Clin. Lab. 2017;63:XXX-XXX

ORIGINAL ARTICLE

High-Throughput Screening of *Escherichia coli* O157:H7 Pathogenic Genes via Pathway Enrichment and Operon Analysis

Guangxin Yuan, Jia Fu, Liping An, Yan Zhuang, Yu Wang, Yunfeng Song, Zijing Qiao, Xiaolong Wang, Yufei Tian, Xiao Han, Peige Du, Guangyu Xu

College of Pharmacy, Beihua University, Jilin, 132013, China

SUMMARY

Background: **About thirty thousand people globally die every day from infectious diarrhea, mostly caused by pathogenic *Escherichia coli* (*E. coli*) O157:H7.**
Methods: **In order to search for clinical diagnostic biomarkers and novel drug targets for infectious diarrhea, we used a bibliometric method to collect pathogenic genes of *E. coli* O157:H7 and performed a functional analysis of the important pathogenic genes by pathway enrichment and operon analysis.**
Results: **We found 364 pathogenic genes which may be involved in infection with *E. coli* O157:H7 including 50 new specific pathogenic genes. It is possible that these newly found pathogenic genes will be of great importance in the treatment of *E. coli* O157:H7 infected diseases and the discovery of novel diagnostic biomarkers.**
Conclusions: **Our findings also lay a theoretical foundation for the control, diagnosis, and prognosis of pathogenic *E. coli* related diseases.**
(Clin. Lab. 2017;63:xx-xx. DOI: 10.7754/Clin.Lab.2017.170314)

Correspondence:
Guangyu Xu
College of Pharmacy
Beihua University, Jilin
Jilin, 132013
China
Phone/Fax: +86 43264608281
Email: xuguangyu2005@163.com

Manuscript accepted May 7, 2017

KEY WORDS

E. coli O157:H7, pathogenic genes, bibliometric method, pathway, operon

INTRODUCTION

Pathogenic *Escherichia coli* (*E. coli*) O157:H7 can cause diarrhea and hemorrhagic colitis, which can initiate many serious complications including hemolytic uremic syndrome and thrombotic thrombocytopenic purpura (also called Moschcowitz disease) [1-4]. Due to the trend for outbreaks and high death rate, *E. coli* O157:H7-infected diseases have become a serious global public health issue and have attracted the attention of governments, health institutions, and scientific research organizations [5,6].

Current studies on *E. coli* O157:H7 are mainly focused on screening and functional mining for new pathogenic genes [7-9]. Following the rapid development of the Microbial Genome Project and high-throughput biocomputing technology, data mining of "multi-omics data" provides a more effective tool for mechanistic analy-

sis of pathogens, which may be significant in the discovery of novel drug targets and clinical diagnostic biomarkers.

In this study, we found 364 pathogenic genes of *E. coli* O157:H7 infected diseases using a bibliometric method and 50 novel pathogenic genes through pathway enrichment and operon analysis. Our findings are a source for novel prognostic markers as well as novel therapeutic targets for *E. coli* O157:H7 infected diseases.

MATERIALS AND METHODS

Bibliometric method

The bibliometric method was used as previously described [10]. We used the keywords "*E.coli*", "O157:H7", "gene" and "pathogenic" to search the publications from 2000 to 2015 in PubMed. Using Epidata3.1, we deleted the repeated and unrelated studies by parallel entry and logical error test. A total of 1251 studies were retrieved and 300 studies were used to analyze the essential gene modules.

KEGG analysis, data sources, and pre-treatment

All pathways used in this study were downloaded from KEGG databases [11]. In total, 78 relevant KEGG pathways were downloaded from http://ftp.genome.jp/pub/kegg/ in September of 2015.

Operon analysis, data sources, and pre-treatment

The Operon database was downloaded from http://csbl.bmb.uga.edu/DOOR/index.php. DOOR^2 (Database of prOkaryotic OpeRons, Version 2.0) is an operon database developed by Computational Systems Biology Lab at the University of Georgia [12]. The operons in this database are predicted based on essential genomic features. The Operon database algorithm is a data-mining classifier. Its features include intergenic distance, neighborhood conservation, phylogenetic distance, information from short DNA motifs, similarity score between Gene Ontology (GO) terms of gene pairs, and length-ratio between a pair of genes.

Analysis on the significance in the function of differential genes

Based on the NCBI gene ontology database, we obtained the GO annotation of the involved genes. Fisher exact test and X^2 test were used to calculate the significance level and misjudgment rate of each GO, and the p-values were calibrated with the misjudgment rate to screen out the significance ($p < 0.05$) of differential genes [13]. Data were artificially analyzed with the European Bioinformatics Institute database.

RESULTS

Pathway enrichment analysis of *E. coli* O157:H7 pathogenic genes

Bibliometric analysis yielded 364 reported *E. coli* O157:H7 pathogenic genes, which accounted for 7% of the total genes available (5360). Among these 364 genes, 137 (37.1%) genes were enriched in 78 relevant pathways.

Genes

ECs4774 and *ECs4841* were enriched in more than 10 related pathways. Twelve genes, including *ECs0783*, *ECs4704*, *ECs4773*, *ECs4850*, *ECs0747*, *ECs0748*, *ECs0753*, *ECs0985*, *ECs4792*, *ECs1712*, *ECS4705*, *ECs2846*, *ECs1847*, and *ECs5343* were enriched in 5 to 10 pathways (Table 1) and there were 121 genes enriched in less than 5 pathways.

Operon analysis of *E. coli* O157:H7 pathogenic genes

Operon information was collected for these 364 pathogenic genes using http://csbl.bmb.uga.edu/DOOR/index.php (detailed in S2) and found that 284 genes (73%) were located on *E. coli* O157:H7 pathogenic operon regions. Among these operons, 7 contained more than 4 genes, while operon 19262 (Figure 1) and operon 19719 contained the largest number of genes. There were 277 operons located by no more than 4 genes (Table 2).

GO annotation of *E. coli* O157:H7 pathogenic genes

Clustering analysis was performed with the 364 pathogenic genes using the information provided by the Database for Annotation, Visualization and Integrated Discovery (DAVID) [14], and functions of these genes were found to be closely related to the bacterial secretion system (16 genes), pathogenic *E. coli* infection (7 genes) [15], purine-pyrimidine metabolism (181 genes) [16], amino acids and proteins (34 genes), as well as RNA degradation (42 genes) [17]. We therefore indicated that the functions of *E. coli* O157:H7 pathogenic genes were mostly correlated with the top two annotated functions. Following this we screened out all pathways and operons that related to bacterial secretion system and pathogenic *E. coli* infection (Table 3). All *E. coli* infection related genes were enriched into pathway ecs05130 and 81% of the genes involved in the bacterial secretion system were enriched into pathway ecs03070. Sixteen bacterial secretion-related genes were distributed into 9 operons, and 7 *E. coli* infection-related genes were distributed into 6 operons.

DISCUSSION

Generally speaking, the genes that cause the disease are harmful and are called "pathogenic genes". At present, more and more studies are aimed at these "pathogenic genes" for designing drugs or as a drug target. So it is of

Table 1. *E. coli* pathogenic genes and enriched pathways.

Genes	Gene ID	Pathway	No.
ECs4774	915128	ecs00071/ecs00280/ecs00281/ecs00310/ecs00362/ecs00380/ecs00410/ecs00640/ecs00650/ecs 00903/ecs00930/ecs01040/ecs01100/ecs01110/ecs01120/ecs01130/ecs01200/ecs01212	18
ECs4841	915060	ecs00010/ecs00030/ecs00051/ecs00052/ecs00680/ecs01100/ecs01110/ecs01120/ecs01130 ecs01200/ecs01230/ecs03018	12
ECs0783	917517	ecs00010/ecs00260/ecs00680/ecs01100/ecs01110/ecs01120/ecs01130/ecs01200/ecs01230	9
ECs4704	915283	ecs00270/ecs00280/ecs00290/ecs00770/ecs01100/ecs01110 ecs01130/ecs01210/ecs01230	9
ECs4773	915130	ecs00071/ecs00280/ecs00281/ecs00362/ecs00592/ecs01110/ecs01120/ecs01130/ecs01212	9
ECs4850	915049	ecs00010/ecs00030/ecs00051/ecs00680/ecs01100/ecs01110 ecs01120/ecs01130/ecs01200	9
ECs0747	917478	ecs00020/ecs00190/ecs00650/ecs01100ecs01110/ecs01120 ecs01130 /ecs01200	8
ECs0748	917480	ecs00020/ecs00190/ecs00650/ecs01100/ecs01110/ecs01120/ecs01130/ecs01200	8
ECs0753	917490	ecs00020/ecs00640/ecs00660/ecs01100/ecs01110/ecs01120/ecs01130/ecs01200	8
ECs0985	917728	ecs00260/ecs00680/ecs00750/ecs01100/ecs01120/ecs01130 ecs01200 /ecs01230	8
ECs4792	915103	ecs00220/ecs00250/ecs00630/ecs00910/ecs01100/ecs01120/ecs01230/ecs02020	8

Table 2. *E. coli* pathogenic genes and located operons.

Operon ID	Genes	No.
19262	ECs1321/ECs1322/ECs1323/ECs1324/ECs1325/ECs1326/ECs1327	7
19719	ECs3713/ECs3714/ECs3715/ECs3716/ECs3717/ECs3718/ECs3719	7
19266	ECs1352/ECs1353/ECs1354/ECs1355/ECs1356/ECs1357	6
19720	ECs3724/ECs3725/ECs3726/ECs3727/ECs3728/ECs3729	6
19091	ECs0456/ECs0457/ECs0458/ECs0459/ECs0460	5
19721	ECs3730/ECs3731/ECs3732/ECs3733/ECs3734	5
19884	ECs4552/ECs4553/ECs4554/ECs4556	4
19046	ECs0237/ECs0238/ECs0239	3
19062	ECs0320/ECs0321/ECs0323	3
19325	ECs1668/ECs1669/ECs1670	3
19330	ECs1697/ECs1698/ECs1699	3
19331	ECs1703/ECs1704/ECs1705	3
19333	ECs1715/ECs1716/ECs1717	3
19911	ECs4703/ECs4704/ECs4705	3

Table 3. Bacterial secretion system and pathogenic *E. coli* infection-related genes, pathways and operons.

	Bacterial secretion system	Pathogenic *E. coli* infection
Genes	ECs3725/ECs3726/ECs3733/ECs3721/ECs4552/ECs3730/ECs3732/ECs3719/ ECs0460/ECs0459/ECs4569/ECs3731/ECs3716/ECs0458/ECs4640/ECs4904	ECs1205/ECs2715/ECs2973/ECs2974/ ECs4559/ECs4561/ECs1206
Pathways	ecs03070/ ecs03060	ecs05130
Operons	19720/19721/1412787/19884/19719/19091/19885/1413065/19952	1412060/1412534/19566/1413035/ 1413037/1412061

Table 4. Pathogenic genes among different pathways.

Pathway	Genes	No.
ecs01100	ECs0566/ECs0035/ECs0668/ECs0708/ECs0746/ECs0993/ECs1712/ECs2446/ECs3045/ ECs3118/ECs3640/ECs4275/ECs4841/ECs0120/ECs0477/ECs0527/ECs0566/ECs3029/ECs0747/ ECs0748/ECs0753/ECs0775ECs0783/ECs0985/ECs1704/ECs1713/ECs1715/ECs1812/ECs1847/ ECs2666/ECs2699/ECs2846/ECs3600/ECs3684/ECs3739/ECs3740/ECs3929/ECs3942/ ECs4498/ECs4672/ECs4704/ECS4705/ECs4735/ECs4792/ECs4850/ECs4910/ECS5022/ ECs5222/ECs5343/ECs4774	49
ecs01120	ECs0747/ECs0748/ECs0753/ECs0783/ECs0985/ECs1712/ECs3029/ECS3660/ECs3739/ECs3740/ECs4773/ ECs4792/ECs4841/ECs4850/ECs5037/ECs5052/ECs4774	17
ecs01130	ECs0527/ECs0747/ECs0748/ECs0753/ECs0783/ECs0985/ECs1712/ECs1713/ECs2846/ECs4672/ECs4704/ ECS4705/ECs4739/ECs4773/ECs4841/ECs4850/ECs4774	17
ecs03070	ECs3716/ECs3718/ECs3721/ECs3724/ECs3725/ECs3726ECs3730/ECs3731/ECs3732/ECs3733/ECs4552/ ECs4569/ECs4640/ECs4774/ECs0458/ECs0459/ECs0477/ECs3347	15
ecs01110	ECs1712/ECs1713/ECs1715/ECs1847/ECs3929/ECs4275/ECs4704/ECs4773/ECs4850/ECs4903/ECs5343/ ECs4774	12
ecs00230	ECs0527/ECs0566/ECs1712/ECs1847/ECs3118/ECs3739/ECs3740/ECs3922/ECs4736/ECs4910/ECs5343	11

Table 5. Virulence and unknown genes among pathways.

Pathway	Virulence genes	Unknown genes
ecs03070	ECs3716/ECs3718/ECs3721/ECs3724/ECs3725/ ECs3726/ECs3730/ECs3731/ ECs3732/ECs3733/ECs4552/ECs4569/ ECs4640 /ECs0458/ECs0459	ECs3923/ECs4575/ECs4573/ECs4583/ECs4582/ ECs4581/ECs4580/ECs4568/ECs4565/ECs0460/ ECs4904/ECs4054/ECs4165/ECs0102/ECs0101/ ECs4313/ECs4487/ECs3473/ECs4766/ECs4767/ ECs4768/ECs0665/ECs0236/ECs0607/ECs2060/ ECs0234/ECs0226/ECs0218/ECs0224/ECs0223
ecs05130	ECs1205/ECs1206/ECs2973/ECs2974/ECs0848/ECs2715/ ECs4559/ECs4561	ECs4590/ECs4550/ECs4562/ECs4564/ECs1812/ ECs1814

Table 6. Statistics of proven and unproven pathogenic genes among the operons containing at least 3 pathogenic genes.

Operon ID	Proven pathogenic genes	Unproven pathogenic genes	Proportion
19884	ECs4552/ECs4553/ECs4554/ECs4556	ECs4551	80%
19046	ECs0237/ECs0238/ECs0239	ECs0236	75%
19062	ECs0320/ECs0321/ECs0323	ECs0319/ECs0322/ECs0324	50%
19325	ECs1668/ECs1669/ECs1670		100%
19330	ECs1697/ECs1698/ECs1699	ECs1693/ECs1694/ECs1695/ECs1696	43%
19331	ECs1703/ECs1704/ECs1705		100%
19333	ECs1715/ECs1716/ECs1717	ECs1718/ECs1719/ECs1720	50%
19911	ECs4703/ECs4704/ECs4705	ECs4702/ ECs4706	60%

great significance to find the new "pathogenic genes". In the current study, we collected 364 *E. coli* pathogenic genes using a bibliometric method and found 50 new, specific pathogenic genes of *E. coli* O157:H7 through pathway enrichment and operon analysis. Our findings may be helpful in the search for clinical diagnostic biomarkers and novel drug targets for infectious diarrhea. Among these 50 newly discovered pathogenic genes, two genes (*ECs4774* and *ECs4841*) [18] were enriched into more than 10 pathways, while there were 14 genes

Table 7. The final list of the 50 predicted pathogenic genes.

Gene	Pathway analysis	Operon analysis
ECs4590	√	
ECs4550	√	
ECs4562	√	
ECs4564	√	
ECs1812	√	
ECs1814	√	
ECs4551		√
ECs0236		√
ECs0319		√
ECs0322		√
ECs0324		√
ECs1693		√
ECs1694		√
ECs1695		√
ECs1696		√
ECs1718		√
ECs1719		√
ECs1720		√
ECs4702		√
ECs4706		√

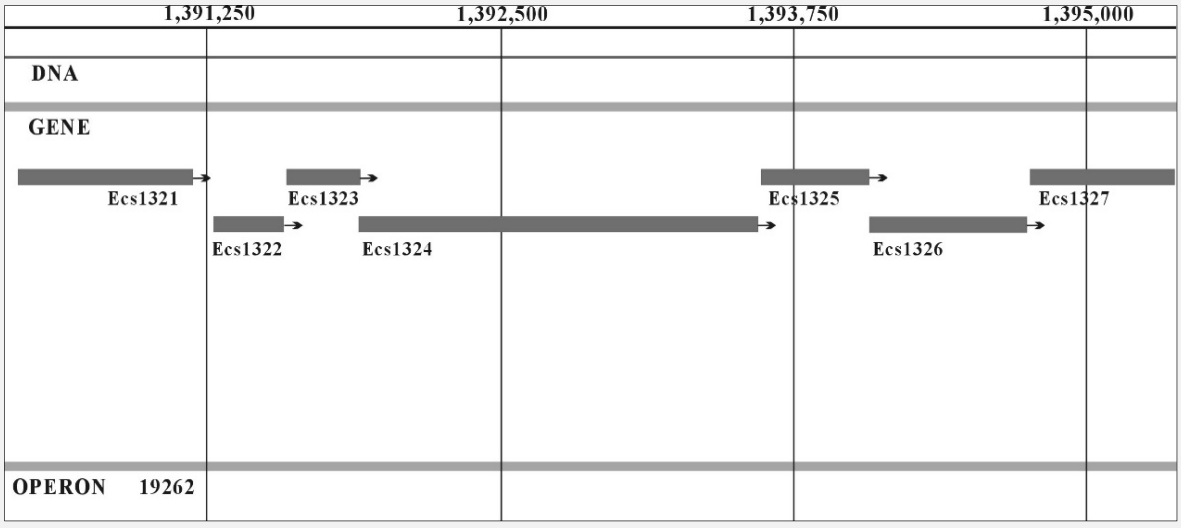

Figure 1. The *E. coli* operon 192628.

The upper parts denote genes, and the lower part denotes operons.

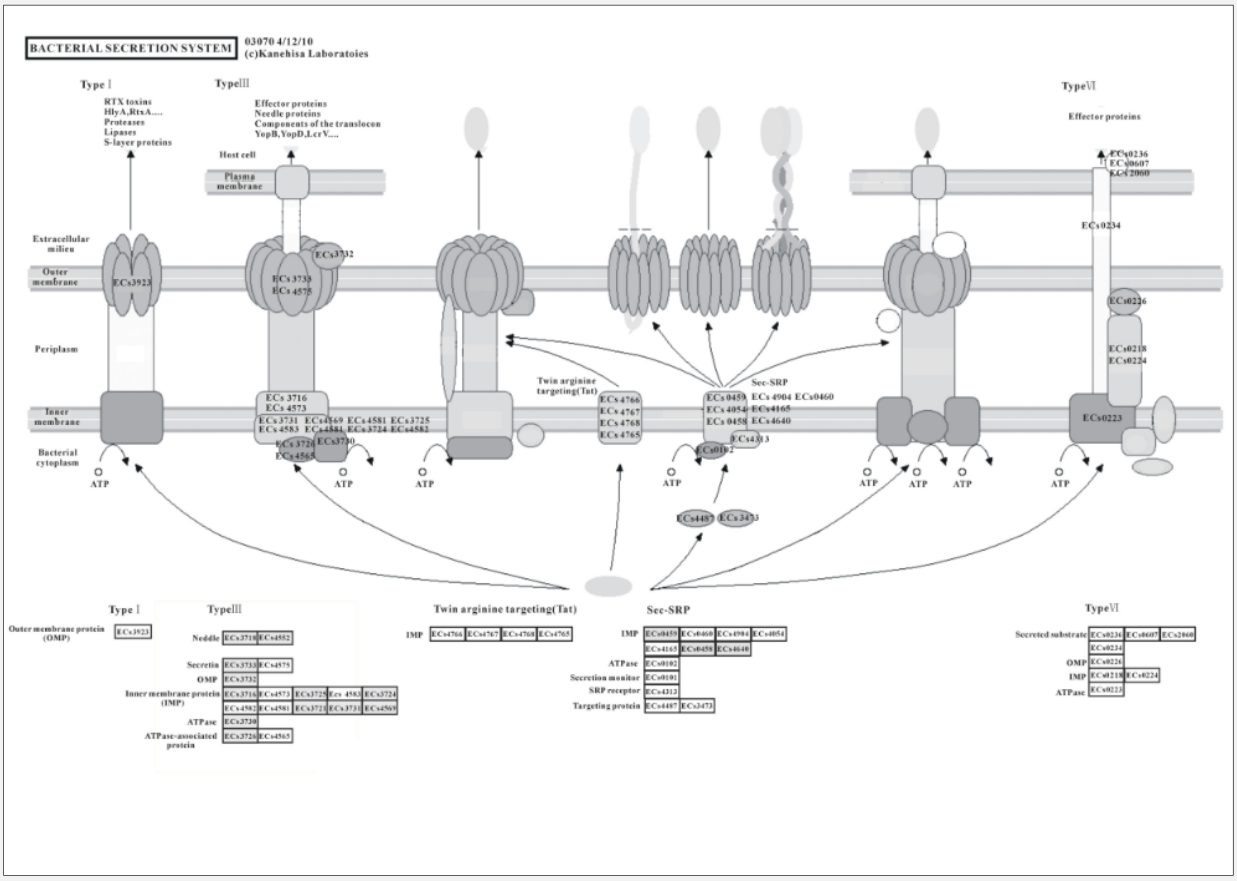

Figure 2. The hpy03070 pathway and its function in the bacterial secretion system.

The grey boxes indicate pathogenic genes of *E. coli*, and the white boxes indicate other genes.

into not less than 5 pathways including *ECs0783*, *ECs4704*, *ECs4773*, *ECs4850*, *ECs0747*, *ECs0748*, *ECs0753*, *ECs0985*, *ECs4792*, *ECs1712*, *ECS4705*, *ECs2846*, *ECs1847*, and *ECs5343* (Table 4). In addition, pathway ecs01100, which is primarily responsible for regulation of metabolism including the Kreb's cycle, was enriched with more than 20 pathogenic genes. Thus, these genes may be involved in the regulation of *E. coli* metabolism. We also inferred possible biological function of these newly discovered genes by pathway enrichment analysis. For example, there were 10 genes affiliated with pathways ecs03070, ecs01120, ecs01130, ecs00230, and ecs01110 which were involved in purine metabolism and biosynthesis of secondary metabolites and antibiotics [19]. Among 45 unknown *E. coli* related genes in pathway ecs03070, associated with the bacterial secretion system (Figure 2), 15 (33.3%) were speculated to be potential pathogenic genes (Table 5). Infection of *E. coli* was closely related with the bacterial secretion system [20], which would cause cytopathy by activating the signaling pathways in host cells. We therefore inferred that the remaining 30 genes with unknown function in pathway ecs03070 may also contain some pathogenic genes. Potential roles and mechanisms of these genes are in need of further verification. In particular, pathway ecs05130 is involved in infection of *E. coli* (Figure 3). In this pathway, 8 (57.1%) of the 14 *E coli* related genes were proposed to be pathogenic and the other 6 had possible involvement of *E. coli* infection due to the background of the regulatory function of ecs05130. At the same time, we found that pathway ecs05130 was also in pathogenic Escherichia coli infection after functional clustering (Table 3), which was consistent with the results obtained by pathway analysis.

We found that the pathogenic genes of *E. coli* were mainly located on the most common metabolic pathways of bacterium, which indicates that the pathogenic

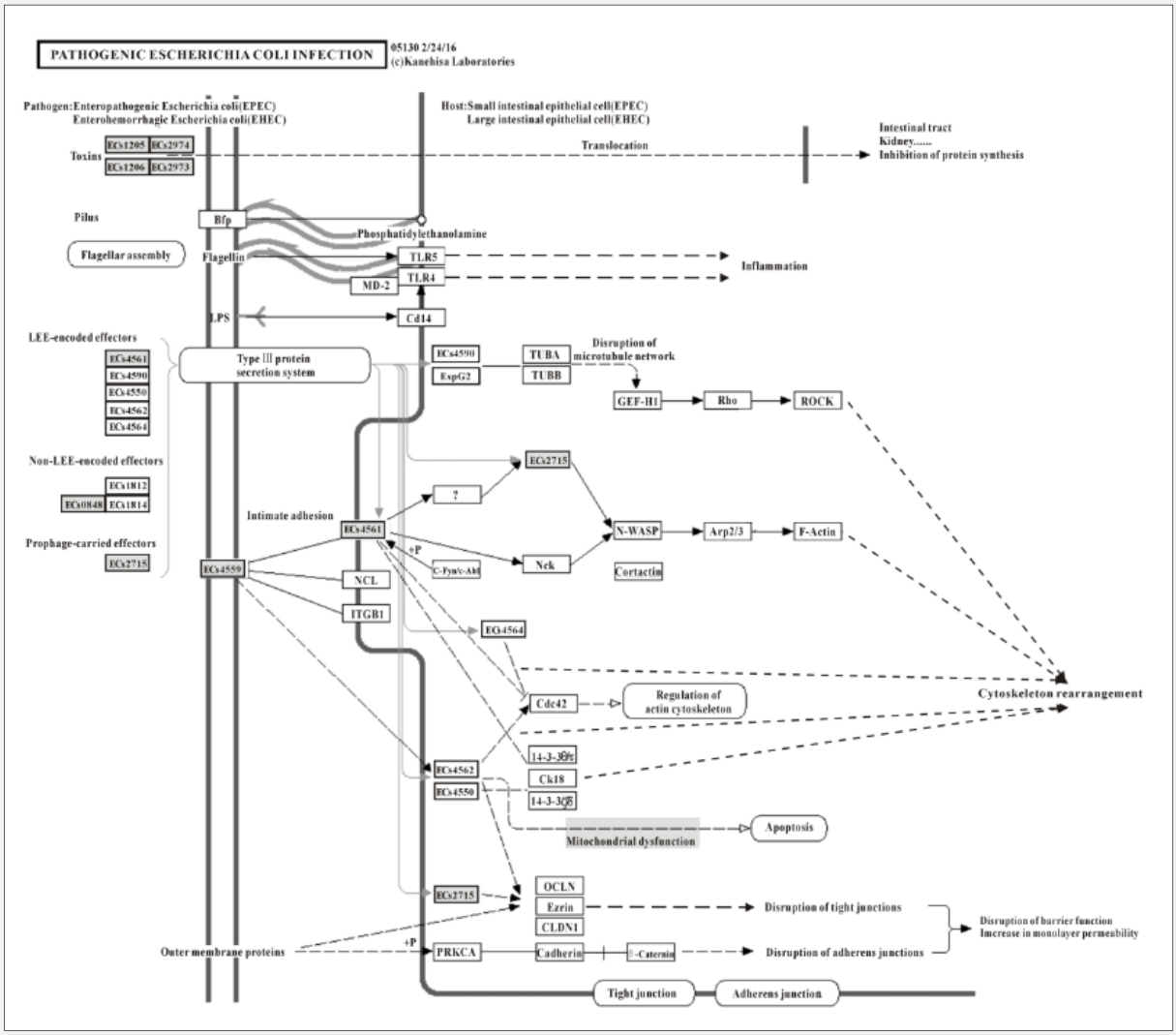

Figure 3. The hpy05130 pathway and its function in pathogenic *E. coli* infection.

The grey boxes indicate pathogenic genes of *E. coli*, and the white boxes indicate other genes.

mechanism of *E. coli* has no specific characteristics. We also identified that pathways ecs03070 and ecs05130 were closely related with the bacterial secretion system. This result was in accordance with our previous studies, and our postulation that the 2 unproven genes may also be pathogenic.

We found that operon19884 in the functional cluster (Table 3) belongs to the bacterial secretion system function, while operon19884 contained a total of five genes, of which four are known pathogenic genes (Table 6). The operon analysis showed that in the operons containing at least 3 pathogenic genes, 82.7% of genes located at these operons were pathogenic (Table 6), while in the operon containing at least 5 pathogenic genes, this proportion was 100%. As the operon is a gene module containing a list of genes with similar functions, it is potential that the unproven genes located at the same operon in this study may also be pathogenic.

CONCLUSION

We collected 364 reported *E. coli* pathogenic genes by bibliometric analysis and identified 50 novel pathogenic genes through pathway enrichment and operon analysis.

These results provide a source for novel prognostic markers as well as novel therapeutic targets for *E. coli* O157:H7 infected diseases.

Acknowledgement:
This work was supported by the National Natural Science Foundation of China (81401712), "University student innovation research project (UIRP)" (201510201 106).

Declaration of Interest:
The authors declare that they have no competing interests.

References:

1. Roussel C, Cordonnier C, Galia W, et al. Increased EHEC survival and virulence gene expression indicate an enhanced pathogenicity upon simulated pediatric gastrointestinal conditions. Pediatr Res 2016 Nov;80(5):734-43 (PMID: 27429202).

2. Zhang D, Coronel-Aguilera CP, Romero PL, et al. The Use of a Novel NanoLuc -Based Reporter Phage for the Detection of Escherichia coli O157:H7. Sci Rep 2016;6:33235 (PMID: 27624517).

3. Schmidt CE, Shringi S, Besser TE. Protozoan Predation of Escherichia coli O157:H7 Is Unaffected by the Carriage of Shiga Toxin-Encoding Bacteriophages. PLoS One 2016;11(1): e0147270 (PMID: 26824472).

4. Eppinger M, Cebula TA. Future perspectives, applications and challenges of genomic epidemiology studies for food-borne pathogens: A case study of Enterohemorrhagic Escherichia coli (EHEC) of the O157:H7 serotype. Gut Microbes 2014;6(3):194-201 (PMID: 25483335).

5. Bonetta S, Pignata C, Lorenzi E, et al. Detection of pathogenic Campylobacter, *E. coli* O157:H7 and Salmonella spp. in wastewater by PCR assay. Environ Sci Pollut Res Int 2016;23(15): 15302-9 (PMID: 27106076).

6. Munns KD, Zaheer R, Xu Y, et al. Comparative Genomic Analysis of Escherichia coli O157:H7 Isolated from Super-Shedder and Low-Shedder Cattle. PLoS One 2016;11(3):e0151673 (PMID: 27018858).

7. Kim HW, Rhee MS. Influence of Low-Shear Modeled Microgravity on Heat Resistance, Membrane Fatty Acid Composition, and Heat Stress-Related Gene Expression in Escherichia coli O157:H7 ATCC 35150, ATCC 43889, ATCC 43890, and ATCC 43895. Appl Environ Microbiol 2016;82(10):2893-901 (PMID: 26944847).

8. Chopyk J, Moore RM, DiSpirito Z, et al. Presence of pathogenic Escherichia coli is correlated with bacterial community diversity and composition on pre-harvest cattle hides. Microbiome 2016;4: 9 (PMID: 27000779).

9. Ravan H, Amandadi M, Sanadgol N. A highly specific and sensitive loop-mediated isothermal amplification method for the detection of Escherichia coli O157:H7. Microb Pathog 2016;91:161-5 (PMID: 26724736).

10. Zhang XC, Huang DS, Li F. Cancer nursing research output and topics in the first decade of the 21st century: results of a bibliometric and co-word cluster analysis. Asian Pac J Cancer Prev 2012;12(8):2055-8 (PMID: 22292650).

11. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27-30 (PMID: 10592 173).

12. Mao X, Ma Q, Zhou C, et al. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res 2013;42 (Database issue):D654-9 (PMID: 24214966).

13. Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 2006;34 (Web Server issue):W293-297 (PMID: 16845012).

14. Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007; 8(9):R183 (PMID: 17784955).

15. Wang L, Dong Y, Wang S, Li L, Kong X, Lv A. First sporadic case of pathogenic Escherichia coli infection in Black swan in China. Microb Pathog 2016;98:32-6 (PMID: 27354206).

16. Vreken P, van Kuilenburg AB, Meinsma R, van Gennip AH. Dihydropyrimidine dehydrogenase deficiency. Identification of two novel mutations and expression of missense mutations in E. coli. Adv Exp Med Biol 1998;431:341-6 (PMID: 9598088),

17. Arluison V, Taghbalout A. Cellular localization of RNA degradation and processing components in Escherichia coli. Methods Mol Biol 2015;1259:87-101 (PMID: 25579581).

18. Hayashi T, Makino K, Ohnishi M, et al. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001; 8(1):11-22 (PMID: 11258796).

19. Ferreira Antunes M, Eggimann FK, Kittelmann M, et al. Human xanthine oxidase recombinant in *E. coli*: A whole cell catalyst for preparative drug metabolite synthesis. J Biotechnol 2016 Oct 10; 235:3-10 (PMID: 27021957).

20. Cao B, Zhao Y, Kou Y, Ni D, Zhang XC, Huang Y. Structure of the nonameric bacterial amyloid secretion channel. Proc Natl Acad Sci USA 2014;111 (50):E5439-44 (PMID: 25453093).