Dear Author,

Here are the final proofs of your article. Please check the proofs carefully.

All communications with regard to the proof should be sent to bmcproductionteam2@spi-global.com.

Please note that at this stage you should only be checking for errors introduced during the production process. Please pay particular attention to the following when checking the proof:

- Author names. Check that each author name is spelled correctly, and that names appear in the correct order of first name followed by family name. This will ensure that the names will be indexed correctly (for example if the author’s name is ‘Jane Patel’, she will be cited as ‘Patel, J.’).
- Affiliations. Check that all authors are cited with the correct affiliations, that the author who will receive correspondence has been identified with an asterisk (*), and that all equal contributors have been identified with a dagger sign (†).
- Ensure that the main text is complete.
- Check that figures, tables and their legends are included and in the correct order.
- Look to see that queries that were raised during copy-editing or typesetting have been resolved.
- Confirm that all web links are correct and working.
- Ensure that special characters and equations are displaying correctly.
- Check that additional or supplementary files can be opened and are correct.

Changes in scientific content cannot be made at this stage unless the request has already been approved. This includes changes to title or authorship, new results, or corrected values.

**How to return your corrections**

*Returning your corrections via online submission:*
- Please provide details of your corrections in the online correction form. Always indicate the line number to which the correction refers.

*Returning your corrections via email:*
- Annotate the proof PDF with your corrections.
- Send it as an email attachment to: bmcproductionteam2@spi-global.com.
- Remember to include the journal title, manuscript number, and your name when sending your response via email.

After you have submitted your corrections, you will receive email notification from our production team that your article has been published in the final version. All changes at this stage are final. We will not be able to make any further changes after publication.

Kind regards,

BioMed Central Production Team 2
The dnd operon for DNA phosphorothioation modification system in Escherichia coli is located in diverse genomic islands

Wing Sze Ho¹, Hong-Yu Ou², Chew Chieng Yeo³ and Kwai Lin Thong¹*

Abstract

Background: Strains of Escherichia coli that are non-typeable by pulsed-field gel electrophoresis (PFGE) due to in-gel degradation can influence their molecular epidemiological data. The DNA degradation phenotype (Dnd⁺) is mediated by the dnd operon that encode enzymes catalyzing the phosphorothioation of DNA, rendering the modified DNA susceptible to oxidative cleavage during a PFGE run. In this study, a PCR assay was developed to detect the presence of the dnd operon in Dnd⁺ E. coli strains and to improve their typeability. Investigations into the genetic environments of the dnd operon in various E. coli strains led to the discovery that the dnd operon is harboured in various diverse genomic islands.

Results: The dndBCDE genes (dnd operon) were detected in all Dnd⁺ E. coli strains by PCR. The addition of thiourea improved the typeability of Dnd⁺ E. coli strains to 100% using PFGE and the Dnd⁺ phenotype can be observed in both clonal and genetically diverse E. coli strains. Genomic analysis of 101 dnd operons from genome sequences of Enterobacteriaceae revealed that the dnd operons of the same bacterial species were generally clustered together in the phylogenetic tree. Further analysis of dnd operons of 52 E. coli genomes together with their respective immediate genetic environments revealed a total of 7 types of genetic organizations, all of which were found to be associated with genomic islands designated dnd-encoding GIs. The dnd-encoding GIs displayed mosaic structure and the genomic context of the 7 islands (with 1 representative genome from each type of genetic organization) were also highly variable, suggesting multiple recombination events. This is also the first report where two dnd operons were found within a strain although the biological implication is unknown. Surprisingly, dnd operons were frequently found in pathogenic E. coli although their link with virulence has not been explored.

Conclusion: Genomic islands likely play an important role in facilitating the horizontal gene transfer of the dnd operons in E. coli with 7 different types of islands discovered so far.

Keywords: DNA degradation, Dnd, Escherichia coli, PFGE, Thiourea, Genetic environment, Genomic islands, Dnd-encoding genomic island

Background

Microbial source tracking (MST) is important to trace the source of a pathogen in an outbreak situation. Pulsed field gel electrophoresis (PFGE) is one of the tools for MST and is the method of choice for a number of bacterial pathogens as it is highly discriminative, reproducible [1] and standardized protocols are available.

However, certain bacterial species such as Escherichia coli, Pseudomonas aeruginosa, Klebsiella pneumoniae and many more are susceptible to DNA degradation, resulting in smeared DNA, which can be improved by the addition of thiourea into the PFGE running buffer [2-4]. The DNA degradation phenotype (designated Dnd) was first observed during an electrophoretic separation of DNA of a Gram-positive bacterium, Streptomyces lividans [5]. A DNA modification system (conferred by a five-gene dnd cluster, dndABCDE) was then found to be responsible for the Dnd phenotype by mediating...
the incorporation of sulphur into the DNA backbone via a process called phosphorothioation [6,7]. The DNA modification renders the DNA susceptible to cleavage by peracetic acid generated during a PFGE run [8,9]. While the 5-gene cluster (consisting of dndA and the dndBCDE operon) is found in bacteria across different families (such as Acidobacteriaceae, Clostridiaceae, and Streptomyces- ceae), Enterobacteriaceae including E. coli only harbour a 4-gene dndBCDE operon, lacking the dndA gene. IscS, a cysteine desulfurase, which has a similar function as DndA, was then identified in E. coli to support DNA phosphorothioation [10,11].

**Results**

Addition of thiourea improved the typeability of Dnd strains by 100%

Without the addition of thiourea to the PFGE run, all 12 Dnd E. coli strains from previous studies [13,14] yielded degraded DNA despite repeated attempts (3 times) (Figure 1(a), lanes S1-B2). However, when thiourea was added to the running buffer, there was a 100% marked improvement in typeability for all the E. coli strains (Figure 1(b), lanes S1-B2). Strains with BlnI-digested chromosomal DNA were also typeable when thiourea was added in the PFGE run (data not shown). Seven zoonotic VTEC strains (lanes S1-S7) and two clinical strains (lanes A1-A2) shared two indistinguishable pulstotypes (i.e., clonal) (Figure 1(b)) while another 3 (lanes A3, B1, B2) were genetically different.

Sequence analysis of dndB-E genes and association of the DNA degradation phenotypes and genotypes

PCR detection of dndc gene using primers described previously [15] revealed that out of 12 Dnd E. coli strains, only 7 zoonotic VTEC (Figure 1, lanes S1-S7) and 2 clinical (Figure 1, lanes B1-B2) Dnd+ strains were positive for the dndC gene. Another 3 Dnd+ strains (A1-A3) were negative for the dndC gene and these strains were obtained from the same medical center and source. The failure to amplify the dndc gene from 3 out 12 Dnd+ strains using previously described primers [15] prompted us to develop a PCR assay targeting the internal regions of dndB, dndC, dndD, and dndE based on the dndBCDE DNA sequences of E. coli available in dndDB [10].

![Figure 1](image.png)

**Figure 1** Effects of thiourea improved the typeability of Dnd+ strains by PFGE. a: PFGE gel obtained by performing PFGE run without the addition of thiourea. b: PFGE gel obtained by performing PFGE run with the addition of thiourea to the agarose gel and running buffer. Lane M: XbaI digested H9812 DNA marker; lanes S1-S7, VTEC isolates from pigs; lanes A1-A3, clinical isolates obtained from medical center A; lanes B1-B2, clinical isolates obtained from medical center B; C, control E. coli isolate which is typeable with and without the addition of thiourea (i.e., clonal). Gel images were captured using GelDoc (BioRad, Hercules, CA) digital gel documentation system.
Dnd⁺ strains yielded the expected PCR amplicons for all 4 genes (dndB, dndC, dndD, and dndE) (using the designed primers listed in Table 1). All Dnd⁻ strains (including E. coli, Shigella sonnei, Salmonella Typhimurium and Salmonella Enteriditis) that did not display degradation phenotype were negative for the dnd genes. Significant correlation was also found between the presence of the dnd gene cluster and the Dnd phenotype (p < 0.05).

At least two sets of the amplified dndB-dndE genes from different strains with a similar pulsotype were sequenced. It was found that the dndB-dndE genes from different strains with the same pulsotype shared identical (100%) nucleotide sequences. Multiple sequence alignment (MSA) of the nucleotide sequences of the dndB-dndE genes with E. coli SE11, 55989 and B7A were also carried out. Using E. coli 55989 as reference, the pairwise alignments showed that E. coli S1-S7, B1 and B2 shared dndB-E sequences that are highly similar to E. coli 55989 (BLASTN identities of dndB-C = 100%; dndD = 99.8% and dndE = 98.5%). Strains A1 and A2 harboured identical dnd genes but are different from E. coli 55989 (dndB = 89.9%; dndC = 85.2%; dndD = 98.2%; dndE = 95.9%). The dndB-E sequences of strain A3 are mostly distantly related to those of E. coli 55989 (dndB = 87.9%; dndC = 89.6%; dndD = 80.6%; dndE = 84%) but are highly similar to E. coli B7A (identical dndB-D sequences with 2 SNPs with respect to dndE) (Additional file 1: Figure S1). Hence, the inability to amplify the dndC genes using primers described previously [15] for strain A1-A3 might be due to the diversity in the sequences of the dnd operon in various E. coli strains. Further in silico analysis also revealed that mismatches at the primer priming sites (n = 2 – 10) were found for E. coli 55989, SE11 and B7A. However, it should be noted that primers described by Wang et al., 2011 [15] for the expression studies of the dnd operon from Salmonella enterica (although closely related but not identical to E. coli). Furthermore, to the best of our knowledge, there are no other appropriate primers available for the detection of the dndBCDE gene cluster in E. coli. It should also be noted that the DNA sequences of the dndB-E genes included in the MSA are partial CDS. The DNA sequences of dndB-dndE genes from each pulsotype have been deposited in GenBank under accession nos. KJ702391-KJ702410.

PCR detection of iscS revealed that all E. coli (both Dnd⁺ and Dnd⁻ strains), Salmonella spp. and Shigella strains were positive for the iscS gene. In fact, iscS is also found in Citrobacter spp., Klebsiella spp., and Morganella spp. using BLASTN indicating that this cysteine desulphurase gene may be ubiquitous in different enterobacterial strains, thus supporting its suggested importance in sustaining fundamental life processes [16]. However, this also indicates that the presence of the iscS gene is not a good genotypic marker for the Dnd⁺ phenotype. Representative iscS sequence obtained from this study have been deposited in GenBank under accession no KC839813.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Oligonucleotide sequence (5' to 3')</th>
<th>Target gene</th>
<th>Amplicon size (bp)</th>
<th>Annealing temperature (°C)</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t1.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table 1 Primers used for detection of iscS genes and genes in dnd operon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t1.3</td>
<td>IscS-CMU F</td>
<td>ATCTGACAACCTGGCCGATCA</td>
<td>iscS</td>
<td>860</td>
<td>44</td>
</tr>
<tr>
<td>t1.4</td>
<td>IscS-CMU R</td>
<td>CTTCCAGTAGTAAGAACGACCT</td>
<td>iscS</td>
<td>780</td>
<td>44</td>
</tr>
<tr>
<td>t1.5</td>
<td>dndC-F</td>
<td>CCGATCGGCTGCGATGTT</td>
<td>dndC</td>
<td>579</td>
<td>51</td>
</tr>
<tr>
<td>t1.6</td>
<td>dndC-R</td>
<td>CTTCCACCGCCGCCGTT</td>
<td>dndC</td>
<td>659</td>
<td>56</td>
</tr>
<tr>
<td>t1.7</td>
<td>dndB-F</td>
<td>CATTGGCAGATAACCGGCC</td>
<td>dndB</td>
<td>207</td>
<td>50</td>
</tr>
<tr>
<td>t1.8</td>
<td>dndB-R</td>
<td>CAGAAAGCGATGAAAGCGC</td>
<td>dndB</td>
<td>532</td>
<td>56</td>
</tr>
<tr>
<td>t1.9</td>
<td>dndC-F</td>
<td>TTCCACCCCCCGGTGTTTC</td>
<td>dndC</td>
<td>605</td>
<td>51</td>
</tr>
<tr>
<td>t1.10</td>
<td>dndC-R</td>
<td>GCCTGCCTCCCGAAATG</td>
<td>dndC</td>
<td>196</td>
<td>48</td>
</tr>
<tr>
<td>t1.11</td>
<td>dndD-F</td>
<td>AGCTTCTCTTATGTAATGTTCCAGA</td>
<td>dndD</td>
<td>605</td>
<td>48</td>
</tr>
<tr>
<td>t1.12</td>
<td>dndD-R</td>
<td>AGCAACTGGATCCTGCTGTTT</td>
<td>dndD</td>
<td>196</td>
<td>48</td>
</tr>
</tbody>
</table>
were thus, incomplete. The accession numbers of all 106 genomes with positive hits were summarized in Additional file 2: Table S1 which also included the *E. coli* strain background information.

A phylogenetic tree was constructed using the multiple sequence alignment of the nucleotide sequences of the 101 *dnd* operons (Figure 2). All *dnd* sequences from *Salmonella* spp. (blue) and *Erwinia amylovora* (green) were grouped into two distinct separate clusters. On the other hand, the *dnd* sequences from the genomes of other Enterobacteriaceae such as *Citrobacter* spp., *Cronobacter* spp., *Enterobacter* spp. and *Serratia* spp. were generally not clustered into any groups (Figure 2). As for *E. coli* (red), 45 out of 53 of them were grouped together.

![Figure 2 Phylogenetic tree of dnd operons from Enterobacteriaceae. Asterisk (*) indicates that the genome contains two copies of dnd operons. Different alphabets (a, b, cl, cII, d, e, f) depict the subgroups based on the immediate genetic environment. Colored genome strain codes facilitate visualization. Maximum likelihood (ML) method was used to construct the phylogenetic tree using MEGAS. Bootstrap confidence values greater than 50% are shown at branches. Nodes with less than 50% bootstrap value are collapsed.](image-url)
into one main cluster. The remaining 8 were placed at different nodes. Another phylogenetic tree was constructed using the nucleotide sequences of 100 dnd operons (E. coli MS545-1 was excluded because the immediate vicinity of its dnd operon were located in the gaps, hence the total number of E. coli genomes in the analysis was reduced to 52) together with their respective immediate vicinity (1 kb) (Additional file 3: Figure S2) to determine whether the immediate genetic environment will cause any changes to the phylogenetic tree. Surprisingly, both trees (Figure 2 and Additional file 3: Figure S2) yielded highly similar results with only slight differences. For example, Cronobacter turicensis z3032 is more closely related to Enterobacter mori LMG 25706 (both forms a sister group) than to Dickeya dadantii 3937 by comparing their dnd operons only (Figure 2). On the other hand, Cronobacter turicensis z3032 is more closely related to Salmonella enterica strains when the dnd operons and their respective immediate vicinity were included in the analysis (Additional file 3: Figure S2). Overall, both trees yielded similar results.

Based on the sequence similarity of the immediate upstream and downstream regions (1 kb) of the 52 dnd operons of E. coli, a total of 7 subgroups were further identified. The 7 subgroups were designated a, b, c, cI, d, e and f and were included in Figure 2. It was observed that genomes harbouring highly conserved dnd operons also shared similar immediate genetic environments and vice versa. Of the 7 subgroups identified, groups cl and cII share similar immediate downstream but different upstream regions while all other groups were very different with respect to their immediate genetic environment. Interestingly, 4 of the 52 dnd operons from E. coli genomes belonged to two E. coli draft genomes (KOEGE 118 and KTE 9). MSA analysis revealed that the two copies of the dnd operon and their associated genetic environment in the same E. coli genome are different and are distantly located from each other in the phylogenetic tree generated (groups cl and e) (Figure 2 and Additional file 3: Figure S2).

**Upstream and downstream regions of 12 Dnd+ E. coli strains resemble three different types of genetic environment**

PCR detection of the immediate upstream and downstream regions of the dnd operon for the 12 Dnd+ E. coli strains revealed three types of immediate genetic environment. Strains S1-S7, B1 and B2 shared dnd operons with associated genetic environment similar to group a; A1 and A2 to cl; and A3 to group f. All three types of immediate genetic environment identified in the Dnd+ E. coli strains using PCR matched to those determined by *in silico* analysis, indicating that there were only 7 types of different immediate genetic environment identified to date for E. coli.

**Diverse dnd-encoding genomic islands were observed for various representative E. coli genomes**

A genome from each of the seven subgroups determined by the sequence similarity of their respective immediate genetic environment was selected for further genomic analysis. Only 1 genome was selected from each subgroup because it was impossible to determine all complete GIs carrying the dnd operon as most of the genomes in the public database were partially sequenced. All 7 dnd operons from the selected E. coli genomes were found to be associated with GIs. The 7 dnd-encoding GIs were then designated based on the naming system described in the Methods section. Thus, the dnd-encoding GI located in E. coli 55989 (CU9281452) was designated “dndI_I E. coli 55989”; KOEGE 7 (16a) acYxM-superccont1.4.C20 (AWAA01000021) as “dndI_I KOEGE 7”; KOEGE 118 (317a) acYxt-superccont1.1.C29 (AWAR01000029.1) as “dndI_I KOEGE 118”; UMEA 3200-1 acYxW-superccont1.1.C4 (AWCH01000041) as “dndI_I UMEA 3200-1”; SE11 (AP009240.1) as “dndI_I SE11”; KTE9 acASr-superccont1.1.C7 (ANVJ01000007.1) as “dndI_I KTE9” whereas in MS45-1 was excluded because the immediate vicinity of the dnd operon was located in the gaps, hence the total number of E. coli genomes in the analysis was reduced to 52. On the other hand, Cronobacter turicensis C496_10 and C496_11 shared distinct immediate genetic environments and are distantly located from each other in the phylogenetic tree. All three types of different immediate genetic environment identified for E. coli 7 (16a) acYxM-superccont1.4.C20 (AWAA01000021) as “dndI_I E. coli 7” whereas in MS45-1 was excluded because the immediate vicinity of the dnd operon was located in the gaps, hence the total number of genomes in the analysis was reduced to 52.

Interestingly, 4 of the 52 dnd operons from E. coli genomes belonged to two E. coli draft genomes (KOEGE 118 and KTE 9). MSA analysis revealed that the two copies of the dnd operon and their associated genetic environment in the same E. coli genome are different and are distantly located from each other in the phylogenetic tree generated (groups cl and e) (Figure 2 and Additional file 3: Figure S2). Overall, both trees yielded similar results.

Based on the sequence similarity of the immediate upstream and downstream regions (1 kb) of the 52 dnd operons of E. coli, a total of 7 subgroups were further identified. The 7 subgroups were designated a, b, c, cI, d, e and f and were included in Figure 2. It was observed that genomes harbouring highly conserved dnd operons also shared similar immediate genetic environments and vice versa. Of the 7 subgroups identified, groups cl and cII share similar immediate downstream but different upstream regions while all other groups were very different with respect to their immediate genetic environment. Interestingly, 4 of the 52 dnd operons from E. coli genomes belonged to two E. coli draft genomes (KOEGE 118 and KTE 9). MSA analysis revealed that the two copies of the dnd operon and their associated genetic environment in the same E. coli genome are different and are distantly located from each other in the phylogenetic tree generated (groups cl and e) (Figure 2 and Additional file 3: Figure S2).

**Upstream and downstream regions of 12 Dnd+ E. coli strains resemble three different types of genetic environment**

PCR detection of the immediate upstream and downstream regions of the dnd operon for the 12 Dnd+ E. coli strains revealed three types of immediate genetic environment. Strains S1-S7, B1 and B2 shared dnd operons with associated genetic environment similar to group a; A1 and A2 to cl; and A3 to group f. All three types of immediate genetic environment identified in the Dnd+ E. coli strains using PCR matched to those determined by *in silico* analysis, indicating that there were only 7 types of different immediate genetic environment identified to date for E. coli.
covering the *dnd* operon and its immediate vicinity) but *dnd*-II<sub>EC KOEHE 118</sub> had an additional 74 kb segment containing a type VI secretion system (T6SS) that was absent from *dnd*-II<sub>EC KTE9</sub> (Figure 4). Thus, in terms of the *dnd* operon and its immediate genetic vicinity, both *dnd*-II<sub>EC KOEHE 118</sub> and *dnd*-II<sub>EC KTE9</sub> were grouped within the same group e (Figure 2 and Additional file 3: Figure S2). It should be noted that not all genomes from the same subgroup shared identical GIs as they are often very large in size, and variation may have occurred within the GIs caused by other mobile genetic elements [17].

*dnd*-encoding genomic islands displayed mosaic structures

Due to the great diversity of the islands harbouring *dnd* operons, three representative islands from *E. coli* strains B7A, 55989 and KOEHE 118 belonging to different groups (Figure 2) were selected for further investigation to determine the evolutionary forces driving the great variation between the *dnd*-encoding GIs. Based on the GC content (Figure 5), the *dnd*-encoding GIs displayed mosaic structures and can be separated into several modules. Further examination on the presence of ORFs from the *dnd*-encoding GIs in other *E. coli* genomes further supported their mosaic characteristics. The *dnd* modification systems, *dpt* restriction systems, together with the hypothetical proteins in between the two systems were collectively found associated with *dnd*-positive genomes only (ORFs depicted by green-coloured bars). On the other hand, majority of the ORFs outside the *dpt* and *dnd*-clusters are present in both *dnd*-positive and *dnd*-negative *E. coli* genomes (brown-coloured bars, Figure 5), with the exception of an integrase gene (*E. coli* B7A) and 6 ORFs at the immediate upstream of *dnd* operon in KOEHE118, which are found associated with *dnd*-positive genomes only. Intriguingly, some hypothetical proteins are strain-specific, which can be observed in *E. coli* strain 55989 (yellow-coloured bar).

![Genetic map of 7 different dnd-encoding genomic islands of E. coli.](image)
Different genetic contents were identified at the *leu* tRNA insertion sites of *dnd*-negative and *dnd*-positive *E. coli* genomes

All *dnd*-encoding GIs determined in our study were associated with tRNA, an integration hotspot in bacterial genomes [18]. This prompted us to investigate the genetic contents at the same tRNA gene site (*leu* tRNA) in *dnd*-negative *E. coli* genome. The tRNA insertion sites of four genomic regions (1 *dnd*-negative and 3 *dnd*-positive) from *E. coli* strains CFT073, UMEA 3176-1, B7A and 55989 were compared (Figure 6). Upstream flanking regions of the genomic islands which were inserted into the *leu* tRNA gene sites for all 4 *E. coli* genomes are highly conserved, indicating similar upstream backbone (Figure 6). While a 15.6 kb non-*dnd*-encoding GI (consists mainly of hypothetical proteins) was found attached to the *leu* tRNA gene site of *E. coli* CFT073, the similar attachment sites in *E. coli* strains B7A, UMEA 3176-1 and 55989 were occupied with *dnd*-encoding island of variable sizes (53.3 kb – 59.5 kb). The downstream flanking regions of genomic islands for *dnd*-negative (*E. coli* CFT073) and *dnd*-positive (*E. coli* UMEA 3176-1 and B7A) genomes are also conserved (Figure 6), indicating that the *leu* tRNA gene site in *E. coli* is a variable region with other types of GIs found in *dnd*-negative strains. Despite the highly conserved upstream and downstream flanking regions for *dnd*-encoding GIs for *E. coli* strain UMEA 3176-1 and B7A, slight variations within the highly similar GIs were still observed with transposases found in the variable regions (Figure 6).

**Discussion**

PFGE is often regarded as the “gold standard” for subtyping *E. coli*. However, Dnd+ *E. coli* strains that were untypeable from our previous studies yielded incomplete molecular epidemiological data. The addition of thiourea to the PFGE running buffer has enabled us to overcome the untypeable problem of all 12 Dnd+ *E. coli* strains. In
in our study, 9/12 Dnd+ strains were represented by two distinctly different pulsotypes [S1-S7 (n = 7); A1-A2 (n = 2)] (Figure 1), suggesting that the Dnd+ phenotype might be an inherent clonal trait for strains with the two respective pulsotypes. No Dnd− strains were found to share similar pulsotypes with the Dnd+ strains when we compared the pulsotypes with previous studies [13,14]. The 9 Dnd+ strains were multidrug resistant with 2 extended-spectrum beta-lactamase (ESBL) producers and 7/9 were pathogenic VTEC and were clonal [13,14]. The resistance and pathogenic state of these Dnd+ strains may therefore pose possible threats to public health, hence improvement in their typeability is important to generate valuable epidemiological data. The pathogenic state of the Dnd+ strains also prompted us to further explore the strain background of all 52 E. coli genomes that carried dnd operons. Surprisingly, majority of them (n = 44/52; 85%) were pathogenic in nature with 26 being diarrheagenic E. coli and 18 extra-intestinal pathogenic E. coli. To the best of our knowledge, no study has reported this finding before and the link between virulence and dnd operons has yet to be explored. Undoubtedly, bias might be present as pathogenic E. coli strains have higher level of public interest, and hence the availability of their genome data might be more compared to non-pathogenic E. coli. Nevertheless, this finding may provide a caveat for future studies on the possible interesting relationship between pathogenicity and the presence of the dnd operon in E. coli.

The dnd clusters of different bacterial species have reportedly a diverse nature [10], not only with regard to their respective sequence similarities, but also the genetic context of the dnd clusters as well. Some bacterial species harboured a 4-gene dndBCDE operon whereas others carried a five-gene dnd cluster (dnda and dndBCDE operon). In fact, dnda is not found in the dnd clusters of all Enterobacteriaceae reported in dndDB (including E. coli), indicating that the gene contents in the dnd clusters for closely related bacteria are similar. This is also the reason

Figure 5  GC content of three representative dnd-encoding genomic islands. GC content of dnd-encoding GIs are shown above their respective genetic maps, with blue and red region showing GC content above and below 50%, respectively. The dnd-encoding GIs consist of several modules, depicted by the coloured bars below their respective genetic maps. ORFs above the brown coloured bars were found in both dnd-positive and dnd-negative E. coli genomes. ORFs that are underlined by green bars are strictly associated with dnd-positive genomes only while ORFs underlined by the yellow bar are strain-specific.
why dndA gene was not included as a target gene for prediction of degradation phenotype of E. coli in our study. Although DndA might be absent in some dnd clusters, studies have revealed that IscA can serve as functional homologues of DndA and may be found in more than one copy in the genome [15,19]. As previously reported [10], the dnd cluster of E. coli belonged to the 4-gene dndBCDE operon. All 12 Dnd⁺ E. coli strains in this study were positive for the dndB-dndE genes and the presence of dndB-dndE genes was significantly associated with the DNA degradation phenotypes (p < 0.5). This indicates that the presence of these genes can be a potential genetic marker to predict the phenotype of DNA degradation. A previous study [12] developed an assay to predict the DNA degradation phenotype of Mycobacterium abscessus based on the fact that the dnd gene cluster was associated with a genomic island of the M. abscessus complex and the genetic environment of the dnd-encoding GI was highly conserved. One pair of primers targeted the dndC gene while another pair confirmed the absence of the dnd operon. If the dnd operon is present, the amplicon will be too large to be amplified. On the other hand, if the dnd-encoding GI is absent, a short bridge amplicon spanning the GI insertion site will be generated for Dnd⁻ M. abscessus strains [12]. However, this approach was not applicable in our study due to the variable genetic environment of the E. coli dnd operons based on MSA analysis. Hence for E. coli, the detection of all 4 dndB-dndE genes would be a better alternative to confirm the presence of the dnd operon.

The dissemination of dnd operons was reportedly facilitated by mobile genetic elements especially GIs [10,19]. Hence, representative E. coli genomes harbouring dnd operons were subjected to genomic island determination. Our data showed that all the dnd operons in the selected genomes were in fact located in GIs. Out of 31 dnd operons described in dndDB, 29 were also reported to be located in chromosomal islands, one was plasmid-encoded while another one is located in a plasmid-derived chromosomal segment. These results showed the potential role of genomic islands in facilitating the horizontal gene transfer of dnd operons. Genomic islands have also been known to be the mechanism of diversification contributing to bacterial evolution [20]. However, it should be noted that majority of the genomes of E. coli in this study were draft genomes. Thus, we were unable to ascertain if the dnd-
encoding GIs identified were chromosomally located or plasmid-encoded. Further analysis would be necessary to verify their location.

The genetic environment of the dnd operon in E. coli is very diverse (Additional file 3: Figure S2 and Figures 3, 4, 5 and 6). On the other hand, the dnd operon together with its immediate vicinity in Erwinia amylovora and Salmonella enterica appeared to be conserved within the same bacterial species (Figure 2 and Additional file 3: Figure S2), indicating that their genetic environment could be similar. Analysis of the representative E. coli genomes has shown that their respective dnd operons were associated with genomic islands that possess diverse genetic context.

Different genetic fragments of the island that corresponded to regions of different types of E. coli genomes (dnd-positive or dnd-negative) indicate that the genetic elements were likely acquired from different origins. Genetic variation can also be observed among highly conserved dnd-encoding GIs due to insertion or deletion of transposases along with other hypothetical proteins (Figure 6). A large cluster of T6SS was also found inserted into the upstream of another dnd-encoding island (Figure 4). Furthermore, marked difference in GC content was also observed across the dnd-encoding GIs, supporting the mosaic composition of the GIs carrying dnd operons. The mosaic structure of the dnd-encoding GIs (Figure 5), presence of phage-like integrases and presence of transposable elements flanking the GIs as well as across the dnd-encoding GIs (Figure 6), suggest multiple integration or recombination events, which has also been reported by He et al., (2007) [19] for the dnd-encoding GI of Streptomyces lividans 66. Insertion and deletion of genetic elements as well as recombination appear to play an essential role in the constant evolution of these GIs.

Besides the diverse genetic context of the dnd-encoding GIs of E. coli, the tRNA genes associated with the islands can also be variable. The 3'-end of tRNA genes frequently served as the insertion sites for foreign DNA fragments [18]. All 9 dnd-encoding GIs described in this study were associated with tRNA genes, but only limited to three types (leu, thr and phe tRNA genes) (Figures 3 and 4), which have also been reported to be among the most frequently targeted tRNA genes for insertion in E. coli [18,21]. One of the possible reasons that particular tRNA genes often served as an integration hotspot is that the integrases of GIs may have specificity for certain tRNA genes [22]. For E. coli KOEGE 118 and KTE 9 which harboured two dnd-encoding GIs within the same genome, their respective dnd-encoding GIs were also associated with different tRNA genes (leu and thr tRNA genes) (Figure 4). However, it should be noted that not all GIs are associated with tRNA genes [23].

This is the first report describing two dnd operons found in the same bacterial genome. However, the benefits to its host, if any, are unknown. Several advantages to bacteria with DNA phosphorothioation modification have been proposed. One of them is that the modified host DNA is able to act as an antioxidant as phosphorothioation gave the modified DNA chemical reduction property. Thus, the modified DNA is resistant to growth inhibition and oxidative double-stranded DNA breakage caused by H_2O_2 [9]. H_2O_2 is generated during aerobic metabolism and studies have revealed that aerobic E. coli can generate high levels of H_2O_2 that are toxic enough to cause damage to their own DNA. On the other hand, exogenous H_2O_2 can also be found in the environment as competing organisms can release H_2O_2 as toxin to suppress the growth of competitors [24]. Bacteria can get killed by high levels of H_2O_2 generated endogenously and exogenously. Hence it is essential to protect the host bacteria against damage caused by peroxide and DNA that has been modified by phosphorothioation was hypothesized as one of the mechanisms that can offer protection to the host [24]. Nevertheless, the biological implication of having two copies of the dnd operons within a genome is currently unknown.

All the dnd-encoding GIs identified in this study also harboured dptFGH, a set of three conserved dnd-linked orthologs which are often found near the dnd operon. E. coli KTE9 and KOEGE118 which harboured 2 dnd-encoding GIs in the same genome also carried the three conserved GIs within the islands. The conserved dnd-linked orthologs have been proven to function as restriction systems in Salmonella enterica serovar Cerr087 [25] and recently, in E. coli B7A [26]. The restriction system is able to restrict foreign DNA (such as phages and plasmids) that lacks the specific phosphorothioation modification when the host DNA is protected by the modification, and hence prevents the invasion of heterologous DNA [25]. Phosphorothioate DNA modifications incorporated by Dnd proteins (DndABCDE) that work together with the DptFGH restriction system function as a restriction-modification (R-M) system, which is evident in E. coli B7A [25]. However, the same study also ruled out other known R-M mechanisms, namely the types I, II, III and IV R-M systems [26], implying that the Dpt-Dnd system is likely a new R-M-type mechanism. These two systems which are located in the same genomic island provide selective advantages to the bacterial host and hence may be stably maintained [12,27]. Nevertheless, the three conserved dnd-linked orthologs were not identified in all dnd encoding GIs described in previous studies [10,19] but only in several diverse bacterial strains such as E. coli, Hahella chejuensis, Oceanobacter sp., and Bacillus cereus. This serves to underline not only the diversity of the genetic environment surrounding the various dnd operons but also their possible roles in their respective hosts. Perhaps they are like toxin-antitoxin systems which have multiple biological functions depending on their location in the
Detection of dnd genes and their respective hosts [28] which have adopted and adapted these genetic modules for cellular function. Intriguingly, previously described dnd-encoding GI of Geobacter uranireducens Rf4 was also found to carry putative toxin-antitoxin loci [29]. The presence of the dnd operons in such varied genomic islands in E. coli and their biological implication will hopefully be revealed in the near future.

Conclusion
The designed PCR assay and the use of thiourea in PFGE may improve the detection and typeability of Dnd E. coli strains. Our study showed that genomic islands not only play a potential pivotal role in facilitating the horizontal gene transfer of dnd operons, but also essential in generating diversity within the genomic islands of closely related bacteria. Different dnd-encoding genomic islands with mosaic compositions were identified within E. coli. Since the dnd operon is often associated with mobile genetic elements, the operon may be integrated into genetically diverse strains followed by dissemination. Therefore the Dnd phenotype can be observed in both genetically closely related and diverse strains. The finding that dnd operons were more often found in pathogenic E. coli may provide caveats for future study on their possible linkage.

Methods
Pulsed-field gel electrophoresis
A total of 12 E. coli strains that were reported to be untypeable from previous studies [13,14] were further analysed in this study. Seven of the Dnd+ strains were verotoxigenic E. coli (VT1EC) isolated from different piglets in the same swine farm. The other five were clinical samples obtained from two medical centers (A and B) and from different sources (stool, blood, urine, and swab). The DNA degradation observed in these strains was thought to be Tris-dependent due to the characteristic smearing pattern and their untypeability was confirmed by repeating the PFGE three times. PFGE was carried out on XbaI-digested and Bln-digested genomic DNA of both Dnd+ and Dnd− strains [13] with and without the addition of 50 μM thiourea (Sigma Aldrich, USA) into the 0.5 × TBE buffer and agarose gels [3]. XbaI-digested Salmonella enterica serovar Braenderup H9812 was used as the DNA size marker.

PCR assay detecting dndB, dndC, dndD and dndE genes
A PCR assay targeting the internal sequences of the dndB, dndC, dndD and dndE genes were developed based on the dnd operons of 3 Dnd+ E. coli genome sequences (B7A, 55989 and SE11 with accession numbers CP005998.1, NC_011748 and NC_011415, respectively) using Primer3 (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) [30]. PCR detection of iscS gene was also carried out using primers described previously [11]. PCR detection of dndB, dndC, dndD, dndE and iscS genes was carried out for 12 Dnd+ (which displayed the degradation phenotype) and 48 Dnd− (without degradation phenotype) strains. Five each of Shigella sonnei, Salmonella serovar Typhimurium and Salmonella serovar Enteriditis strains that do not display DNA degradation phenotype were also included as negative controls. The presence of all 4 dnd genes was indicative of the presence of the dnd operon. Strain that did not yield any of the expected amplicons was considered as negative for Dnd genotypes. PCR amplified dndB, dndC, dndD, and dndE genes were sequenced at a commercial facility (First BASE Laboratories).

Association of the DNA degradation phenotypes and genotypes
Fisher’s exact test was used to determine the association of the DNA degradation phenotypes and genotypes using R (version 2.12.2, Vienna, Austria, [www.R-project.org]) [31]. A significance level of p < 0.05 was considered as statistically significant.

Comparative analysis of dnd operons with 1kb of their respective immediate vicinity
The three complete dnd operons of Dnd+ E. coli strains (B7A, 55989 and SE11) available from dndDB (dnd database, http://db-mm.ultimate.edu.cn/dndDB) were used as query sequences because the dnd operons as well as their association with genomic islands (GI) have been reported before. BLASTN against GenBank databases of nr/nt and WGS was carried out with Enterobacteriaceae set as ‘organism of selection’ [32]. The sequences of dnd operons of all positive hits were then extracted.

The immediate vicinity of the dnd operon was also included by using a uniform distance of 1000 bp upstream and downstream from the dnd operon. A distance of 1000 bp was selected because PCR can then be employed to examine the immediate upstream and downstream of the dnd operon of the studied strains. Furthermore, the 1000 bp region should be sufficient to determine the neighboring gene(s) of the dnd operon since an ORF of E. coli generally has an average length of approximately 1000 bp [33]. All positive hits were considered as dnd operon positive although the Dnd phenotypes for most hits have not been reported before. Genetic relationships of the dnd operons and together with their immediate genetic environment were determined by constructing phylogenetic trees using maximum-likelihood (ML) method with MEGA 5. The dnd operons of E. coli genomes were further subgrouped based on their respective immediate upstream and downstream regions as different genetic environments may indicate different evolutionary origins.
Identification, designation and analysis of dnd-encoding genomic islands in E. coli

One E. coli genome was selected from each of the subgroups (based on their respective immediate genetic environment) for genomic island identification. In cases where only draft genomes were available in the subgroup(s), the genome was selected based on the longest assembled contig where the dnd operon is located. Genomic islands harbouring the dnd operon were determined using the VRprofile server (http://bioinfo-mml.sjtu.edu.cn/VRprofile/) and mGenomeSubtractor tool [34]. The common features of a genomic island such as organism-atypical G + C contents, integration into the 3' end of tRNA genes, and integrase-encoding sequences were also analyzed for the putative laterally acquired dnd-encoding GIs. A standardized naming system for dnd-encoding GIs was also proposed in this study to avoid confusion. Using "dndI EC 55989" as an example, "dndI" means "dnd-encoding genomic island"; "EC" means first dnd-encoding GI identified in the respective strain" while "EC 55989" depicts the "strain code". The genetic maps, GC contents of representative genomic regions and dnd-encoding GIs were illustrated using Easyfig [35]. BLASTN comparison between the dnd-encoding GIs were performed with default settings using Easyfig [35]. Homologues search against the GenBank databases of nr/nt and WGS using BLASTN (http://blast.ncbi.nlm.nih.gov/blast) were carried out for ORFs of representative dnd-encoding genomic islands, to determine their match to dnd-negative or dnd-positive bacterial genomes.

PCR determination of immediate genetic environment of dnd operons for 12 Dnd+ E. coli strains

Primers targeting at the immediate upstream and downstream of the dnd operon were also designed based on the 52 sequences extracted from the E. coli genomes. A total of 13 pairs of primers were designed based on the 7 types of genetic environment found for E. coli (Table 2). The

<table>
<thead>
<tr>
<th>Primer</th>
<th>Oligonucleotide sequence (5' to 3')</th>
<th>Target gene</th>
<th>Amplicon size (bp)</th>
<th>Annealing temperature (°C)</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp_a-US-F</td>
<td>AACAACCTGAGGTCAGCGGA</td>
<td>upstream of</td>
<td>489</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_a-US-R</td>
<td>AGGCCTATCGTGCTGCCCAGA</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp_b-US-F</td>
<td>TGGGGTTGGATGCTGAGTCAT</td>
<td>upstream of</td>
<td>723</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_b-US-R</td>
<td>TCAGATGGCTATCGTGGCTG</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp_c-US-F</td>
<td>GAAACCAACCCTTTTACGTC</td>
<td>upstream of</td>
<td>823</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_c-US-R</td>
<td>GAAACCAAGACGTGGCTCAT</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp_d-US-F</td>
<td>AGCAGGAAGCCATGTTTTG</td>
<td>upstream of</td>
<td>905</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_d-US-R</td>
<td>TGCCACTTCAGTGCTGACA</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp_e-US-F</td>
<td>AACGGCAATAGACGCTGTCA</td>
<td>downstream of</td>
<td>518</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_e-US-R</td>
<td>AGCAACCGTCTTCGTTCTGT</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp_f-US-F</td>
<td>ACATCTGCTGGCTACGCTTT</td>
<td>downstream of</td>
<td>715</td>
<td>S2</td>
<td>this study</td>
</tr>
<tr>
<td>gp_f-US-R</td>
<td>TGTTCATGCGGTCTTAGCA</td>
<td>dnd operon</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
primers are summarized in Table 2 and were used to determine the immediate genetic environment for the 12 DndE E. coli studied strains.

Additional files

Additional file 1: Figure S1. Multiple sequence alignment of dnd genes using Clustal W. a) dndB gene; b) dndC gene; c) dndD gene; d) dndE gene. A dot (.) denotes a position at which all query sequences have the same exact nucleotide. Nucleotide differences in reference and query sequences are indicated as underlined and uppercase alphabets, respectively. Numbers indicate the nucleotide sequence positions. Nucleotide sequences for E. coli S599891, SE115 and B7A were retrieved from dndDB [GenBank: NC_011748, NC_011415, CP0005988.1, respectively]. PFPN8-EC12, EC1003-8, EC1004-8, EC062-10 and EC075-10 represent strains in lanes S3, A1, A3, B1 and B2 in Figure 1.

Additional file 2: Table S1. Strains included in phylogenetic analysis of dnd operon. E. coli strain background information is also included.

Additional file 3: Figure S2. Phylogenetic tree of dnd operons together with their respective immediate genetic environment (1kb) from Enterobacteriaceae. Asterisks (*) indicate that the genome contains two copies of dnd operons. Different alphabets (a, b, c, d, e, f, g) depict the subgroups based on the immediate genetic environment. Maximum likelihood (ML) method was used to construct the phylogenetic tree using MEGAS. Bootstrap values have been calculated using 1000 replicates. Bootstrap confidence values greater than 50% are shown at branches. Nodes with less than 50% bootstrap value are collapsed.

Competing interests

The authors declare that they have no competing interest.

Authors’ contributions

WSH designed the study, analysed and interpreted data and wrote the manuscript. WH and HYO performed genomic analysis. WH, HYO, CCY and KLT helped in drafting and critically reviewed the manuscript and contributed important intellectual output. HYO and KLT provided funding for the project. CCY and KLT co-supervised the project while KLT also designed, interpreted and supervised the project. All authors contributed to editing of the manuscript and all authors read and approved the final manuscript.

Acknowledgements

We thank University of Malaya for facilities and support. This work was partially supported by Postgraduate Research Fund (PS318/2013B), High Impact Research Grant-Molecular Genetics (reference UMC/625/1/HIR/MOHE/D02 [A00002-5000] from University of Malaya and MOSTI [GOA03-2013], National Natural Science Foundation of China (9177082, 31130068) and the 973 program, Ministry of Science and Technology, China (2015CB553702). Weng Se Ho is supported by full time research assistantship from HI of University of Malaya.

Author details

1 Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. State Key Laboratory for Microbial Metabolism and School of Life Sciences & Biotechnology, Shanghai Jiaotong University, 200030 Shanghai, China. Biomedical Research Centre, Faculty of Medicine, Universiti Sultan Zainal Abidin, 20400 Kuala Terengganu, Malaysia.

Received: 18 June 2014 Accepted: 28 February 2015

References


Author Query Form

Journal: BMC Genomics
Title: The dnd operon for DNA phosphorothioation modification system in Escherichia coli is located in diverse genomic islands
Authors: Wing Sze Ho, Hong-Yu Ou, Chew Chieng Yeo, Kwai Lin Thong
Article: 1421

Dear Authors,

During production of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions. If you intend to annotate your proof electronically, please refer to the E-annotation guidelines. We recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

<table>
<thead>
<tr>
<th>Query No.</th>
<th>Query</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://www.ncbi.nlm.nih.gov/tools/primer-blast/">http://www.ncbi.nlm.nih.gov/tools/primer-blast/</a></td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://www.R-project.org">www.R-project.org</a></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://db-mml.sjtu.edu.cn/dndDB">http://db-mml.sjtu.edu.cn/dndDB</a></td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://bioinfo-mml.sjtu.edu.cn/VRprofile/">http://bioinfo-mml.sjtu.edu.cn/VRprofile/</a></td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://blast.ncbi.nlm.nih.gov/blast">http://blast.ncbi.nlm.nih.gov/blast</a></td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://www.intechopen.com/books/dna-replication-current-advances/phosphorothioation-an-unusual-post-replicative-modification-on-the-dna-backbone">http://www.intechopen.com/books/dna-replication-current-advances/phosphorothioation-an-unusual-post-replicative-modification-on-the-dna-backbone</a></td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://www.R-project.org">www.R-project.org</a></td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>URL: Please check that the following URLs are working. If not, please provide alternatives: <a href="http://www.ncbi.nlm.nih.gov/books/NBK21136/">http://www.ncbi.nlm.nih.gov/books/NBK21136/</a></td>
<td></td>
</tr>
</tbody>
</table>