Whole-Genome Sequencing and Phylogenomic Analysis of "Cryptic" *Escherichia* Strains Reveals Novel Species- and Subspecies-Level Clades

bei RESOURCES
SUPPORTING INFECTIOUS DISEASE RESEARCH

Poster: 977
Date: June 9, 2018

ATCC®

Phone: 800.638.6597
Email: Tech@atcc.org

Marco A. Riojas, Ph.D.[1,2], Brian J. Cantwell, Ph.D.[1,2], Manzour Hernando Hazbón, Ph.D.[1,2]
[1]ATCC, Manassas, VA 20110; [2]BEI Resources, Manassas, VA 20110

Abstract

Background. Previously described environmental, animal, and human *Escherichia* isolates were found to be monophyletic but were referred to as "cryptic" because they were found to be phenotypically indistinguishable from representative *E. coli* strains.
Methods. The whole-genome sequence of these strains was obtained, *de novo* assembled, and compared to type/reference strains using digital DNA-DNA hybridization (dDDH).
Results. A phylogenomic analysis shows the existence of distinct clades that indicate both multiple novel subspecies of *E. coli* and novel species of *Escherichia*.
Conclusion. At the genomic level, these strains form cohesive phylogenomic clusters and are sufficiently distinguishable from existing taxa that they warrant possible formal reclassification into novel species and/or subspecies.

Background

Despite the long history of study of *Escherichia coli* as a model system and pathogen, there have been relatively few species of the genus *Escherichia* identified to date. Currently, five species have been recognized: *E. coli*[1], *E. fergusonii*[2], *E. albertii*[3], *E. marmotae*[4], and *E. hermanii*[2]. Several additional lineages have been previously classified as *Escherichia*, but subsequent phylogenetic analysis has led to the reclassification of the species previously known as *E. adecarboxylata*, *E. blattae*, and *E. vulneris* to *Leclercia adecarboxylata*[5], *Shimwella blattae*[6], and *Pseudescherichia vulneris*[7], respectively.

While most isolates of *Escherichia* are associated with the gastrointestinal tract of humans or domestic animals, surveys of environmental isolates and similar population studies of *Escherichia* have been conducted. One such study was conducted in 2005 in the laboratory of Thomas Whittam and this work uncovered a number of novel clades of *Escherichia*. This work utilized extended multi-locus sequence typing of internal sequences from 22 housekeeping genes to produce a phylogenetic analysis of strains from humans, wild animals, birds, and environmental and water sources from Australia, Asia, and North America. Due to the fact that no identifiable biochemical differences were identified, these five new clades were labeled "cryptic" *Escherichia*.[8] In a 2015 review, Dr. Seth Walk added additional support for these clades through a phylogenetic tree analysis using the original 22 multilocus sequence typing (MLST) loci and a genomic analysis coupled with a literature review to identify potential unifying features of some of these clades, including identification of virulence-associated genes in Clade I (most closely related to *E. coli*) and adaptation to an environmental niche outside the gastrointestinal tract for Clades II–V.[9]

Because MLST requires the selection of specific loci for phylogenetic comparison, it is by nature inherently vulnerable to selection bias. On the other hand, comparative techniques that use the entirety of the genome are unlikely to suffer from the same sort of bias. In this manner, a phylogenomic comparison is likely to provide more complete and unbiased results than a more limited phylogenetic comparison like MLST. Here, we present the results of such a phylogenomic analysis on a set of 89 Escherichia strains that include the previously described "cryptic" strains.

Materials and Methods

Bacterial strains and DNA extraction. Eighty-nine strains of environmental, animal, and human *Escherichia* isolates were deposited into the BEI Resources (BEIR) collection. These strains were cultured via standard methods, and DNA was extracted via proprietary methods.

Whole genome sequencing. The Illumina® MiSeq® system and a v2 (2×250) flow cell were used to obtain the whole-genome sequence (WGS) of the BEIR *Escherichia* strains. The CLC Genomics Workbench software package (QIAGEN®) was used to trim the sequence reads and to perform *de novo* assembly.

Genomic analysis. The WGS of the BEIR *Escherichia* strains were compared to the WGS of 11 previously sequenced type/reference strains of *Escherichia* and closely related genera available in GenBank. These genomes include the type strains of *E. albertii*, *E. coli*, *E. fergusonii*, *E. hermanii*, *E. marmotae*, *P. vulneris*, *L. adecarboxylata*, *Shimwellia blattae*, and *Shigella flexneri*; reference (non-type) strains included *E. coli* K-12 MG1655 and *Yersinia* (outgroup). WGS-based genomic analysis was performed with digital DNA-DNA hybridization (dDDH) via the Genome-to-Genome Distance Calculator (GGDC) v2.1; formula 2 was used for analysis.[10-13] The genome-to-genome distances (GGDs) were interpreted and were used to build a phylogenomic tree that was visualized via iTOL, as previously described.[14]

Results

GenBank genomes. The 11 type/reference strains clade into nine distinct groups that show dDDH values <70% (range: 20.2–43.7%) from each other, indicating different species-level groups. These strains are shown in blue text in Figure 1 and served as a comparative reference for the dDDH analysis of the BEIR strains. Three of these genomes—the *E. coli* type strain, *E. coli* K-12, and the *Shigella flexneri* type strain—show dDDH values between 70-80% of each other (range: 74.5–84.9%), indicating that these three genomes belong to the same species. Further, the type strain of *S. flexneri* clades with *E. coli* K-12 at a dDDH distance of 84.9%, indicating that they belong to the same subspecies. This replicates previously published results.[12]

BEIR strain genomes. The 89 BEIR strains formed clades with the type/reference strains as follows. Two strains (B090 and B156) claded with the type strain of *E. albertii* (dDDH: 88.6–91.3%). Two strains (B646 and E1118) clade with the type strain of *E. marmotae* (dDDH: 92.3–94.5%). One strain (B253) clades with the type strain of *E. fergusonii* (dDDH: 90.6%). Two strains (H605 and TA004) clade together without an existing type strain (dDDH: 70.7%). Similarly, four strains (E1492, H442, M863, and TW10509) clade together without an existing type strain (dDDH: 88.6–91.3%). The remaining 78 strains fall within the circumscription (dDDH: 67.3–99.3%) of the type strain of *E. coli* (as well as with *E. coli* K-12 and *S. flexneri*).

The 78 strains that fall within the circumscription of E. coli clade into five distinct subspecies-level groups. Two strains (B093 and H299) each form a single-strain subspecies-level clade. A group is formed by 21 strains that clade with the type strain of E. coli (dDDH: 84.5–99.1%). Ten strains form a distinct subspecies-level clade (dDDH: 81.9–90.1%) within E. coli but separate from any type/reference strains. Forty-five strains form a subspecies-level clade with E. coli K-12 and the type strain of S. flexneri (dDDH: 74.9–99.3%).

Conclusions

- These cryptic strains form cohesive and distinct phylogenomic clusters and are sufficiently distinguishable from existing taxa to warrant possible formal reclassification into novel species and/or subspecies.
- Strains E1492, H442, M863, and TW10509 form a species-level clade that likely represents a novel species very closely related to, but distinct from, *E. coli*.
- Strains H605 and TA004 clade into a novel *Escherichia* species most closely related to *E. marmotae*. The dDDH distance of these two strains to each other indicates that they represent two different subspecies of the same novel species.
- The *E. coli* type strain and the 21 BEIR strains that clade with it make up a subspecies-level clade that, according to the rules of nomenclature[15], should be assigned the name *E. coli* subsp. *coli*.
- *E. coli* K-12 and *S. flexneri* clade with 45 BEIR strains and were labeled with the unofficial name *E. coli* subsp. K-12/shigella (Figure 1).
- Three novel subspecies-level clades were identified and were labeled with the unofficial names *E. coli* subsp. New1, New2, and New3 (Figure 1).

References

1. Castellani A, Chalmers AJ. Manual of Tropical Medicine, 3rd ed., William Wood and Co., New York, 1919.
2. Farmer JJ, *et al*. *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of *Enterobacteriaceae* isolated from clinical specimens. J Clin Microbiol 21: 77-81, 1985.
3. Huys G, *et al*. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. Int J Syst Evol Microbiol 53: 807-810, 2003.
4. Liu S, *et al*. *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. Int J Syst Evol Microbiol 65: 2130-2134, 2015.
5. Tamura K, et al. *Leclercia adecarboxylata* Gen. Nov., Comb. Nov., formerly known as *Escherichia adecarboxylata*. Curr Microbiol13; 179-184, 1986.
6. Priest FG, Barker M. Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. Int J Syst Evol Microbiol 60: 828-833, 2010.
7. Alnajar S, Gupta RS. Phylogenomics and comparative genomic studies delineate six main clades within the family *Enterobacteriaceae* and support the reclassification of several polyphyletic members of the family. Infect Genet Evol 54: 108-127, 2017.
8. Walk S, *et al*. Cryptic Lineages of the Genus *Escherichia*. Appl Environ Microbiol 75: 6534-6544, 2009.
9. Walk S. The "Cryptic" *Escherichia*. EcoSal Plus 6(2): doi: 10.1128/ecosalplus.ESP-0002-2015, 2015.
10. Auch AF, *et al*. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2(1):117-34, 2010.
11. Auch AF, Klenk HP, Goker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci 2(1):142-8, 2010.
12. Meier-Kolthoff JP, *et al*. Complete genome sequence of DSM 30083^{T}, the type strain (U5/41^{T}) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic Sci 9:2, 2014.
13. Meier-Kolthoff JP, *et al*. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14:60, 2013.
14. Riojas MA, *et al*. Phylogenomic analysis of the species of the *Mycobacterium tuberculosis* complex demonstrates that *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium microti* and *Mycobacterium pinnipedii* are later heterotypic synonyms of *Mycobacterium tuberculosis*. Int J Syst Evol Microbiol. 2018 Jan;68(1):324-332.
15. Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol [Epub ahead of print], 2015.

Figure 1. A phylogenomic tree of the "cryptic" BEIR *Escherichia* strains shown in comparison to type and reference species of *Escherichia* and related species. The type/reference strain genomes are shown in blue text and include the GenBank accession number. The BEIR *Escherichia* strains are labeled with the strain name given upon deposit followed by the BEI Resources catalog number.

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 Web: www.atcc.org