BRIEF REPORT

REVISED Large scale enterohemorrhagic *E coli* population genomic analysis using whole genome typing reveals recombination clusters and potential drug target [version 3; peer review: 2 approved]

Previously titled: Pangenome guided pharmacophore modelling of enterohemorrhagic *Escherichia coli sdiA*

DJ Darwin Bandoy

Department of Veterinary Paraclinical Sciences, University of the Philippines Los Baños, Los Baños, Laguna, 4031, Philippines

v3 **First published:** 09 Jan 2019, **8**:33
https://doi.org/10.12688/f1000research.17620.1

Second version: 01 Oct 2019, **8**:33
https://doi.org/10.12688/f1000research.17620.2

Latest published: 01 Sep 2020, **8**:33
https://doi.org/10.12688/f1000research.17620.3

Abstract

Enterohemorrhagic *Escherichia coli* continues to be a significant public health risk. With the onset of next generation sequencing, whole genome sequences require a new paradigm of analysis relevant for epidemiology and drug discovery. A large-scale bacterial population genomic analysis was applied to 702 isolates of serotypes associated with EHEC resulting in five pangenome clusters. Serotype incongruence with pangenome types suggests recombination clusters. Core genome analysis was performed to determine the population wide distribution of sdiA as potential drug target. Protein modelling revealed nonsynonymous variants are notably absent in the ligand binding site for quorum sensing, indicating that population wide conservation of the sdiA ligand site can be targeted for potential prophylactic purposes. Applying pathotype-wide pangenomics as a guide for determining evolution of pharmacophore sites is a potential approach in drug discovery.

Keywords
pangenome, pharmacophore, EHEC, Escherichia coli

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 3 (revision) 01 Sep 2020		✓ report
version 2 (revision) 01 Oct 2019	✓ report	? report
version 1 09 Jan 2019	✗ report	? report

1. **Kerry K. Cooper**, University of Arizona, Tucson, USA

2. **Olivier Tenaillon**, French Institute of Health and Medical Research (INSERM), Paris, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: DJ Darwin Bandoy (drbandoy@up.edu.ph)

Author roles: Bandoy DD: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was funded by the University of the Philippines Enhanced Creative Work and Research Grant.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Bandoy DD. **Large scale enterohemorrhagic *E coli* population genomic analysis using whole genome typing reveals recombination clusters and potential drug target [version 3; peer review: 2 approved]** F1000Research 2020, **8**:33 https://doi.org/10.12688/f1000research.17620.3

First published: 09 Jan 2019, **8**:33 https://doi.org/10.12688/f1000research.17620.1

REVISED **Amendments from Version 2**

Version 3 contains the McDonald-Kreitman test calculations to determine the nonsynonymous to synonymous mutations ratios.

Any further responses from the reviewers can be found at the end of the article

Introduction

One of the more prominent strains of *Escherichia coli* is the enterohemorrhagic *E. coli* (EHEC) pathotype associated with global outbreaks of bloody diarrhea and hemolytic uremic syndrome (HUS) usually by consumption of undercooked beef[1]. Within the cattle reservoir, sdiA gene is required by *E. coli* to survive within the acidic rumen environment. SdiA is used by *E. coli* to sense acyl homoserine in a quorum sensing system[2]. However, it is considered as an orphan as the cognate acyl homoserine synthase is absent, and hence sdiA is considered an environmental sensor to sense the nearby microbial community. SdiA is stabilized by acyl homoserine lactone and acts as transcription factor glutamate decarboxylase needed for survival in the acidic environment. Hence blocking the ability of EHEC to survive the acidic ruminal environment is a proposed mechanism to control shedding in the cattle reservoir.

Whole genome sequencing of bacterial pathogens, particularly EHEC, is quickly transforming the workflows of epidemiological investigations. However, most bioinformatic pipelines used in clinical investigation perform data reduction of genomes and artificially reduce diversity due to comparison of a limited number of housekeeping genes[3]. While wgMLST attempts to increase the number of genes for analysis, the assignment of a single reference genome appears to be inadequate in light of the pangenome. Various studies have shown that a significant number of genes that are present to the entire universe of genes within a species are missed for variant calling if only a single reference gene is used[4]. In this study, a multi-scale approach was applied to generate genome wide clustering using the entire pangenome, composed of the core genome and the accessory genome via variable k-mers[5]. This approach allows differentiation between clusters as well as within serotypes, which is a limitation of using low resolution techniques like MLST.

The concept of the pangenome, which represents the entirety of the genes that are present within a species, which can also be adjusted to the pathotype level, was applied in this particular study. The EHEC pangenome represents the combination of genes seen in the EHEC pathotype. While a prior pangenome of *E. coli* contained 17 genomes, I generated and updated EHEC pangenome with 702 genomes, representing the largest population wide whole genome comparison to date[6]. The pangenome enables clustering of isolates using gene presence and absence. Targetting the core genome, represented in this study by sdiA, enables integration of population genomics with drug discovery target identification. This strategy enables to capture the pangenome wide variation and ensures all conserved variants are targeted by the drug discovery pipeline coupling the pangenome to pharmacophore modelling.

Methods

EHEC population

EHEC associated serotypes are defined based on a previous study[7]. This study defined EHEC strains as subgroup of Shiga-toxin producing *E. coli* and are belonging to the following serotypes (O26:H11,O45:H2,O103:H2,O111:H8,O121: H19, O145:H28, and O157:H7). Whole genome sequences with the associated EHEC metadata was downloaded from Enterobase 1.1.2 using the keyword search of the respective serotypes within the *E. coli* species[8]. This search yielded 702 genomes from environmental, animal and clinical samples. (Underlying data: Metadata from Enterobase 1.1.2 of EHEC pangenome[9]). As this genomes are different from version 1 of this paper, previous Figure 1 was deleted and new Figure 1A was generated reflecting the expanded genomes used in the analysis.

EHEC pangenome

Whole genome typing in the context of the pangenome was performed using PopPUNK (POPulation Partitioning Using Nucleotide Kmers) 1.1.6.[5]. The genomes were annotated with Prokka 1.13.3 as per published protocol[10]. Gff files were extracted as input for the pangenome pipeline Roary 3.11.2 using the following parameters for not splitting paralogs (roary -s -p 32 *.gff) and the resulting presence absence matrix together with the accessory genome phylogeny visualized in Phandango 1.3.0 and is represented as Figure 1B[11]. Each blue bar represents an individual gene and solid blue blocks represent gene clusters. Previous Figure 1B was deleted and new version of Figure 1B was regenerated integrating the new genomes.

Allelic variant calling

Snippy variant calling pipeline 4.3.5 was used to determine the synonymous and nonsynonymous protein mutations using sdiA of *Escherichia coli* O157:H7 str. Sakai as reference. The –contigs option was added to the standard commandline (snippy –outdir –ref sdiA_sakai.gbk). The resulting individual variants of sdiA was merged into EHEC *E. coli* sdiA variant calling data (Underlying data[9]). Previous Figure 3 in version was removed as the new data was better represented by a new Table 2. McDonald-Kreitman test was done using the Snippy output containing data on synonymous and nonsynonymous mutations[12].

In silico sdiA protein modelling

SdiA genes were extracted from the pangenome output of Roary and protein *in silico* modelling performed using SWISS-MODEL[13–17]. SdiA protein sequences were used as targets to search for protein templates within the SWISS-MODEL library. Model selection was based on the template with the highest quality prediction by the target-template alignment.

Results and discussion

Pangenome based clustering integrated the core and accessory elements was applied on 702 whole genomes sequences from serotypes associated with EHEC from diverse sources in the environment as well as animal and human hosts capture the evolutionary space. The majority of the available sequences are from O157 H7 representing 68.5% (481 out of 702) and the rest from the other major non-O157 serotype designated as the "big six", with O45 H2 1.9% (13 out of 702), O103 H2 10.7% (77 out of 702), O26 H11 1.3% (9 out of 702), O111 H8 6.0% (42 out of 702), O121 H19 8.1% (57 out of 702) and O145 H28 3.2% (23 out of 702). The variable-length k-mer analysis and comparison software (PopPUNK) enables scalable, annotation and alignment free approach to large scale population genomics[5]. The accessory genome details the recent acquisition of mobile elements via horizontal gene transfer conveying metabolic, virulence and antibiotic resistance properties which cannot be captured by classical approaches. Eliminating an integral property of recombigenic organism underestimates the diversity and artificially creates similarity and relatedness. The analysis yielded five major pangenomic clusters of EHEC associated isolates. Cluster I is represented by O157 with three genomic subclusters, cluster two contains serotypes O103 and O45, cluster III contains serotype O121, cluster IV contains serotypes O26 and O111 and cluster V contains serotype O145 (Figure 1A). This updated analysis expanded the genomes from version 1 of this paper with 152 genomes into 702 which necessitates the regeneration

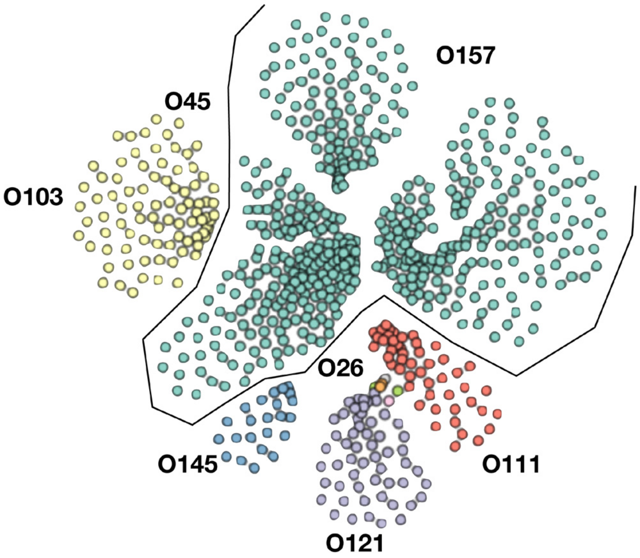

Figure 1A. Pangenome wide clustering using k-mers. There are three clusters within the 0157 serotype, 026 is clustered with O111 as well as 103 with O45. Previous Figure 1A was replaced to reflect the increase in genomes analyzed.

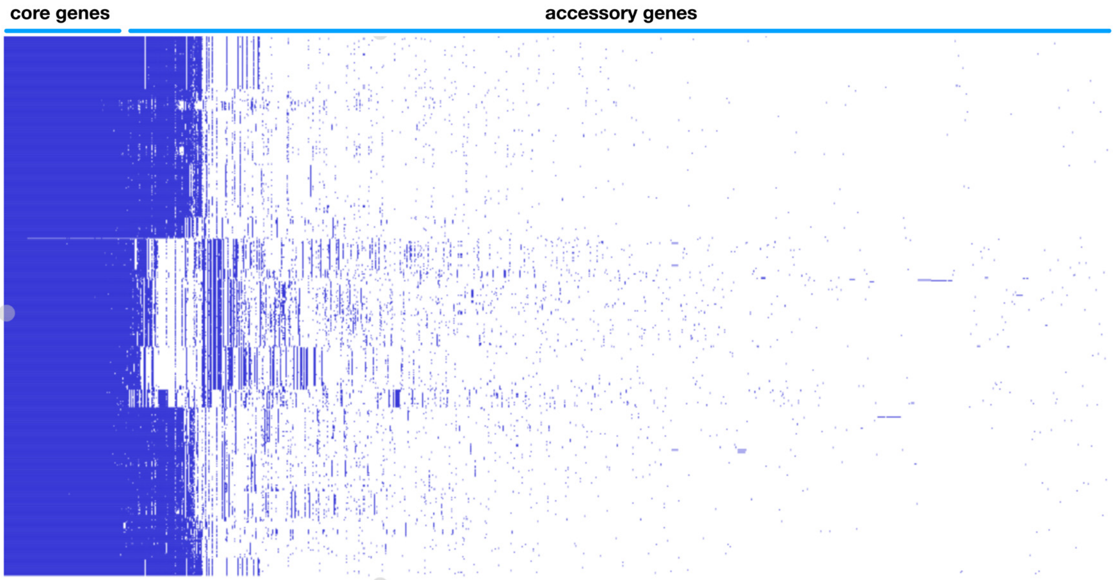

Figure 1B. EHEC pangenome showing genomic diveristy with the gene presence absence variation matrix. Previous Figure 1B was replaced to reflect the increase in genomes analyzed.

of Figure 1. A better visualization of the pangenome cluster was also utilized. Clusters containing several serotypes like cluster II and IV indicate that recombination events blur the genomic boundary resulting to being meshed together in a gradient of dots visually. This novel genome wide framework allows a greater resolution of comparison, as it is now possible to compare similar organisms within the same serotype and determine specific lineages integrating the accessory genome. The acquisition of genomic islands unique to individual isolates are well defined in the pangenome gene presence absence matrix (Figure 1B). The core genome is 2966 (Table 1) and total gene count within the EHEC pangenome is 27774, exceeding previous estimates of total *E. coli* pangenome 22,000. This enormous difference between the core gene and total gene highlights the variation between the different isolates, which can be strain specific and individual isolate specific as indicated by the pangenome data. However, further analysis is limited due to the incompleteness of the metadata entry with regards to the pertinent parameters such specific geolocation, organ of isolation, severity of clinical signs and others.

SdiA is a core gene found across the EHEC pangenome clusters based on the genome wide pangenome analysis, indicating that it can be a suitable interventional target. Considering the huge diversity between pangenome clusters, sdiA homology was analyzed and compared. Remarkably, pangenome cluster I showed highly conserved sdiA structure across global spatial and temporal range (30 years), in spite of cluster I diverging to three separate subsclusters. Divergence from the canonical sdiA structure is more prominent in other genomic clusters. Pangenome cluster II yielded the most number of nonsynonymous mutations (50%) in sdiA gene (Table 2). The percentage distribution for the rest of the pangenome clusters are as follows: 22% for cluster IV, 21% for cluster III and 4% for cluster V. The topological relevance of the predominant mutations was further contextualized by protein modelling.

The impact of the most prevalent nonsynonymous mutations were analyzed with protein modelling using sdiA of *Escherichia coli* O157:H7 str. Sakai as template. The most ranked nonsynonymous mutation is asparagine to serine at amino acid position 101 with 39.1% (210/536 located adjacent to η-4 phenylalanine which is associated with the ligand docking (Figure 2B). This is followed by 24.4% (131/536) of the nonsynonymous mutation is due to conversion of arginine to lysine at position 189 of sdiA (Figure 2A). This amino acid is located with the α-6 domain, adjacent to the amino acid clusters associated with sdiA dimerization. Previous protein modelling determined the role of guanidinium group of arginine which enables interactions in three different directions enabling a more complex electrostatic interaction versus lysine as well as the higher pKa value in arginine that can yield a more stable ionic interaction compared to lysine[18]. β-5 domain alanine to threonine change at amino acid position 140 is the third ranked nonsynonymous mutation with 34.9% (187/536) (Figure 2C).

Table 1. Pangenome metrics.

	Percentage Occurence	
Core genes	(99% <= strains <= 100%)	2966
Soft core genes	(95% <= strains < 99%)	301
Shell genes	(15% <= strains < 95%)	2889
Cloud genes	(0% <= strains < 15%)	21618
Total genes	(0% <= strains <= 100%)	27774

Table 2. Nonsynonymous mutations summary integrating the pangenome clusters.

EHEC Pangenome Cluster	Serotype	Nonsynonymous mutation position		
		101_240	140_240	189_240
II	O103H2	77	77	77
IV	O111H8	40	40	40
III	O121H19	55	55	
V	O145H28	23		
I	O157H7	2	2	1
II	O45H2	13	13	13
	Total	210	187	131

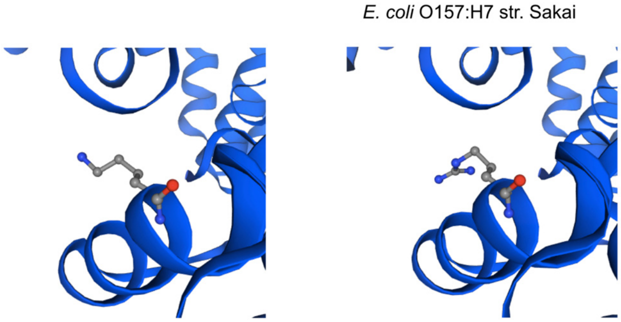

Figure 2A. Protein model of the nonsynonymous variant at amino acid position 189.

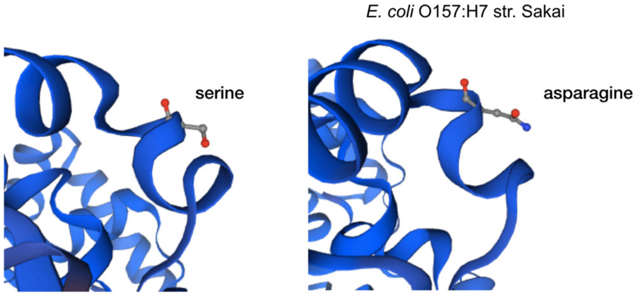

Figure 2B. Protein model of the nonsynonymous variant at amino acid position 101.

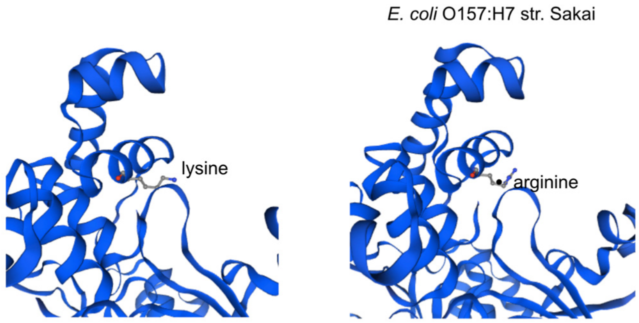

Figure 2C. Protein model of the nonsynonymous variant at amino acid position 140.

None of the highly ranked nonsynonymous mutations impact the ligand interaction, indicating the conservation of the sdiA motif across the population in geographic and temporal distribution, which suggests the possibility of targeting sdiA for quorum sensing inhibition. Mutational analysis using McDonald-Kreitman test indicate differential selection pressures between serotypes. Serotypes O103:H2,O45:H2 and O111:H8 have slightly higher between group nonsynonymous/synonymous ratios (0.42,0.45,0.43 respectively) than within species nonsynonymous/synonymous ratios (0.375 using O157:H7 as within species group). Serotypes O145:H28, O121:H19, O26:H11 have lower values compared to the within species values (0.33, 0.22,0 respectively).

Conclusion

While EHEC pangenome is remarkably diverse, the allelic variants of sdiA, particularly nonsynonymous mutants, indicate the conservation of quorum sensing domain, indicating that targeting this structure can be effective across the different lineages of EHEC pathotype.

Data availability

All underlying and extended data available from Open Science Framework: Supplemental Data for Pangenome guided pharmacophore modelling of enterohemorrhagic *Escherichia coli* sdiA, https://doi.org/10.17605/OSF.IO/BNZ85[9]

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Underlying data

Table 1 Metadata from Patric Database of EHEC *E. coli* pangenome, version 1 replaced with the updated 702 genomes

Table 2 EHEC *E. coli* pangenome presence absence matrix, version 1 replaced with the updated 702 genomes

Table 3 EHEC *E. coli* sdiA variant calling data, version 1 replaced with the updated 702 genomes

Extended data

SWISS-MODEL Homology Modelling Report available at osf.io/bnz85.

References

1. Rohde H, Qin J, Cui Y, *et al.*: **Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4.** *N Engl J Med.* 2011; **365**(8): 718–724.
PubMed Abstract | Publisher Full Text

2. Sperandio V: **SdiA sensing of acyl-homoserine lactones by enterohemorrhagic *E. coli* (EHEC) serotype O157:H7 in the bovine rumen.** *Gut Microbes.* 2010; **1**(6): 432–435.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Maiden MC, Bygraves JA, Feil E, *et al.*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A.* 1998; **95**(6): 3140–3145.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Méric G, Yahara K, Mageiros L, *et al.*: **A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*.** *PLoS One.* 2014; **9**(3): e92798.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Lees JA, Harris SR, Tonkin-Hill G, *et al.*: **Fast and flexible bacterial genomic epidemiology with PopPUNK.** *Genome Res.* 2019; **29**(1): 304–316.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Rasko DA, Rosovitz MJ, Myers GS, *et al.*: **The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates.** *J Bacteriol.* 2008; **190**(20): 6881–6893.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Delannoy S, Beutin L, Fach P: **Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes.** *J Clin Microbiol.* 2013; **51**(10): 3257–3262.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Alikhan NF, Zhou Z, Sergeant MJ, *et al.*: **A genomic overview of the population structure of *Salmonella*.** *PLoS Genet.* 2018; **14**(4): e1007261.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Bandoy D: **Supplemental Data for Pangenome Guided Pharmacophore Modelling of Enterohemorrhagic Escherichia Coli sdiA.** 2019.
http://www.doi.org/10.17605/OSF.IO/BNZ85

10. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics.* 2014; **30**(14): 2068–2069.
PubMed Abstract | Publisher Full Text

11. Page AJ, Cummins CA, Hunt M, *et al.*: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics.* 2015; **31**(22): 3691–3693.
PubMed Abstract | Publisher Full Text | Free Full Text

12. McDonald JH, Kreitman M: **Adaptive protein evolution at the *Adh* locus in *Drosophila*.** *Nature.* 1991; **351**(6328): 652–654.
PubMed Abstract | Publisher Full Text

13. Waterhouse A, Bertoni M, Bienert S, *et al.*: **SWISS-MODEL: homology modelling of protein structures and complexes.** *Nucleic Acids Res.* 2018; **46**(W1): W296–W303.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Bienert S, Waterhouse A, de Beer TA, *et al.*: **The SWISS-MODEL Repository-new features and functionality.** *Nucleic Acids Res.* 2017; **45**(D1): D313–D319.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Bertoni M, Kiefer F, Biasini M, *et al.*: **Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology.** *Sci Rep.* 2017; **7**(1): 10480.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Benkert P, Biasini M, Schwede T: **Toward the estimation of the absolute quality of individual protein structure models.** *Bioinformatics.* 2011; **27**(3): 343–350.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Guex N, Peitsch MC, Schwede T: **Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective.** *Electrophoresis.* 2009; **30 Suppl 1**: S162–173.
PubMed Abstract | Publisher Full Text

18. Sokalingam S, Raghunathan G, Soundrarajan N, *et al.*: **A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein.** *PLoS One.* 2012; **7**(7): e40410.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 3

Reviewer Report 03 November 2020

https://doi.org/10.5256/f1000research.29079.r70557

✓

Olivier Tenaillon
IAME (Infection Antimicrobials Modelling Evolution), UMR 1137, French Institute of Health and Medical Research (INSERM), Paris, France

Table 2 is not properly labelled. Positions should be presented without the _240 and eventually the precise mutation named. It should be mentionned that the numbers refers to number of strains carrying the mutation.

Results from the MK test indicate low conservation. Which outgoup has been used? If the outgroup is too close too few mutations will be present to have reliable value. More details should be given on how the tests have been preformed

Apart from these minor comments the manuscript is fine.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 06 November 2019

https://doi.org/10.5256/f1000research.22447.r54557

✓ **Kerry K. Cooper**

School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA

While I disagree with the author that "The inclusion of non-EHEC is necessary as outgroup comparison groups and does not debunk the validity of the analysis pipeline". As including non-EHEC genomes in an EHEC core genome analysis does alter the results of the output EHEC core genome, which is evident by the author removing those genomes from the analysis.

However, the author has done an excellent job of incorporating a huge amount of only EHEC serotype genomes into the analysis, and as a result has generated a much stronger study. The second version of this manuscript is tremendously better, and the author's additional work has made it significantly higher quality paper.

The only comments are extremely minor:

- In the results and discussion section: Amino acid position 140 is ranked third with 34.9%, but the position 189 is 2nd with 24.4% frequency, these should be switched.

- I would also recommend switching the Figure 2A,B, and C around, so that they are introduced in order of A, B and C.

Otherwise, no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am an expert in foodborne bacterial genomics, epidemiology and pathogenesis, particularly E. coli, Salmonella, Campylobacter, and Listeria.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 29 October 2019

https://doi.org/10.5256/f1000research.22447.r54556

? **Olivier Tenaillon**

IAME (Infection Antimicrobials Modelling Evolution), UMR 1137, French Institute of Health and Medical Research (INSERM), Paris, France

I think this version is better than the previous one, but I still think the connection to sidA could be made stronger and the relevance of the focus on that gene and this subgroup also. Any conserved gene could be a target for drug. So is this gene more interesting than others in that group, the question should be more precise from the beginning.
The focus on some nonsynymous mutations is interesting, but if these mutations are neutral there

is not much interest in showing a detailed structure. A precise Ka/Ks study should be done to tell if these few non synonymous are more than expected.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 15 May 2019

https://doi.org/10.5256/f1000research.19267.r46991

?

Olivier Tenaillon
IAME (Infection Antimicrobials Modelling Evolution), UMR 1137, French Institute of Health and Medical Research (INSERM), Paris, France

The present manuscript presents a state-of-the-art pan genome analysis of EHEC strains and a subsequent analysis of the variation in sdiA.

The analysis of sdiA could have been completed with simple KA/Ks analysis and compared to that of the core genome. For now, there is no connection between the two analysis, the authors could have simply performed the analysis of sdiA.

Some mutants have frame shifts in the gene, this is not discussed.

I think the author could try to connect more the two parts of the analysis.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbial genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 May 2019

DJ Darwin Bandoy, University of the Philippines Los Baños, Los Baños, Philippines

I thank the reviewer for the effort in doing the review. I accept all the suggestions and will add the population genetic analysis in the next version of the paper.

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 April 2019

https://doi.org/10.5256/f1000research.19267.r47169

❌

Kerry K. Cooper
School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA

I would state the biggest issue with the manuscript is the work is not technically sound, because upon examining the metadata file from Patric, numerous strains included in the EHEC pangenome were in fact not EHEC strains. Many of the isolates are UPEC, STEC or EPEC strains that should not be included in the analysis and are resulting in potential errors in the results and conclusions of the manuscript. All O55 strains are EPECs, O104 from the Germany outbreak is a STEC not EHEC, O127 is an EPEC, and CFT073 is an UPEC strain, just as some examples.

Furthermore, the analysis also includes O157:H7 strains that are wild type and mutant strains, and the mutants should not be included in the analysis. Additionally, the authors mention three clades include O157 (Clade I), O26 (Clade II) and O80 (Clade III), however these clades are formed because the vast majority of the strains used in the analysis come from those three serotypes. The analysis is missing three serotypes from the "big 6" serotypes, including O45, O121 and O145, and only include one or two representatives of two of the other serotypes O111, O103. There are numerous genomes available for each of these serotypes through NCBI that should be included in this analysis. Particularly as the "big 6" serotypes represent >50% of the infections, and in the United States represent adulterants in ground beef or other meat products. Therefore, they are a vital aspect for the development of pharmacophore modelling of *sdiA* to prevent colonization in cattle.

Additionally, several genomic studies by Ogura et al (2009)[1] and Cooper et al (2014)[2] have shown that many of these "big 6" serotypes arise along different evolutionary pathways or split from O157 at different time points thus acquire different genes. It would vital to include these in analysis to see if these different pathways impacted the conservation of *sdiA*. The author should also provide a much cleaner version of the metadata as a separate tab in the spreadsheet that includes only those strains that were included in the analysis. Unfortunately, the above-mentioned issue means that all of the results in the manuscript are potential erroneous and need to be completely re-done with the elimination of non-EHEC strains and the inclusion of additional "big 6" genomes to provide a scientifically sound analysis.

It would also be helpful to include in the methods section the date of the search, as the database is constantly changing making reproduction a little bit easier by other researchers. Upon the new analysis it would be helpful to include a brief table or statement of the serotype breakdown included in the EHEC pangenome that would also eliminate some of the above-mentioned issues and make it easier for readers to get a sense of those serotypes included in the analysis.

There are a number of grammatical errors or poor phrasing in the manuscript that should be reviewed and corrected. Such as there is only one author and no indication of other researchers on the manuscript, yet the manuscript keeps stating we instead of I.

Finally, there are several points made in the introduction and discussion that do not have references. For example, in the introduction the author mentions "pangenome of *E. coli* was published in 2008 contained 8 genomes" but does not reference the paper. Additionally, the author mentions "serogroup O80 has aside from Shiga toxin, an extra-intestinal virulence plasmid (pS88), is currently emerging in France" but do not reference anything indicating the emergence in France. I have also provided the citations for Ogura et al.[1] and Cooper et al.[2] for the author to review and potentially cite in the manuscript.

References

1. Ogura Y, Ooka T, Iguchi A, Toh H, et al.: Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli.*Proc Natl Acad Sci U S A*. 2009; **106** (42): 17939-44 PubMed Abstract | Publisher Full Text
2. Cooper KK, Mandrell RE, Louie JW, Korlach J, et al.: Comparative genomics of enterohemorrhagic Escherichia coli O145:H28 demonstrates a common evolutionary lineage with Escherichia coli O157:H7.*BMC Genomics*. 2014; **15**: 17 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am an expert in foodborne bacterial genomics, epidemiology and pathogenesis, particularly E. coli, Salmonella, Campylobacter, and Listeria.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 15 May 2019

DJ Darwin Bandoy, University of the Philippines Los Baños, Los Baños, Philippines

I appreciate the work of the reviewer for going through the metadata. The inclusion of non-EHEC is necessary as outgroup comparison group and does not debunk the validity of the analysis pipeline. In light of this clarification, since this is the sole reason for the non-approval, I appeal for an approval with reservations as the analysis is technically and scientifically sound.

I am in the process of including the big 6 serotypes in the analysis based on the reviewer's comments, as well as the additional references which are very constructive additions to the paper. Again, as the reviewer sees the value in redoing a more inclusive analysis of EHEC serotypes, this is another justification to approve with reservation the paper submitted.

I beg to disagree with the comment of using "inclusive we" in place of I as a grammatical error. The use of royal or inclusive we in lieu of I is a matter of preference. This is the only part of the review I do not agree with.

Competing Interests: No competing interests were disclosed.