Computational prediction and experimental validation of novel markers for detection of STEC O157:H7

Guo-Qing Wang, Feng-Feng Zhou, Victor Olman, Ying-Ying Su, Ying Xu, Fan Li

Guo-Qing Wang, Ying-Ying Su, Fan Li, Department of Pathogenobiology, Norman Bethune Medical College of Jilin University, Changchun 130000, Jilin Province, China
Guo-Qing Wang, Feng-Feng Zhou, Victor Olman, Ying Xu, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, United States
Ying Xu, Department of Bioinformatics, College of Computer Science and Technology, Jilin University, Changchun 130000, Jilin Province, China

Author contributions: Wang GQ designed and performed the research; Zhou FF and Olman V analyzed the data and generated the genomic barcode tool; Su YY performed the experiments; Li F and Xu Y designed the study and provided the financial support for this work; Wang GQ wrote the manuscript. Supported by National Natural Science Foundation of China, No. 30872415, in part by National Science Foundation, No. DBI-0354771, ITR-IIS-0407204, DBI-0542119, CCF0621700, National Institutes of Health, No. 1R01GM075331 and 1R01GM081682, and grant from BioEnergy Science Center, US Department of Energy BioEnergy Research Center supported by the Office of Biological and Environmental Research in DOE Office of Science

Correspondence to: Fan Li, Professor, Norman Bethune Medical College of Jilin University, Changchun 130021, Jilin Province, China. lifan@jlu.edu.cn
Telephone: +86-431-85619574 Fax: +86-431-85619574
Received: October 15, 2010 Revised: February 12, 2011 Accepted: February 19, 2011 Published online: April 14, 2011

Abstract

AIM: To identify and assess the novel makers for detection of Shiga toxin producing Escherichia coli (STEC) O157:H7 with an integrated computational and experimental approach.

METHODS: High-throughput NCBI blast (E-value cutoff e-5) was used to search homologous genes among all sequenced prokaryotic genomes of each gene encoded in each of the three strains of STEC O157:H7 with complete genomes, aiming to find unique genes in O157: H7 as its potential markers. To ensure that the identified markers from the three strains of STEC O157:H7 can serve as general markers for all the STEC O157:H7 strains, a genomic barcode approach was used to select the markers to minimize the possibility of choosing a marker gene as part of a transposable element. Effectiveness of the markers predicted was then validated by running polymerase chain reaction (PCR) on 18 strains of O157:H7 with 5 additional genomes used as negative controls.

RESULTS: The blast search identified 20, 16 and 20 genes, respectively, in the three sequenced strains of STEC O157:H7, which had no homologs in any of the other prokaryotic genomes. Three genes, wzy, Z0372 and Z0344, common to the three gene lists, were selected based on the genomic barcode approach. PCR showed an identification accuracy of 100% on the 18 tested strains and the 5 controls.

CONCLUSION: The three identified novel markers, wzy, Z0372 and Z0344, are highly promising for the detection of STEC O157:H7, in complementary to the known markers.

© 2011 Baishideng. All rights reserved.

Key words: Shiga toxin producing Escherichia coli O157: H7; Diagnosis; Marker genes; Infectious diseases

Peer reviewer: Shiu-Ming Kuo, MD, University at Buffalo, 15 Farber Hall, 3435 Main Street, Buffalo, NY 14214, United States

Wang GQ, Zhou FF, Olman V, Su YY, Xu Y, Li F. Computational prediction and experimental validation of novel markers for detection of STEC O157:H7. World J Gastroenterol 2011; 17(14): 1910-1914 Available from: URL: http://www.wjgnet.com/1007-9327/full/v17/i14/1910.htm DOI: http://dx.doi.org/10.3748/wjg.v17.i14.1910
INTRODUCTION

Shiga toxin producing Escherichia coli (STEC) O157:H7 is a food-borne pathogen that can cause both epidemic outbreaks and sporadic cases of diarrhea, hemorrhagic colitis, hemolytic-uremic syndrome and thrombotic thrombocytopenic purpura[1]. In recent years, epidemic outbreaks of STEC O157:H7 occurred in the United States, Japan and other industrial countries as well as in developing nations, thus posing a serious threat to human health and economic developments[2]. The effectiveness of current treatment remains frustratingly limited with major side effects, antibiotics-based treatment of patients with STEC O157:H7 infection increases the risk of hemolytic uremic syndrome, especially in children and seniors[3]. The potential for large-scale outbreaks of STEC O157:H7 infection and the lack of effective treatment have inspired intensive researches on the early detection of O157:H7.

A number of methods have been developed for the detection of STEC O157:H7. Morphological analysis and serotype identification are time-consuming, laborious and not always reliable[4]. A fast, highly sensitive and reliable technique, polymerase chain reaction (PCR) assay, has been employed to detect the specific target genes associated with STEC O157:H7[5]. A number of virulence genes can be used in detecting STEC O157:H7, such as representative virulence gene (eaeA) and stx[6]. However, these marker genes have unacceptably high false positive and negative rates[7]. It is, therefore, urgently necessary to identify novel and more effective diagnostic markers for the detection of STEC O157:H7 with a high sensitivity and reliability.

One of the key reasons for the sub par performance of existing markers is that studies that identified the current markers have not fully taken the advantages of available genomic sequence data of STEC O157:H7 and hundreds of other prokaryotes with complete genomes. In this paper, we present an integrated study that combined large-scale genome sequence comparisons, sequence feature analysis and PCR-based experimental validation of marker identification, and report three marker genes based on our sequence feature analysis and experimental validation. These genes represent the promising complements to the known marker genes.

MATERIALS AND METHODS

Genome sequence data

Seven hundred and fifty completely sequenced prokaryote genomes, including 3 strains of STEC O157:H7, were downloaded from the NCBI Prokaryotic Genome Database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) in January 2009.

Computational identification of candidate marker genes

NCBI blast was used to identify the candidate marker genes across any of the 750 prokaryotic genomes analyzed for each gene encoded by each of the 3 strains of STEC O157:H7 with complete genomes as previously described[8]. A gene of STEC O157:H7 was considered a potential marker gene if it did not have a blast hit with E-value < 10^-5 and identity > 95%, which was identical in the 3 strains of STEC O157:H7.

Stability analysis of candidate marker genes

The following strategy was employed to predict the instability of a gene. A gene was considered stable in STEC O157:H7 if the flanking region (1500 bps on each side of the gene) had a higher sequence identity than 50% with the corresponding flanking region of its orthologous genes in the other two strains of STEC O157:H7. A stable gene should also have no transposons[9,10] or phages[11,12] in the flanking region. RNA was closely associated with the pathogenicity islands[13], thus genes within 3000 bps of RNA genes on the safe side were excluded.

A genomic barcode scheme was developed for visualizing a genome, which demonstrated that genomic barcodes can effectively identify “abnormal” genes[14]. A key step of the approach was to calculate the 4-mer frequencies of each 4-mer together with its reverse complement, representing these 4-mer frequencies as a vector of the number of combined 4-mer frequencies and their corresponding reverse complements, with 136 arranged in the alphabetical order. A key interesting observation was that the majority of fragments in a genome had highly similar 4-mer frequencies calculated throughout a genome. Sequence fragments with distinct 4-mer frequencies often indicate horizontal gene transfers. The distance between two vectors of 4-mer frequencies was expressed as the Euclidean distance between the two vectors.

Validation of predicted marker genes by PCR

Ten STEC O157:H7 isolates were obtained from the University of Maryland and 8 STEC O157:H7 isolates were obtained from the Center for Disease Control of China (Table 1). The isolates were cultured on
sorbitol-substituted MacConkey agar and serologically typed for O and H antigens. Shiga toxin genes (stx1 and stx2) of all the isolates were detected by PCR as previously described[13]. Five non-STEC O157 isolates were used as negative controls in PCR, namely Escherichia coli (E. coli) w3110, E. coli ATCC25922, S. aureus ATCC29523, P. aeruginosa ATCC27853, and K. pneumonia ATCC700603, which were obtained from Clinical Test Center of Ministry of Public Health, China.

PCR was carried out in a 50 μL reaction mixture containing 5 × Flexi buffer, 25 mmol/L MgCl2, 10 mmol/L dNTP, Taq DNA polymerase (Promega, USA), 10 pmol of each primer (IDT, USA) and 3 μL bacterial lysates. PCR amplification conditions were optimized to obtain the optimal reaction parameters. The PCR amplification products were visualized by separation on a 2% agarose gel stained with ethidium bromide and by UV transillumination. The primer, annealing temperature and expected product size for each gene are listed in Table 2.

**RESULTS**

**Candidate marker genes for Shiga toxin producing E. coli O157:H7**

The three sequenced genomes of STEC O157:H7 were scanned with 20, 16 and 20 genes identified in STEC O157:H7 Sakai, STEC O157:H7 EDL933 and STEC O157:H7 EC4115, respectively (Table 3), and no whole-gene homology was observed in any of the 750 prokaryotic genomes as detected by blast with E-value < 5 and sequence identity > 95% presented in the 3 STEC O157:H7 genomes. Functional analyses, based on PfamScan[14] and Blast2GO[15], indicated that most of these genes encoded the hypothetical proteins except for wzy (O antigen polymerase). These genes could potentially serve as markers for STEC O157:H7. Virulence marker genes, such as eaeA, stx1 and stx2A, were not included in these genes, because most of them are part of (horizontally transferred) phages, plasmids and pathogenicity islands, with homologs in other prokaryotic genomes.

**Assessment of instability of candidate marker genes**

The instability of candidate marker genes was assessed if each predicted marker overlapped a known mobile genetic element. In addition, whether the flanking region (1500 bps on each side) of each gene is well conserved across the flanking regions of its orthologous genes in the other two strains of STEC O157:H7 was evaluated using sequence identity 50% as the cutoff, which showed that four genes, wzy, Z0372, Z0344 and Z1539 (Table 3), are probably not part of any mobile genetic elements.

**Genomic barcode analysis**

A conserved approach was taken by removing candidate marker genes with substantially different nucleotide compositions from the rest of the genome, measured with the genomic barcode. The distance distribution between the k-mer compositions of each gene in the genome and the average k-mer compositions of the whole genome is shown in Figure 1, which reveals that among the four candidate genes, Z1539 had a large k-mer composition-based distance to the average k-mer compositions of the genome, while the other three candidate genes had highly similar k-mer compositions to those of the whole genome, thus Z1539 was removed from our list of the candidate markers.

In preparation of our manuscript, the genome of a new strain of STEC (TW14359) was available with the 3 marker genes as we predicted here.

**PCR validation**

PCR was performed using the three predicted markers, wzy, Z0372, Z0344, and one STEC O157:H7 representative virulence gene (eaeA), on 18 STEC O157:H7 strains and 5 control organisms, including two non-O157 E. coli strains. The 18 STEC O157:H7 strains were detected using the three genes, with an accuracy of 100% compared to 77.8% using the existing marker eaeA (Figure 2). The virulence genes stx1 and stx2 were present in the 18 STEC

---

**Table 2** Oligonucleotide primer sequences and parameters used for detection Shiga toxin producing Escherichia coli O157:H7

| Gene | sequence (5’ to 3’) | Expected length (bp) | PCR condition Ta/ Te (°C) |
|------|---------------------|----------------------|--------------------------|
| eaeA | F-AAGCCGACTGGTGCACCT | 450 | 55/60 |
|       | R-ACGCTGCTCACTAGATGT |       |          |
| wzy  | F-GAACGATTTCTTTCCGACACC | 276 | 50/60 |
|       | R-GCGCAATATTAGCGACTATG |       |          |
| Z0372| F-AGAATCTCATCCTCGCATTT | 342 | 52/60 |
|       | R-TCTCGGATTTCCGACATTAT |       |          |
| Z0344| F-ATTGTCAGGGAAATTAGCGTG | 121 | 51/60 |
|       | R-TGTGCTTTAAGGTGTTGAACCGA |       |          |

F: Forward; R: Reverse; Ta: Annealing temperature; Te: Elongation time; PCR: Polymerase chain reaction.

---

**Figure 1** Distance distribution histogram for the barcode k-mer compositions of each gene compared with the average k-mer compositions of the whole genome. The x-axis is the k-mer genomic barcode distance, and the y-axis is the frequency. The bar indicates the frequency of different ranges of k-mer distance throughout the whole genome. The curve is the fitting curve to the histogram.
O157:H7 strains. Both sets of markers did equally well without any false predictions on the control samples.

**DISCUSSION**

The accelerated production of microbial genome sequences provides a unique opportunity for the early diagnosis of pathogenic microbes based on comparative genomic studies. Due to the extremely complex and varied STEC genotype, the current diagnostic targets cannot meet the demand for rapid, accurate diagnosis of this pathogen. Furthermore, a number of multi-drug resistant bacteria strains survive in the natural environment and pose a great threat to human health, and the multi-drug resistance results from the horizontally transferable genes. So it is essential to identify these genes specific to a group of closely related pathogenic microbes, and the microbes should be early diagnosed by detecting these marker genes.

In the present study, such gene islands in STEC O157:H7 were identified with three completely sequenced strains of STEC O157:H7. Through a large-scale existence scanning of the genes in the three genomes, 20, 16 and 20 genes in the three strains were identified, respectively. Our previous study\[14\] suggested that genes with barcodes significantly different to the host genomes are usually acquired.

| EDL933 | Sakai | EC4115 | 3000 bp flanking regions | Function |
|--------|-------|--------|--------------------------|----------|
| EaeA   | ECs2844 | ECH74115_2973 | 78 | − | − | − | O antigen polymerase |
| Z0372  | ECs0334 | ECH74115_0348 | 56 | − | − | − | Hypothetical protein |
| Z0344  | ECs0307 | ECH74115_0324 | 50 | − | − | − | Hypothetical protein |
| Z1539  | ECs1281 | ECH74115_1278 | 50 | − | − | − | Hypothetical protein |
| Z3621  | ECs3239 | ECH74115_3589 | 13 | − | − | − | Hypothetical protein |
| Z3271  | ECs2809 | ECH74115_3086 | 0 | − | − | − | Hypothetical protein |
| Z0948  | ECs0804 | ECH74115_0880 | 9 | − | − | + | Hypothetical protein |
| Z1328  | ECs1061 | ECH74115_1143 | 32 | − | − | + | Hypothetical protein |
| Z1430  | ECs1165 | ECH74115_3572 | 0 | − | − | + | Hypothetical protein |
| Z3348  | ECs2979 | ECH74115_3543 | 0 | + | − | + | Hypothetical protein |
| Z3118  | ECs2755 | ECH74115_2802 | 0 | − | − | + | Hypothetical protein |
| Z0244  | ECs0212 | ECH74115_0250 | 38 | − | + | + | Hypothetical protein |
| Z1153/Z1592 | ECs5413 | ECH74115_1531 | 50 | − | − | − | Hypothetical protein |
| Z2107/Z2378/Z6055/Z3108 | ECs1960/ECs2748 | ECH74115_2190/ECH74115_2260/ECH74115_2579 | 0 | + | + | + | Hypothetical protein |
| Z1782/Z6064 | ECs2271 | ECH74115_3158 | 2 | − | − | + | Hypothetical protein |

"Ident %" is the identity percentage of 3000-bp flanking regions among the three genomes. "+" and "−" in column "rRNA", "Tpase" and "phage" represent if the 3000-bp flanking regions contain any rRNA, transposase or phage. The gene names and their functions were retrieved from the NCBI Prokaryotic Genome Database, as described in materials and methods.
and may easily excise the host genomes. These genes were removed from the further analysis in our study. Our computational pipeline reached 3 genes for each of the three strains. The wet laboratory PCR experiments confirmed their existence in 18 clinically retrieved STEC O157:H7 strains but not in 5 control pathogenic microbial samples. We believe that these three marker genes can complement the current detection technique of STEC O157:H7.

We are also working on the identification of marker genes for other human pathogenic microbes, especially for multi-drug resistant strains.

COMMENTS

Background
Shiga toxin producing Escherichia coli (STEC) O157:H7 is an important food-borne pathogen of human gastrointestinal diseases. The potential for large-scale outbreaks of STEC O157:H7 and the lack of effective treatments have inspired intensive researches on the early detection of this pathogen.

Research frontiers
Traditional morphological analysis and serotype identification for detecting STEC O157:H7 are time-consuming, laborious and not always reliable. Polymerase chain reaction (PCR) is a highly desirable method to detect specific target genes associated with O157:H7. However, existing marker genes have unacceptably high false-positive and negative rates. The advantages of available genomic sequence data of O157:H7 and hundreds of other prokaryotes with complete genomes will provide us a great opportunity to select diagnostic markers for rapid and reliable detection of STEC O157:H7.

Innovations and breakthroughs
To the best of the authors’ knowledge, this is the first study to use bioinformatics approach for high-throughput screen of diagnostic markers for detection of pathogens. Furthermore, the authors combined computational biology and molecular biology to solve biological problems, which will provide us a new vision for the prevention of infectious diseases.

Applications
The authors identified and validated three novel and highly promising markers, wzy, Z0372 and Z0344, which may outperform the existing markers for rapid and reliable detection of STEC O157:H7.

Terminology
Genomic barcode is a computational technique, representing the k-mer nucleotide sequence frequency distributions across a whole genome as a 2-D image. In this paper, the authors used this technique to visualize the genome, showing that parts of the genome may have foreign origins.

Peer review
This manuscript is a well-planned, executed study. Although the work was not carried out in gastrointestinal tissue or cells, it is a good example of using bioinformatics approach for the prevention and management of intestinal diseases.

REFERENCES

1 Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001; 8: 11-22
2 Orskov F, Orskov I. Escherichia coli serotyping and disease in man and animals. Can J Microbiol 1992; 38: 699-704
3 Wong CS, Brandt JR. Risk of hemolytic uremic syndrome from antibiotic treatment of Escherichia coli O157:H7 colitis. JAMA 2002; 288: 3111
4 Belanger SD, Boissinot M, Ménard C, Picard FJ, Bergeron MG. Rapid detection of Shiga toxin-producing bacteria in feces by multiplex PCR with molecular beacons on the smart cycler. J Clin Microbiol 2002; 40: 1436-1440
5 Cui S, Schroeder CM, Zhang DY, Meng J. Rapid sample preparation method for PCR-based detection of Escherichia coli O157:H7 in ground beef. J Appl Microbiol 2003; 95: 129-134
6 Chen S, Xu R, Yee A, Wu KY, Wang CN, Read S, De Grandis SA. An automated fluorescent PCR method for detection of shiga toxin-producing Escherichia coli in foods. Appl Environ Microbiol 1998; 64: 4210-4216
7 Paton AW, Paton JC. Detection and characterization of Shiga toxicigenic Escherichia coli by using multiplex PCR assays for stx1, stx2, eaeA, enterohemorrhagic E. coli hlyA, rfbO111, and rfbO157. J Clin Microbiol 1998; 36: 598-602
8 McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004; 32: W20-W25
9 Chen Y, Zhou F, Li G, Xu Y. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. Gene 2009; 436: 1-7
10 Zhou F, Olman V, Xu Y. Insertion Sequences show diverse recent activities in Cyanobacteria and Archaea. BMC Genomics 2008; 9: 36
11 Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequence: current status, policy and new initiatives. Nucleic Acids Res 2009; 37: D32-D36
12 Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 2008; 24: 863-865
13 Ranquet C, Geiselman J, Toussaint A. The tRNA function of SsrA contributes to controlling repression of bacteriophage Mu prophage. Proc Natl Acad Sci USA 2001; 98: 10220-10225
14 Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. BMC Bioinformatics 2008; 9: 546
15 Li F, Zhao C, Zhang W, Cui S, Meng J, Wu J, Zhang DY. Use of ramification amplification assay for detection of Escherichia coli O157:H7 and other E. coli Shiga toxin-producing strains. J Clin Microbiol 2005; 43: 6086-6090
16 Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008; 36: D281-D288
17 Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 2008; 36: 3420-3435

S-Editor Tian L, E-Editor Wang XL, L-Editor Ma WH