Escherichia Coli: What Is and Which Are?

Marta Cobo-Simón ,* Rowan Hart, and Howard Ochman

Department of Molecular Biosciences, University of Texas at Austin, Austin, TX

***Corresponding author:** E-mail: marta.cobo@outlook.com.

Associate editor: Brandon Gaut

Abstract

***Escherichia coli* have served as important model organisms for over a century—used to elucidate key aspects of genetics, evolution, molecular biology, and pathogenesis. However, defining which strains actually belong to this species is erratic and unstable due to shifts in the characters and criteria used to distinguish bacterial species. Additionally, many isolates designated as *E. coli* are genetically more closely related to strains of *Shigella* than to other *E. coli*, creating a situation in which the entire genus of *Shigella* and its four species are encompassed within the single species *E. coli*. We evaluated all complete genomes assigned to *E. coli* and its closest relatives according to the biological species concept (BSC), using evidence of reproductive isolation and gene flow (i.e., homologous recombination in the case of asexual bacteria) to ascertain species boundaries. The BSC establishes a uniform, consistent, and objective principle that allows species-level classification across all domains of life and does not rely on either phenotypic or genotypic similarity to a defined type-specimen for species membership. Analyzing a total of 1,887 sequenced genomes and comparing our results to other genome-based classification methods, we found few barriers to gene flow among the strains, clades, phylogroups, or species within *E. coli* and *Shigella*. Due to the utility in recognizing which strains constitute a true biological species, we designate genomes that form a genetic cohesive group as members of *E. coli*$_{\mathbf{BIO}}$.**

***Key words:* *Escherichia coli*, *Shigella*, enteric bacteria, recombination, speciation.**

Introduction

When initially isolated, *Escherichia coli* was designated *Bacillus coli communis*, a latinization describing its prominent characteristic as a "common colon bacterium" that could be readily cultured in a variety of substrates. The original specimen, as first described in 1885, was distinguished by its colony and cellular morphology, and its ability to ferment glucose, produce acid, and sour milk (Escherich 1885). Upon its rechristening in 1,919 to acknowledge its discoverer, and in the decades that ensued, features used for assignment to this species were expanded to include a suite of characters that distinguish *E. coli* from other enteric species (Koser 1923; Kauffmann 1944). Most notably, *E. coli* are lactose, catalase, and indole positive, and oxidase, urease, and citrate negative, although there is a low level of polymorphism for many of these properties.

Genetic and genomic features entered into the classification of *E. coli* in the 1960s with the application of DNA–DNA hybridization (DDH) procedures (Marmur et al. 1963). By this method, strains were considered as members of *E. coli* if they displayed ≥70% DNA similarity to the reference strains (Brenner et al. 1972)—noting that although DDH percentages do not match the actual amount of DNA identity between strains (Rosselló-Mora 2006), this method pioneered a threshold-based approach for defining bacterial species. Subsequently, other nucleic-acid-based cutoffs were applied to the delineation of bacterial species, such as ≥97% (Tindall et al. 2010; Yarza et al. 2014) and more recently ≥99% (Edgar 2018) 16S RNA sequence identity, or ≥95% average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005) for the core set of genes shared among strains (Jain et al. 2018). Naturally, there is a certain circularity to this approach since sequence-identity thresholds were ascertained from strains that were already assigned to *E. coli* based on metabolic, morphological, or biochemical features, thereby constraining the genetic cutoffs to species boundaries that were already established. And unfortunately, hybridization and sequence-identity thresholds are convenient rather than universal, their biological basis remains unclear.

Phylogenetic analysis of *E. coli* strains that were considered to span the diversity in the species at large defined six main clades (A, B1, B2, D, E, and F) and several rarer clades (Herzer et al. 1990; Chaudhuri and Henderson 2012). However, expanding the set to include strains from additional animal and environmental sources yielded five "cryptic" clades (termed CI to CV) that were all more closely related to *E. coli* than to its sister species *Escherichia fergusonii* (Walk et al. 2009; Luo et al. 2011). The taxonomic status of these five unclassified clades remains uncertain: they cannot be differentiated from *E. coli* based on phenotypic characters, but they are genetically divergent, which led to a proposal that a least some of these clades (e.g., Clades III + IV and Clade V) might represent distinct species (Walk 2015).

As additional full genomes were integrated into the analyses, the phylogenetic structure and evolutionary relationships of *E. coli* became more refined, with recognition of increased numbers of subspecific groups (Lu et al. 2016; Abram et al. 2021) and suggestions that some might represent actual or incipient species (Didelot et al. 2012; Kang et al. 2021). To accommodate the burgeoning numbers of sequenced strains in all taxa, the Genome Taxonomy Database (GTDB; gtdb.ecogenomic.org/) recommended the application of a genome-wide identity threshold (analogous to ANI) to define bacterial species (Parks et al. 2018). Imposing their metrics, strains currently classified as *E. coli* would be split into six species—*E. coli*, *E. coli*_E, *Escherichia ruysiae*, *Escherichia marmotae*, *Escherichia sp001660175*, and *Escherichia sp005843885*—with the majority consigned to *E. coli* (Parks et al. 2021).

Classification of *E. coli* has also been confounded by the intransigence of *Shigella* as a separate genus. Every strain assigned to *Shigella* appears to fall within the variation spanned by *E. coli* (Brenner et al. 1973; Ochman et al. 1983), and the four *Shigella* species originated independently, and multiple times, from within *E. coli* (Rolland et al. 1998; Pupo et al. 2000; Lan and Reeves 2002). Clearly, the taxonomy of *E. coli* is idiosyncratic and often supports conflicting results. To resolve these incongruencies, and to apply consistent and objective criteria to identifying species boundaries, we analyze and classify a comprehensive set of *Escherichia* and *Shigella* genomes according to the biological species concept (BSC) (Mayr 1942), a universally accepted procedure that circumscribes species based on homologous gene exchange. Although asexuality is often assumed to render bacteria immune to classification by the BSC (Donoghue 1985; Rosselló-Mora and Amann 2001; Costechareyre et al. 2009), the patterns of recombination in *E. coli* and related enteric bacteria provide a consistent and robust signal for species assignment (Brenner and Falkow 1971; Shen and Huang 1986; Dykhuizen and Green 1991; Lawrence and Retchless 2009; Didelot et al. 2012) and allow the application of a single biological feature to define species across all branches of the Tree of Life.

Results

To establish the species boundaries of *E. coli* and determine which sequenced genomes should be assigned to this species, we implemented three genome-based methods, including two (*ConSpeciFix* and PopCOGenT) that adhere to the precepts of the BSC. We considered 1,635 complete genomes designated as *E. coli* in the National Center for Biotechnology Information (NCBI) database (www.ncbi.nlm.nih.gov/) as of August 2020. To ensure that the breadth of variation in the species at large is represented, we included the genomes of strains classified to the five *Escherichia* phylogroups (Clades I–V) described by Walk et al. (2009), to the *E. coli* phylogroups resolved by Abram et al. (2021), and to the newly designated *Escherichia* species proposed by the GTDB (*E. albertii*, *E. coli*, *E. coli*_E, *E. fergusonii*, *E. marmotae*, *E. ruysiae*, *Escherichia* sp001660175, *Escherichia* sp002965065, *Escherichia* sp004211955, and *Escherichia* sp005843885) (gtdb.ecogenomic.org). Additionally, we analyzed all other fully sequenced genomes assigned to the genus *Escherichia* as well as representatives of the four designated species of *Shigella*, which are known to have originated from within *E. coli* but have mostly maintained their status as a separate genus for historical reasons.

ConSpeciFix

Using gene flow as a condition for species membership, this method calculates recombination based on homoplasies in genes common to the strains under consideration (Bobay et al. 2018).

(i) Applying *ConSpeciFix* to assess species status of strains designated as *E. coli* in NCBI. Genomes classified as *E. coli* in the NCBI database form more than one species, with 12 strains that are sexually isolated from the rest of the *E. coli* genomes. All remaining genomes designated as *E. coli* in the NCBI database constitute a true biological species (hereafter referred to as *E. coli*$_{\mathrm{BIO}}$ to identify membership as a biological species) and serve as a reference to evaluate genomes classified by other means. The numbers and distribution of strains that are members of this biological species, and those that are reproductively isolated and excluded from *E. coli*$_{\mathrm{BIO}}$, are shown in figures 1 and 2.

(ii) Applying *ConSpeciFix* to assess species status of phylogroups defined by Walk et al. (2009) and Abram et al. (2021). All the studied strains classified to the *E. coli* phylogroups of Abram et al., and to the *E. coli* taxonomic group and Clade I specified by Walk et al., which together include strains classified as *E. coli* and *Escherichia* sp. in NCBI, are members of *E. coli*$_{\mathrm{BIO}}$ based on *ConSpeciFix*. Genomes in their remaining clades (Clades II–V) constitute different species by *ConSpeciFix*, with the exception of one genome in Clade IV (*E. coli* B49-2 serovar O157:H7) and one in Clade V (*E. coli* strain E620 serovar ON5), both of which are members of *E. coli*$_{\mathrm{BIO}}$ (supplementary table S1, Supplementary Material online).

(iii) Applying *ConSpecFix* to assess the status of *E. coli* species defined by GTDB. The GTDB recognizes *E. coli* and nine additional species (*E. ruysiae*, *E. marmotae*, *E. coli*_E. *E. fergusonii*, *Escherichia albertii*, *Escherichia* sp001660175, *Escherichia* sp002965065, *Escherichia* sp004211955, *and Escherichia* sp005843885). Of these nine additional species, only *Escherichia coli*_E, one strain of *Escherichia* sp001660175 (based on ANI), three strains of *Escherichia* sp005843885 (one of them based on ANI), and three strains of *E. ruysiae* (two of them based on ANI) were classified as *E. coli* in the NCBI database (supplementary table S1, Supplementary Material online). Although listed as differently by the GTDB, one strain of *E. albertii* and one strain of

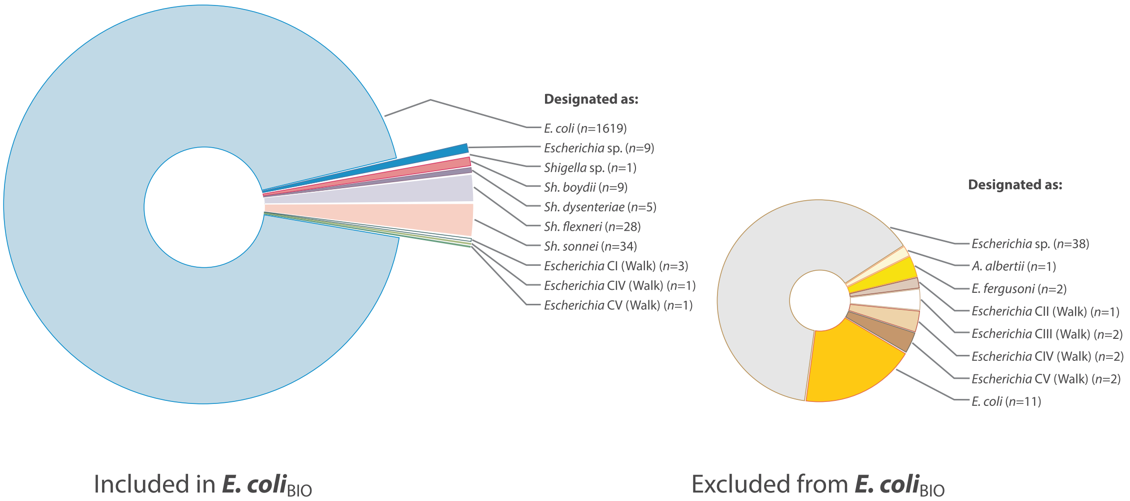

Fig. 1. Assignment of sequenced genomes to the biological species *E. coli*$_{\text{BIO}}$. Wedges are labeled according to their taxon designation in the NCBI database or their assignment to an *Escherichia* phylogroup by Walk et al., with the number of genomes in each taxon indicated. Note that genomes that are both assigned to an *Escherichia* phylogroup (CI–CV) and taxonomically defined in the NCBI are excluded from counts of NCBI genomes. For example, one of the genomes assigned to *E. coli* by NCBI but excluded from *E. coli*$_{\text{BIO}}$ belongs to Walk Clade IV and was therefore excluded from the count of *E. coli*.

E. marmotae are classified as *E. coli* in the NCBI. Of the 12 genomes assigned to *E. coli* in the NCBI but excluded from *E. coli*$_{\text{BIO}}$, nine were also excluded from *E. coli* by the GTDB (supplementary table S1, Supplementary Material online). *ConSpeciFix* assigned the GTDB species *Escherichia* sp001660175 ($n = 1$), sp004211955 ($n = 2$), and sp005843885 ($n = 38$) to a separate biological species. No members of these three GTDB species belong to *E. coli*$_{\text{BIO}}$ when used as test lineages, but they form a biological species distinct from *E. coli*$_{\text{BIO}}$ when *Escherichia* sp005843885 is used as a reference lineage. None of the other GTDB species is a member of either *E. coli*$_{\text{BIO}}$ or this new species.

(iv) Other enteric species. All tested genomes of the four *Shigella* species are members of *E. coli*$_{\text{BIO}}$. In contrast, none of the genomes currently classified to any of the other *Escherichia* species [*E. albertii* ($n = 1$), *E. fergusonii* ($n = 2$), *E. marmotae* ($n = 1$)] or to any of the other enteric genera considered [*Proteus* ($n = 2$), *Citrobacter* ($n = 2$), *Cronobacter* ($n = 2$), *Salmonella* ($n = 111$), *Enterobacter* ($n = 6$), and *Klebsiella* ($n = 4$)] is a member of *E. coli*$_{\text{BIO}}$. Genomes from genera other than *Escherichia* were included as controls.

PopCOGenT

PopCOGenT is an alternate method for grouping genomes based on gene flow (Arevalo et al. 2019). For the representative set of genomes evaluated by this method ($n = 128$), there were a total of 21 species-groups, of which 10 contained strains designated as *E. coli* in the NCBI database (supplementary table S1, Supplementary Material online). The phylogenetic relationships of a dereplicated subset of these genomes, along with their nomenclature, strain and species designations in different databases, and species-groupings based on several metrics, are presented in figure 2.

(i) Applying *PopCOGenT* to assess species status of clades defined by Walk et al. (2009) and Abram et al. (2021). Genomes from the *E. coli* taxonomic group specified by Walk are assigned to *PopCOGenT* species-groups 0 and 1 (fig. 2; supplementary table S1, Supplementary Material online), and Clades I, II, and III of Walk are each classified as different *PopCOGenT* species-groups (4, 5, and 6, respectively). Genomes from Walk Clade IV assort into two species-groups: one of which contains only Clade IV genomes, and another that contains genomes from both Clade V and the canonical *E. coli and Shigella flexneri* taxonomic groups specified by Walk. Similarly, genomes from Clade V of Walk segregate into two species-groups—the aforementioned one that contains genomes from Clade IV and the canonical *E. coli* and *S. flexneri* taxonomic groups, and a unique species-group (19) that contains only Clade V genomes. Several of the *E. coli* phylogroups defined by Abram et al. were distinguished as different species-groups by *PopCOGenT*.

(ii) Applying *PopCOGenT* to assess the status of *E. coli* species defined by GTDB. The **five** GTDB-recognized species within *E. coli* (*E. coli*, *E. coli*_E, *E. ruysiae*, *Escherichia* sp001660175, and *Escherichia* sp005843885) and the **five** other *Escherichia* species (*E. albertii*, *E. fergusonii*, *E. marmotae*, *E.* sp002965065, and *E.* sp004211955) were classified to multiple species-groups by *PopCOGenT* (supplementary table S1, Supplementary Material online). Each of the *Escherichia* species recognized by the GTDB forms

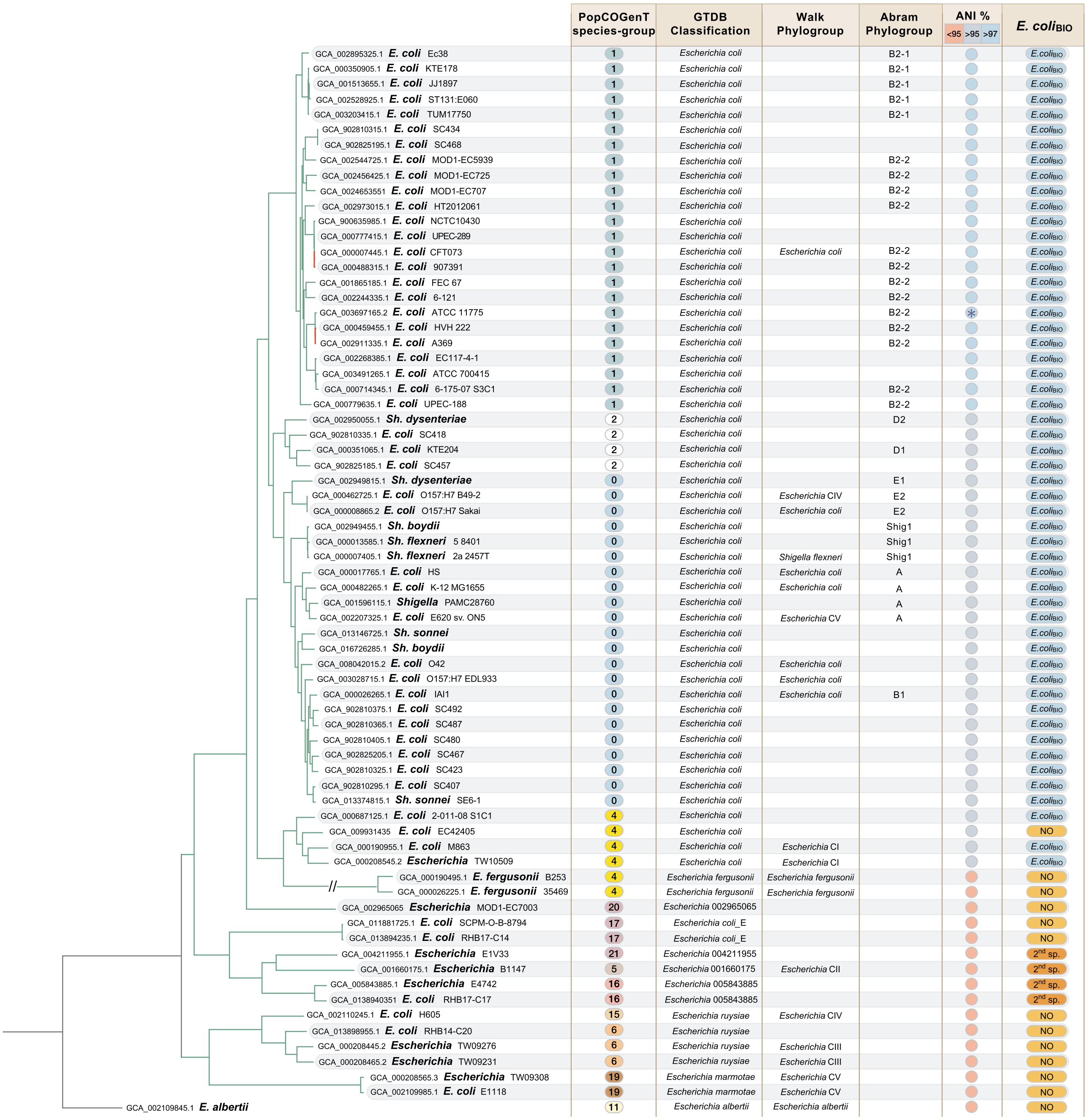
PopCOGenT species-group
GTDB Classification
Walk Phylogroup
Abram Phylogroup
ANI %
<95 >95 >97
E. coli$_{BIO}$
GCA_002895325.1 E. coli Ec38 1 Escherichia coli B2-1 E.coli$_{BIO}$
GCA_000350905.1 E. coli KTE178 1 Escherichia coli B2-1 E.coli$_{BIO}$
GCA_001513655.1 E. coli JJ1897 1 Escherichia coli B2-1 E.coli$_{BIO}$
GCA_002528925.1 E. coli ST131:E060 1 Escherichia coli B2-1 E.coli$_{BIO}$
GCA_003203415.1 E. coli TUM17750 1 Escherichia coli B2-1 E.coli$_{BIO}$
GCA_902810315.1 E. coli SC434 1 Escherichia coli E.coli$_{BIO}$
GCA_902825195.1 E. coli SC468 1 Escherichia coli E.coli$_{BIO}$
GCA_002544725.1 E. coli MOD1-EC5939 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002456425.1 E. coli MOD1-EC725 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_0024653551 E. coli MOD1-EC707 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002973015.1 E. coli HT2012061 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_900635985.1 E. coli NCTC10430 1 Escherichia coli E.coli$_{BIO}$
GCA_000777415.1 E. coli UPEC-289 1 Escherichia coli E.coli$_{BIO}$
GCA_000007445.1 E. coli CFT073 1 Escherichia coli Escherichia coli B2-2 E.coli$_{BIO}$
GCA_000488315.1 E. coli 907391 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_001865185.1 E. coli FEC 67 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002244335.1 E. coli 6-121 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_003697165.2 E. coli ATCC 11775 1 Escherichia coli B2-2 * E.coli$_{BIO}$
GCA_000459455.1 E. coli HVH 222 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002911335.1 E. coli A369 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002268385.1 E. coli EC117-4-1 1 Escherichia coli E.coli$_{BIO}$
GCA_003491265.1 E. coli ATCC 700415 1 Escherichia coli E.coli$_{BIO}$
GCA_000714345.1 E. coli 6-175-07 S3C1 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_000779635.1 E. coli UPEC-188 1 Escherichia coli B2-2 E.coli$_{BIO}$
GCA_002950055.1 Sh. dysenteriae 2 Escherichia coli D2 E.coli$_{BIO}$
GCA_902810335.1 E. coli SC418 2 Escherichia coli E.coli$_{BIO}$
GCA_000351065.1 E. coli KTE204 2 Escherichia coli D1 E.coli$_{BIO}$
GCA_902825185.1 E. coli SC457 2 Escherichia coli E.coli$_{BIO}$
GCA_002949815.1 Sh. dysenteriae 0 Escherichia coli E1 E.coli$_{BIO}$
GCA_000462725.1 E. coli O157:H7 B49-2 0 Escherichia coli Escherichia CIV E2 E.coli$_{BIO}$
GCA_000008865.2 E. coli O157:H7 Sakai 0 Escherichia coli Escherichia coli E2 E.coli$_{BIO}$
GCA_002949455.1 Sh. boydii 0 Escherichia coli Shig1 E.coli$_{BIO}$
GCA_000013585.1 Sh. flexneri 5 8401 0 Escherichia coli Shig1 E.coli$_{BIO}$
GCA_000007405.1 Sh. flexneri 2a 2457T 0 Escherichia coli Shigella flexneri Shig1 E.coli$_{BIO}$
GCA_000017765.1 E. coli HS 0 Escherichia coli Escherichia coli A E.coli$_{BIO}$
GCA_000482265.1 E. coli K-12 MG1655 0 Escherichia coli Escherichia coli A E.coli$_{BIO}$
GCA_001596115.1 Shigella PAMC28760 0 Escherichia coli A E.coli$_{BIO}$
GCA_002207325.1 E. coli E620 sv. ON5 0 Escherichia coli Escherichia CV A E.coli$_{BIO}$
GCA_013146725.1 Sh. sonnei 0 Escherichia coli E.coli$_{BIO}$
GCA_016726285.1 Sh. boydii 0 Escherichia coli E.coli$_{BIO}$
GCA_008042015.2 E. coli O42 0 Escherichia coli Escherichia coli E.coli$_{BIO}$
GCA_003028715.1 E. coli O157:H7 EDL933 0 Escherichia coli Escherichia coli E.coli$_{BIO}$
GCA_000026265.1 E. coli IAI1 0 Escherichia coli Escherichia coli B1 E.coli$_{BIO}$
GCA_902810375.1 E. coli SC492 0 Escherichia coli E.coli$_{BIO}$
GCA_902810365.1 E. coli SC487 0 Escherichia coli E.coli$_{BIO}$
GCA_902810405.1 E. coli SC480 0 Escherichia coli E.coli$_{BIO}$
GCA_902825205.1 E. coli SC467 0 Escherichia coli E.coli$_{BIO}$
GCA_902810325.1 E. coli SC423 0 Escherichia coli E.coli$_{BIO}$
GCA_902810295.1 E. coli SC407 0 Escherichia coli E.coli$_{BIO}$
GCA_013374815.1 Sh. sonnei SE6-1 0 Escherichia coli E.coli$_{BIO}$
GCA_000687125.1 E. coli 2-011-08 S1C1 4 Escherichia coli E.coli$_{BIO}$
GCA_009931435 E. coli EC42405 4 Escherichia coli NO
GCA_000190955.1 E. coli M863 4 Escherichia coli Escherichia CI E.coli$_{BIO}$
GCA_000208545.2 Escherichia TW10509 4 Escherichia coli Escherichia CI E.coli$_{BIO}$
GCA_000190495.1 E. fergusonii B253 4 Escherichia fergusonii Escherichia fergusonii NO
GCA_000026225.1 E. fergusonii 35469 4 Escherichia fergusonii Escherichia fergusonii NO
GCA_002965065 Escherichia MOD1-EC7003 20 Escherichia 002965065 NO
GCA_011881725.1 E. coli SCPM-O-B-8794 17 Escherichia coli_E NO
GCA_013894235.1 E. coli RHB17-C14 17 Escherichia coli_E NO
GCA_004211955.1 Escherichia E1V33 21 Escherichia 004211955 2nd sp.
GCA_001660175.1 Escherichia B1147 5 Escherichia 001660175 Escherichia CII 2nd sp.
GCA_005843885.1 Escherichia E4742 16 Escherichia 005843885 2nd sp.
GCA_0138940351 E. coli RHB17-C17 16 Escherichia 005843885 2nd sp.
GCA_002110245.1 E. coli H605 15 Escherichia ruysiae Escherichia CIV NO
GCA_013898955.1 E. coli RHB14-C20 6 Escherichia ruysiae NO
GCA_000208445.2 Escherichia TW09276 6 Escherichia ruysiae Escherichia CIII NO
GCA_000208465.2 Escherichia TW09231 6 Escherichia ruysiae Escherichia CIII NO
GCA_000208565.3 Escherichia TW09308 19 Escherichia marmotae Escherichia CV NO
GCA_002109985.1 E. coli E1118 19 Escherichia marmotae Escherichia CV NO
GCA_002109845.1 E. albertii 11 Escherichia albertii Escherichia albertii NO

Fig. 2. Maximum-likelihood phylogenetic tree of selected *Escherichia* genomes. Genomes were selected to represent the extent of diversity present in the genus and have <99.8% ANI to their nearest relative. For each genome, strain accession number and designation in the NCBI database is followed, from left to right, by *PopCOGenT* species-group, GTDB v207 classification, Walk et al. phylogroup (Clades I–V) and Abram et al. phylogroups wherever possible, ANI to *E. coli* ATCC 11775, and membership status in *E. coli*$_{BIO}$. (Note that GCA_002109985_1, *E. marmotae* E1118, is labeled as *E. coli* in the PATRIC database). *PopCOGenT* species-groups are distinguished by number, and ANI % denotes extent of sequence identity to the reference genome (marked with asterisk): the first 24 genomes in the tree present an ANI > 97%; the following 30 an ANI between 95 and 97%, and the last 16 an ANI <95%. All branches have bootstrap support values >90% except for the strains GCA_000007445.1, GCA_000488315.1, GCA_000459455.1 and GCA_002911335.1, which are <60%. Seven of the 12 NCBI-classified *E. coli* strains that were excluded from *E. coli*$_{BIO}$ are NCBI pathogen detection assemblies (i.e., surveillance genomes) and were not classified by the GTDB classification: those genomes lacking taxonomic assignation in GTDB were classified to the same species as their closest relative having an ANI >95%.

a unique *PopCOGent* species-group, except 1) *E. ruysiae*, whose members were distributed into two *PopCOGent* species-groups (6, 15), 2) *E. coli*, whose members were distributed into four *PopCOGent* species-groups (0, 1, 2, 4), and 3) *E. coli*_E, whose members were distributed into two *PopCOGent* species-groups (8, 17) (fig. 2; supplementary table S1, Supplementary Material online).

(iii) Other enteric species. Whereas *PopCOGenT* separated the NCBI-designated strains of *E. coli* into 10 species-groups, all *Shigella* genomes considered by *PopCOGenT*, except *Sh. dysenteriae* (accession

GCA_002950055.1), were classified as members of species-group 0. Most strains that were assigned to *Escherichia* species other than *E. coli* (or whose species status went unassigned) were deemed separate species by *PopCOGenT*, although many partitioned in species-groups that also contained members of *E. coli*. Though not included in figure 2, *PopCOGenT* distinguished *Salmonella enterica*, *Salmonella bongori*, *Enterobacter cloacae*, *Enterobacter carcerogenus*, and *Proteus mirabilis* as distinct species.

FastANI

This metric is based on sequence-identity thresholds (typically 95%) to delineate strains that constitute a species (Jain et al. 2018).

(i) Applying ANI to assess species status of clades defined by Walk et al. (2009) and Abram et al. (2021). Genomes from *S. flexneri* and the *E. coli* taxonomic group specified as Clade I by Walk, and one genome each from Clades IV and V, are all classified as members of the same species based on 95% ANI to the reference genome, *E. coli* ATCC 11775. Applying this 95% ANI threshold, the studied phylogroups distinguished by Abram et al. are also included in this species (supplementary table S1, Supplementary Material online). All genomes in this ANI species were originally designated as *E. coli* or *S. flexneri* in NCBI, except in the case of one genome classified as *Escherichia* sp. All remaining members of Clades IV and V, and all other members of the other clades defined by Walk et al., are sufficiently distant from *E. coli* ATCC 11775 and are not considered members of the species at this ANI threshold.

(ii) Applying ANI to assess the status of *E. coli* species defined by GTDB. Because the GTDB circumscribes species based on sequence-identity thresholds, the majority of the genomes assigned to *E. coli* have an ANI > 95% to the *E. coli* ATCC 11775 reference genome; however, there are a few exceptions due to the normalization applied by this database (supplementary table S1, Supplementary Material online).

(iii) Other enteric species. applying a sequence-identity threshold of 95% to *E. coli* ATCC 11775, all tested genomes of the four *Shigella* spp. are members of *E. coli*. None of the genomes classified to other *Escherichia* species (*E. albertii* and *E. fergusonii*) or to any of the other enteric genera (*Proteus, Citrobacter, Cronobacter, Salmonella, Enterobacter,* and *Klebsiella*) is a member of *E. coli* at this sequence-identity threshold.

Applying the many-to-many option in ANI returned results that were virtually identical to those recovered with the one-to-many comparisons to the single reference genome. For example, all other enteric species yielded ANI values <95% to members of both *E. coli* and *Shigella*. With regard to the biological species (*E. coli*$_{\mathrm{BIO}}$) defined *ConSpeciFix*, most genomes displayed ANI values <95% using the many-to-many option; however, the minimum ANI of 93.90% occurred between two strains having 97% and 98% ANI with the reference genome.

Maximum-Likelihood (ML) Phylogeny

To examine the evolutionary relationships among strains, we constructed a phylogeny on the dereplicated set of 70 genomes having <99.8% ANI to one another.

(i) Applying an ML phylogeny to assess species status of clades defined by Walk et al. (2009) and Abram et al. (2021). Our results broadly confirm the phylogroups distinguished by Walk et al. (2009), which is not surprising given that their phylogroups represent phylogenetically resolved clades. All *E. coli* and *Shigella* genomes that they defined were monophyletic, and Clades I, II, III, IV, and V each formed monophyletic groups, with the exception of one strain from each of Clade IV and Clade V, which grouped with *E. coli* and *Shigella*. In our tree, the clade containing *E. coli* and *Shigella* is most closely related to Walk Clade I, which is a sister group to *E. fergusonii*, and Walk Clades III, IV, and V together form a separate clade. In addition, each of the phylogroups resolved Abram et al. (2021) is monophyletic (fig. 2).

(ii) Applying an ML phylogeny to assess the status of *E. coli* species defined by GTDB. The clades defined in the ML phylogeny are consistent with the species distinguished by GTDB, and each is monophyletic. The only exceptions are the two strains classified as *E. fergusonii*, which reside on a very long branch, have low ANI (<95%) to the *E. coli* reference strain, and are not members of *E. coli*$_{\mathrm{BIO}}$ based on *ConSpeciFix* (fig. 2; supplementary table S1, Supplementary Material online). The high bootstrap support of this branch suggests an ancient separation followed by limited recombination with divergent members of *E. coli*, as exemplified by the inclusion of *E. fergusonii* genomes in *PopCoGenet* species-group 4.

(iii) Other enteric species. Based on the ML phylogeny, the only members of other *Escherichia* species that occur in the monophyletic group that contains *E. coli* and *Shigella* are the two aforementioned strains of *E. fergusonii*.

Discussion

Bacterial strains were originally typed as *E. coli* based on their growth characteristics and possession of specific metabolic properties, and, more recently, based on their sequence similarity to one another or to a canonical strain. In addition, there are sufficiently high levels of recombination among strains, despite their asexual mode of reproduction, to warrant the classification of strains to this species based on the BSC. Using homologous exchange as the sole criterion for species assignment, we found that the vast majority of strains currently designated as *E. coli*, or as any of the species of *Shigella*, are all members of a single biological species, which we term *E. coli*$_{\mathrm{BIO}}$. Species-level definitions for the genus *Escherichia* have already been described by Walk (2015) and by Denamur et al. (2021), who have recently proposed a dichotomy between *E. coli* sensu stricto and *E. coli* sensu *lato*. However, such a classification scheme is inadequate because it does not have a biological basis and it can be universally applied (and, moreover, the new species, *E. coli* sensu *lato*, does not include *Shigella*, which belongs to the same species based on all genetic-based methods).

The species boundaries of *E. coli*$_{\mathrm{BIO}}$, which are based solely on homologous recombination within the set of core genes shared by all strains, largely agree with the classifications proposed by other schemes. For example, all methodologies, except *PopCOGenT*, consider the *E. coli* phylogroups of Abram et al. (2021) as comprising a single species, whereas *PopCOGenT* separates them into multiple species. That *PopCOGenT*, which also uses gene flow to delineate species, distinguishes more species than *ConSpeciFix* is due to the fact that *PopCOGenT* considers entire genomes when assigning species membership and can include horizontally transferred regions that are confined to subsets (or even pairs) of strains. Given that events of horizontal gene transfer occur over broad phylogenetic distances (and even between organisms classified to different domains or kingdoms), we chose exclude regions that are sporadically distributed among genomes and to confine analyses to core genes present in all genomes considered.

Strains typed to *Shigella* have been viewed as distinct from *E. coli* because they exhibited certain defining characteristics, including the absence of motility (due to a deletion in the *fliF* operon or insertion in the *flhD* operon) (Al Mamun et al. 1997) and an inability to ferment lactose (due to the lack of one or more *lac* fermentation or permease genes) (Luria and Burrous 1957; Khot and Fisher 2013). Moreover, the four species of *Shigella* are conventionally distinguished from one another by their O serotypes (Wheeler and Stuart 1946; Lan and Reeves 2002) because many of the other diagnostic properties, such as the utilization of mannitol and decarboxylation of ornithine, can be shared among species. However, the traits used to discriminate species of *Shigella*, and *Shigella* from *E. coli*, are often observed in enteroinvasive *E. coli*, which blurs the distinction between these species and genera.

In actuality, *E. coli* and *Shigella* were initially assigned to the same genus due to their similarities but to different species to distinguish pathogenic and nonpathogenic forms (*Bacillus dysenteriae* and *B. coli*, respectively) (Shiga 1898). But as chronicled in figure 3, due to their medical significance, pathogenic strains were elevated to a separate genus in the following decades despite their resemblance to enteroinvasive *E. coli* (Ewing et al. 1952). The close genetic relationship between *E. coli* and *Shigella* was initially recognized in the 1950s based on their ability to reciprocally recombine (Luria and Burrous 1957), but because *Shigella* recombined with *E. coli* at lower frequencies than observed among strains of *E. coli*, each taxon maintained its status as a separate genus. However, subsequent analyses of genetic and genomic characters by DNA hybridization (Brenner et al. 1969, 1972), multilocus enzyme electrophoresis (Ochman et al. 1983), and chromosomal and plasmid gene phylogenies (Pupo et al. 2000; Lan and Reeves 2002) all indicated that strains typed as *Shigella* fall within the variation observed in *E. coli*.

The fact that *Shigella* remains classified as a distinct genus, despite its genetic and phenotypic overlap with *E. coli*, is further complicated by the fact that other named species within the genus *Escherichia* (e.g., *E. albertii* or *E. fergusonii*) do not recombine with *E. coli*, and can be differentiated based on such metabolic characters as 1) the lack of acid production from D-xylose, melibiose, L-rhamnose, and dulcitol for *E. albertii* (Hinenoya et al. 2019) and (2) an incapacity to ferment sorbitol and lactose, coupled with the ability to ferment adonitol, amygdalin, and cellobiose for *E. fergusonii* (Farmer et al. 1985). Taken together, this creates a situation in which the genus *Escherichia* contains multiple distinguishable species, whereas the four named species of *Shigella* should be subsumed within *E. coli*.

To mitigate confusion that might stem from abolishing the genus *Shigella*, Brenner et al. (1973) proposed the use of two separate nomenclatures—one for diagnostic purposes and one for genetic purposes—though it is difficult to see how this serves as an improvement. Lan and Reeves (2002) regarded the species of *Shigella* as serotypes within *E. coli* and removed the generic name, referring to them simply as Boydii, Sonnei, Flexneri, and Dystenteriae. Meier-Kolthoff et al. (2014) proposed including the four *Shigella* species as subspecies of *E. coli*, with nomenclature following guidelines of the Bacteriological Code (Lapage et al. 1992): In this system, for example, *Shigella dysenteriae* would be renamed as *E. coli* subsp. *dysenteriae*, and current members of *E. coli* as *E. coli* subsp. *coli*. Along similar lines, Parks et al. (2020) suggested including the four species of *Shigella* within the genus *Escherichia*, creating *E. sonnei*, *E. boydii*, *E flexneri*, and *E. dysenteriae*. However, based on DNA similarity threshold that they routinely use to define species, these newly named *Escherichia* species should remain within *E. coli* (Parks et al. 2021).

To circumvent issues surrounding the elimination or amendment of species names, we propose that conspecifics defined by the BSC be classified under the heading of a single biological species, as denoted by a subscripted suffix "BIO" adjoined to the latin biome. This procedure would place strains of *E. coli* and *Shigella* under the umbrella of a single biological species, in this case, *E. coli*$_{\mathrm{BIO}}$, but would retain their full names to maintain clinically and historically relevant information. As such, *S. dysenteriae* would be labeled *Ecoli*$_{\mathrm{BIO}}$ *S. dysenteriae*, and current members of *E. coli* as *E. coli*$_{\mathrm{BIO}}$ followed by their strain designation. This resolution mimics the nomenclature developed for serovars of *S. enterica* and does not impose a taxonomic revision but is nevertheless useful in indicating which strains are members of the same biological species.

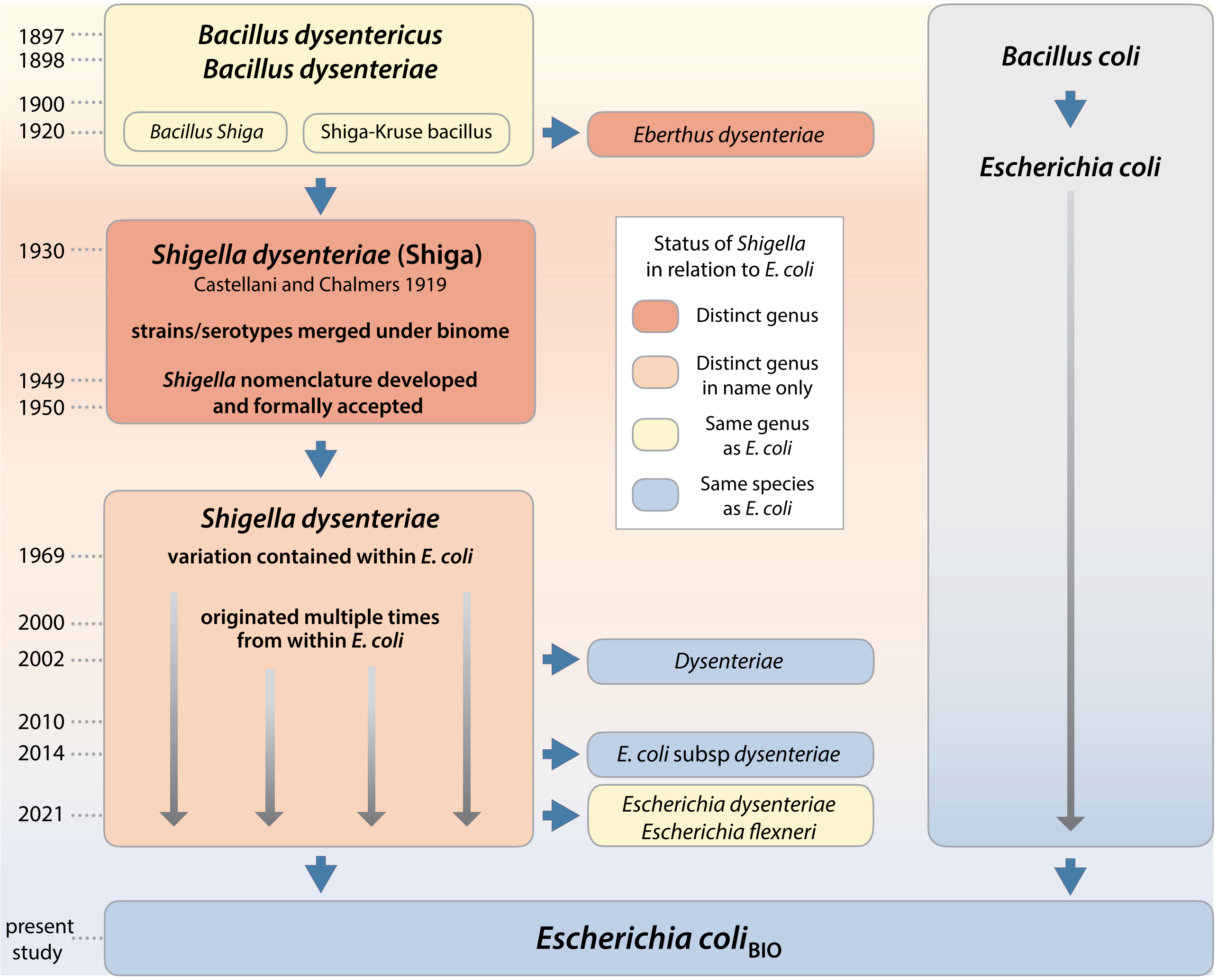

Fig. 3. Chronological changes in the *S. dysenteriae* and *E. coli* nomenclature. References used to produce this figure are listed in Supplementary Material online.

The retention of strain appellations in the proposed scheme maintains consistency with the traditional nomenclature and avoids conflict with clinical identification and applications.

Despite the ability of *E. coli* and other bacteria to acquire genes from distant sources, recombination between shared homologs occurs primarily among sequences with high levels of similarity (Shen and Huang 1986; Rayssiguier et al. 1989; Roberts and Cohan 1993; Matic et al. 1995; Zawadzki et al. 1995; Majewski and Cohan 1999). This feature enables a natural classification of bacteria into species based on their propensity for homologous exchange, a biological criterion that can be applied to all lifeforms. To assure the universality of species definition, it is, therefore, necessary to confine analyses of recombination to the core set of genes shared among genomes. Those sequences with rare or sporadic distributions, as might originate from infrequent or independent events of horizontal gene transfer between taxa, occur in eukaryotes as well as bacteria (Akanni et al. 2015; Husnik and McCutcheon 2018; Wu et al. 2022), and can involve very distant taxa. Thus, such genes are best excluded from consideration when delineating species boundaries

Species, when defined by their capacity for gene flow, constitutes the only taxonomic rank based on a biological process rather than an arbitrary or subjective criterion (Bapteste and Boucher 2009; Lawrence and Retchless 2009). The recent availability of genome sequence data now allows the application of the same parameters for delineating species boundaries to asexual lifeforms (bacteria, archaea, viruses), all of which were previously considered as not amenable to classification based on the BSC (Donoghue 1985; Rosselló-Mora and Amann 2001; Costechareyre et al. 2009). This uniformity in defining species has implications beyond taxonomic classification in that the formation of equivalently defined species allows comparisons of evolutionary processes across all lifeforms (Staley 2009) and more accurate inferences about the rates and patterns of speciation in different groups of organisms.

The ANI divergence between strains in *E. coli*$_{\mathrm{BIO}}$ can be as much as 6.1%. This relatively high level of divergence between members of the same species is evident at other taxonomic ranks: for example, between *E. coli*$_{\mathrm{BIO}}$ and other species of *Escherichia* (*E. albertii*, *E. fergusonii*, and Clades II, III, IV, and V), sequence divergence ranges from 8% to 12%, and between *Escherichia* and its sister genus, *Salmonella*, the divergence among shared genes averages 15%. This degree of variation within and among species sharply contrasts the situation in, say, humans, in which the sequence divergence between homologs from two individuals is a mere 0.1% (Lek et al. 2016), and there is only a 0.5% difference to our sister species *Homo neanderthalensis* (Noonan et al. 2006) and 1.2% difference to our sister genus *Pan* (Carroll 2003).

The genetic approach to bacterial identification and classification, which began in the 1960s (Marmur et al. 1963), is more instructive than metabolic typing, which relies on a subjective set of diagnostic features (which themselves can originate by different means within and across species, and are often not discrete) (Priest et al. 1993). Moreover, a genetic delineation of biological species divulges the actual extent of phenotypic variation that is present in a species. For example, *E. coli* is traditionally distinguished from *S. enterica* as being Lac-positive and Citrate-negative; however, many members of *E. coli*, including most *Shigella* and many pathogenic strains, are lactose nonfermenters, and citrate-positive strains of *E. coli* have been reported (Ishiguro et al. 1978) and evolved (Blount et al. 2012). All of the classification methods that we evaluated indicate that the majority of *E. coli* and *Shigella* represent a single species; however, our analyses, based on the propensity for homologous exchange, provide the genetic basis for this conclusion.

Reports that strains within some phylogenetic clades of *E. coli* recombine at higher frequencies within one another than with members of other clades—as might be expected if homologous exchange relied wholly on the degree of sequence similarity—has been interpreted as evidence of incipient speciation (Didelot et al. 2012; Kang et al. 2021). However, applying the principles of the BSC, we established the genetic boundaries of *E. coli*, termed *E. coli*$_{\mathrm{BIO}}$, which was found to include all members of the genus *Shigella*, exclude only 12 genomes currently classified as *E. coli* in the NCBI database, and to be distinct from the other named species within the genus. Aside from its utility in classification and systematics, applying a universal species concept and identifying populations that readily engage in gene flow is valuable for studying novelty and diversity within species, and the mechanisms by which bacterial species form.

Materials and Methods

Genomes Analyzed

We downloaded a total of 1,635 genome sequences classified as *E. coli* by the NCBI database (www.ncbi.nlm.nih.gov/), which included representatives of the species within *E. coli* recognized by the GTDB v207 (April 8, 2022; gtdb.ecogenomic.org/) (Parks et al. 2018) and the *E. coli* phylogroups of Abram et al. (2021). To maximize core-genome size, we restricted our analyses to all complete, ungapped genomes available at the time of analysis. Additionally, we retrieved complete genome sequences for *Escherichia* species other than *E. coli* (*E. albertii*, $n = 1$; *E. fergusonii*, $n = 2$; and 53 *Escherichia* strains not assigned to species), the five *Escherichia* phylogroups (CI–CV) described by Walk et al. (2009) ($n = 12$), the four named species of *Shigella* (*S. flexneri*, $n = 28$; *S. boydii*, $n = 9$; *S. dysenteriae*, $n = 5$; *S. sonnei*, $n = 34$), one unassigned strain of *Shigella*, *S. enterica* ($n = 106$), *S. bongori* ($n = 5$), *E. cloacae* ($n = 3$), *Klebsiella pneumoniae* ($n = 3$), *P. mirabilis* ($n = 2$), and one strain each of *Citrobacter koseri*, *Citrobacter rodentium*, *Cronobacter sakazakii*, *Cronobacter turicensis*, *Enterobacter cancerogenus*, *Enterobacter lignolyticus*, *Enterobacter* sp., and *Klebsiella variicola*. Accession numbers, strain, and species assignments and nomenclature in the NCBI and GTDB databases (and Walk et al. and Abram et al. phylogroups, where applicable), and taxonomic classification based on the schemes implemented in this study, are presented in supplementary table S1, Supplementary Material online.

Initial assignment of genomes to a named species followed the nomenclature designated in the NCBI database. Currently, the NCBI database uses an ANI metric to assign genomes to species, with species-level assignments representing strains having >95% ANI for at least 90% of the shared portions of their genomes (Ciufo et al. 2018). Assignments to bacterial genera do not rely on fixed ANI cutoffs, and accommodations are made for certain genera, such as *Shigella*, which is known to be polyphyletic and contained within *E. coli*. The GTDB also defines species based on >95% ANI to a representative strain, except in cases in which representatives from different species, as obtained from cross-referencing the LPSN, BacDive, StrainInfo, and NCBI databases, are very closely related and a higher threshold must be applied.

Classification Methods and Detecting Gene Flow Among Strains

Complete genomes were partitioned into sets according to their nomenclature, phylogenetic groupings, or degree of DNA similarity. For each selected set, we evaluated the extent of recombination among genomes and the consistency among the taxonomic assignments based on different methods and criteria. We applied and compared the following methodologies for species-level classification:

(*i*) Average Nucleotide Identity (ANI). We calculated ANI, a whole-genome metric for evaluating the degree of DNA sequence identity, using FastANI (Jain et al. 2018). When assigning strains to *E. coli* by this approach, we applied the "one-to-many" option and used the type strain *E. coli* ATCC 11775 (https://lpsn.dsmz.de/species/escherichia-coli), which was fully sequenced in 2019 (Wadley et al. 2019), as the species representative to which all other genomes were compared. As such, all genomes with an ANI ≥ 95% to ATCC 11775 would be designated members of *E. coli*. We also applied the "many-to-many" option in FastANI employing the same DNA identity threshold.

(ii) *ConSpeciFix*. To identify species boundaries according to the precepts of the BSC, we used the *ConSpeciFix* v1.3.0 pipeline (Bobay et al. 2018), which recognizes genomes as belonging to the same species based on their capacity for gene flow. In *ConSpeciFix*, gene flow is estimated by assessing the extent of homologous recombination among genes in the core genome. The core genome is built with single-copy orthologs that occur in at least 85% of all strains considered, with single-copy orthologs aligned in MAFFT v7 (Katoh and Standley 2013) and merged into a single concatenate. Based on the core-genome phylogeny, *ConSpeciFix* calculates the number of homoplastic alleles (*h*, recombinant sites, i.e., those not related by vertical ancestry) relative to the number of nonhomoplastic alleles (*m*, vertically transmitted mutations), using a distance-based approach, with higher h/m ratios indicative of more recombination (Bobay et al. 2018).

To calculate h/m ratios, which estimates the limits of recombination among genomes, a representative of a different species or phylogenetic clade (the "test lineage") is included in a set of genomes previously determined by *ConSpeciFix* to recombine with one another (the "reference lineages"). Disruptions or reductions in h/m values caused by the inclusion of the test lineage indicate that the test lineage does not recombine with the reference lineages and, thus, belongs to a different species based on the BSC. Analyses were extended to include different combinations of reference genomes and test lineages in order to define species boundaries.

To define the set of *E. coli* genomes that constitute the reference lineages, we initially examined the 1,635 complete genomes available in the NCBI database. Because it was computationally infeasible to run the entire set of genomes through the *ConSpeciFix* pipeline as a single group, we randomly subdivided those strains designated as *E. coli* into subgroups of 150 genomes and analyzed each subgroup separately. These analyses identified 12 genomes that were reproductively isolated from the rest of the *E. coli* genomes and, therefore, removed from the set of NCBI-designated *E. coli* strains that were randomly sampled to produce new sets of reference lineages for assessing recombination with test lineages. Within the *ConSpeciFix* pipeline, we also tested the extent of gene flow between *E. coli* and representative genomes of other species of *Escherichia*, the four species of the genus *Shigella*, and several non-*Escherichia* species of *Enterobacteriaceae* (supplementary table S1, Supplementary Material online).

(iii) *PopCOGenT*. Another approach for defining bacterial species based on gene exchange, *PopCOGenT* (Arevalo et al. 2019), uses the presence of anomalously similar regions to infer events of gene transfer between genomes. Unlike *ConSpeciFix*, which deduces the source and ancestry of each polymorphic site, *PopCOGenT* is based on the premise that SNPs occur more frequently in vertically inherited genes than in recently transferred regions, and that genomes engaging in gene exchange will have longer and more frequent stretches of identical regions. In both *ConSpeciFix* and *PopCOGenT*, genomes connected through gene exchange are considered members of the same species, but the methods differ in criteria for identifying recombination, and possibly, species boundaries.

We applied *PopCOGenT* to a total of 128 genomes, including many that were originally assigned to *E. coli* but whose species status has been questioned or changed. In addition to strains consistently classified as *E. coli*, this set included strains labeled as *E. ruysiae* and *E. marmotae* by the GTDB, the 12 *E. coli* genomes from the NCBI database recognized by *ConSpeciFix* as being reproductively isolated from the rest of the species, representatives of the CI–CV phylogroups (Walk et al. 2009) as well as representatives of other species (*E. albertii*, *E. fergusonii*, *Shigella* sp. PAMC 28760, *Shigella sonnei*, *S. flexneri*, *S. dysenteriae*, *Shigella boydii*, *S. enterica*, and *S. bongori*) and the phylogroups of Abram et al. (2021) (supplementary table S1, Supplementary Material online).

To compare the species assignments and nomenclature of *Escherichia* and *Shigella* strains across different classification schemes, we first dereplicated the dataset, retaining only a single representative strain for cases in which multiple genomes averaged >99.8% nucleotide identity. This dereplication yielded a reduced, but otherwise identical, phylogeny that was used for *ConSpeciFix*. The maximum-likelihood phylogeny of the 70 genomes remaining after dereplication was generated with RaxML (Stamatakis, 2014) using the analysis tool PhaME (Shakya et al. 2020). To generate this phylogeny, PhaME uses nucmer2 (Delcher et al., 2002) to first aligns each genome against itself in order to identify and eliminate the repeated regions within a genome, and then aligns the repeat-free genomes against the selected reference genome. The RaxML phylogeny of the aligned genomes with associated bootstrap branch-support values was built using the evolutionary model GTR and the rate heterogeneity model GAMMA with an estimation of invariable sites (GTRGAMMAI).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Kim Hammond for assistance with figures. This work was supported by the National Science Foundation Dimensions (grant number 1831730) and the National Institutes of Health (R35GM118038) to H.O.

Author Contributions

H.O. conceived the study; H.O. and M.C.S. supervised the research activity planning and execution; M.C.S. and R.H. analyzed and interpreted the data; M.C.S., R.H., and H.O. wrote the paper, read, and approved the final manuscript.

Data Availability Statement

All complete genomes used in this analysis are available from the NCBI database (https://www.ncbi.nlm.nih.gov/) using accession numbers listed in supplementary table S1, Supplementary Material online.

References

Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, Ussery DW. 2021. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun Biol*. **4**:1–12.

Akanni WA, Siu-Ting K, Creevey CJ, McInerney JO, Wilkinson M, Foster PG, Pisani D. 2015. Horizontal gene flow from Eubacteria to Archaebacteria and what it means for our understanding of eukaryogenesis. *Philos Trans R Soc B Biol Sci*. **370**: 20140337.

Al Mamun AA, Tominaga A, Enomoto M. 1997. Cloning and characterization of the region III flagellar operons of the four *Shigella* subgroups: genetic defects that cause loss of flagella of *Shigella boydii* and *Shigella sonnei*. *J Bacteriol*. **179**:4493–4500.

Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. 2019. A reverse ecology approach based on a biological definition of microbial populations. *Cell*. **178**:820–834.

Bapteste E, Boucher Y. 2009. *Epistemological impacts of horizontal gene transfer on classification in microbiology*. Clifton, New Jersey: Humana Press.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*. **489**:513–518.

Bobay LM, Ellis BSH, Ochman H. 2018. Conspecifix: classifying prokaryotic species based on gene flow. *Bioinformatics*. **34**: 3738–3740.

Brenner DJ, Falkow S. 1971. C. Molecular relationships among members of the Enterobacteriaceae. *Adv Genet*. **16**:81–118.

Brenner DJ, Fanning GR, Johnson KE, Citarella RV, Falkow S. 1969. Polynucleotide sequence relationships among members of Enterobacteriaceae. *J Bacteriol*. **98**:637–650.

Brenner DJ, Fanning GR, Miklos GV, Steigerwalt AG. 1973. Polynucleotide sequence relatedness among *Shigella* species. *Int J Syst Evol Microbiol*. **23**:1–7.

Brenner DJ, Fanning GR, Skerman FJ, Falkow S. 1972. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J Bacteriol*. **109**:953–965.

Carroll SB. 2003. Genetics and the making of Homo sapiens. *Nature*. **422**:849–857.

Chaudhuri RR, Henderson IR. 2012. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol*. **12**:214–226.

Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, DiCuccio M. 2018. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol*. **68**:2386.

Costechareyre D, Bertolla F, Nesme X. 2009. Homologous recombination in Agrobacterium: potential implications for the genomic species concept in bacteria. *Mol Biol Evol*. **26**:167–176.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. **30**: 2478–2483.

Denamur E, Clermont O, Bonacorsi S, Gordon D. 2021. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol*. **19**:37–54.

Didelot X, Méric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics*. **13**:1–15.

Donoghue MJ. 1985. A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist*. **88**: 172–181.

Dykhuizen DE, Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol*. **173**:7257–7268.

Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*. **34**:2371–2375.

Escherich T. 1885. Die Darmbakterien des Neugeborenen und Säuglings. *Fortschr Med*. **3**:515–522.

Ewing WH, Hucks MC, Taylor MW. 1952. Interrelationship of certain *Shigella* and *Escherichia* cultures. *J Bacteriol*. **63**:319–325.

Farmer JJ, Fanning GR, Davis BR, O'Hara CM, Riddle C, Hickman-Brenner FW, Asbury MA, Lowery VA, Brenner DJ. 1985. *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of Enterobacteriaceae isolated from clinical specimens. *J Clin Microbiol*. **21**:77–81.

Herzer PJ, Inouye S, Inouye M, Whittam TS. 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *J Bacteriol*. **172**:6175–6181.

Hinenoya A, Ichimura H, Awasthi SP, Yasuda N, Yatsuyanagi J, Yamasaki S. 2019. Phenotypic and molecular characterization of *Escherichia albertii*: further surrogates to avoid potential laboratory misidentification. *Int J Med Microbiol*. **309**:108–115.

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. **16**:67–79.

Ishiguro N, Oka C, Sato G. 1978. Isolation of citrate-positive variants of *Escherichia coli* from domestic pigeons, pigs, cattle, and horses. *Appl Environ Microbiol*. **36**:217–222.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. **9**:1–8.

Kang Y, Yuan L, Shi X, Chu Y, He Z, Jia X, Lin Q, Ma Q, Wang J, Xiao J, *et al.* 2021. A fine-scale map of genome-wide recombination in divergent *Escherichia coli* population. *Brief Bioinform*. **22**:bbaa335.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**:772–780.

Kauffmann F. 1944. Zur Serologie Der Coli-Gruppe. *Acta Pathol Microbiol Scand*. **21**:20–45.

Khot PD, Fisher MA. 2013. Novel approach for differentiating *Shigella* species and *Escherichia coli* by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J Clin Microbiol*. **51**:3711–3716.

Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol*. **187**:6258–6264.

Koser SA. 1923. Utilization of the salts of organic acids by the colon-aerogenes group. *J Bacteriol*. **8**:493–520.

Lan R, Reeves PR. 2002. *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes Infect*. **4**:1125–1132.

Lapage SP, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, Clark WA. 1992. *International code of nomenclature of bacteria bacteriological code, 1990 revision*. Washington (DC): ASM Press.

Lawrence JG, Retchless AC. 2009. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. In: Gogarten MB, Gogarten JP, Olendzenski LC, editors. *Horizontal gene transfer. Methods in molecular biology*. Vol. 532. Pittsburgh: Humana Press. p. 29–53.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**:285–291.

Lu S, Jin D, Wu S, Yang J, Lan R, Bai X, Liu S, Meng Q, Yuan X, Zhou J, *et al.* 2016. Insights into the evolution of pathogenicity of

Escherichia coli from genomic analysis of intestinal *E. coli* of Marmota himalayana in Qinghai-Tibet plateau of China. *Emerg Microbes Infect*. **5**:1–9.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A*. **108**:7200–7205.

Luria SE, Burrous JW. 1957. Hybridization between *Escherichia coli* and *Shigella*. *J Bacteriol*. **74**:461–476.

Majewski J, Cohan FM. 1999. DNA sequence similarity requirements for interspecific recombination in Bacillus. *Genetics*. **153**:1525–1533.

Marmur J, Falkow S, Mandel M. 1963. New approaches to bacterial taxonomy. *Annu Rev Microbiol*. **17**:329–372.

Matic I, Rayssiguier C, Radman M. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell*. **80**:507–515.

Mayr E. 1942. *Systematics and the origin of species*. Cambridge: Harvard University Press.

Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, Rohde C, Rohde M, Fartmann B, Goodwin LA, *et al.* 2014. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci*. **9**:2.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, *et al.* 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science*. **314**:1113–1118.

Ochman H, Whittam TS, Caugant DA, Selander RK. 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *Microbiology*. **129**:2715–2726.

Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*. **38**:1079–1086.

Parks DH, Chuvochina M, Reeves PR, Beatson SA, Hugenholtz P. 2021. Reclassification of *Shigella* species as later heterotypic synonyms of *Escherichia coli* in the Genome Taxonomy Database. *bioRxiv*. preprint bioRxiv:2021.09.22.461432.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. **36**:996–1004.

Priest FG, Tsubota K, Austin B. 1993. *Modern bacterial taxonomy*. Berlin, Heildelberg: Springer Science & Business Media.

Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A*. **97**: 10567–10572.

Rayssiguier C, Thaler DS, Radman M. 1989. The barrier to recombination between *Escherichia coli* and Salmonella typhimurium is disrupted in mismatch-repair mutants. *Nature*. **342**:396–401.

Roberts MS, Cohan FM. 1993. The effect of DNA sequence divergence on sexual isolation in Bacillus. *Genetics*. **134**:401–408.

Rolland K, Lambert-Zechovsky N, Picard B, Denamur E. 1998. *Shigella* and enteroinvasive *Escherichia coli* strains are derived from distinct ancestral strains of *E. coli*. *Microbiology*. **144**:2667–2672.

Rosselló-Mora R. 2006. DNA–DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In: Stackebrandt E, editor. *Molecular identification, systematics, and population structure of prokaryotes*. Berlin, Heidelberg: Springer. p. 23–50.

Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. *FEMS Microbiol Rev*. **25**:39–67.

Shakya M, Ahmed SA, Davenport KW, Flynn MC, Lo CC, Chain PSG. 2020. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Sci Rep*. **10**:1723.

Shen P, Huang H V. 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics*. **112**:441–457.

Shiga K. 1898. Ueber den Erreger der Dysenterie in Japan. *Zentralbl Bakteriol Mikrobiol Hyg (Vorläufige Mitteilung)*. **23**:599–600.

Staley JT. 2009. Universal species concept: pipe dream or a step toward unifying biology? *J Ind Microbiol Biotechnol*. **36**: 1331–1336.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**: 1312–1313.

Tindall BJ, Rosselló-Móra R, Busse HJ, Ludwig W, Kämpfer P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol*. **60**:249–266.

Wadley TD, Jenjaroenpun P, Wongsurawat T, Ussery DW, Nookaew I. 2019. Complete genome and plasmid sequences of *Escherichia coli* type strain ATCC 11775. *Microbiol Resour Announc*. **8**: e00046-19.

Walk ST. 2015. The “Cryptic” *Escherichia*. *EcoSal Plus*. **6**. doi: 10.1128/ecosalplus.ESP-0002-2015

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus *Escherichia*. *Appl Environ Microbiol*. **75**:6534–6544.

Wheeler KM, Stuart CA. 1946. The mannitol-negative *Shigella* group. *J Bacteriol*. **51**:317–325.

Wu F, Speth DR, Philosof A, Crémière A, Narayanan A, Barco RA, Connon SA, Amend JP, Antoshechkin IA, Orphan VJ. 2022. Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. *Nat Microbiol*. **7**:200–212.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*. **12**:635–645.

Zawadzki P, Roberts MS, Cohan FM. 1995. The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics*. **140**:917–932.