INTRODUCTION

Transcription elongation is punctuated by pauses that serve important functions in permitting correct folding of structural RNA, efficient coupling of transcription and translation, and ensuring efficient transcription termination at the correct site (Saba et al., 2019). Whilst most pausing events serve an important function, on occasion RNA polymerase (RNAP) is unable to restart transcription and must be removed from the DNA to prevent damaging collisions with the DNA replication machinery or other transcription complexes (Adelman & Lis, 2012; Gupta et al., 2013; Pomerantz & O’Donnell, 2008, 2010; Rocha, 2004). Several systems used to resolve stalled transcription complexes have been characterized; for example, Mfd has been shown to bind to stalled transcription complexes and prevent RNAP from sliding back to the DNA to restart transcription (Adelman & Lis, 2012; Gupta et al., 2013; Pomerantz & O’Donnell, 2008, 2010; Rocha, 2004).

Multiple classes and isoforms of the RNA polymerase recycling motor protein HelD

Joachim S. Larsen1, Michael Miller1, Aaron J. Oakley2, Nicholas E. Dixon2, Peter J. Lewis1,2

1School of Environmental and Life Sciences, University of Newcastle, Callaghan, New South Wales, Australia
2School of Chemistry and Molecular Bioscience, University of Wollongong and Illawarra Health and Medical Research Institute, Wollongong, New South Wales, Australia

Correspondence
Peter J. Lewis, School of Environmental and Life Sciences, University of Newcastle, Callaghan, NSW 2308, Australia.
Emails: peter.lewis@newcastle.edu.au; lewisp@uow.edu.au

Funding information
Australian Research Council, Grant/Award Number: DP210100365

Abstract

Efficient control of transcription is essential in all organisms. In bacteria, where DNA replication and transcription occur simultaneously, the replication machinery is at risk of colliding with highly abundant transcription complexes. This can be exacerbated by the fact that transcription complexes pause frequently. When pauses are long-lasting, the stalled complexes must be removed to prevent collisions with either another transcription complex or the replication machinery. HelD is a protein that represents a new class of ATP-dependent motor proteins distantly related to helicases. It was first identified in the model Gram-positive bacterium Bacillus subtilis and is involved in removing and recycling stalled transcription complexes. To date, two classes of HelD have been identified: one in the low G+C and the other in the high G+C Gram-positive bacteria. In this work, we have undertaken the first comprehensive investigation of the phylogenetic diversity of HelD proteins. We show that genes in certain bacterial classes have been inherited by horizontal gene transfer, many organisms contain multiple expressed isoforms of HelD, some of which are associated with antibiotic resistance, and that there is a third class of HelD protein found in Gram-negative bacteria. In summary, HelD proteins represent an important new class of transcription factors associated with genome maintenance and antibiotic resistance that are conserved across the Eubacterial kingdom.

KEYWORDS

gene expression regulation, helicases, phylogenetic analysis, RNA polymerase

1 | INTRODUCTION

Transcription elongation is punctuated by pauses that serve important functions in permitting correct folding of structural RNA, efficient coupling of transcription and translation, and ensuring efficient transcription termination at the correct site (Saba et al., 2019). Whilst most pausing events serve an important function, on occasion RNA polymerase (RNAP) is unable to restart transcription and must be removed from the DNA to prevent damaging collisions with the DNA replication machinery or other transcription complexes (Adelman & Lis, 2012; Gupta et al., 2013; Pomerantz & O’Donnell, 2008, 2010; Rocha, 2004). Several systems used to resolve stalled transcription complexes have been characterized; for example, Mfd has been shown to bind to stalled transcription complexes and prevent RNAP from sliding back to the DNA to restart transcription (Adelman & Lis, 2012; Gupta et al., 2013; Pomerantz & O’Donnell, 2008, 2010; Rocha, 2004).
complexes (either a stochastic pause during transcription of structured RNA or at a site of DNA damage), physically removing it from the DNA or restarting it via a RecG-like ATPase motor domain (Ghodke et al., 2020; Ho et al., 2018; Kang et al., 2021; Le et al., 2018; Ragheb et al., 2021; Shi et al., 2020; Westblade et al., 2010). In _B. subtilis_ RNaseJ1 clears stalled RNAP using a torpedo mechanism (5′–3′ exonuclease activity followed by RNAP displacement) (Sikova et al., 2020), and in _Escherichia coli_ the helicase protein RapA is important in recycling RNAP (Liu et al., 2015). UvrD/PcrA in concert with Gre factors has been reported to act on RNAP stalled at a DNA lesion, binding to the complex and using the energy of ATP hydrolysis to backtrack away from the lesion to allow repair systems access to the damaged DNA (Epshtein et al., 2014; Hawkins et al., 2019), although it now appears that the role of these helicases is in preventing the formation of, and resolving, R-loops (RNA-DNA hybrids) that can have a detrimental effect on DNA replication (Urrutia-Irazabal et al., 2021).

An additional system identified in Gram-positive bacteria required for recycling stalled transcription complexes involves the action of the motor protein HelD (Wiedermannova et al., 2014). The designation of HelD (also called helicase IV) was originally made for a protein identified in _E. coli_ as a weakly processive 3′–5′ DNA helicase (Wood & Matson, 1987). To avoid confusion with the separate classes of HelD proteins that are the focus of this work, the _E. coli_ protein will be referred to as helicase IV. Based on conserved sequence motifs Helicase IV is a superfamily 1 (SF1) helicase, related to housekeeping helicase UvrD/PcrA (Figure 1). The _B. subtilis_ gene _yvgS_ was assigned the name _helD_ based on limited protein sequence conservation to helicase IV (Wiedermannova et al., 2014), although the proteins differed with respect to domain organization (Koval et al., 2019; Wiedermannova et al., 2014) (Figure 1). Little functional, and no structural information is available for helicase IV, although a model generated by AlphaFold2 (Jumper et al., 2021) enables tentative comparison of UvrD/PcrA, helicase IV, and _B. subtilis_ HelID (Figure 1). Helicase IV and HelID show similarity with UvrD/PcrA around the well-defined 1A and 2A helicase domains (blue and orange, respectively, Figure 1a), but not in other structural motifs associated with helicase activity (UvrD/PcrA domains 1B and 2B). Both helicase IV and HelID have N-terminal domains not present in UvrD/PcrA helicases, and helicase IV has a putative 1B domain which may account for its reported helicase activity, whilst in the equivalent 1B domain position HelID contains an unrelated sequence that folds into a novel clamp-arm (CA) structure important in transcription recycling (Newing et al., 2020; Wiedermannova et al., 2014). Whilst UvrD/PcrA and helicase IV have helicase activity, HelID shows none suggesting it has evolved from an SF1-type helicase into a transcription recycling factor that utilises the energy from ATP hydrolysis catalysed by its helicase motifs for its transcription-related activity.

Studies on HelID from low G+C ( _Bacillus subtilis_ ) and high G+C ( _Mycobacterium smegmatis_ ) Gram-positives revealed that there are two distinct classes of the enzyme, confirmed by phylogenetic and structural analyses (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020). Class I HelID was described from _B. subtilis_, whilst the structurally distinct Class II enzyme was identified in _M. smegmatis_ (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020). Class I and II HelIDs have similar motor domains but differ in the structure of their arms and the mechanism by which these arms perform the mechanical activity of removing nucleic acids and recycling RNAP (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020).

---

**FIGURE 1** Relationship between UvrD/PcrA and helicase IV/HelID proteins. The left side shows scaled linear representations of the domain organization of superfamily 1 (SF1) helicase UvrD/PcrA (top), _Escherichia coli_ helicase IV (middle), and _B. subtilis_ HelID (bottom). A scale bar (amino acids) is shown at the bottom. The right-hand side shows structures, aligned via their 1A and 2A domains, with domains colored corresponding to the left panels. Top, UvrD (PDB ID 3LFU); middle, helicase IV (AlphaFold2 model, AF2); bottom, HelD (taken from RNAP-HelID complex PDB ID 6WVK). 1A, 2A, and 2B refer to conserved SF1 helicase domains. NTD, SCA, and CA refer to the AlphaFold2 modeled N-terminal domain of helicase IV and the secondary channel arm and clamp arm of HelID, respectively.
The recent structures of HelD from *B. subtilis* and *M. smegmati*s bound to core RNAP (α₂ββ′ω) (Kouba et al., 2020; Newing et al., 2020) are shown in Figure 2a and b, along with the Class I *B. subtilis* (Figure 2c) and Class II *M. smegmatis* (Figure 2d) enzymes. HelD has an unusual mode of action dependent on two arms (CA and SCA, Figure 2c and d) attached to the central UvrD-like ATPase motor domain (Head and Torso, Figure 2c and d), in which nucleic acids are pushed out of the active site whilst the DNA binding clamp and RNA exit channels are simultaneously opened, leading to the release of the stalled RNAP (Newing et al., 2020). This recycling activity is powered by ATP hydrolysis and the mechanical action of the two arms that flank the motor domain. In the Class I HelD, the long SCA (Figure 2a and c) can physically remove nucleic acids from the active site (dotted circle in Figure 2a), whereas in the Class II HelD the SCA is too short, and instead nucleic acid removal is performed by a CA insert called the PCh-loop (Figure 2b and d) (Kouba et al., 2020; Newing et al., 2020). Recent reports also suggest that some Class II HelDs (from *M. abscessus* and *Streptomyces venezuelae*) can confer rifampicin resistance through removal of rifampicin by the PCh-loop (Hurst-Hess et al., 2021; Surette et al., 2021). Recent reports also suggest that some Class II HelDs (from *M. abscessus* and *Streptomyces venezuelae*) can confer rifampicin resistance through removal of rifampicin by the PCh-loop (Hurst-Hess et al., 2021; Surette et al., 2021).

In this work, we take advantage of the recent structural information to compile a detailed phylogenetic analysis of HelD showing that many organisms contain more than one (up to 5) different versions of HelD, that the genes encoding these enzymes are all expressed, that HelD is likely to have been acquired by horizontal gene transfer in Gram-negative *Bacteroides* and Gram-positive *Coriobacteria* and *Acidimicrobia*, and that there is a third Class of HelD found in the Gram-negative *Deltaproteobacteria*.

**FIGURE 2** The two known structural classes of HelD. Panel A shows the structure of the *B. subtilis* RNAP-Class I HelD complex (PDB ID 6WVK). Panel B shows the *M. smegmatis* RNAP-Class II HelD complex (PDB ID 6YYS; state II). RNAP subunits and HelDs are colored identically in both panels with the transparency of the β′ subunit set at 50% so that HelD structures adjacent to the RNAP active site region (dashed circles) can be more easily visualized. Panels C and D show HelD structures in the same orientation as in Panels A and B, with the ATP binding site colored in blue and the PCh-loop from *M. smegmatis* HelD colored in yellow (see text for details).

**2 | EXPERIMENTAL PROCEDURES**

**2.1 | Sequence retrieval and analysis**

The sequence of *B. subtilis* 168 HelD (UniProtKB ID: O32215) was used to search for homologues on 11/11/2020 using the NCBI Conserved Domain Architecture Retrieval Tool (Geer et al., 2002), which identified 13,781 sequences, which were trimmed to 11,821 to remove partial sequences (<600 aa). To aid subsequent analyses, particularly for the study of multiple copies of helD genes, the original sequences were used to search complete reference genomes from the KEGG (https://www.kegg.jp) and JGI (https://jgi.doe.gov) databases. HelD and RpoB sequences retrieved from these complete genomes were used for subsequent phylogenetic studies.

**2.2 | Construction of phylogenetic trees**

Selected sequences were aligned using MAFFT (Katoh et al., 2002, 2019) with default settings. Sequence alignments were then trimmed using Gblocks (https://gblocks.tangerine.fr). The best-fitting model (LG) was determined using ProtTest 3 (Darriba et al., 2011) and phylogenetic trees were constructed using MrBayes 3.2 (Huelsenbeck & Ronquist, 2001; Ronquist et al., 2012), which were run until the standard deviation was below 0.01. Phylogenetic trees were also made on MEGA-X (Kumar et al., 2018), using the Maximum Likelihood statistical method with 1000 bootstrap replications, and...
using RAxML (Stamatakis, 2006) using default settings. All trees had the same topology. Trees were visualised using iTol (Letunic & Bork, 2019).

### 2.3 Transcriptome data and analysis

Gene expression data were obtained from datasets deposited in the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) and were: *B. subtilis* 168 (Revilla-Guarinos et al., 2020); *B. cereus* F837/76 (Jessenberger et al., 2019); *Clostridium perfringens* 13 (Soncini et al., 2020); *Streptomyces coelicolor* A3(2) (Jeong et al., 2016); *Mycobacterium smegmatis* MC2-155 (Feng et al., 2020); *Mycobacterium fringens* DK1622 (SRA accession code: PRJNA516475); *Bacteroides vulgatus* ATCC8482 (SRA accession code: PRJNA473003). Reads were mapped to the respective reference genome sequences, and gene expression levels were calculated in Genious Prime 2020.2.3 (https://www.geneious.com). Transcript per million (TPM) values were used for comparison of *helD* expression levels *cf. rpoB*, and *pcrA/uvrD* (for *S. coelicolor* A3(2)).

### 2.4 Structure modeling

RNAP RpoB (β) and RpoC (γ′) subunits from *M. xanthus* DK1622 were modeled in SWISS-MODEL (Waterhouse et al., 2018) using *E. coli* RNAP, PDB ID: 6ALF (Kang et al., 2017) as a defined template. The *M. xanthus* HelD structure was modeled using i-Tasser (Yang et al., 2015) with output model 1 (C-score ~0.48) selected for presentation in this work. Structural images used in this work were prepared in ChimeraX (Pettersen et al., 2021).

### 3 RESULTS AND DISCUSSION

#### 3.1 Distribution and phylogeny of HelD

Searching for HelD-like sequences using the conserved domain architecture retrieval tool (CDART; NCBI) portal identified >13,000 hits. Additional searches using NCBI BLASTP suggest that there are substantially more sequences in the database, but many of these are from incomplete genomes and/or metagenomic sequencing projects, making systematic identification and classification of sequences unfeasible, particularly in cases where an organism carries more than one *helD* gene (see below). Nevertheless, it is clear that HelD is widely distributed in the eubacteria, especially in the Firmicutes and Actinobacteria phyla of the Gram-positive eubacterial domain. To date, we have not detected HelD-like sequences in Archaea or Eucarya. Previously, Newing et al. (Newing et al., 2020) showed that HelD sequences fall into two classes, which was confirmed at the structural and functional level in comparing HelD proteins from the Firmicutes and Actinobacteria (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020). Using a wider range of carefully curated sequences from complete genomes identified from the initial CDART search, an unrooted phylogenetic tree was constructed to enable a more detailed understanding of HelD distribution and phylogeny which was compared against the RNAP RpoB (β) subunit (Figure 3; note different tree scales).

Four features are clear from this tree (Figure 3a): 1. HelD is also present in Gram-negative bacteria; 2. The third class of HelD is present in the Deltaproteobacteria; 3. In some organisms HelD has been ancestrally acquired by horizontal gene transfer; 4. Many organisms contain more than one *helD* gene, with the Firmicutes, Clostridia, Acidimicrobia, and Deltaproteobacteria having up to three, and the Actinobacteria up to five.
FIGURE 3 Unrooted phylogenetic trees of HelD (A) and RpoB (B) sequences constructed by Bayesian analysis. Tree scale representing amino acid substitutions per site, and bootstrap probability values (red least, to green most, probable) are on the left. Note that the scales are different for HelD and RpoB trees. The HelD class into which sequences fall is indicated in the outer circles as Class I, II, and III. Colored arcs indicate the bacterial classes into which the HelD sequences fall; teal, Firmicutes; pale green, Actinobacteria; purple, Clostridia; orange, Bacteroidia; red, Deltaproteobacteria; brown, Coriobacteria; pale yellow, Acidimicrobilia. Individual organisms and HelD sequences are numbered (largest to smallest) and color-coded starting clockwise from Bacillus subtilis. Organism numbers with one HelD are numbered in black; two, blue; three, red; four, orange; five, green and are listed as follows with gene identifiers and protein length (aa) in brackets: 1 Bacillus subtilis 168 (BSU_33450, 774aa). 2 Bacillus licheniformis ATCC 14580 (b1_00699, 776aa). 3 Bacillus megaterium DSM 319 (BMD_3869, 772aa). 4 Bacillus cereus ATCC10987 (#1 BCE_3516, 768 aa; #2 BCE_2839, 689 aa). 5 Bacillus anthracis Ames (#1 BA_1040, 776 aa; #2 BA_2814, 689 aa). 6 Bacillus cereus AH187 (#1 BCAH187_A1206, 777 aa; #2 BCAH187_A2861, 689 aa). 7 Bacillus cereus ATCC14579 (#1 BA_1040, 777 aa; #2 BA_2814, 689 aa). 8 Bacillus thuringiensis Bt407 (#1 btg_c11000, 778aa; #2 btg_c29280, 691aa). 9 Lactobacillus plantarum WCFS1 (#1 lpl_0432, 762aa; #2 lpl_0910, 768aa). 10 Lactobacillus rhamnosus GG (#1 lrh_01975, 763aa; #2 lrh_02619, 762aa). 11 Leuconostoc lactis W1K1m40 (lfc_04535, 788aa). 12 Lactobacillus acidophilus NCFM (lac_1676, 687aa). 13 Carnobacterium bicornis subsp. subsp. Xyli CTCB07 (Lxx_20770, 787aa).

Coriobacteria class, typified by Olsenella uli that is associated with gingivitis, are all located to the Class I branch of the tree (numbers 16–20 branch). Branch divergence and clustering of sequences to regions of the tree comprising Lactobacilli (numbers 14, 15, 21–24; Figure 3) and Clostridia (numbers 25–29; Figure 3) indicate that an ancestral Coriobacteria likely acquired helD genes by horizontal gene transfer from these organisms (Appendix 1; Figure A1). That Coriobacteria is isolated from human crevice, gastrointestinal and genital tracts (Clavel et al., 2014) is consistent with this proposition. The length of the branches suggests this horizontal transfer event occurred long ago but after the evolution of the mammalian hosts that provide environments with co-localised Lactobacilli, and that helD genes have been stably inherited and co-evolved within the Coriobacteria. In addition to the helD gene from Adlercreutzia equilfaciens DSM 19450 (AEQU_1689, number 20.1; Figure 3)
that clusters with those of the other Coriobacteria. *A. equillicens* contains a second *helD* gene (AEQU_0484, number 20.2; Figure 3) that clusters with *Clostridia*. The fact that *Lactobacilli*, *Rhodothermia* with additional representation in the classes *Bacteroides* shows *helD* is widely distributed throughout the class *Bacteroides* with additional representation in the classes *Rhodothermia* and *Ignavibacteria* (Appendix 1; Figure A3). Phylogenetically, many of the Bacteroidal *helD* sequences clustered close to *helD* sequences from *Clostridiales* (Figure 3a and Appendix 1 Figure A4; sequences 27–29 *C. difficile*, 30–35 Bacteroides). Extended analysis indicated that *helD* sequences from *Bacteroides* and *Parabacteroides* (family *Porphyromonadaceae*) clustered closest to those from *Firmicutes* that are strict gut anaerobes from the order *Clostridiales* (Appendix 1; Figure A5). These bacteria were from cluster IV (*Ruminococcaceae*) and XIVa (*Lachnospiraceae*) that are abundant gut microbes as - associated with many aspects of good health, and the cluster XI gut pathogen *C. difficile* (Lopetuso et al., 2013; Lozupone et al., 2012; Milani et al., 2017). Since the *Bacteroides* and *Parabacteroides* are also abundant obligate gut anaerobes, this clustering suggested that *helD* was horizontally transferred from an anaerobic gut *Firmicute*, most likely from the order *Clostridiales* (Appendix 1; Figure A5). Analysis of the genome context of *helD* genes indicated they were not (or are no longer) located in mobile genetic elements, except for *B. thetaiotaomicron*, and along with their widespread distribution in

![Figure 4](image-url)
**3.3 A novel HelD class in Gram-negative bacteria**

The analysis presented in this work also shows that there is a third class of HelD proteins encoded by the *Deltaproteobacteria* (Class III, Figures 3 and 4; see below). Newing et al. (Newing et al., 2020) identified Class I and II HelD proteins based on the conservation of twelve sequence motifs. These motifs (labeled I-XII, Appendix 1; Figure A6) are all conserved in Class III proteins (exemplified by *Myxococcus xanthus* HelD), despite the low overall levels of sequence similarity found in HelD proteins (Newing et al., 2020). A model of *M. xanthus* HeILD was also generated from an unbiased screen of the protein structure database (Figure 4; see Materials and Methods). As seen with Class I and II proteins, there is a HeILD-specific N-terminal domain of ~50–150 amino acids that has a long antiparallel α-helical structure (secondary channel arm, SCA, Figure 4b) that is required to anchor HeILD in the secondary channel of its cognate RNAP (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020), and the 1A helicase domain is split by the insertion of an arm-like structure (clamp arm, CA, Figure 4b and S6) that is used to bind within the primary channel of RNAP, forcing it open to aid the release of bound nucleic acids (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020).

An absolutely conserved DWR (Asp-Trp-Arg) sequence motif was identified in the unique N-terminal domain of all HeILD sequences, and determination of the structures of HeILD showed that the conserved Trp residue resides within a hydrophobic pocket called the Trp-cage, important in stabilizing the interaction between the N-terminal domain wedged deep into the secondary channel of RNAP and the helicase 1A domain (Newing et al., 2020). In most HelD sequences identified to date, the DWR motif is extended to DWR[XA/SP], but in *Deltaproteobacterial* HeILDs there is an additional amino acid inserted in this motif following the R residue, i.e., DWRX[XA/SP], which is a key defining feature of a Class III
HelD (Appendix 1; Figure A6). This additional amino acid does not appear to be highly conserved, the motif being DWRFAP in M. xanthus. Mycoplasma mycoides (MycMyco), Mycoplasma capricolum (MycCapr), Mycoplasma pneumoniae (MycPneu), Mycoplasma marinum (MycMari), Erysipelatoclostridium coccaceum (ClosCoc), Erysipelatoclostridium inoccuum (ClosInno), Clostridioides difficile (ClosDiff), Clostridium botulinum (ClosBotu), Clostridium perfringens (ClosPerf), Clostridium sartagoforme (ClosSart), Clostridium beijernickii (ClosBeij), Bacillus subtilis (BacSu), Staphylococcus lentus (StapLent), Staphylococcus equorum (StaphEquo), Staphylococcus saprophyticus (StapSapr), Staphylococcus aureus (StapAure), Staphylococcus felis (StapFeli), Staphylococcus agnetis (StapAgne), Staphylococcus rosti (Staprost), Staphylococcus pseudointermedius (StapPseud), Staphylococcus delphini (StapDelp), Enterococcus gallinarum (EnteGall), Enterococcus cecorum (EnteCeco), Enterococcus plantarum (EntePlan), Enterococcus munditii (EnteMundt), Enterococcus faecalis (EnteFaec), Enterococcus faecium (EnteFium), Streptococcus pneumoniae (StrepPneu), Streptococcus marmotae (StrepMarmo), Streptococcus suis (StrepSuis), Streptococcus thermophilus (StrepTherm), Streptococcus agalactiae (StrepAgala), Streptococcus pyogenes (StrepPyog), Streptococcus canis (StrepCanis), Lactobacillus sakei (LactSake), Lactococcus rodentium (LactRode), Lactobacillus johnsonii (LactJohn), Lactobacillus salivarius (LactSali), Lactobacillus plantarum (LactPlan), Lactobacillus fermentum (LactFerm), Lactobacillus brevis (LactBrev), Leuconostoc pseudomesenteroides (LeucPseud), Leuconostoc mesenteroides (LeucMese), Leuconostoc sp. (Leuconos), and Leuconostoc lactis (LeucLact)

FIGURE 6 Phylogenetic tree of RpoB with respect to the distribution of HelD and the d subunit of RNAP. Tree scale and bootstrap values are shown on the left. Organisms that contain the d subunit (DELTA) are shown in red, just HelD (blue) and both d and HelD (black). Mycoplasma mycoides (MycMyco), Mycoplasma capricolum (MycCapr), Mycoplasma pneumoniae (MycPneu), Mycoplasma marinum (MycMari), Erysipelatoclostridium coccaceum (ClosCoc), Erysipelatoclostridium inoccuum (ClosInno), Clostridioides difficile (ClosDiff), Clostridium botulinum (ClosBotu), Clostridium perfringens (ClosPerf), Clostridium sartagoforme (ClosSart), Clostridium beijernickii (ClosBeij), Bacillus subtilis (BacSu), Staphylococcus lentus (StapLent), Staphylococcus equorum (StaphEquo), Staphylococcus saprophyticus (StapSapr), Staphylococcus aureus (StapAure), Staphylococcus felis (StapFeli), Staphylococcus agnetis (StapAgne), Staphylococcus rosti (Staprost), Staphylococcus pseudointermedius (StapPseud), Staphylococcus delphini (StapDelp), Enterococcus gallinarum (EnteGall), Enterococcus cecorum (EnteCeco), Enterococcus plantarum (EntePlan), Enterococcus munditii (EnteMundt), Enterococcus faecalis (EnteFaec), Enterococcus faecium (EnteFium), Streptococcus pneumoniae (StrepPneu), Streptococcus marmotae (StrepMarmo), Streptococcus suis (StrepSuis), Streptococcus thermophilus (StrepTherm), Streptococcus agalactiae (StrepAgala), Streptococcus pyogenes (StrepPyog), Streptococcus canis (StrepCanis), Lactobacillus sakei (LactSake), Lactococcus rodentium (LactRode), Lactobacillus johnsonii (LactJohn), Lactobacillus salivarius (LactSali), Lactobacillus plantarum (LactPlan), Lactobacillus fermentum (LactFerm), Lactobacillus brevis (LactBrev), Leuconostoc pseudomesenteroides (LeucPseud), Leuconostoc mesenteroides (LeucMese), Leuconostoc sp. (Leuconos), and Leuconostoc lactis (LeucLact)
than that of *M. smegmatis* (HelD<sub>MS</sub>) but shorter than the *B. subtilis* protein (HelD<sub>BS</sub>). The tip of the SCA of HelD<sub>MX</sub> does not reach the active site (catalytic Mg<sup>2+</sup>, green sphere; compare dashed circles in Figure 5c–f) but would clash with the bridge helix in RNAP (teal, Figure 5d and f), potentially causing it to distort and displace the template DNA strand as seen with HelD<sub>BS</sub> (Newing et al., 2020). The RNAP trigger loop contains a large insertion in the *Deltaproteobacteria* (δ′ln6, Figure 5b) similar to that seen in *Gammaproteobacteria*, and it was assumed this (and the δln4 insertion, Figure 5b) would sterically interfere with HelD binding to RNAP in Gram-negative bacteria. Although the trigger loop in the modeled *M. xanthus* RNAP–HelD complex does clash with HelD<sub>MX</sub> (Figure 5e and f), this is not extensive, and given the inherent flexibility in this domain, small conformational changes would readily enable binding as seen in Gram-positive bacteria (Kouba et al., 2020; Newing et al., 2020; Pei et al., 2020). The CA of HelD<sub>MX</sub> is similar in size to that of HelD<sub>MS</sub> (although it does not contain a PCh domain; Figure 4b). The CA domain is required for clamp opening and DNA release in the Gram-positive systems, and likely will serve a similar function in Class III HelDs.

Examination of sequences retrieved from the CDART search indicated *helD* genes may be even more widely distributed in the *Proteobacteria* (including the *Gammaproteobacteria*), although this could not be verified by searches of complete genomes in databases such as KEGG and may represent misclassification from metagenomic sequencing projects. For example, BLASTP searches suggest hits reported as being from *E. coli* and *Vibrio vulnificus* identified from metagenomic data are in fact from *Bacteroides* and *Bacillus*, respectively (Poyet et al., 2019), and NCBI SRA accession code: PRJNA523266). Nevertheless, *helD* genes may be more widely distributed in *Proteobacteria*.

### 3.4 RNAP δ subunit and HelD

The *Firmicutes* have the smallest multi-subunit RNAPs currently known (Lane & Darst, 2010a, 2010b), as well as auxiliary subunits δ and ε that are not found in other bacteria (Keller et al., 2014; Weiss & Shaw, 2015). In the original work characterizing the function of HelD as a transcription complex recycling factor, it was shown that although δ or HelD on their own enhanced recycling, there was a synergistic relationship between them in *B. subtilis* transcription recycling assays (Wiedermannova et al., 2014). Structural analysis of RNAP recycling complexes shows that δ and HelD interact, as well as providing clues as to how δ could enhance the recycling activity of HelD by augmenting clamp opening (Pei et al., 2020). These structural studies also provided insights into how δ could facilitate transcription recycling in the absence of HelD (Miller et al., 2021). Genome searches indicated that not all *Firmicutes* contained both *helD* and *rpoE* (encoding the δ subunit) genes, and an analysis was performed based on the rpoB gene to establish whether there is segregation of genes amongst orders and/or based on the natural environment (Figure 6).

In the bulk of cases, the *Bacilli, Lactobacilli, Leuconostoc*, and *Enterococci* contained genes for both HelD and δ, and if the gene for one protein was missing, the other was present (Figure 6). The *Staphylococci* were heterogeneous with species such as *S. rostri* containing both *helD* and *rpoE* genes, whereas *S. aureus* only contained the gene for the δ subunit. There is a segregation of species containing both *helD* and *rpoE* cf. *rpoE* only, with *rpoE* only present in the *S. saprophyticus* and *S. aureus* clusters (Takahashi et al., 1999).

Species that fall within the *S. hyicus-intermedius* cluster (e.g., *S. rostri*) contained both *helD* and *rpoE*, but there were exceptions such as *S. felis*, which only contained *rpoE* (Figure 6). The *Streptococci* (order *Lactobacillales*) only contained the *rpoE* gene (Figure 6), whereas the *Clostridia*, except for *C. Erysipelatoclostridium*) cocleatum and *inocua*, only contained *helD* genes (Figure 6). Thus, it appears that in the *Firmicutes*, especially class *Bacillus*, the default situation is for both *rpoE* and *helD* to be present, but the absence of one gene is compensated for by the presence of the other to ensure the ability to recycle stalled transcription complexes is retained.

### 3.5 Many bacteria contain multiple *helD* genes

A striking observation made in the preliminary phylogenetic analysis of HelD was that some organisms contain more than one *helD* gene (Newing et al., 2020). This preliminary analysis has now been extended and it is clear that the presence of >1 *helD* is common and is found in both Gram-positive and -negative organisms (Figure 3a). Using complete genome sequences, up to 5 genes encoding HelD have been identified (e.g. *Nonomuraea sp.* ATCC55076 [organism 55]; Figure 3a and Appendix 1; Figure A7), and organisms have been identified with 1, 2, 3, 4, or 5 *helD* genes. Although most contain a single *helD* gene, low G+C Gram-positives and Gram-negatives were not found with >3, and high G+C Gram-positive *Actinobacteria* such as *Streptomyces*, *Nonomuraea*, and *Frankia* were identified with 24 *helD* genes. A simple assumption is that these multiple genes are the product of amplification through recombination, and this may well be the root of their original source, but phylogenetic analysis indicates each gene is unique, and organisms with more than one *helD* gene tend to encode both large (~740–850 aa) and small (~680–720 aa) variants. The variation in sequence length is due to differences in the flanking SCA and CA domains (arms) with the core 1A and 2A helicase domains all being of similar size. This suggests the motor function of these proteins is conserved, but the function of large vs small HelD variants may differ depending on the size of the SCA and CA arms. The multiple *helD* genes also segregate to Class I, -II, or -III according to the organism in which they are found; Class I sequences are found in *Firmicutes*, whereas *Actinobacteria* all have Class II sequences (except for the *Coriobacterium* Adlercreutzia equilifaciens, above), and Class III sequences are found in *Deltaproteobacteria*. Of the *Bacteroides/Parabacteroides* analyzed to date, all encode only a single Class I *helD* gene.

Some or all of the additional *helD* sequences might have represented cryptic genes that are not expressed under any conditions,
or that they are differentially expressed during different growth phases or conditions, which might provide clues to potential functions. Transcriptomics data were retrieved from the Sequence Read Archive (SRA) for selected organisms containing 1 or >1 helD representative of all three classes of HelD, and expression levels compared relative to rpoB (RNAP β subunit) and another housekeeping gene (SF1 helicase pcrA/uvrD). In all cases, all of the helD genes were expressed, often at an approximately similar level to pcrA/uvrD (Figure 7). The RNA-seq data of B. subtilis helD and pcrA obtained from experiments by Revilla-Guarinos et al. (Revilla-Guarinos et al., 2020) to examine changes in gene expression in a model soil organism on exposure to the antifungal agent amphotericin B produced by Streptomyces closely matched that of the oligonucleotide hybridization transcriptomics data of Nicolas et al. (Nicolas et al., 2012) and showed the level of helD expression was not influenced by amphotericin B and was ~3% that of rpoB (Figure 7a). This is also consistent with proteomics analysis indicating HeID is present at ~6% the level of RNAP (Delumeau et al., 2011). B. cereus contains two helD genes and the data set from strain F837/76 (Jessberger et al., 2019) grown in the presence and absence of mucin that can influence toxin production shows that both copies (one large, one small variant) are expressed, albeit at low levels, and expression is not significantly affected on exposure to mucin (Figure 7b). C. perfringens also contains two Class I helD genes, labeled CPE_0599 (small; 706 aa) and CPE_1619 (large; 763 aa) in strain 13, and expression levels were determined from datasets of cells grown in brain heart infusion (BHI) and a rich medium developed for the optimal growth of fastidious anaerobes, fastidious anaerobe broth +glucose (FABG) medium (Soncini et al., 2020). Both genes were expressed at levels comparable to helD in B. subtilis, and their cognate pcrA/uvrD, although CPE_0599 expression increased ~3-fold and CPE_1619 expression decreased in FABG medium compared to BHI medium (Figure 6c).

S. coelicolor A2(3) contains four Class II helD genes, two encoding large (SCO_2952 744 aa, and SCO_5439 755 aa) and two encoding small (SCO_4195 680 aa, and SCO_4316 681 aa) variants. Data from a study on growth phase-dependent changes in gene expression (Jeong et al., 2016) were obtained from the SRA for analysis of helD expression and compared with rpoB and pcrA. All four helD genes
were expressed with relative levels changing ~2-fold dependent on the growth phase (Figure 7d). Expression levels were generally highest during mid-log and transition, and lowest during late and stationary phases, with modest changes between the ratios of expression of the different gene copies at all stages. The RNA-seq data set for *M. smegmatis* comparing changes in gene expression on the deletion of the transcript cleavage factor GreA that is important in rescuing back-tracked RNAP (Feng et al., 2020) showed that expression of the single *helD* gene was substantially higher than in most other organisms, at about 25% the level of *rpoB* suggesting HelD may be particularly abundant in the *Mycobacteria* (Figure 7e). The expression levels of *helD* were similar in the presence and absence of *greA* indicating each factor acts on stalled transcription complexes independently of each other.

Analysis of RNA-seq data showed *helD* genes were also expressed in Gram-negative *M. xanthus* and *B. vulgatus* (Figure 7f and g), showing that despite the structural differences adjacent to the HelD interaction sites in the β and β′ subunits of RNAP from these organisms, HelDs are expressed and likely able to bind and functionally interact with their cognate RNAPs. The data for *M. xanthus* were obtained to examine changes in gene expression during the development of fruiting bodies and spores. It is interesting to note that expression of *helD* in *M. xanthus* increases during the development of spores (not to be confused with sporulation in the *Firmicutes*) and may point to a role in the storage of inactive RNAP during dormancy as has been proposed for *B. subtilis* HelD (Pei et al., 2020). The study in *B. vulgatus* was designed to investigate the effect on gene expression of exogenous thiamine that may be important in niche establishment in the gut. Therefore, in most/all organisms that contain *helD* gene(s), it/they are expressed. The reason why one organism contains a single gene and closely related species contain more than one (e.g. *B. subtilis* and *B. cereus*, Figure 6a and b) is currently not clear, but the expression data would suggest that each isoform has a functional role to play in the cell, and there is not a significant difference in the expression of large vs small *helD* variants.

4 | CONCLUSIONS

In this work, we have examined the phylogenetic distribution and classification of the transcription recycling factor HelD in detail and have identified a new class restricted to the *Deltaproteobacteria*. In addition, it appears *helD* genes have been acquired by horizontal transfer on at least three occasions; *Bacteroides* have acquired *helD* from the *Clostridiales*, whereas the *Coriobacteria* have acquired it from the *Lactobacillales* and *Clostridiales*. The gut microbiome is known as an environment conducive to horizontal gene transfer, especially with respect to the distribution of antibiotic resistance genes (McInnes et al., 2020), and given that *Bacteroides*, *Lactobacillales*, *Clostridiales*, and *Coriobacteria* are all common in the gut microbiome, it appears *A. equolifaciens* has acquired *helD* genes from gut microorganisms on two separate occasions. Indeed, an unusual feature of *helD* genes is that many organisms contain multiple paralogues and that all versions are expressed. Why some organisms have a single gene for *helD* while a closely related species has multiple expressed copies is unclear, and this will make a fascinating avenue for future research. It is interesting to note that actinobacteria, such as *Streptomyces*, *Frankia*, and *Nonomuraea* (numbers 50, 51, 54, and 55; Figure 3) that are known producers of valuable bioactive compounds used as antibiotics and anti-cancer drugs contained the largest number of *helD* genes (4–5). The 5 *helD* genes in *Nonomuraea* (number 55, Figure 3), which is a known producer of DNA-intercalating agents (Sungthon & Nakaew, 2015) may be involved in genome maintenance through recycling stalled transcription complexes during the production of these compounds. *Nonomuraea* and other *Actinomycetales* sometimes have a second *rpoB* gene that confers resistance of RNAP to compounds such as rifampicin and sorangicin that is induced by stress and is associated with the production of secondary metabolites (D’Argenio et al., 2016). The combination of multiple HelD isoforms with drug-resistant RNAP may be important in this proposed genome maintenance activity. In some organisms, such as *M. abcessus* and *S. venezuelae* *helD* expression is induced in the presence of the antibiotic rifampicin, conferring resistance, and this is associated with the presence of a DNA sequence called the Rifamycin Associated Element (RAE) found upstream of the gene (Hurst-Hess et al., 2021; Surette et al., 2021). It is proposed that the tip of the PCh loop can physically remove rifampicin bound to the RNAP β subunit in a pocket close to the active site. In *S. venezuelae* (organism #50, Figure 3) that has five *helD* genes, only one (SVEN_6029, #50.3) is induced in the presence of rifampicin and has an upstream RAE (Surette et al., 2021). It is interesting to note that despite encoding a rifampicin-resistant RNAP β subunit, *Nonomuraea* also has an RAE located directly upstream of *helD* NOA_42280 (#55.3; Appendix 1; Figures A7 and A8).

Investigation of the distribution of *helD* genes with upstream RAEs revealed they were clustered to two sub-branches of the *Actinobacteria* (Appendix 1; Figure A8) that may be considered the HelR grouping based on the nomenclature of these proteins by (Hurst-Hess et al., 2021; Surette et al., 2021). It should be noted that clearly identifiable RAEs could not be found upstream of all the genes in the HelR group, including for *Frankia alni*, *Nocardia brasiliensis*, or *Mycolicibacterium phlei* (54.2, 56.2, and 64, respectively; Figure 3 and Appendix 1 Figure A2). Rifampicin has also been observed to induce *helD* expression in the low G+C Gram-positive *B. subtilis*, but this induction does not confer resistance to the drug (Hutter et al., 2004). Nevertheless, the ability of naturally produced antibiotics to induce the expression of *helD* genes suggests HelD proteins have a potentially important role in preserving genome integrity and gene expression in the bacteria in which they are found.

An additional area of future research should include functional and structural studies of HelD from Gram-negative bacteria, as due to the location of lineage-specific inserts in the β and β′ subunits of RNAP in Gram-negatives it was assumed HelD-like proteins would bind poorly or be sterically inhibited from binding. HelD proteins represent a new class of motor enzymes involved in transcription
complex recycling that are widely distributed in bacteria that make an important contribution to our understanding of the multiple different mechanisms used to resolve potentially lethal stalled transcription complexes.

Finally, it is important that genome annotation databases are updated as helD genes are often classified as pcrA, uvrD, or helicase IV-ATPase. Correct annotation of helD genes will enable a more detailed understanding of the distribution, evolution, and function of this fascinating new category of transcription factors.

ETHICS STATEMENT
None required.

ACKNOWLEDGEMENTS
The authors appreciate the constructive comments from Brett Neilan, Leanne Pearson-Neilan, Caitlin Romanis, and Karl Hassan during the preparation of this article. This work was funded by the Australian Research Council grant DP210100365 (PL, ND, and AO).

CONFLICT OF INTEREST
None declared.

AUTHOR CONTRIBUTIONS
Joachim S. Larsen: Formal analysis (supporting); Methodology (supporting); Software (supporting); Writing-review & editing (supporting). Michael Miller: Formal analysis (supporting); Writing-review & editing (supporting). Aaron J. Oakley: Formal analysis (supporting); Funding acquisition (equal); Writing-review & editing (supporting). Nicholas E. Dixon: Formal analysis (supporting); Funding acquisition (equal); Writing-review & editing (supporting). Peter Lewis: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (equal); Investigation (lead); Methodology (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing-original draft (lead); Writing-review & editing (lead).

DATA AVAILABILITY STATEMENT
All data are provided in full in the results section of this paper and all sequences used are available from the NCBI at https://www.ncbi.nlm.nih.gov

ORCID
Joachim S. Larsen https://orcid.org/0000-0001-7270-8081
Michael Miller https://orcid.org/0000-0001-7990-3381
Aaron J. Oakley https://orcid.org/0000-0002-4764-014X
Nicholas E. Dixon https://orcid.org/0000-0002-5958-6945
Peter J. Lewis https://orcid.org/0000-0002-1992-062X

REFERENCES
Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans. Nature Reviews Genetics, 13, 720–731.
Clavel, T., Lepage, P., & Charrier, C. (2014). The family coirobacteraceae. In: E. Rosenberg, E. F. Delong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), The prokaryotes (pp. 201–238). Springer.
D’Ardenne, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase complexes.

Epilepsy Research, 10, e62117.
Sungthong, R., & Nakaew, N. (2015). The genus Nonomuraea: A review of a rare actinomycete taxon for novel metabolites. *Journal of Basic Microbiology*, 55, 554–565.

Surette, M. D., Waglechner, N., Koteva, K., & Wright, G. D. (2021). *HelR is a helicase-like protein that protects RNA polymerase from rifamycin antibiotics*. bioRxiv 2021.05.10.443488.

Takahashi, T., Satoh, I., & Kikuchi, N. (1999). Phylogenetic relationships of 38 taxa of the genus *Staphylococcus* based on 16S rRNA gene sequence analysis. *International Journal of Systematic Bacteriology*, 49(Pt 2), 725–728.

Urrutia-Irazabal, I., Ault, J. R., Sobott, F., Savery, N. J., & Dillingham, M. S. (2021). Analysis of the PcrA-RNA polymerase complex reveals a helicase interaction motif and a role for PcrA/UvrD helicase in the suppression of R-loops. *eLife*, 10, e68829. http://dx.doi.org/10.7554/elife.68829

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumieny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46, W296–W303.

Weiss, A., & Shaw, L. N. (2015). Small things considered: The small accessory subunits of RNA polymerase in Gram-positive bacteria. *FEMS Microbiology Reviews*, 39, 541–554.

Westblade, L. F., Campbell, E. A., Pukhrambam, C., Padovan, J. C., Nickels, B. E., Lamour, V., & Darst, S. A. (2010). Structural basis for the bacterial transcription-repair coupling factor/RNA polymerase interaction. *Nucleic Acids Research*, 38, 8357–8369.

Wiedermannova, J., Sudzinova, P., Koval, T., Rabatinova, A., Sanderova, H., Ramaniuk, O., Rittich, S., Dohnalek, J., Fu, Z., Halada, P., Lewis, P., & Krasny, L. (2014). Characterization of HelD, an interacting partner of RNA polymerase from *Bacillus subtilis*. *Nucleic Acids Research*, 42, 5151–5163.

Wood, E. R., & Matson, S. W. (1987). Purification and characterization of a new DNA-dependent ATPase with helicase activity from *Escherichia coli*. *Journal of Biological Chemistry*, 262, 15269–15276.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, 12, 7–8.

**How to cite this article:** Larsen, J. S., Miller, M., Oakley, A. J., Dixon, N. E., & Lewis, P. J. (2021). Multiple classes and isoforms of the RNA polymerase recycling motor protein HelD. *MicrobiologyOpen*, 10, e1251. [https://doi.org/10.1002/mbo3.1251](https://doi.org/10.1002/mbo3.1251)
FIGURE A1  Acquisition of helD genes by Coriobacteria from Firmicutes and Clostridia. The phylogenetic tree from Figure 1 is shown on the left side with the region boxed expanded on the right side. Bacterial classes are colored, species numbered, number of helD genes colored as in Figure 1: Bacilli, teal; Clostridia, purple; Coriobacteria, brown. 14 Enterococcus faecium Aus0004 (#2 EFAU004_00387, 711 aa). 15 Enterococcus faecium DO (#2 HMPREF0351_10397, 711 aa). 16 Olsenella uii DSM 7084 (OLS_0501, 731aa). 17 Atopobium parvulum DSM 20469 (Apar_0360, 736aa). 18 Slackia heliotrinireducens DSM 20476: (Shel_05840 (698aa). 19 Eggerthella lenta DSM 2243(Elen_2835, 716aa). 20 Adlercreutzia equolifaciens DSM 19450 (#1 AEQU_1689, 761aa; #2 AEQU_0484, 733aa). 21 Vagococcus teuberi (vte_03205, 717aa). 22 Enterococcus faecalis V583 (EF_0933, 732 aa). 23 Enterococcus faecalis DENG1 (DENG_00988, 732 aa). 24 Enterococcus faecalis OG1RF (OG1RF_10660, 740 aa). 25 Clostridium beijerinckii NCIMB 8052 (#2 cbe_2724, 745aa). 26 Epulopiscium sp. N.t. morphotype B (EPU_RS03295, 735aa). 27 Clostridioides difficile 630 (CD630_04550, 704 aa). 28 Clostridioides difficile RM20291 (CDR20291_0396, 704 aa). 29 Clostridioides difficile CD196 (CD196_0410, 704 aa). One helD gene, black; two, blue; three, red
Acquisition of helD genes by Acidimicrobiia from Actinobacteria. The phylogenetic tree from Figure 1 is shown on the left side with the region boxed expanded on the right side. Bacterial classes are colored, species numbered, number of helD genes colored as in Figure 1: Actinobacteria, pale green; Acidimicrobiia, pale yellow. 48 Acidobacterium ferrooxidans (Afer_1829, 706aa). 49 Cutibacterium acnes KPA171202 (PPA0733, 753aa). 50 Streptomyces venezuelae (#1 SVEN_2719, 779aa; #2 SVEN_5092, 747aa; #3 SVEN_6029, 722aa; #4 SVEN_4127, 675aa; #5 SVEN_3939; 665aa). 51 Streptomyces coelicolor A3(2) (#1 SC05439, 755 aa; #2 SCO2952, 744 aa; #3 SCO4316, 681 aa; #4 SCO4195, 680 aa). 52 IImotobacter coccineus (#1 aym_09360, 715aa; #2 aym_20540, 654aa). 53 Frankia casuarinae Cc13 (#1 fra_0952, 829aa; #2 fra_2397, 727aa). 54 Frankia alni ACN14a (#1 fal_1589, 939aa; #2 fal_4723, 877aa; #3 fal_3805; 866aa; #4 fal_4811, 751aa). 55 Nonomuraea sp. ATCC55076 (#1 NOA_23645, 772 aa; #2 NOA_16240, 762 aa; #3 NOA_42280, 715 aa; #4 NOA_08745, 660 aa; #5 NOA_48960, 655 aa). 56 Nocardi brasilienensis O31_020410 (#1 nbr_012985, 776aa; #2 nbr_020410, 731aa; #3 nbr: O3I_005870, 699aa). 57 Kineococcus radiotolerans SRS30216 (#1 kra_3607, 759aa; #2 kra_0164, 684aa). 58 Microbacterium sp. PAMC 28756 (mip_00070, 717aa). 59 Mycobacterium hominis SJTG1 (mhos_01135, 744aa). 60 Nocardia farcinica IFM10152 (#1 NFA_19060, 765aa; #2 NFA_44160, 726aa). 61 Mycobacterium smegmati MC2 155 (MSMEG_2174, 736aa). 62 Rhodococcus sp. 008 (#1 rhod_26990, 760aa; #2 rhod_09075, 731aa). 63 Mycobacterium sp. JS5623 (Mycsm_03949, 732aa). 64 Mycolicibacterium phlei (MPHL_03003, 726aa). 65 Mycobacteroides abscessus ATCC 19977 (MAB_3189c, 753aa). 66 Rhodococcus equi 103S (#1 REQ_25070, 759aa; #2 REQ_15310, 739aa). 67 Nocardia asteroides NCTC11293 (#1 nad_03000, 753; #2 nad_04408, 735aa). 68 Leifsonia xyleli subsp. Xyli CTCB07 (Lxx_20770, 787aa). 69 Bifidobacterium longum NCC2705 (BLO_1314, 759aa). 70 Bifidobacterium bifidum PRL2010 (bfp_0546, 759aa). 71 Brevibacterium linens BS258 (bly_10570, 743aa). 72 Brevibacterium flavum ZL-1 (bfv_07580, 755aa). 73 Corynebacterium glutamicum ATCC13031 (CG_1555, 755aa). 74 Corynebacterium diptheriae NTCC13129 (DIP_1156, 770aa). 75 Rhodococcus rhodochrous NCTC10210 (rrt_02795, 772aa). One helD gene, black; two, blue; three, red; four, orange; five, green.
Distribution of helD genes in the phylum Bacteroidota. HelD sequences from Bacteroidota RefSeq genomes were retrieved from a BLASTP search and mapped to individual species within the phylum Bacteroidota using Annotree (Mendler et al., 2019). Bacteroidotal classes are shown in the colored outer ring with Bacteroidia in pink, Rhodothermia in grey, Chlorobia in light grey, UBA10030 in lime green, Kryptonia in pale green, Ignavibacteria in cyan, Kapabacteria in pale blue, and SZUA-365 in blue. Individual species are shown as lines radiating out from the circular dendrogram with species containing HelD sequences highlighted in bright blue.
Acquisition of helD genes by Bacteroides from Clostridia. The phylogenetic tree from Figure 1 is shown on the left side with the region boxed expanded on the right side. Bacterial classes are colored, species numbered, number of helD genes colored as in Figure 1: Bacilli, teal; Clostridia, purple; Bacteroides, orange; Coriobacteria, brown. 20 Adlercreutzia equolifaciens DSM 19450 (#2 AEQU_0484, 733aa). 25 Clostridium beijerinckii NCIMB 8052 (#2 cbe_2724, 745aa; #3 cbe_4782, 724aa). 26 Epulopiscium sp. N.I. morphotype B (EPU_RS03295, 736aa). 27 Clostridoides difficile 630 (CD630_04550, 704 aa). 28 Clostridoides difficile RM20291 (CDR20291_0396, 704 aa). 29 Clostridium difficile CD196 (CD196_0410, 704 aa). 30 Bacteroides vulgatus ATCC 8482 (BVU_3010 (671aa). 31 Bacteroides caccae ATCC 43185 (CGC64_00555, 683aa). 32 Bacteroides cellulosilyticus WH2 (BcelWH2_01491, 693aa). 33 Bacteroides thetaiotaomicron VPI-5482 (BT_1890, 686aa). 34 Bacteroides ovatus ATCC 8483 (Bovatus_02598 (687aa). 35 Bacteroides xylanisolvens XB1A (BXY_17560, 687aa). 36 Staphylococcus delphini NCTC12225 (sdp_01978, 681aa). 37 Clostridium botulinum A ATCC3502 (#3 CBO_3341, 709 aa). 38 Clostridium botulinum A ATCC19377 (#3 CLB_3399, 709 aa). 39 Clostridium botulinum B1 Okra (#2 CLD_1179, 709 aa). 40 Clostridium perfringens 13 (#2 CPE_0599, 706 aa). 41 Clostridium perfringens ATCC13124 (#2 CPF_0580, 706 aa). 42 Clostridium perfringens SM101 (#2 CPR_0566 706 aa). One helD gene, black; two, blue; three, red.
**FIGURE A5** Phylogenetic tree of HelD sequences from Bacteroides and Clostridia. Tree scale and bootstrap values are shown at the top and left, respectively. Colored boxes denote cluster IV and XIVa Clostridia (yellow), cluster IX Clostridia (purple), and Bacteroides (red). The numbers in parentheses correspond to the organisms used in Figure 1. Roseburia intestinalis (Rosebur intest), *Blautia* sp. SG-772 (Blaut SG772), *Blautia* sp. N6H1-15 (Blaut N6H1-15b), *Pseudoflavonifractor* sp. BSD2780061688st1 E11 (PseudoflavonB), *Ruminococcus lactaris* (Rumino lact), *Anaerotruncus* sp. 1XD22-93 (Anaerotruncus 1XD22-93), *Ruminococcus gnarus* (Rumino gnarus), *Coprococcus comes* (Copro comes), *B. vulgatus* ATCC 8482 (BVU 3010), *B. caccae* ATCC 43185 (CGC64 00555), *B. cellulosilyticus* WH2 (BcelWH2 01491), *B. ovatus* ATCC 8483 (Bovatus 02598), *B. xylanisolvens* XB1A (BXY 17560), *C. difficile* 630 (CDif 630), *C. difficile* RM20291 (CDif R20291), *C. difficile* CD196 (CDif CD196), *Faecalitcatenia contorta* (Faecal cont), *Caproiciproducens* sp. NJN-50 (Caproiciproducens NJN-50), *Eubacterium uniforme* (Eubact uncl), *Hungatella hathewayi* (Hungatel hath), *Eubacterium limosum* (Eubact limo), *Faecalicatena orotica* (Faecal orot), *Blautia marasmi* (Blaut maras), *Enterocloster bolteae* (Enteroclo bolA and B), *Clostridium symbiosum* (Clost symbio), and *Enterocloster citroniae* (Enterocli citro)
FIGURE A6  Conserved HeID sequence motifs. Panel A shows a schematic of B. subtilis HeID domain organization with conserved sequence motifs adapted from Newing et al., (Newing et al., 2020), with panel B showing the equivalent sequence motifs from M. xanthus HeID. Appendix Table A1 shows the conserved sequence motifs with sequence numbers referring to the B. subtilis HeID sequence. X corresponds to a poorly conserved sequence (any amino acid) and h to a conserved hydrophobic residue. Residues colored red are specific to class I and green to class II sequences. The HeID motifs from the Class III M. xanthus HeID (Class IIIMX) are shown in the right column with absolutely conserved motif residues shown in purple (blue for the ATP binding motifs) and the Class III defining residue (F in the case of M. xanthus) that is inserted in the DWRAP motif shown in grey (see text for more details).
FIGURE A7  Distribution of the five helD genes from Nonomuraea sp. ATCC55076. The phylogenetic tree from Figure 1 is shown unannotated apart from boxing the region corresponding to the Actinobacteria pale green, and indicating the location of the Nonomuraea helD genes: #1 NOA_23645, 772 aa; #2 NOA_16240, 762 aa; #3 NOA_42280, 715 aa; #4 NOA_08745, 660 aa; #5 NOA_48960, 655 aa
**Figure A8** The helR group of helD variants that (potentially) confer rifampicin resistance. The phylogenetic tree from Figure 1 is shown on the left side with the region boxed expanded on the right side. Bacterial classes are colored, species numbered, the number of helD genes colored as in Figure 1. Only the (potential) helR variants are shown: 50 *Streptomyces venezuelae* (#3 SVEN_6029), 55 *Nonomuraea* sp. ATCC55076 (#3 NOA_42280, 715 aa), 58 *Microbacterium* sp. PAMC 28756 (mip_00070, 717aa), 59 *Mirobacterium hominis* SJTG1 (mhos_01135, 744aa), 60 *Nocardia farcinica* IFM10152 (#2 NFA_44160, 726aa), 61 *Mycobacterium smegmatis* MC2 155 (MSMEG_2174, 736aa), 62 *Rhodococcus* sp. 008 (#2 rhod_09075, 731aa), 63 *Mycobacterium* sp. SJTG1 (mhos_01135, 744aa), 64 *Mycobacteroides abscessus* ATCC 19977 (MAB_3189c, 735aa). 65 *Rhodococcus equi* 103S (#2 REQ_15310, 739aa). 66 *Nocardia asteroides* NCTC11293 (#2 nad_04408, 735aa). One helD gene, black; two, blue; five, green.

**Table A1** Comparison of conserved class I and II HelD motifs with those from class III *M. xanthus* HelD

| Motif | Position (B. subtilis numbering) | Sequence | Class IIImx |
|-------|----------------------------------|----------|-------------|
| I     | 098-102                          | YXFX     | PYFAH       |
| II    | 118-123                          | YXG      | LLGR        |
| III   | 135-146                          | hXDXWRYX `X X X `Y | VIDWRAPVARVFY |
| IV    | 209-222                          | `X'T `Q XEQ `X `X FY Y | VTAMLDAEQYAES |
| V     | 233-240                          | GXXGKT   | Walker A site |
| VI    | 244-255                          | hXRXA`X A`Y LL X X `K | LHRLEKLFDDP |
| VII   | 279-285                          | YLPXLGXS  | LLAPEGL    |
| VIII  | 543-550                          | hYD`Q'X`X | Walker B site |
| IX    | 568-576                          | TXGDX`Q'X | TLGIDEMQ    |
| X     | 603-610                          | LXXYR`X`X`K | LVQSYRCP   |
| XI    | 713-718                          | KG`X`Y`D`X | KGLED     |
| XII   | 740-747                          | V`X`X`X`R`X`X | HVAVTRTS   |