西北农业学报 2014,23(12):17-24
Acta Agriculturae Boreali-occidentalis Sinica doi: 10.7606/j.issn.1004-1389.2014.12.004

网络出版日期:2014-12-31
网络出版地址:http://www.cnki.net/kcms/detail/61.1220.S.20141231.1040.026.html

基因本体论在福氏志贺菌转录组研究中的应用

朱 阵,刘翠翠,王 婧,张继瑜,魏小娟,周绪正,李 冰,郭 肖
(中国农业科学院 兰州畜牧与兽药研究所,农业部兽用药物创制重点实验室,兰州 730050)

摘 要 利用RNA-seq技术进行环丙沙星敏感和耐药的福氏志贺菌2457T株的序列测定与组装,对其碱基序列借助基因本体(Gene Ontology,GO)在生物进程、细胞组分、分子功能三方面进行注释和生物学分析。研究经过耐药诱导的转录组差异表达情况,并探讨差异表达基因在耐药性产生过程中的作用及相互关系。结果表明,通过耐药诱导有3 641个基因差异表达,其中961个GO条目注释到39个功能类别。同时发现硫酸盐、硝酸盐氧化还原相关酶系,电子/H^+传导基因,膜转运蛋白,ABC外排泵家族4类耐药相关的功能基因组分。

关键词 RNA-seq;福氏志贺菌;转录组;基因本论
中图分类号 Q78 **文献标志码** A **文章编号** 1004-1389(2014)12-0017-08

Application of Gene Ontology Annotations in *Shigella flexneri* Transcriptome Study

ZHU Zhen, LIU Cuicui, WANG Jing, ZHANG Jiyu,
WEI Xiaojuan, ZHOU Xuzheng, LI Bing and GUO Xiao
(Key Laboratory for Veterinary Drug Innovation, Ministry of Agriculture, Lanzhou Institute of Husbandry and Pharmaceutical Science, Chinese Academy of Agriculture Sciences, Lanzhou 730050, China)

Abstract The sequence and assemble the transcriptome of ciprofloxacin-sensitive and -resistant *Shigella flexneri* 2457T strain was conducted by RNA-seq and Gene Ontology was used to annotate and analyze the base sequence in biological processes, cellular components, molecular functions. After the study of drug-induced transcriptome expression differences, and the role of genes differentially expressed in the process of drug resistance and their relationship was explored. Results showed that there were 3 641 genes differentially expressed, of which 961 GO annotation entries to 39 functional categories. At the same time, drug-related functional gene components were found, including sulfates and nitrates redox-related enzymes, electronic / H^+ spread genes, membrane transport proteins and ABC efflux pump family.

Key words RNA-seq; *Shigella flexneri* ; Transcriptome; GO annotations

志贺氏菌病(Shigellosis)又称细菌性痢疾,是一种具有高度传染性和严重危害性的急性肠道传染病,严重危害人类的健康和生命安全,目前尚无有效的预防措施,通常主张抗菌治疗。世界卫生组织(WHO)推荐菌痢的首选治疗药物为环丙沙星[1]。然而伴随着环丙沙星的临床广泛应用,耐药菌株迅速出现。并呈现产生周期短、耐药率高及多药耐药等特点,给临床治疗带来极大困难。本研究将利用梯度剂量环丙沙星对福氏志贺菌2a 2457T菌株诱导完成耐药菌株的构建。然后

收稿日期:2014-04-15 **修回日期**:2014-07-15
基金项目:国家自然科学基金(31272603、31101836)。
第一作者:朱 阵,男,在读硕士,从事兽医药理与毒理学的研究。E-mail:zhuzhen234@yeah.net
通信作者:张继瑜,男,博士,研究员,主要从事新兽药的研究与开发工作。E-mail:infzjy@sina.com

提取标准 2457T 菌株与耐药菌株的总 RNA，利用 RNA-seq 技术对转录组进行测序，并对转录组序列进行 GO 注释。

基因本体论(Gene Ontology，GO)起源于本体论，是关于基因和蛋白质知识的标准词汇，近年来，在生物信息科学领域得到广泛应用[2]，同时其功能和重要性也被广泛认可[3]。它是由基因本体联合会(Gene Ontology Consortium)所建立的适用于各物种基因的注释，对基因和蛋白质功能划分和描述的数据库[4]，为相关基因数据的统一、数据转换和挖掘奠定基础。由于多物种基因先后解码，并且大量的 ESTs(Expressed sequence tags)和基因表达谱信息的日益积累，增加了基因注释的复杂程度，同时一些物种中的基因或者蛋白质的生物信息可以应用于其他物种，但是这些繁琐的功能信息需要在文献积累中寻找，然而对于某个功能不同文献可能有不同的描绘，这就造成基因功能检索、注释同一性的障碍。GO 注释则基于建立的特定词汇集合来描绘生物学功能进而实现基因注释的同一性[5]。

GO 提供了一系列的语义(Terms)用来描述基因、基因产物的特性，且不具有物种特异性。目前，GO 数据库中已建立起 3 大独立的本体论词汇体系：细胞组分(Cellular component)用于描述亚细胞结构、位置和大分子复合物，如核仁、端粒和识别起始的复合物；分子功能(Molecular function)用于描述基因、基因产物个体的功能，如与碳水化合物结合或 ATP 水解酶活性等；生物进程(Biological process)用于描述分子功能的有序组合，达成更为广泛的生物功能，如有丝分裂或嘌呤代谢等，在这 3 个体系下又独立出不同的亚层次，层层向下将本体论词条串联构成树型结构，可在不同层次注释基因。因此基因本体论形成具有三级结构的标准语言[4]。

1 材料与方法

1.1 材 料

1.1.1 菌 株 福氏志贺菌 2a 2457T 株，由军事医学科学院生物工程研究所惠赠；标准质控菌株 *E. coli* ATCC25922，购自国家菌种保藏中心。

1.1.2 试 剂 环丙沙星，购自中国药品生物制品鉴定所；LB 培养基，麦康凯培养基，购自广东环凯微生物科技有限公司；SV total RNA Isolation System，购自 Promega 公司；2457T 菌株鉴定引物 R002，由上海生工生物工程有限公司合成；*Taq* 酶、dNTP、*Taq* Buffer、DL2000 DNA Maker、Loading Buffer、琼脂糖均购自 TaKaRa；其他试剂为实验室常规试剂。

1.2 方 法

1.2.1 耐药菌株构建 根据美国临床和实验室标准协会(CLSI)的规定使用微量肉汤稀释法测定环丙沙星对福氏志贺菌 2457T 标准菌株最低抑菌质量浓度(MIC)，标准质控菌为大肠杆菌 ATCC29522 株。然后采用梯度剂量药物诱导细菌耐药性，即依次采用 1/4 MIC、1/2 MIC、MIC 等梯度药物剂量的环丙沙星对福氏志贺菌进行诱变，每诱导一代，则复壮一代。每一代诱导菌用 R002 引物进行 PCR 验证，最后测定环丙沙星对诱导菌的 MIC，如果 MIC≥4 μg/mL 则可以停止诱导。

1.2.2 total RNA 提取 采用 Promega 公司的 SV total RNA Isolation System 提取福氏志贺菌标准菌株(Y1)与耐药菌株(N1)总 RNA，取培养 4 h 的菌液 1 mL 置于 1.5 mL 的离心管中，14 000 r/min 离心 5 min。去除上清液，尽量吸干净。加入 100 mL 的溶菌酶和 TE 混合物，涡旋 30 s 混合。室温孵育 3～5 min。加入 75 μL RNA Lysis Buffer、350 μL RNA Dilution Buffer，倒转混合，勿离心。转移上清至新的离心管，尽量不要碎片。加 200 μL φ= 95%乙醇，吹打 3～4 次混合，转移至离心柱和收集管中，12 000～14 000 r/min 离心 1 min。弃液体，加入 600 μL RNA Wash Solution，12 000～14 000 r/min 离心 1 min。弃液体，轻轻吹打混匀，不可涡旋。新鲜配制以下液体：40 μL Yellow Core Buffer、5 μL 0.09 mol/L $MnCl_2$、5 μL DNase Ⅰ Enzyme，将此 50 μL 混合液加于膜上。20～25 ℃孵育 15 min，然后加入 200 μL DNase Stop Solution，12 000～14 000 r/min 离心 1 min。加入 600 μL RNA Wash Solution，12 000～14 000 r/min 离心 1 min。弃液体，加入 250 μL RNA Wash Solution，高速离心 2 min。将离心柱转移于 1.5 mL 离心管，加 100 μL 无 RNA 酶水于膜中央，12 000～14 000 r/min 离心 1 min，所得即为 RNA 溶液，置于－80 ℃ 冻存，并将提取的 RNA 样品运用紫外分光光度计检测其在 260 nm、280 nm 处的吸收值，估计 RNA 的纯度和质量浓度，同时用10 g/L 的琼脂糖凝胶电泳检测所提 RNA

的质量。

1.2.3　转录组测序　用 Dnase 酶去除 RNA 样品中的 DNA;然后使用试剂盒去除总 RNA 中的 rRNA,加入 fragmentation buffer 将 mRNA 随机打成短片段,以 mRNA 为模板用六碱基随机引物合成 cDNA 第 1 链,加入缓冲液、dNTPs、RNase H 和 DNA 聚合酶Ⅰ合成 cDNA 第 2 链,在经过 QiaQuick PCR 试剂盒纯化并加 EB 缓冲液洗脱之后做末端修复、加 polyA 并连接测序接头,然后用琼脂糖凝胶电泳进行片段大小选择,最后进行 PCR 扩增,采用 Illumina HiSeq™2000 测序平台对福氏志贺菌 2 个样品的转录组进行测序。

1.2.4　GO 注释　基因注释主要基于氨基酸序列比对,将基因的氨基酸序列比对到各数据库中,从而获得相应的功能注释信息。由于一个基因的编码蛋白可以在多水平进行定义,即比对结果不具有唯一性,GO 注释的原理就是利用计算机程序建立基因产物和本体论词条之间的关联,最终保留最佳的比对结果作为注释。所有的注释均使用 BLAST 软件结合各个数据库特点完成[4]。

GO 功能分析的过程包含 2 个方面:①差异表达基因的 GO 功能分类注释,给出具有某个 GO 功能的基因列表及基因数目统计;②差异表达基因的 GO 功能显著性富集分析,给出与基因组背景相比,在差异表达基因中显著富集的 GO 功能条目,从而给出差异表达基因与哪些生物学功能显著相关。该分析首先把所有差异表达基因向 Gene Ontology 数据库(http://www.geneontology.org/)的各个 Term 映射,计算每个 Term 的基因数目,然后应用超几何检验,找出与整个基因组背景相比,在差异表达基因中显著富集的 GO 条目,其计算公式为:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

其中,N 为所有基因中具有 GO 注释的基因数目;n 为 N 中差异表达基因的数目;M 为所有基因中注释为某特定 GO Term 的基因数目;m 为注释为某特定 GO Term 的差异表达基因数目。计算得到的 P-value 通过 Bonferroni 校正之后,以 corrected P-value≤0.05 为阈值,满足此条件的 GO Term 定义为在差异表达基因中显著富集的 GO Term。通过 GO 功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

本研究将根据以上原则分别针对细胞组分、分子功能、生物进程 3 大本体论体系进行注释。

2　结果与分析

2.1　转录组组装结果分析

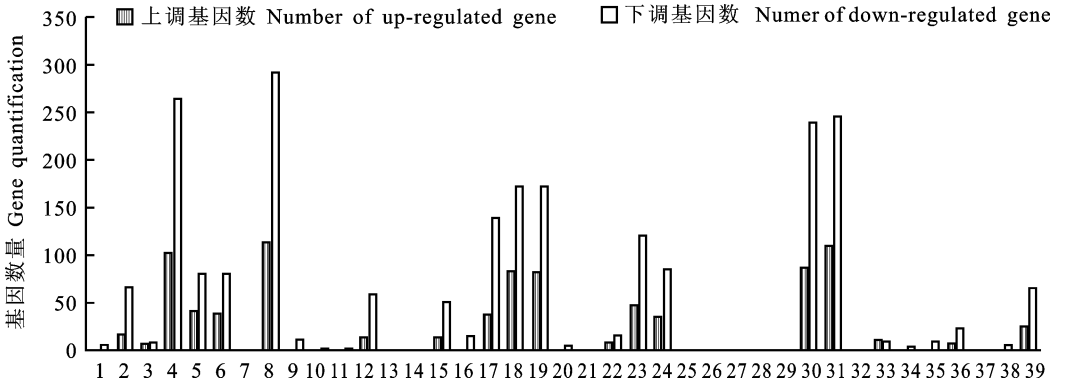
测序得到的原始图像数据中,某些原始序列带有 Adaptor 序列,或含有少量低质量序列。因此要经过一系列数据处理以去除杂质数据,大致过程包括:去除含 Adaptor 的 Reads、去除 N 的比例大于 10%的 Reads、去除低质量 Reads,最后获得 Clean reads。然后使用短 Reads 比对软件 SOAPaligner/soap 2 将 2 个样品的 Clean reads 比对到福氏志贺菌 2a 2457T 参考基因组,结果显示 Y1 和 N1 中分别有 25 529 196 和 23 936 216 个 Unique match reads 匹配到参考基因组,共覆盖 3 641 个福氏志贺菌基因,有超过 98%的基因覆盖度大于 50%,这表明 RNA-seq 测序技术具有较高的灵敏度,可以覆盖绝大多数福氏志贺菌基因组,同时保证转录组注释的准确性。

2.2　GO 注释结果

依据基因编码的产物蛋白的性质将其归属到其中一类或者多类中。通过 GO 数据库注释,依据基因在不同本体论词汇体系中注释情况判断可能的功能。GO 注释 3 大分支统计结果如图 1。图 1 显示对差异表达基因进行 GO 注释统计的情况,在鉴定的 3 641 个福氏志贺菌基因中,有 961 个 GO 条目注释到 39 个功能类别,其中最主要的是细胞进程、代谢过程、生物调节、细胞及细胞组分、膜与膜组分、催化及转运活性等方面,表明这些差异基因在 2 个样品的转录库是非常丰富的,而且能够编码多种与结构、监管和代谢相关的蛋白质。

2.2.1　GO 注释　对于样品 Y 和样品 N 的生物进程、细胞组分、分子功能 3 大体系注释结果中差异表达基因中显著富集的 GO term(P≤0.05)结果见表 1。

2.2.2　GO 注释中差异基因显著性富集分析　词条归属分布是以有向非循环图(Directed acyclic graphs)形式层层向下,将 GO 词条分配给基因序列,形成串联的树状结构图。对于生物进程、细胞组分、分子功能注释的树状图如图 2～4 所示。图中不同的颜色表示不同的 P-value 值,当 P≤0.05 时表明差异表达显著。



1. 生物粘附 Biological adhesion;2. 生物调节 Biological regulation;3. 组织或生物起源细胞组件 Cellular component organization or biogenesis;4. 细胞的过程 Cellular process;5. 建立本地化 Establishment of localization;6. 本地化 Localization;7. 运动 Locomotion;8. 代谢过程 Metabolic process;9. 多生物过程 Multi-organism process;10. 负调控的生物过程 Negative regulation of biological process;11. 积极的生物过程的调控 Positive regulation of biological process;12. 生物过程的调控 Regulation of biological process;13. 繁殖 Reproduction;14. 生殖过程 Reproductive process;15. 刺激反应 Response to stimulus;16. 信号 Signaling;17. 单一的生物过程 Single-organism process;18. 细胞 Cell;19. 细胞部分 Cell part;20. 细胞外区域 Extracellular region;21. 细胞外区域部分 Extracellular region part;22. 大分子复合物 Macromolecular complex;23. 膜 Membrane;24. 膜部分 Membrane part;25. 类核 Nucleoid;26. 细胞器 Organelle;27. 病毒粒子 Virion;28. 病毒体部分 Virion part;29. 抗氧化活性 Antioxidant activity;30. 缠绕 Binding;31. 催化活性 Catalytic activity;32. 通道调节活性 Channel regulator activity;33. 电子载体活性 Electron carrier activity;34. 酶调节活性 Enzyme regulator activity;35. 分子传感器活性 Molecular transducer activity;36. 核酸绑定转录因子的活性 Nucleicacid binding transcription factor activity;37. 蛋白结合转录因子的活性 Protein binding transcription factor activity;38. 受体的活性 Receptor activity;39. 运输活性 Transporter activity

图 1 样品 Y 和样品 N 转录组基因功能注释 GO 功能分类图

Fig. 1 GO function classification figure of transcriptome gene function annotation for sample Y and N

表 1 Y1-VS-N1 GO 功能条目统计($P \leqslant 0.05$)

Table 1 Statistics of the Gene Ontology(GO) terms($P \leqslant 0.05$)for Y1-VS-N1

GO 组分 GO component	差异基因数目及比例 Number and ratio of differential genes	注释到的所有基因及比例 Number and ratio of annotated genes	校正值 Adjusted value	GO ID	注释到 GO 上的基因 Number of GO annotated genes
生物进程 Biological process					
硫化氢代谢 Hydrogen sulfates metabolism	6 out of 509 genes, 1.2%	6 out of 3 097 genes, 0.2%	0.010 78	GO:007081	S2966, S2967, S2968, S2971, S2972, S2973
硫化氢生物合成 Hydrogen sulfates biosynthesis	6 out of 509 genes, 1.2%	6 out of 3 097 genes, 0.2%	0.010 78	GO:0070814	S2966, S2967, S2968, S2971, S2972, S2973
细胞组分 Cell components					
细胞周质间隙 Periplasmic space	40 out of 300 genes, 13.3%	121 out of 1 810 gene, 6.7%	0.000 14	GO:0042597	S0156, S0274, S0291, S0432, S0482, S0512, S1064, S1140, S1196, S1205, S1381, S1600, S1805, S1884, S1917, S2114, S2364, S2417, S2420, S2423, S2454, S2464, S2519, S2520, S2521, S2574, S2626, S3152, S3178, S3269, S3597, S3598, S4045, S4159, S4223, S4224, S4273, S4463, S4469, S4820
分子功能 Molecular function					
氧化还原酶的活性 Oxidoreductase activity	20 out of 527 genes, 3.8%	29 out of 3 127 gene, 0.9%	1.94×10^{-7}	GO:0016661	S0869, S1311, S1312, S1313, S1314, S1471, S1809, S1884, S2003, S2004, S2416, S2419, S2420, S2421, S3597, S3598, S3599, S3600, S4378, S4379
硝酸还原酶活性 Nitrate reductase activity	9 out of 527 genes, 1.7%	12 out of 3 127 gene, 0.4%	0.004 93	GO:0008940	S1311, S1312, S1313, S1314, S1884, S2416, S2419, S2420, S2421
血红素结合 Heme-binding	9 out of 527 genes, 1.7%	13 out of 3 127 gene, 0.4%	0.013 66	GO:0020037	S1066, S2004, S2416, S2771, S2972, S3597, S3598, S3727, S4379
离子跨膜转运蛋白活性 Anion transmembrane transporter activity	33 out of 527 genes, 6.3%	100 out of 3 127 gene, 3.2%	0.016 13	GO:0008509	S0111, S0438, S0803, S0825, S1310, S1745, S1762, S1868, S2455, S2517, S2518, S2519, S2520, S2521, S2623, S2624, S2625, S2626, S2649, S2875, S3336, S3419, S3594, S3630, S3762, S3853, S4018, S4042, S4043, S4044, S4045, S4233, S4559

2.3 注释结果分析

据表 1 所示,GO 注释到的差异显著基因($P \leqslant 0.05$)有 93 个,其中 84 个为明确功能的差异显著基因,9 个为假定蛋白(Hypothetical protein),在这些差异显著基因中有 30 个基因具有多定义性,在 2 个及以上体系中被注释到,说明基因功能并不是静态独立发挥作用的,而是在协同各生命活动共同发挥作用。对其基因比对可以确定差异基因的功能和机理预测。生物进程中注释到的 6 个差异显著基因为 cys 型,可表达硫酸腺苷激酶和亚硫酸盐还原酶相关亚基,参与硫化氢的合成与代谢,并在此过程中通过氧化还原反应

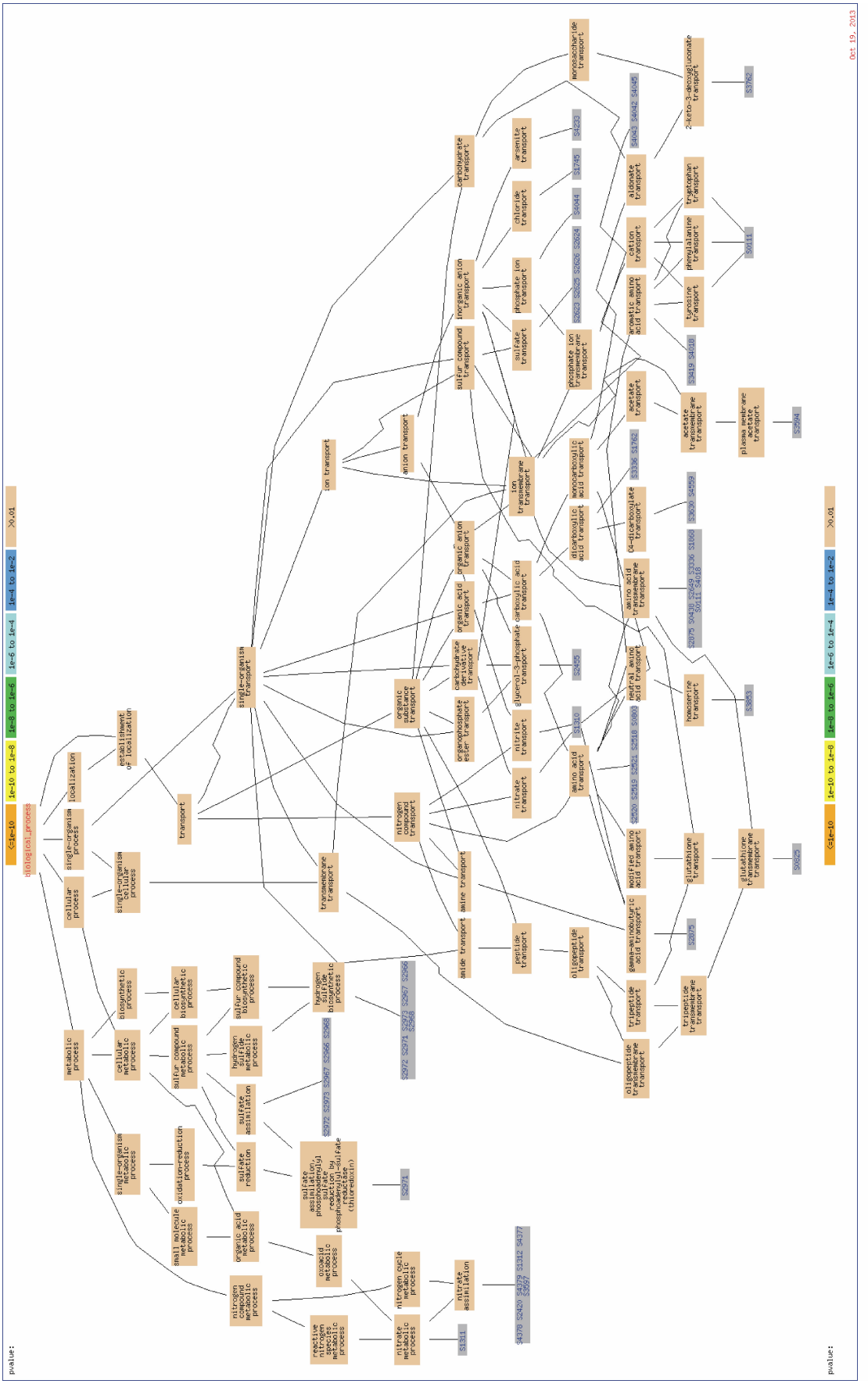


图 2　Y-VS-N生物进程 GO注释树状图

Fig.2　GO annotation tree of biological process for Y-VS-N

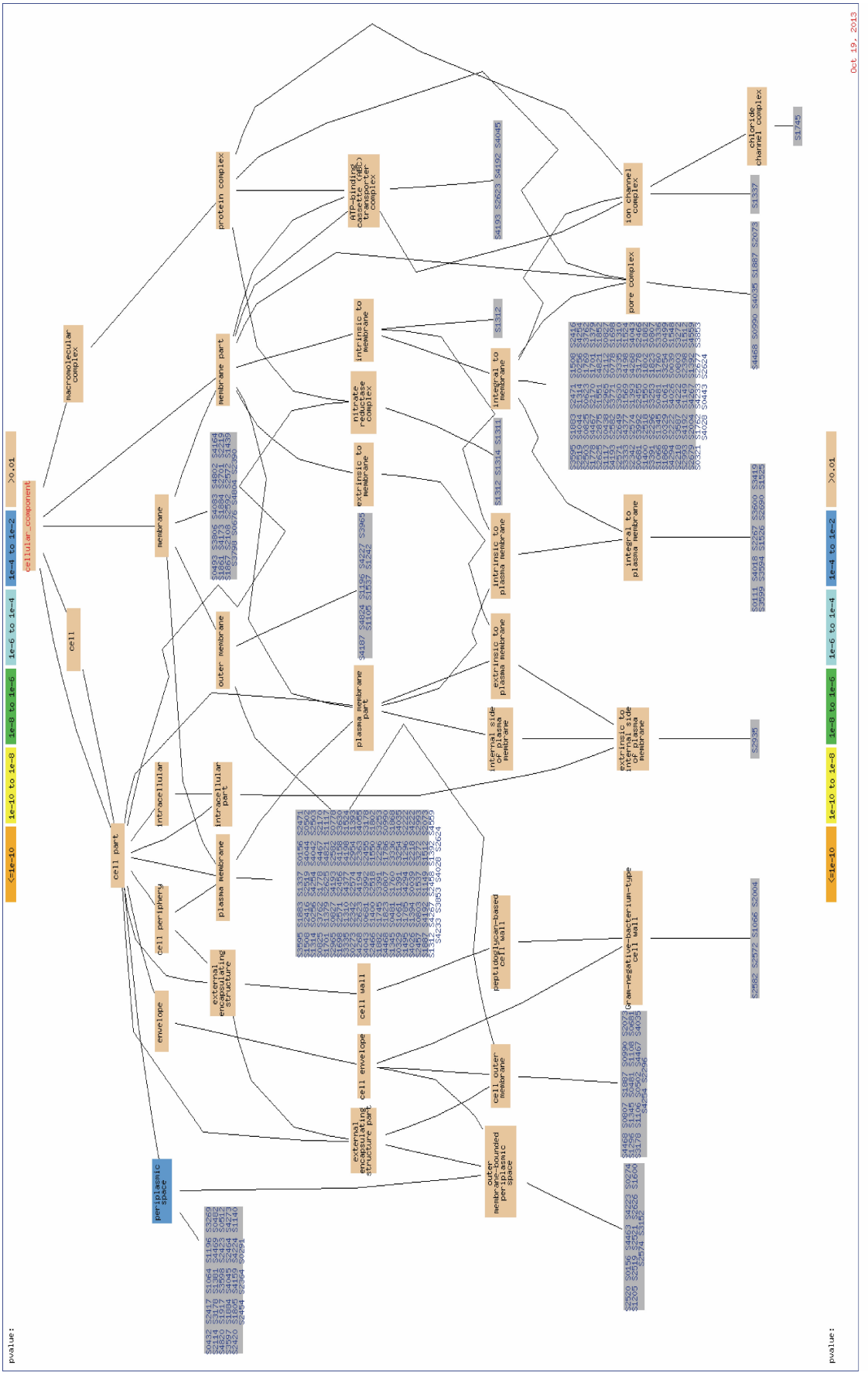


图 3　Y1-VS-N1细胞组分 GO注释树状图

Fig.3　Molecular function GO annotation tree for Y1-VS-N1

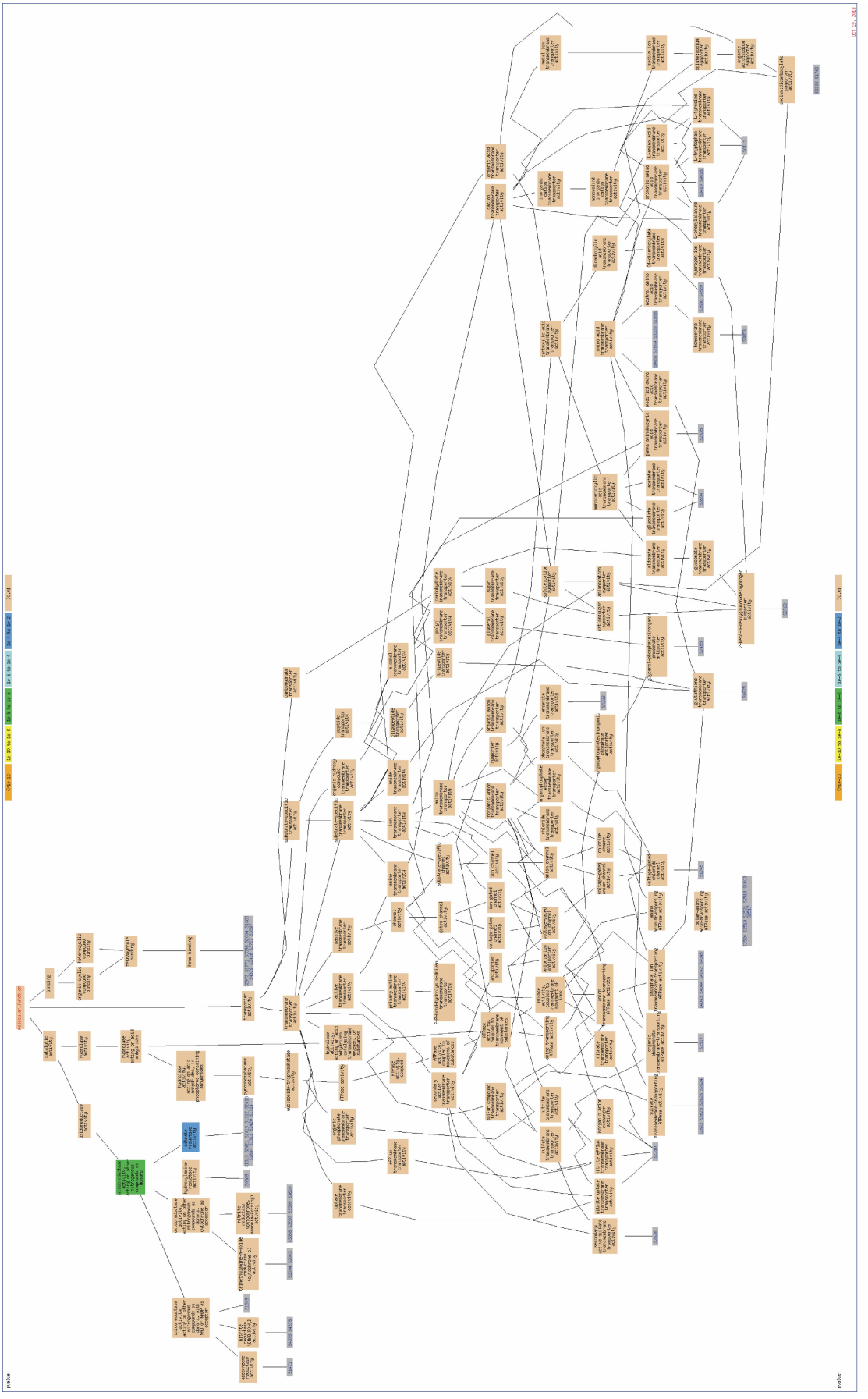


图 4　Y1-VS-N1分子功能 GO注释树状图

Fig.4　Cellular component GO annotation tree for Y1-VS-N1

生成 ATP 和 GTP，提供能量；细胞组分中显著差异基因表现在细胞周质间隙或质膜间隙中以 torT、sodC、nrfB 为主的氧化还原酶，以 potD 和 pstS 为代表的 ABC(ATP binding cassette)外排泵体系及转运蛋白 mglB 等，同时还发现周质保护基因 *fimC* 和引起应激反应基因 *spy*，在周质体系中这些基因通过氧化还原反应为 ABC 外排系统提供能量；分子功能分支显著差异主要表现在氧化还原酶活性、硝酸还原酶活性、血红色结合体系、离子跨膜运输 4 方面，前 3 者为各种氨基酸代谢过程中的酶类(nar、nir、nrf 等)和 H^+/电子传导体系基因，离子跨膜转运主要是氨基酸转运蛋白(his、aroP 等)，硫酸盐转运(cys)，通透酶(hisQ、tnaB 等)和 ABC 转运体系(pst、glnP)，起到氨基酸代谢作用。

3 讨 论

伴随着生物信息学和二代测序的逐步兴起，测序数据的分析与简化程序也被人们广泛关注。语义相似度已成为生物信息学重要的手段，因为它综合实验和序列信息，从而提供定量关联基因的方法[6]。GO 注释作为最主要的基因注释手段之一，已经在基因分类中成为一个标准。GO 大型数据库集可通过高水平本体论词汇将基因产物按功能进行分类[7]，可以揭示某些特性群体有着相似表达模式的原因[8]，从而在海量遗传信息中找到富集特定性能的基因功能类[9]。同时还可对候选基因进行分类和筛选。

试验中经 GO 注释到的差异显著基因可归纳为 4 大类：硫酸盐、硝酸盐氧化还原相关酶系，电子/H^+传到基因，膜转运蛋白，ABC 外排泵家族。ABC(ATP binding cassette)[10]家族外排泵依赖 ATP 为能量源发挥多耐药功能。然而，革兰氏阴性菌中主要的 RND[11]外排泵家族却未见显著差异。依据 GO 注释的差异显著基因可以判定环丙沙星对福氏志贺菌 2a 2457T 株的诱导过程中膜蛋白发生改变，使得进入菌体内抗生素通量减少，而营养物质增多，以便自身代谢；对于进入菌体内的抗生素可由 ABC 外排泵排出体外，所需能量由电子/H^+转运过程和氧化还原反应提供，同时应激反应基因和保护物质表达基因的表达也可对菌体起到保护作用。在耐药菌株中众多机制协同作用造成了耐药性的产生。

目前，GO 数据库并未完善[12]，并且 GO 发展具有一定的片面性，各种疾病相关基因被广泛关注，造成一些基因未能得到注释分类，需要多个数据库多重注释，来完善数据的注释分析。

Reference（参考文献）：

[1] Traa B S, Walker C L, Munos M, *et al*. Antibiotics for the treatment of dysentery in children[J]. Int J Epidemiol, 2010, 39(1): 70-74.

[2] Kagaya Y, HoboT, Murata M, *et al*. Abscisic acid-induced transcription is mediated by phosphorylation of an abscisic acid response element binding factor, TRAB1[J]. Plant Cell, 2002, 14(12): 3177-3189.

[3] Berners L T, Hendler I, Lassila O. The semantic Web[J]. Scientific American, 2001, 284(5): 34- 43.

[4] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium[J]. Nat Genet, 2000, 25(1): 25-29.

[5] Harris M A, Clark I, lrcland A, *et al*. The Gene Ontology (GO) database and informatics resources[J]. Nucleic Acid Rcs, 2004, 32: 258-261.

[6] Caniza H, Romero A E, Heron S, *et al*. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology[J]. Bioinformatics, 2014, 22: 1-2.

[7] Wang X, Gorlitsky R, Almeida J S. From XML to RDF: how semantic web technologies will change the design of 'omic' standards[J]. Nat Biotechnol, 2005, 23(9): 1099-1103.

[8] Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies[J]. Genome Biol, 2005, 6: R46.

[9] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems[J]. Bioinformatics, 2005, 21(18): 3587-3595.

[10] Van Veen H W, Konings W N. The ABC family of multidrug transporters in microorganisms[J]. Biochim Biophys Acta, 1998, 1365(1-2): 31-36.

[11] Nikaido H. Antibiotic resistance caused by gram-negtive multidrug efflux pumps[J]. Cli Effect Di, 1998, 27(1): 532-541.

[12] Rhee S Y, Wood V, Dolinski K, *et al*. Use and misuse of the gene ontology annotations[J]. Nat Rev Genet, 2008, 9(7): 509-515.