释。将新基因与 COG 数据进行比对，并进行功能注释与分类，结果发现共有 22 个类别里的 258 个新基因得到了注释(图 6)，其中，一般功能(general function prediction only)的基因占总体的 20.93%，所占比例最大；复制、重组和修复(replication, recombination and repair)、碳水化合物的运输和代谢(carbohydrate transport and metabolism)、氨基酸的运输和代谢(amino acid transport and metabolism)3 个类别共居第 2 位，占总体的 7.75%；转录(transcription)占总体的 7.36%，其余分类的基因数较少，其中核结构(nuclear structure)、细胞运动(cell motility)、真核细胞的细胞外结构(extracellular structures)3 个类别里均无新基因出现。

2.2.8 部分新基因的克隆分析。以秋丰 N 中肠组织的 cDNA 为模板，随机挑选 3 个新基因设计引物扩增目的基因，获得特异性片段，与预期片段大小相符，经测序结果与参考序列比对后发现编码序列高度相似。

3 讨论

该研究基于所选参考基因组序列，共发掘 788 个新基因，通过生物信息学软件将发掘的新基因与 NR、Swiss-Prot、GO、COG 及 KEGG 数据库进行序列比对，共获得 746 个新基



注:A.细胞组分;B.分子功能;C.生物学过程

Note:A.Cellular components;B.Molecular function;C.Biological processes

图 3　注释基因在 GO 中的分类情况

Fig.3　Classification of annotated genes in GO

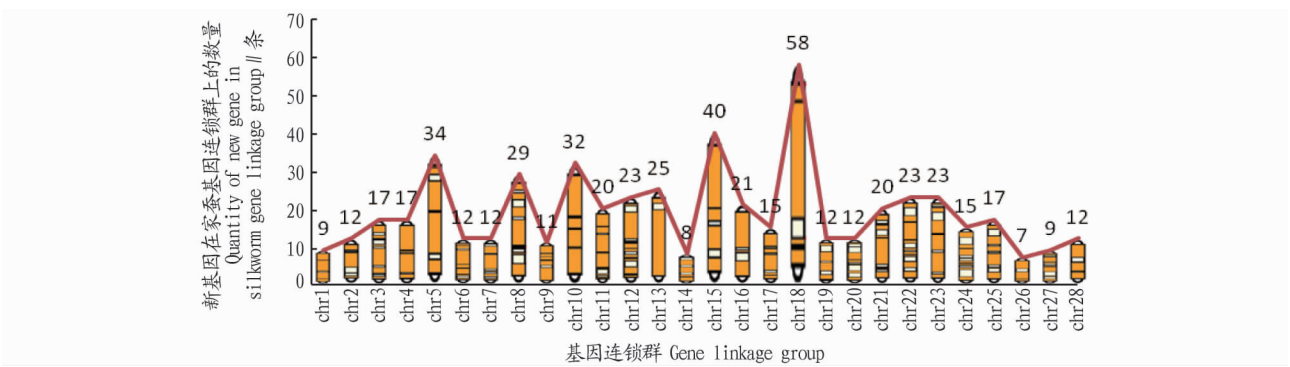


图 4　新基因在家蚕基因连锁群的分布

Fig.4　Distribution of new gene in silkworm gene linkage group

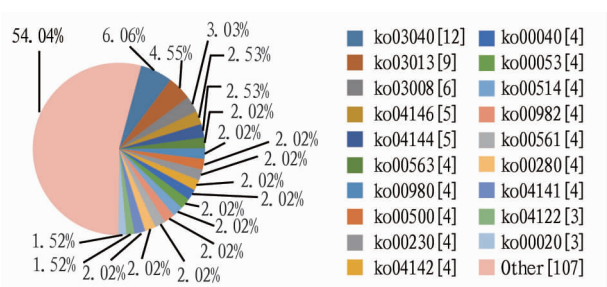


图 5　KEGG 通路富集分布

Fig.5　KEGG pathways distribution

因的注释信息;746 个注释的基因中,其中 611 条(81.90%)基因与家蚕序列同源;利用 Blast2Go 软件对筛选到的基因进行 GO 富集分析,结果显示,基因主要注释到细胞组分、分子功能和生物学过程 3 个分支中,分别有 947、551 和 1 777 个,注释到细胞部分(cell part)、黏合(binding)及代谢过程(metabolic process)的基因数目最多;将新基因的 Locus 与家蚕基因连锁群进行比对发现,有 545 个新基因分布在不同的染色体上,且在 28 条染色体上都有分布;通过 KEGG 分析对新基因进行通路富集分析发现,共有 198 个新基因注释到 KEGG

数据库中,分布于 85 条已知的通路中;将新基因与 COG 数据进行比对,并进行功能注释与分类,结果发现共有 22 个类别里的 258 个新基因得到了注释。新基因与各个数据的序列比对结果,进一步证实了新基因的存在,该研究对新基因的功能做了初步分析,关于新基因具体的功能还需要进一步研究。

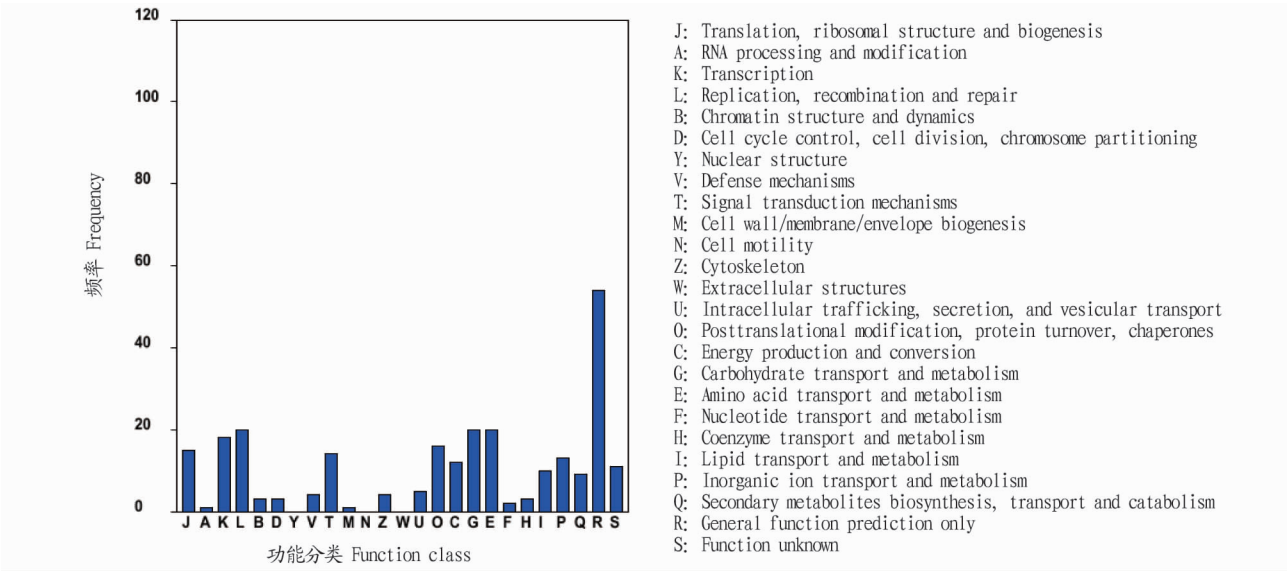


图 6 COG 数据库中功能注释的 unigenes 分类

Fig.6 COG function classification of unigenes

参考文献

[1] 周华,张新,刘腾云,等.高通量转录组测序的数据分析与基因发掘[J].江西科学,2012,30(5):607-611.

[2] WANG Z,GERSTEIN M,SNYDER M.RNA-Seq:A revolutionary tool for transcriptomics[J].Nature reviews genetics,2009,10(1):57-63.

[3] 李智奕,宁维,陈利平,等.新一代测序技术及其在植物转录组研究中的应用[J].河南农业科学,2013,42(12):1-5.

[4] COSTA V,ANGELINI C,DE FEIS I,et a1.Uncovering the complexity of transcriptomes with RNA-Seq[J].Joumal of biomedicine and bioteehnology,2010(5757):1-19.

[5] 韩昆鹏,段炼,李婷婷,等.京海黄鸡卵巢转录组研究:基因结构分析与新基因发掘注释[J].中国畜牧兽医,2016,43(4):854-861.

[6] LI G,QIAN H Y,LUO X F,et al.Transcriptomic analysis of resistant and susceptible *Bombyx mori* strains following BmNPV infection provides insights into the antiviral mechanisms[J].International journal of genomics,2016(2):1-10.

[7] 邓泱泱,荔建琦,吴松锋,等.nr 数据库分析及其本地化[J].计算机工程,2006,32(5):71-73,76.

[8] APWEILER R,BAIROCH A,WU C H,et al.UniProt:The Universal Protein knowledgebase[J].Nucleic acids research,2004,32:115-119.

[9] ASHBURNER M,BALL C A,BLAKE J A,et al.Gene ontology:Tool for the unification of biology[J].Nature genetics,2000,25(1):25-29.

[10] TATUSOV R L,GALPERIN M Y,NATALE D A,et al.The COG database:A tool for genome-scale analysis of protein functions and evolution[J].Nucleic acids research,2000, 28(1):33-36.

[11] KANEHISA M,GOTO S,KAWASHIMA S,et al.The KEGG resource for deciphering the genome[J].Nucleic acids research,2004,32:277-280.

[12] ALTSCHUL S F,MADDEN T L,SCHÄFFER A A,et a1.Gapped BLAST and PSI-BLAST:A new generation of protein database search programs[J].Nuclei acids research,1997,25(17):3389-3402.

(上接第 54 页)

环境质量不是固定不变而是一个动态变化的过程,在公园内应该建立年度宏观监测,同时提高环境监测能力,这就需要相关部门加大财政资金、技术和设备的投入力度[6]。

3.2.3 加强治理。森林公园开发旅游后,旅游者的快速增加也带来了环境污染[6],建议当地区委、区政府加强协调沟通,进一步强化对园区的监督管理,控制园区规模和企业污染物的排放。

参考文献

[1] 化国强,肖靖,黄晓军,等.基于全极化 SAR 数据的玉米后向散射特征分析[J].江苏农业科学,2011,39(3):562-565.

[2] 翁俊.洪泽湖古堰森林公园的植物种类及应用[J].黑龙江农业科学,2018(1):96-99.

[3] 李友元.以生态学理论为指导建设和管理好森林公园[J].湖南林业科技,1993(3):34-38.

[4] 吴楚材,黄艺,刘云国,等.张家界国家森林公园环境质量评价[J].中国园林,1994(3):32-38.

[5] 林轶.中国少数民族地区家庭旅馆的发展研究:以龙胜平安寨家庭旅馆发展为例[D].南宁:广西大学,2004.

[6] 郭盛才,彭威雄,边俊景.大岭山森林公园环境质量监测与评价[J].林业调查规划,2011,36(4):115-118.

[7] 夏赞才.张家界现代旅游发展史研究[D].长沙:湖南师范大学,2004.

[8] 胡海辉.可持续发展的庐山风景区旅游规划方法与实践研究[D].哈尔滨:东北林业大学,2007.

[9] 刘仁芳.自然保护区生态旅游规划研究及其在六峰湖规划中的应用[D].哈尔滨:东北农业大学,2004.

[10] 贾志军.环境监测管理信息系统的设计与实现[D].成都:电子科技大学,2012.

[11] 丁志.毓秀山国家森林公园旅游资源质量评价与开发对策研究[D].南昌:江西农业大学,2012.

[12] 宗妍.森林公园景观规划设计:以山东泗水泉林国家森林公园为例[D].北京:中国林业科学研究院,2015.

[13] 王崑.东北东部林区生态旅游的研究[D].哈尔滨:东北林业大学,2004.

[14] 项小清.水质监测的监测对象及技术方法综述[J].低碳世界,2013(6):70-71.

[15] 杨勇.包头市土默特右旗敕勒川湿地公园规划建设研究[D].杨凌:西北农林科技大学,2009.

[16] 范欣芳.浚县某镇污水处理厂工程初步设计[D].郑州:郑州大学,2016.

[17] 刘慧慧.千家坪国家森林公园森林旅游产品开发设计[D].西安:陕西师范大学,2010.

[18] 强晓鸣.陕西牛背梁国家级自然保护区生态旅游资源与评价[D].杨凌:西北农林科技大学,2006.

[19] 杨尚英.秦岭北麓森林公园空气负离子资源的开发利用探讨[J].生态经济,2003(10):138-139.