Insufficiency of PCR diagnostics for Detection of Diarrhoeagenic *Escherichia coli* in Ibadan, Nigeria

Olabisi C. Akinlabi¹, Rotimi A. Dada², El-shama Q. Nwoko¹, and Iruka N. Okeke¹.

1. Department of Pharmaceutical Microbiology, Faculty of Pharmacy, University of Ibadan, Oyo, Nigeria.

2. Medical Laboratory Science Programme, College of Health Sciences, Bowen University Iwo and Department of Pharmaceutical Microbiology, Faculty of Pharmacy, Ahmadu Bello University, Zaria.

*Correspondence to* iruka.n.okeke@gmail.com.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Understanding the contribution of different diarrhoeagenic *Escherichia coli* pathotypes to disease burden is critical to mapping risk and informing vaccine development. Targeting select virulence genes by PCR is the diagnostic approach of choice in high-burden, least-resourced African settings. We compared the performance of a commonly-used multiplex protocol to whole genome sequencing (WGS).

PCR was applied to 3,815 *E. coli* isolates from 120 children with diarrhoea and 357 healthy controls. Three or more isolates per specimen were also Illumina-sequenced. Following quality assurance, ARIBA and Virulencefinder database were used to identify virulence targets. Root cause analysis of deviant PCR results was performed by examining target sensitivity using BLAST, Sanger sequencing false-positive amplicons, and identifying lineages prone to false-positivity using in-silico multilocus sequence typing and a Single Nucleotide Polymorphism phylogeny constructed using IQTree.

The sensitivity and positive predictive value of PCR compared to WGS ranged from 0-77.8% while specificity ranged from 74.5-94.7% for different pathotypes. WGS identified more enteroaggregative *E. coli* (EAEC), fewer enterotoxigenic *E. coli* (ETEC) and none of the Shiga toxin-producing *E. coli* detected by PCR, painting a considerably different epidemiological picture. Use of the CVD432 target resulted in EAEC under-detection, and enteropathogenic *E. coli* eae primers mismatched more recently described intimin alleles common in our setting. False positive ETEC were over-represented among West Africa-predominant ST8746 complex strains.
PCR precision varies with pathogen genome so primers optimized for use in one part of the world may have noticeably lower sensitivity and specificity in settings where different pathogen lineages predominate.
Introduction

Diarrhoea is a principal contributor to morbidity and mortality in young children, particularly in Africa and parts of Asia. A range of pathogens can produce diarrhoea, including diarrhoeagenic *Escherichia coli* (DEC). *E. coli* is the most common proteobacterial intestinal commensal and this presents a challenge for delineating potentially pathogenic strains, which harbor virulence genes, from those that are beneficial. There are several DEC pathotypes, five of which have been validated in human volunteer challenge studies and/or significant outbreaks: enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), Shiga toxin-producing *E. coli* (STEC) (which includes enterohaemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC) and enteroaggregative *E. coli* (EAEC) (1). Diffusely-adherent *E. coli* (DAEC), cell-detaching *E. coli* (CDEC) and cytolethal distending toxin-producing *E. coli* (CLDTEC) are among pathotypes that have been proposed but require more evidence to validate them. Within each pathotype, there is considerable genetic diversity and different lineages evolved convergently. There are also hybrid strains that carry virulence genes corresponding to multiple pathotypes (2–7).

Traditionally, DEC pathotype identification uses tissue-culture or vertebrate animal-based diagnostics, which are onerous, highly specialized and therefore unsuitable for routine clinical use (8). The Gold Standard for adherent DEC pathotypes (EPEC, EAEC, DAEC and CDEC) has been a tedious epithelial adherence assay, which few labs can perform, while ETEC, STEC and CLDTEC Gold Standards depend on detection of the toxins that these pathotypes produce. EIEC can be identified by detecting entry into epithelial cells or conjunctivitis produced in rodent models. Today, the genes encoding defining phenotypes for each pathotype are known, making it possible to detect DEC using DNA-based methods that offer greater throughput and more
widespread application. PCR for DEC in particular is a low-cost method that can leverage
instrumentation and expertise initially acquired for other purposes and therefore present in most
reference and research laboratories, and in some smaller clinical laboratories. While automated
PCR-based protocols are rapidly gaining ground in high-income settings, the low cost for manual
PCR set up, particularly for end-point PCR, and the flexibility around its consumable procurement
mean that it remains an approach of choice in low-income settings, including Nigeria where we
work. Because of the large number of targets needed to cover the most prominent DEC pathotypes,
and the need to integrate DEC diagnosis into a larger search for diarrhoeal pathogens, multiplex
PCR protocols are favored, even though these are notoriously challenging to optimize (9–11). PCR
is also perceived to have a lower false-positive rate than nucleotide hybridization, which was
popular in the twentieth century but requires somewhat costly and often inaccessible detection
methods. Moreover, hybridization can pick up similar but not identical genes and therefore
confound DEC diagnoses (12,13)

The range of known virulence gene alleles and pathogenic lineages, as well as the absence of a
precise genetic definition for EAEC and newer pathotypes, mean that the more targets that can be
detected, the more sensitive a testing method becomes. For all of these reasons, whole genome
sequencing offers the promise of reliably detection any DEC pathotype, without the need of several
cumbersome assays. Due to the high cost and limited availability of WGS, most labs continue to
use PCR-based tests that were developed and validated in a very limited number of settings against
small panels of strains. We compared the performance of a popular PCR-based protocol that has
been in use for almost two decades to WGS. We then identified root causes of PCR under-
performance in our setting thereby uncovering more general principles for PCR diagnostics failure for pathogens that show considerable diversity.

Materials and methods

Ethical considerations: This study was part of a larger investigation of the epidemiology of diarrhoeagenic *E. coli* among children with diarrhoea presenting to primary health clinics in southwestern Nigeria with ethical approval from the University of Ibadan/University College Hospital ethics committee (UI/EC/15/0093). Patient parents or guardians provided written informed consent to participate. According to the approved protocol, neither PCR nor WGS results, which were performed for surveillance and not obtained in a timely manner, were used to inform patient care.

Review of multiplex PCR methods: We systematically searched Pubmed https://pubmed.ncbi.nlm.nih.gov for “diarrhoeagenic *Escherichia coli*” and “PCR method”. Abstracts retrieved were read to determine whether they included development of a protocol, and if so, where the protocol was developed. Where this was the case, the number of papers citing the reference in question and the location of cited studies was retrieved using Google Scholar https://scholar.google.com.

Strains: Strains used in this study were collected as part of an epidemiological survey for diarrhoeagenic *E. coli* (6). Up to ten isolates were obtained per stool specimen and all isolates were subjected to PCR screening. Up to three isolates showing different biochemical and/or
antimicrobial susceptibility profiles were whole genome-sequenced. Control strains, which are
well-studied and genome-sequenced DEC from other parts of the world, are listed in Table 1.
Table 1: PCR oligonucleotides from the Aranda et al protocol that were used in this study

<table>
<thead>
<tr>
<th>Primer Designation</th>
<th>Function encoded by target gene</th>
<th>Primers (5' to 3')</th>
<th>Target gene or probe</th>
<th>Amplicon size (bp)</th>
<th>Positive control strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>eae f</td>
<td>Intimin of EHEC and EPEC</td>
<td>CTGAACGGCGATTACGCGAA</td>
<td>eae</td>
<td>917</td>
<td>E2348/69</td>
</tr>
<tr>
<td>eae r</td>
<td></td>
<td>CCAGACGATACGATCCAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfp f</td>
<td>Bundle-forming pilus of EPEC</td>
<td>AATGGTGCTTGCGCTTGCTGC</td>
<td>bfpA</td>
<td>326</td>
<td>E2348/69</td>
</tr>
<tr>
<td>bfp r</td>
<td></td>
<td>GCCGCTTTATCCAACCTGGTA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVD432 f</td>
<td>Dispersin transporter</td>
<td>CTGGCGAAAGACTGTATCAT</td>
<td>aatA (CVD432)</td>
<td>630</td>
<td>042,17-2,60A</td>
</tr>
<tr>
<td>CVD432 r</td>
<td></td>
<td>CAATGTATAGAAATCCGCTGTT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT f</td>
<td>Heat liable toxins of ETEC</td>
<td>GGCAGACAGATTATAACGTGC</td>
<td>elt (LT)</td>
<td>450</td>
<td>H10407</td>
</tr>
<tr>
<td>LT r</td>
<td></td>
<td>CGGTCTCTATATTCCCTGTT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST f</td>
<td>Heat stable toxins of ETEC</td>
<td>ATTTTTMTTTCTGTATTRTCTT</td>
<td>est (ST)</td>
<td>190</td>
<td>H10407</td>
</tr>
<tr>
<td>ST r</td>
<td></td>
<td>CACCCGGTACARGCAGGATT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Primer 1</td>
<td>Primer 2</td>
<td>Length</td>
<td>Strain</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>ipaH</td>
<td>invasion-associated locus of EIEC and Shigella</td>
<td>GTTCCTTGACCGCCTTTCCGATACC</td>
<td>GTC</td>
<td>600</td>
<td>E137</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCCGGTCAGCCACCCTCTGAGAGTA</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stx1</td>
<td>Shiga toxin 1 and 2 of EHEC and STEC</td>
<td>ATAAATCGCCATTCGTTGACTAC</td>
<td>AGAACGCCCACTGAGATCATC</td>
<td>244</td>
<td>EDL933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGCATGTCTGAAACTGCTCC</td>
<td>TCGCCAGTTATCTGACATTCTG</td>
<td>190</td>
<td>EDL933</td>
</tr>
</tbody>
</table>

**Notes:**
- CC-BY 4.0 International license
- It is made available under a license which was not certified by peer review.
- The copyright holder for this preprint is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
- This version posted January 9, 2023. 
- doi: 10.1101/2023.01.06.23284276
Virulence Gene Detection by PCR: Molecular diagnosis of DEC was performed by gene amplification using PCR. The *E. coli* genomic DNA was screened by multiplex PCR by targeting the genes as described by Aranda *et al.*, 2004 (9) (Table 1). However, the protocol was modified due to the closeness of the band sizes in Aranda’s PCR2 targets of *est* and *stx2* genes with band sizes 180 and 190 bp respectively (14). The specific modification splits PCR2 two separate reactions, designated PCR3 and PCR 4 (Table 2). Likewise, because of previously observed non-specific products, the annealing temperature of the PCR reactions was increased and the annealing time was reduced for each reaction to avoid previously detected non-specific products (14). The conditions used are as outlined in Table 2. All multiplex PCR was performed using the illustra™ PuReTaq Ready-To-Go™ PCR Beads (GE Healthcare UK.). These contain *Taq* polymerase, BSA, stabilizers, dNTPs, 1.5mM MgCl₂, 50mM KCl and reaction buffer. Each beads is reconstituted to 25 µL by adding 2.5 µl of each forward and reverse primer solution, 20 ng/µl of *E. coli* DNA and nuclease-free water to complete the final volume. Well known prototypical strains were used as positive controls (Table 1) and DH5α and CFT073 as negative controls in the PCR reactions.

Amplicons were visualized on TAE agarose gels containing 1% agarose for PCR 1 and PCR 3 and 1.5% for PCR 4. The volume of agarose used provided 4 mm depth in the electrophoresis tray. Gels were visualized and photographed using the UVP® GelMax Imager which utilizes a built-in midrange 302 nm ultraviolet (UV) transilluminator.
### Table 2: Polymerase Chain Reaction Cycle

<table>
<thead>
<tr>
<th>Step</th>
<th>PCR 1</th>
<th>PCR 2</th>
<th>PCR 3</th>
<th>PCR 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 (9)</td>
<td>2004 (9)</td>
<td>2004 (9)</td>
<td>2004 (9)</td>
<td></td>
</tr>
<tr>
<td>Odetoyin <em>et al.</em>, 2016 (14)</td>
<td>Odetoyin <em>et al.</em>, 2016 (14)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016 (14)</td>
<td>2016 (14)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target genes</th>
<th>aatD, eae and elt, est, ipaH, bfpA, stx1 and stx2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hot start</td>
<td>94°C 3min</td>
</tr>
<tr>
<td>2. Denaturing</td>
<td>94°C 30 sec</td>
</tr>
<tr>
<td>3. Annealing</td>
<td>60°C 30 sec</td>
</tr>
<tr>
<td>4 Extension</td>
<td>72°C 1 min</td>
</tr>
<tr>
<td>5. Cycling</td>
<td>Steps 2-4</td>
</tr>
<tr>
<td></td>
<td>repeated 40 times</td>
</tr>
<tr>
<td>6 Terminal extension</td>
<td>72°C 7 min</td>
</tr>
</tbody>
</table>
Whole genome Sequencing: E. coli genomic DNA were extracted using the Wizard Genomic Extraction kit (Promega) according to the manufacturer’s instructions and they were shipped and bulk sequenced in Sanger using the illumina platform as described previously (6). Raw data were assembled by SPAdes assembler and assembly quality was determined using Quality Assessment Tool for Genome Assemblies (QUAST), and CheckM. ARIBA and Virulencefinder database, were used to identify virulence genes. Intimin allele calling was performed by comparing database prototypes to assembled genomes using BLAST. Sequence data were submitted to ENA and are available from ENA https://www.ebi.ac.uk/ena/browser/home and Genbank https://www.ncbi.nlm.nih.gov/genbank/ as Bioproject PRJEB8667.

PCR amplicon sequencing: Selected isolates that amplified the virulence genes by PCR but did not harbor such genes by WGS were reamplified and amplicons were Sanger sequenced at Inqaba Biotec (Ibadan, Nigeria and Pretoria, South Africa). Returned sequences yielding single peaks were compared to the NCBI nucleotide database using BLAST to retrieve the match and flanking sequences. Primer sequences were then compared for within amplicon/ flanking sequence matches.

Mapping PCR diagnostic results onto phylogenies.

All sequenced isolate reads were mapped to EAEC_042 reference genome (accession number: FN554766) to create a multi-fasta alignment file. Using the multi-fasta alignment file, Single Nucleotide Polymorphism (SNP) were called using SNP-sites and the SNP file was used to construct a phylogenetic tree using IQTree software (15). Microreact (16) was used to visualize phylogenetic trees, STs and loci identified by PCR and WGS.
**Statistical Analysis:** Epi Info version 7 software (Centers for Disease Control and Prevention, Atlanta, GA, USA) was used for statistical calculations. Fisher’s exact test and Chi square test were used to test the association of test-defined pathotypes with disease. p-values < 0.05 were considered significant, with Bonferroni corrections used where applicable.

**Results**

**Multiplex PCR for DEC is rarely developed in Africa but commonly used to identify DEC in African settings:** A search for “diarrheagenic Escherichia coli” and “PCR method” on Medline (Pubmed) initially yielded 207 hits, which when reviewed uncovered 16 studies that developed multiplex PCR protocols for diarrhoeagenic *E. coli* (Table 1). In contrast to some other diagnostic protocols, where one or a few methods gain prominence over time, there are a wide variety of published PCR protocols for DEC with 7 of those published before 2010 having 1,132 or more citations (Table 3). A similar number of studies developed after 2010 are accumulating citations as well. Five protocols were developed and published from Europe, four protocols were developed by groups working in Asian countries and three by South American groups. Only two of the protocols were published by groups based in Africa (both in Egypt) and none in Nigeria, West Africa or elsewhere in the WHO Africa region (Table 3). While target choice has been compared across different protocols (17), we were not able to find studies that included a validation on isolates from the WHO-AFRO region nor any studies comparing different PCR protocols with one another.
Table 3: Application of published DEC PCR identification protocols on the African continent

<table>
<thead>
<tr>
<th>Protocol (Reference)</th>
<th>Developed in</th>
<th>#Google Scholar citations as at 1 Dec 2022</th>
<th># Citations from African use study</th>
<th>Location of African studies citing the protocol*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toma et al 2003(10)</td>
<td>Japan</td>
<td>297</td>
<td>11</td>
<td>Cote d’Ivoire, Kenya, Nigeria, Tanzania, South Africa, Uganda</td>
</tr>
<tr>
<td>Aranda et al 2004(9)</td>
<td>Brazil</td>
<td>294</td>
<td>18</td>
<td>Ghana, Kenya, Nigeria, South Africa</td>
</tr>
<tr>
<td>Vidal et al 2005(18)</td>
<td>Chile</td>
<td>301</td>
<td>24</td>
<td>Bukina Faso, Chad, Ethiopia, Egypt, Ghana, Kenya, Nigeria, South Africa and multi-country</td>
</tr>
<tr>
<td>Muller et al 2007(19)</td>
<td>Germany</td>
<td>56</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Metwally et al 2007(20)</td>
<td>Egypt</td>
<td>22</td>
<td>2</td>
<td>Egypt, Nigeria</td>
</tr>
<tr>
<td>Antikainen et al 2009(8)</td>
<td>Finland</td>
<td>93</td>
<td>15</td>
<td>Bukina Faso, Chad, Malawi, South Africa</td>
</tr>
<tr>
<td>Tobias et al 2012(17)</td>
<td>Sweden</td>
<td>69</td>
<td>2</td>
<td>Ghana, South Africa</td>
</tr>
<tr>
<td>Hegde et al 2012(21)</td>
<td>India</td>
<td>70</td>
<td>4</td>
<td>Burkina Faso, Ethiopia, Nigeria, Sudan</td>
</tr>
<tr>
<td>Fialho et al 2013(22)</td>
<td>Brazil</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Country</td>
<td>Cases</td>
<td>Malaria Countries</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------</td>
<td>-------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>Helmy et al 2013</td>
<td>Egypt</td>
<td>7</td>
<td>Nigeria</td>
<td></td>
</tr>
<tr>
<td>Al Talib 2014</td>
<td>Malaysia</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oh et al 2014</td>
<td>Korea</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Onori et al 2014</td>
<td>Italy</td>
<td>80</td>
<td>Ghana, South Africa, Sudan</td>
<td></td>
</tr>
<tr>
<td>Naziri et al 2015</td>
<td>Iran</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sjoling et al 2015</td>
<td>Sweden</td>
<td>37</td>
<td>Burkina Faso, Cote D’Ivoire, Ghana, Kenya, South Africa</td>
<td></td>
</tr>
<tr>
<td>Zhang et al 2020</td>
<td>China</td>
<td>15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within Africa, we found clear geographic bias in protocol selection that could not, from the aforementioned, be attributed to validation, among studies seeking DEC in human, food or environmental specimens (Table 3). Some well-cited protocols (17,19,26,29) have rarely or never been used in Africa. The Vidal et al 2005 (18) protocol has been used in the most African countries but not more than two or three times in any single country. This study evaluated the Aranda et al (2004) (9) protocol, which uses a minimal primer set to identify all five pathotypes that have been validated through human volunteer studies and/or outbreaks. This was the most commonly employed protocol in Nigeria and some of its primer pairs are employed in other popular multiplex protocols (17,22,30). The protocol originally presented by Aranda et al (2004) (9) consists of two multiplex PCR reactions that detected a total of eight targets, all validated DEC pathotypes. Noting that the amplicon sizes for \textit{stx1} (180 bp) \textit{elt} (190 bp), we previously separated PCR2 into two reactions to avoid errors of interpretation (31).

The epidemiological picture for the same specimen set is radically different when DEC are identified by PCR and WGS: The Aranda et al (2004)(9) PCR performed well for all positive and negative control strains, which were included in every batch, reproducibly giving the expected result every time. \textit{E. coli} from 413 specimens obtained from children in Ibadan, totaling 3,815 isolates and averaging eight isolates per specimen, were tested by PCR. All PCR positives were sequenced and in instances where there were <2 such positives from any specimen, three isolates with different biochemical profiles were selected for sequencing. Thus, due to resource limitation, a smaller number of 1,230 (averaging three per specimen) were tested by whole genome sequencing but all pathotypes identified by PCR were sequenced for validation. As shown in Table
Table 4 shows that PCR detected all five pathotypes and identified ETEC 150 (31.4%) as the most common pathotype (31 est-positive strains, 206 elt-positive strains and 67 strains having both targets), followed by CVD432-positive EAEC 99(20.8%), STEC positive stx1 and/or stx2 46(9.6%), eae and/or bfp-positive EPEC 43(9%) and EIEC from which ipaH amplified 33(6.9%). Thus PCR detected all five DEC pathotypes as shown in Figure 1 and ETEC with est and elt genes or elt alone were significantly associated with diarrhoea (p< 0.05), based on PCR results. The stx1 and stx2 genes were detected in 12 and 63 isolates, respectively, by PCR. WGS did not verify any of the E. coli strains as harboring a Shiga toxin gene. Thus, PCR identified 70 isolates that could be classified as STEC but none of the PCR-identified ‘STEC’ were verified by whole genome sequencing. No EHEC (eae-positive STEC) were identified by either method.

We compared sensitivity and specificity of PCR to WGS for each of the pathotypes. The CVD432 (aatA) gene was targeted for EAEC, eae and bfp for EPEC, elt and est for ETEC, ipaH for EIEC, stx1 and stx2 for STEC, Table 5 shows that PCR sensitivity for DEC identification was 77.8% for ETEC, 61.7% for EAEC, 45% for EPEC and 0% for both EIEC and STEC. PCR did show better specificity for EAEC (91.6%), EPEC (94.7%), EIEC (93.7%), STEC (90.1%). ETEC specificity was considerably lower (74.5%) (Table 5).
Table 4: DEC pathotypes identified by WGS and PCR from the case and control specimens

<table>
<thead>
<tr>
<th>Pathotypes</th>
<th>Cases (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=120</td>
</tr>
<tr>
<td></td>
<td>Controls (%)</td>
</tr>
<tr>
<td></td>
<td>n=357</td>
</tr>
<tr>
<td></td>
<td>Total specimens</td>
</tr>
<tr>
<td></td>
<td>n=477</td>
</tr>
<tr>
<td></td>
<td>P-values</td>
</tr>
<tr>
<td></td>
<td>PCR</td>
</tr>
<tr>
<td>EAEC (CVD432)</td>
<td></td>
</tr>
<tr>
<td>25(20.8) 35(29.2)</td>
<td></td>
</tr>
<tr>
<td>EHEC/STEC</td>
<td></td>
</tr>
<tr>
<td>9(7.5) 0(0)</td>
<td></td>
</tr>
<tr>
<td>stx2</td>
<td>8(6.7)</td>
</tr>
<tr>
<td>stxl</td>
<td>1(0.8)</td>
</tr>
<tr>
<td>stx1, stx2</td>
<td>1(0.8)</td>
</tr>
<tr>
<td>EPEC</td>
<td></td>
</tr>
<tr>
<td>15(12.5) 7(5.8)</td>
<td></td>
</tr>
<tr>
<td>eae+ bfp+</td>
<td>1(0.8)</td>
</tr>
<tr>
<td>eae</td>
<td>1(0.8)</td>
</tr>
<tr>
<td>bfp</td>
<td>6(5)</td>
</tr>
<tr>
<td>ETEC</td>
<td></td>
</tr>
<tr>
<td>41(34.2) 8(6.7)</td>
<td></td>
</tr>
<tr>
<td>elt+est+</td>
<td>12(10)</td>
</tr>
<tr>
<td>elt</td>
<td>26(21.7)</td>
</tr>
<tr>
<td>est</td>
<td>5(4.2)</td>
</tr>
<tr>
<td>EIEC/Shigella (ipaH)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0(0)</td>
</tr>
</tbody>
</table>
Table 5: Evaluation of PCR performance for identification of diarrhoeagenic *E. coli* compared with sequence data.

<table>
<thead>
<tr>
<th></th>
<th>EAEC</th>
<th>ETEC</th>
<th>EPEC</th>
<th>EIEC</th>
<th>STEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Sensitivity</td>
<td>61.7</td>
<td>77.8</td>
<td>45.0</td>
<td>0.0</td>
<td>ND*</td>
</tr>
<tr>
<td>% Specificity</td>
<td>91.6</td>
<td>74.5</td>
<td>94.7</td>
<td>93.7</td>
<td>90.1</td>
</tr>
<tr>
<td>Positive predictive value (%)</td>
<td>73.9</td>
<td>10.7</td>
<td>27.3</td>
<td>0.0</td>
<td>ND*</td>
</tr>
<tr>
<td>Negative predictive value (%)</td>
<td>86.1</td>
<td>98.8</td>
<td>97.5</td>
<td>99.8</td>
<td>100</td>
</tr>
</tbody>
</table>

*There were no STEC true positives*
Low sensitivity of EAEC detection by PCR is due to too few targets: The Aranda et al (2004) protocol, like most other DEC multiplex protocols, seeks only one EAEC target, CVD432. Whole genome sequencing on the other hand allows the investigator to screen strains for a total of 14 loci associated with EAEC (6,32), allowing for greater sensitivity and accounting for the extremely low sensitivity of PCR. When we compared the prevalence of CVD432 detected by PCR to that from WGS among isolates that were tested for both. PCR identified 260 (21%) EAEC isolates while WGS identified 355(28.7%).

Reduced specificity for bfpA PCR and reduced sensitivity of eae primers: A prominent red flag in the PCR results was the 32 strains that repeatedly tested positive for bfpA but were eae negative (Figure 1). Typical EPEC strains harbor the chromosomal locus of enterocyte effacement (LEE) marked by eae and the EPEC adherence factor (EAF) plasmid-borne bfp genes whilst atypical EPEC have eae on the LEE without bfp and the EAF plasmid. Strains carrying bfp without eae are rarely reported. In this study, all 40 isolates from which bfpA was amplified did not have the gene as determined by WGS. This includes seven eae-positive strains that were misclassified as EPEC by PCR. A total of 19 atypical EPEC strains (eae-positive and bfpA-negative) were not identified by PCR. WGS and screening by VirulenceFinder revealed that 17 of these strains carried all other LEE markers. We therefore hypothesized that they may carry intimin alleles that are not amplified by the eae primer pair. We used the eae gene from prototypical EPEC strain E2348/69 to retrieve eae from each of the assembled genomes of these strains using BLAST. BLAST was then used to type each intimin allele against database prototypes (Table 6). We additionally examined each allele for complementarity to the eae1 and eae2 primers. The results of these analyses are summarized in Table 6 and show that while seven of the strains sequenced carried
intimin alpha, kappa and lambda alleles with no mismatches in the primer annealing sites, and therefore should have amplified, ten strains carried the iota, rho and theta intimin alleles, which have between one and three mismatches with the primers, which have the possibility of reducing amplification at conditions optimal for other alleles. Of note is the fact that in addition to the Aranda et al (9) protocol, these eae primers also feature in multiplex PCRs for DEC by Nazari et al, Puno-Sarmiento et al, and Dias et al (33–35) and so this problem is likely to occur with some other multiplex protocols.
Table 6: Intimin alleles of LEE-positive EPEC strains not detected by PCR.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Intimin allele</th>
<th>Prototype accession number</th>
<th>Sequence of \textit{eae1} and \textit{eae2} priming sites (mis-matched bases are not in bold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2348/69 (control)</td>
<td>alpha (28)</td>
<td>AF022236.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LKD064A</td>
<td>rho</td>
<td>DQ523613.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLH22C</td>
<td>alpha</td>
<td>AF022236.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLH35B</td>
<td>alpha</td>
<td>AF022236.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLH012D</td>
<td>kappa</td>
<td>KT591304.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLD100C</td>
<td>iota2</td>
<td>AB647371.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLH332A, LLH332B, LLH332E</td>
<td>epsilon</td>
<td>AB647573.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LL197, LLH161I, LLH161F</td>
<td>lambda</td>
<td>AJ715409.1</td>
<td>CTGAACGGCGATTACGCGAA\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
<tr>
<td>LLH336B, LLH336E, LLD104C, MNH109F, LLHO33E, LLH261C1</td>
<td>theta2</td>
<td>FM872418.1</td>
<td>TG\ldots\text{880 nt}\ldots\text{CTGGATCGTATCGTCTGG}</td>
</tr>
</tbody>
</table>
False positivity with elt primers is due to non-specific amplification in unfamiliar lineages:
The ETEC pathotype is well-defined with highly conserved virulence targets, the elt and est genes. It is also a vaccine priority and there is therefore considerable interest in understanding ETEC epidemiology in lesser surveilled parts of the world, including Nigeria. In this study, false positives for most genes were scattered across the phylogeny (Figure 2). This was true to some extent for elt. However, more elt false-positivity was seen within phylogroup D and in particular in the ST720 and ST69 complexes complex. ST69 is a well-characterized ST complex that includes extraintestinal and enteroaggregative E. coli strains. False-positive ETEC were detected in ST 69 (10/31 isolates) and ST394 (4/12 isolates). The false positivity rate of 14/50 was significantly greater then the overall false positivity rate (p=0.03). ST720 complex is less well studied and in this study comprised isolates belonging to ST 8746 (n=51), ST 720 (n=13), ST8749 (n=8), ST8131 and ST1136 (4 each), with the highest number of isolates and greatest false-positive rates in ST8646 (17/51) and ST720 (6/13). Altogether 17/51 ST8746 strains gave a false-positive elt result, as compared to 213 (92.6%) overall (p= 0.0006) and the false-positivity rate in ST720 (6/13) was even greater. Of the 17 ST8746 complex strains that were erroneously identified as ETEC by PCR, all but three were in fact EAEC strains (Figure 2). Thus, the PCR method for this same lineage simultaneously overestimated the prevalence of ETEC, underestimated EAEC and hid from existence a previously unreported EAEC lineages. STs are numbered chronologically with STs in the Wirth et al (2006) (36) Achtman E. coli scheme above 1500 recorded after the year 2010. Thus, most STs within the complex, including ST8746, are relatively new to science. Of interest is that ST8746 itself was the fourth commonest ST we detected in this study, showing that it is very common in our setting. One strain in the complex, an ST720 strain from Germany, had been recorded in the database as at 2010. The relatively recent registration of other isolates in the
ST complex is likely due to the very few sequence-based studies that take place in Nigeria and nearly locales. Other ST8746 strains have been reported from Ghana (n=1) and United Kingdom (n=1) (37,38) but the largest existing collection is from our own study. The data point to the likelihood that false positives with the elt primers will be common in Nigeria and other settings where ST8746 organisms predominate but likely less so in areas where this lineage is absent or less prominent. They also illustrate the weakness of PCR methods primarily used in West Africa that are not developed here.

The stx multiplex produces a confounding non-specific amplicon due to mis-priming

In previous studies, we have de-complexed Aranda et al (2004)’s second multiplex, which was originally designed to contain five primer pairs (14). Thus, we use one multiplex to detect the elt, est and ipaH genes for ETEC and EIEC and a second duplex PCR for stx1 and stx2. The stx duplex PCR reaction identified 12 stx1- and 63 stx2-positive strains. Upon WGS analysis, none of these isolates, nor any of the other isolates sequenced in the study carried these genes. We sought to understand why PCR yielded a right-sized product in the stx duplex and attempted Sanger sequencing of a few amplicons, most of which failed due to multiple peaks suggesting that each band consisted of multiple amplicons. We did get a product for the stx duplex PCR4 that was presumed to be the right size for stx1 from two strains. Flanking a 192 bp region of the amplified sequence are low complexity regions with good partial matches for the stx1-forward and stx2-reverse primers providing a rational explanation of how a ‘190 bp’ amplicon could be produced in a multiplex reaction. The matching sequence itself represents part of the gene encoding a DNA injecting protein of a temperate bacteriophage Gally Phage LF82_763 (Accession number OV696608), for which several E. coli in the database are lysogenized.
Discussion

PCR performance for DEC identification is poor

Diarrhoeagenic *E. coli* (DEC) are endemic in Africa where a wide range of lineages produce life-threatening infections in young children (39,40). The majority of clinical and research laboratories in West Africa are unable to detect these pathogens, therefore available evidence of their importance comes largely from short-term epidemiological surveys. Of those labs that do diagnose DEC infections, or perform surveillance, multiplex PCR is typically the technique of choice with multiple protocols published in the literature. PCR remains popular for DEC identification here and elsewhere in the world and most of the available information on DEC epidemiology comes from studies employing different PCR methodologies (8,9,42–47,10,11,17,18,29–31,41), PCR precision may vary with pathogen genome so that primers optimized for use in one part of the world may not have sufficient precision in geographies where different pathogen lineages predominate. However, there does not appear to be evidence that any have been optimized for use in West Africa where the pathogen lineage repertoire may differ from the rest of the world.

In this study, we found that the Aranda et al (2004) (9), which performed well in its initial assessment on isolates from Brazil, has been used extensively in Nigeria, including by us, and has limited sensitivity and specificity. This is one of several multiplex PCR protocols, some of which have been well validated, but not on local African pathogenic lineages. It also includes primer pairs for EAEC, ETEC and EPEC that are used in other multiplex protocols. In our 2016-2019 Ibadan, Nigeria setting, the Aranda et al (2004) (9) protocol under-estimated the prevalence of EAEC to some degree because of lowered sensitivity of the CVD432 primer for their target and to a much larger degree because too few EAEC targets are included in the multiplex. PCR
overestimated the prevalence of ETEC and inferred an association with disease because of non-
specificity of ETEC primers and resulted in a locally predominant EAEC lineage being
misclassified as ETEC. PCR led to misidentification of strains belonging to other pathotypes or
commensals as Shiga toxin-producing \textit{E. coli}, which are reportable, again because of primer non-
specificity. Typical EPEC numbers were also inflated, and atypical EPEC numbers diminished
due to mis-priming. Overall PCR methodology offers reasonable sensitivity or specificity, but not
both, for identifying some DEC. This has implications for the use of this method in clinical
diagnosis and for interpreting the results of epidemiological surveys (48). As PCR deficits are
misinforming on specific lineages (such as ST720 complex) or strains carrying specific virulence
gene alleles (e.g. intimin theta), the overall importance of these clades or factors is likely to be
overlooked and could be significant. For example, intimin theta is among newer alleles reputed to
confer enhanced invasion on atypical EPEC (49).

That PCR is not as sensitive as WGS is not a surprise. However, when the departure in results that
PCR generates is widely deviant from reality, the value of PCR testing must be questioned. We
note with interest that under-performance of PCR could be attributed to different root causes for
the different pathotype targets. The insensitivity diagnostic protocols that target CVD432 alone is
a generalized problem with the pathotype and its current definition. The CVD432 locus, now
known as \textit{aatD} encodes part of a secretion system for the antiaggregation protein, Aap, and is
present in only some EAEC. It is common among EAEC strains from some locations, for example
the Chile and India strain sets on which that the “EAEC probe” Baudry et al (1990) was validated
(50). However subsequent data has shown that CVD432-negative EAEC strains predominate in
some geographies, including Nigeria (6,51). The insensitivity of CVD432-based PCR protocols
and could be resolved, at least in part by using additional targets as Czeuczulin et al and Boisen et al (52,53) have recommended. Including more primer pairs will add to the cost of detecting EAEC and, if they are multiplexed, could add risk of mis-identifications unless rigorous validation is performed.

Sensitivity was even lower for EPEC than EAEC, although EPEC are less common overall. In this case, the number of targets cannot be faulted since EPEC strains that have other LEE genes typically have eae. However, we find primer mismatches in some EPEC intimin alleles that were described after the original report of intimin allelic variation by Adu-Bobie et al (1998)(28). While the divergence of these alleles from other intimins has been reported previously (54), potential mispriming at consensus eae priming sites has not.

Unlike EAEC, for which the absence of a precise genetic definition requires multiple targets to be used to identify strains from this pathotype (6,32), ETEC and STEC have genetic definitions based on the presence of toxin genes that encode the virulence factors producing the disease. ETEC are vaccine priority organisms and STEC are reportable. Therefore, respective ETEC and STEC specificities in this study of 74.5% and 90.1%, respectively, are below what is needed to inform interventions and contain these pathogens.

The major advantage of whole genome sequencing is that most DEC pathotypes can be predicted with a high degree of accuracy from sequence data (55). The obvious deterrent for routine use is cost. When isolate-based investigation of pathogens in diarrhoea stools is conducted, multiple genomes will have to be sequenced per specimen, thus WGS DEC studies are particularly
expensive. It is important for resource-constrained labs in endemic areas to be able to pick up key
pathogens and enabling this while sacrificing some sensitivity and specificity is justifiable. Imperfect diagnostics can be useful for identification of sporadic and outbreak cases and, to some
degree for epidemiologic assessments. However, it is clear that some benchmarking or validation
is necessary before reliable conclusions can be drawn from DEC PCR. Some popular protocols
have been validated on less than 10 strains per pathotype (18), which is far lower than the number
of lineages known at the time the test was developed, a number that continues to grow. DEC
identification by PCR is inexpensive and popular in resource limited settings in Africa. However,
an established DEC multiplex PCR protocol is under-performing in our setting. Based on the
results of this study, we would argue that the complete lack of specificity for the vaccine prioritized
pathotype ETEC and misidentification of STEC and EHEC, which are reportable foodborne
pathogens make this methodology, by itself, cost-inefficient. Molecular microbiology methods
that are performed simply because they can be, but are unable to achieve the aims for which they
were developed outside of their originally developed context are rightly currently being phased
out (56). Unfortunately, there are few alternatives to end-point PCR-based DEC diagnostics that
can be used broadly across West Africa, at least until WGS becomes more affordable and better
disseminated. In the interim when existing multiplex PCR protocols are applied in settings
different from those used to develop them, they be used alongside other methodology to verify
outcome.

**Conclusion**

Pathotype lineages vary with geographic location, thus PCR sensitivity may vary based on
geographic depending on the predominating pathotype lineage. Multiplex PCR is an inexpensive
option for identifying DEC pathotypes but cannot be relied upon to perform optimally in settings
where lineages different from those on which protocols are optimized for are predominant. Where
possible, epidemiological surveys for DEC should be whole-genome sequence-based. In the
instances that resource constraints do not permit this additional validation steps will be required to
ensure that target loci for DEC identification are appropriate and that presumptive DEC identified
by PCR indeed harbor the target genes used to define them.

Acknowledgment

We thank David A. Kwasi, Anderson O. Oaikhena, Taiwo Badejo, Anthony Underwood, Ayorinde
Afolayan, Erkison Odih, Catherine Ladipo and A Oladipo Aboderin for technical assistance and
helpful comments.

This work was supported by an African Research Leader’s Award to INO, GD and NRT jointly
funded by the UK Medical Research Council (MRC) and the UK Department for International
Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2
programme supported by the European Union (Award #: MR/L00464X/1). INO is a Calestous
Juma Science Leadership Fellow supported by the Bill and Melinda Gates Foundation (INV-
036234).


43. Jerse AE, Kaper JB. The eae gene of enteropathogenic Escherichia coli encodes a 94-kilodalton membrane protein, the expression of which is influenced by the EAF plasmid. Infect Immun [Internet]. 1991 Dec;59(12):4302–9. Available from:
44. Onanuga A, Igbeneghu O, Lamikanra A. A study of the prevalence of diarrhoeagenic
Escherichia coli in children from Gwagwalada, Federal Capital Territory, Nigeria. Pan Afr

45. Ali MMM, Ahmed SF, Klena JD, Mohamed ZK, Moussa TA, Ghenghesh KS.
Enteroaggregative Escherichia coli in diarrheic children in Egypt: molecular
characterization and antimicrobial susceptibility. J Infect Dev Ctries. 2014 May
14;8(05):589–96.

and Molecular Characterisation of Pathogenic Escherichia coli. Pathog Escherichia coli

47. Okeke IN, Ojo O, Lamikanra A, Kaper JB. Etiology of Acute Diarrhea in Adults in

48. Jesser KJ, Levy K. Updates on defining and detecting diarrheagenic Escherichia coli
https://journals.lww.com/10.1097/QCO.0000000000000665

Invasiveness as a putative additional virulence mechanism of some atypical
Enteropathogenic Escherichia coli strains with different uncommon intimin types. BMC
Microbiol [Internet]. 2009 Dec 21;9(1):146. Available from:

Probe to Identify Enteroaggregative Escherichia coli, a Recently Discovered Diarrheal


Figure Legends

Figure 1: A selection of multiplex PCR gels (a) Multiplex PCR1 with lane L, 1-kb size ladder, Lane 2, EAEC 042 reference control, Lane N, negative control, lane 3-4 E. coli strains INOLLD97B and INOLLH029A positive for eae gene, lane 6, 11, 12, positive isolated E. coli strains with bfp gene, Lane 7, isolated E. coli strains with the bfp and eae genes, lane 10, EPEC E2348/69 reference control with the bfp and eae genes, lane 13 and 14 E. coli strains INOLWH056C and INOLLH297C positive for CVD432 gene (b) Multiplex PCR3 with lane L, 1-kb size ladder, Lane N, negative control, Lane 1, EIEC E137 reference control with ipaH gene, lane 2 ETEC H10407 reference control with the elt and est genes, lane 3-4 isolates INOMNH199C and INOLLD018D E. coli strains positive for est gene. (c) Multiplex PCR4 with EHEC EDL933 reference controls (Lane C). Lane 1-5, strains INOMND61D, INOLLH045A, INOLLH030F, INOMNH077E and INOLKH217A with stx2 genes, Lane L, 1-kb size ladder.

Figure 2: Maximum likelihood SNP tree of genomes used in this evaluation showing sequence type, results for stx1, stx2, stx1+2 PCR positivity; elt, est and elt+est PCR and WGS positivity, and CVD432 PCR and WGS positivity. The figure can be engaged interactively at https://microreact.org/project/pWbLpY3qhV6Npa9q8Rpgrk-wgs-n-pcr. While there was some concordance for CVD432, there was little for the STEC and ETEC toxins. Nodes for the ST720 and ST69 complexes in which elt false positives are over-represented are indicated with arrows. (b) False and true positives, and negatives, for elt overall and for ST720 and ST69 complexes.

Figure 3: Amplicon sequence with matches >6 contiguous base pairs on any stx primer highlighted. (b) Corresponding highlighted regions in stx primers
Figure 1
Figure 2

A

ST720 complex
ST69 complex
All genomes

0% 20% 40% 60% 80% 100%

elt true positives  elt false positives
elt true negatives  elt false negatives
Figure 3

A
>stxmultiplexamplicon [LWD15C]
cattctgaacctgctcaagacgtttcctgcagcagcagcaattctgtgagttttcgtttaat
getgagttcagatatagtgcggtagttgcttactcagcaaggcattttccaccaagtgg
ctttctcgtctctgtgggttttgccttctcatccgccttgtgcgagccagccagggatattttctcg
tagtcgcggccatt

B
Stx1F ATAAATCGCCATT CGTTGACTAC
Stx1F In reverse complement GTAGTCAACGAGATGCGATTAT
stx2R TGCCAGTTATCTGACATTCTG