Discovery of the *Pseudomonas* Polyyne Protegencin by a Phylogeny-Guided Study of Polyyne Biosynthetic Gene Cluster Diversity

Alex J. Mullins, Gordon Webster, Hak Joong Kim, Jinlian Zhao, Yoana D. Petrova, Christina E. Ramming, Matthew Jenner, James A. H. Murray, Thomas R. Connor, Christian Hertweck, Gregory L. Challis, Eshwar Mahenthiralingam

*Microbiomes, Microbes and Informatics Group, Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom*

*Department of Biomolecular Chemistry, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Jena, Germany*

*Department of Chemistry, University of Warwick, Coventry, United Kingdom*

*Warwick Integrative Synthetic Biology Centre, University of Warwick, Coventry, United Kingdom*

*Molecular Biosciences Division, School of Biosciences, Cardiff University, Cardiff, United Kingdom*

*Faculty of Biological Sciences, Friedrich Schiller University Jena, Jena, Germany*

*Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia*

*ARC Centre of Excellence for Innovations in Peptide and Protein Science, Monash University, Clayton, Victoria, Australia*

Alex J. Mullins, Gordon Webster, and Hak Joong Kim made equal contributions to this article. The study was initiated and led by Cardiff University with Alex J. Mullins and Gordon Webster in the first and second positions, and bioinformatic analysis carried out by Alex J. Mullins in the first position. Hak Joong Kim made a parallel characterization of protegencin and demonstrated the essentiality of the conserved desaturases.

**ABSTRACT** Natural products that possess alkyne or polyyne moieties have been isolated from a variety of biological sources and possess a broad range of bioactivities. In bacteria, the basic biosynthesis of polyyynes is known, but their biosynthetic gene cluster (BGC) distribution and evolutionary relationship to alkyne biosynthesis have not been addressed. Through comprehensive genomic and phylogenetic analyses, the distribution of alkyne biosynthesis gene cassettes throughout bacteria was explored, revealing evidence of multiple horizontal gene transfer events. After investigation of the evolutionary connection between alkyne and polyyne biosynthesis, a monophyletic clade was identified that possessed a conserved seven-gene cassette for polyyne biosynthesis that built upon the conserved three-gene cassette for alkyne biosynthesis. Further diversity mapping of the conserved polyyne gene cassette revealed a phylogenetic subclade for an uncharacterized polyyne BGC present in several *Pseudomonas* species, designated *pgn*. Pathway mutagenesis and high-resolution analytical chemistry showed the *Pseudomonas protegens pgn* BGC directed the biosynthesis of a novel polyyne, protegencin. Exploration of the biosynthetic logic behind polyyne production, through BGC mutagenesis and analytical chemistry, highlighted the essentiality of a triad of desaturase proteins and a thioesterase in both the *P. protegens pgn* and *Trinickia caryophylli* (formerly *Burkholderia caryophylli*) caryoyncins pathways. We have unified and expanded knowledge of polyyne diversity and uniquely demonstrated that alkyne and polyyne biosynthetic gene clusters are evolutionarily related and widely distributed within bacteria. The systematic mapping of conserved biosynthetic genes across the available bacterial genomic diversity proved to be a fruitful method for discovering new natural products and better understanding polyyne biosynthesis.

**IMPORTANCE** Natural products bearing alkyne (triple carbon bond) or polyyne (multiple alternating single and triple carbon bonds) moieties exhibit a broad range of important biological activities. Polyyne metabolites have been implicated in important ecological roles such as cepacin mediating biological control of plant pathogens and...
caryoyencin protecting Lagriinae beetle eggs against pathogenic fungi. After further phylogenetic exploration of polyyne diversity, we identified a novel gene cluster in *Pseudomonas* bacteria with known biological control abilities and proved it was responsible for synthesizing a new polyyne metabolite, protegencin. The evolutionary analysis of polyyne pathways showed that multiple biosynthetic genes were conserved, and using mutagenesis, their essentiality was demonstrated. Our research provides a foundation for the future modification of polyyne metabolites and has identified a novel polyyne, protegencin, with potential bioactive roles of ecological and agricultural importance.

**KEYWORDS** *Pseudomonas*, biosynthetic gene clusters, natural products, phylogenetics, polyynes

Bacteria and fungi are an unparalleled source of structurally and functionally diverse metabolites with important applications in medicine and agriculture. Different classes of natural products can possess common structural features. One such moiety is the carbon-carbon triple (alkyne) bond. More than 65 alkyne-containing natural products have been isolated from marine bacteria and possess biotechnologically exploitable spectra of biological activity (1). Other metabolites possess elongated chains of alternating carbon-carbon single and triple bonds (polyynes). Polyynes have been isolated from diverse sources, including plants, fungi, bacteria, and even insects (2). The first bacterial polyyanes, cepacins A and B, were discovered from the bacterium *Burkholderia diffusa* (formerly *Pseudomonas cepacia*) (3). However, the biosynthetic origin of the cepacins was only defined recently in the closely related species *Burkholderia ambifaria*, where these metabolites were shown to function in the biocontrol of damping off disease caused by the oomycete *Globisporangium ultimum* (4). The timeline of bacterial polyyne discovery is interesting, with multiple studies characterizing molecular diversity and different ecological roles (Fig. 1). Following the discovery of cepacins A and B in 1984 (3), several other polyynes were identified in Proteobacteria. Caryoyencin was isolated from *Trinickia caryophylli* (formerly *Burkholderia caryophylli*) (5) and *Burkholderia gladioli* (6). Alongside other antifungal compounds biosynthesized by *B. gladioli*, Lagriinae beetles exploit caryoyencin in a symbiotic relationship to protect their eggs from fungal attack (7). Collimonins were discovered from *Collimonas fungivorans* and displayed antifungal activity (8, 9), and ergoynes were found in the marine grass endophyte *Gynuella sunshinyii* (10) (Fig. 1). For the polyyne Sch 31828, isolated from *Actinobacteria* (11), and *fischerellins A and B*, isolated from *Cyanobacteria* (12, 13), the associated biosynthetic gene clusters (BGCs) remain unknown. While alkyne (14) and polyyne (6) biosynthetic mechanisms have been investigated, the evolution of polyyne biosynthesis, its relationship to alkyne biosynthesis, and overall polyyne diversity have yet to be established.

The influx of bacterial genomic assemblies over the last decade has revolutionized our understanding of bacterial evolution and enhanced our ability to discover natural products through multiple genome mining techniques (15). Common approaches for identifying the metabolic products of novel BGCs discovered by genome mining include comparative metabolic profiling following mutagenesis of target BGCs, activation/inactivation of cluster-situated regulators, and heterologous expression (15, 16). Alternative methods fueled by the increasing availability of genomic data include analyzing the evolutionary diversity of bacteria to identify lineages talented in specialized metabolite biosynthesis (15). A second, phylogeny-based mining strategy exploits the diversity of biosynthetic genes to discover natural product derivatives of known metabolites (15). Such an approach has the advantage of gleaning insight into the horizontal transfer of genes from BGCs by comparing biosynthetic gene trees to evolutionary phylogenies.

Considering the limited insights into polyyne evolution despite evidence of an evolutionarily broad distribution (4, 11) (Fig. 1), we sought to integrate existing knowledge and expand our understanding of the distribution of these structurally intriguing moieties. Here, we show their evolutionary history, by examining the co-occurrence of
Phylogenetics Leads to Pseudomonas Polyyne Discovery

RESULTS

Distribution of alkyne biosynthesis and emergence of polyyne biosynthesis. A phylogenetic tree based on 4,990 protein sequences of the alkyne biosynthetic fatty acyl-AMP ligase, JamA, was constructed to assess the distribution of alkyne biosynthesis in bacteria (Fig. 2). Phylogenies were also constructed based on the corresponding gene, jamA, alongside the protein and gene sequences of the alkyne fatty acid desaturase JamB/jamB, and acyl carrier protein JamC/jamC (see Fig. S1 in the supplemental material).

The ability to biosynthesize alkynes was widely distributed across Proteobacteria, occurring in the Alpha-, Beta-, Delta-, and Gammaproteobacteria, and represented 95.5% of available sequences (4,868 of 4,990). Within the Proteobacteria, Betaproteobacteria were the most dominant representatives at 96.6% (4,704 of 4,868 Proteobacteria) and occurred in multiple deep-branching lineages, potentially indicating several acquisition events into the phylum (Fig. 2), which is also supported by the additional phylogenies of alkyne biosynthetic genes and proteins (Fig. S1). However, the rearrangement of the branchpoints observed in the JamABC/jamABC protein and gene phylogenies confounds the ability to determine the number of horizontal gene transfer events that have occurred (Fig. S1). Despite these phylogenetic limitations, all six phylogenies (Fig. 2; Fig. S1) supported a similar overarching topology. Most sequences (80% [4,013 of 4,990]) occurred in a basal clade composed entirely of Burkholderia species, including B. pseudomallei, B. thailandensis, and B. ubonensis (Fig. 2), while the opposing end of the unrooted phylogeny

alkyne and polyyne biosynthetic cassettes (a minimum gene collection to biosynthesize a specific structural moiety), and their distribution, through a phylogeny-guided genome mining approach. This approach involved constructing a phylogeny based on genes associated with the alkyne and polyyne cassettes, identifying phylogenetic clades of interest, and mining representative genomes from these clades for uncharacterized polyyne biosynthetic gene clusters. Mixed evolutionary lineages within the alkyne phylogeny provided further evidence of their highly promiscuous nature. A distinct, monophyletic clade composed of polyyne biosynthetic gene clusters was observed within the broader alkyne gene cassette distribution. By examining subclade architecture, we identified a previously unexplored Pseudomonas polyyne clade that resulted in the characterization of a novel polyyne BGC, pgn, and its associated metabolite, protegencin.
consistently encompassed 779 sequences with a congruent topology (Fig. 2; Fig. S1). Outside of the Proteobacteria, examples of the alkyne cassette were found in members of the Cyanobacteria (29 genomes), Planctomycetes (2 genomes), and the candidate phylum Tectomicrobia uncultivated sponge symbiont "Candidatus Entotheonella" (1 genome).

Construction of the phylogeny of the biosynthetic fatty acyl-AMP ligase JamA also highlighted a discrepancy in the literature regarding the previously characterized B. pseudomallei alkyne biosynthetic locus (14). Inclusion of the purported JamA homologue alongside the JamA homologue identified during this analysis confirmed the latter to be the genuine JamA homologue (see Fig. S2 in the supplemental material). Annotation of the biosynthetic locus revealed the genuine fatty acyl-AMP ligase was encoded downstream of the previously characterized JamA protein (Fig. S2).
To understand the broader relationship between bacterial alkyne and polyyne biosynthesis, a comparison of characterized polyyne biosynthetic gene clusters was performed. Analysis of the gene content and architecture of four characterized/published polyyne BGCs (for cepacins, collimonins, caryoynencin, and ergoynes) identified seven common genes (Fig. 3). In addition to the three genes encoding the alkyne biosynthetic cassette, jamABC (14), genes encoding two additional fatty acid desaturases, a thioesterase, and rubredoxin were found in all BGCs (Fig. 3). Using this knowledge, we screened DNA sequences flanking the jamABC alkyne biosynthetic cassettes for the presence of the remaining four genes. This revealed a monophyletic clade in the alkyne phylogenies (Fig. 2; Fig. S1) where the 779 corresponding genomes possessed the conserved polyyne gene cassette (Fig. 3), with a few exceptions. Three discrepancies were observed within the monophyletic polyyne clade: B. gladioli strain 3848s-5 and three Streptomyces strains appeared to lack the colocalized thioesterase and rubredoxin genes with the remaining polyyne core biosynthetic genes, but manual inspection of these genomes revealed the BGCs were split across two contigs. A subset of 10 actinobacterial genomes appeared to have the thioesterase- and rubredoxin-encoding genes replaced by a gene encoding a cytochrome P450. These 10 genomes represented three genera (Streptomyces, Micromonospora, and Amycolatopsis) and were confined to a single subclade in the monophyletic polyyne clade. The final discrepancy included two representatives of the family Mycobacteriaceae that lacked the rubredoxin gene.

To understand the broader relationship between bacterial alkyne and polyyne biosynthesis, a comparison of characterized polyyne biosynthetic gene clusters was performed. Analysis of the gene content and architecture of four characterized/published polyyne BGCs (for cepacins, collimonins, caryoynencin, and ergoynes) identified seven common genes (Fig. 3). In addition to the three genes encoding the alkyne biosynthetic cassette, jamABC (14), genes encoding two additional fatty acid desaturases, a thioesterase, and rubredoxin were found in all BGCs (Fig. 3). Using this knowledge, we screened DNA sequences flanking the jamABC alkyne biosynthetic cassettes for the presence of the remaining four genes. This revealed a monophyletic clade in the alkyne phylogenies (Fig. 2; Fig. S1) where the 779 corresponding genomes possessed the conserved polyyne gene cassette (Fig. 3), with a few exceptions. Three discrepancies were observed within the monophyletic polyyne clade: B. gladioli strain 3848s-5 and three Streptomyces strains appeared to lack the colocalized thioesterase and rubredoxin genes with the remaining polyyne core biosynthetic genes, but manual inspection of these genomes revealed the BGCs were split across two contigs. A subset of 10 actinobacterial genomes appeared to have the thioesterase- and rubredoxin-encoding genes replaced by a gene encoding a cytochrome P450. These 10 genomes represented three genera (Streptomyces, Micromonospora, and Amycolatopsis) and were confined to a single subclade in the monophyletic polyyne clade. The final discrepancy included two representatives of the family Mycobacteriaceae that lacked the rubredoxin gene.

To investigate the diversity of the monophyletic clade, a separate phylogeny was constructed based on one of the polyyne-associated desaturase proteins (Fig. 4). This phylogeny was rooted using the basal branches of the clade of interest from both the JamA and JamB phylogenies (Fig. 2): a Gammaproteobacteria subclade and Betaproteobacteria subclade. Within the resulting phylogeny, we defined five major clades representing three Betaproteobacteria clades, one Gammaproteobacteria clade, and an Actinobacteria clade (Fig. 4). Each of the four previously characterized polynes corresponded to a different clade, with collimonins, caryoynencin, and cepacins localized to the three distinct Betaproteobacteria clades (Fig. 4). The ergoynes, biosynthesized by G. sunshinyii, were in the Gammaproteobacteria clade, but with deep branching separating G. sunshinyii from the remainder of the clade members (Fig. 4). Each Proteobacteria clade was dominated by a single genus and mainly structured with relatively shallow branching. In comparison, the Actinobacteria clade possessed deep branching and contained representatives...
**FIG 4** Desaturase protein-based phylogeny of polyne-producing bacteria. Homologues of the cepacin desaturase CnnN (protegencin PgnH) were extracted from bacterial genomes represented in the monophyletic alkyne clade as polyne producers. The four Proteobacteria clades, their composite genera, and associated polyne metabolites are indicated, in addition to the Actinobacteria phylum clade. The Gammaproteobacteria clade was used as the root based on the topologies of alkyne gene phylogenies. Known polyne producers are indicated with asterisks, and the specific strains are labeled. Bootstrap values are indicated for splits between the 5 major clades. The scale bar represents the number of substitutions per position.

- **Betaproteobacteria**
  - *Burkholderia* x396
  - *Caballeronia* x1

- **Actinobacteria**
  - *Streptomycetes* x35
  - *Amycolatopsis* x1
  - *Actinomadura* x4
  - *Streptosporangium* x1
  - *Micromonospora* x2
  - *Rhodococcus* x1
  - *Mycobacterium* x2

- **Betaproteobacteria**
  - *Collimonas* x6
  - *Massilia* x1
  - *Collimonins* (Kai et al. 2018)

- **Betaproteobacteria**
  - *Burkholderia* x258
  - *Trinickia* x3

- **Betaproteobacteria**
  - *Caryoynencin* (Jones et al. 2021)

- **Betaproteobacteria**
  - *Caryoynencin* (Ross et al. 2014)

- **Gammaproteobacteria**
  - *Protegencin* (This study)
  - *Ergynoes* (Ueoka et al. 2018)

- **Polynes detected without bacterial genome sequences**
  - Cepacins: *Burkholderia diffusa* LMG 29043
  - Caryoynencin: *Trinickia caryophylli* NIAES 1192
  - Sch 31828: *Microbispora* sp.
  - Fischerellins: *Fischerella muscicola* UTEX 1829
of seven genera, including Micromonospora, Actinomadura, and Rhodococcus, but was dominated by Streptomyces species. This analysis identified the cepacin BGC in several species that were previously not known to carry the gene cluster (Fig. 4), including B. contaminans, B. vietnamiensis and Caballeronia peredens.

**Exploration of the Gammaproteobacteria clade reveals an uncharacterized polyyne.** Aside from the single representative of the Gynuella genus, the Gammaproteobacteria clade was dominated by Pseudomonas. However, this genus is not known to produce polynes. Evidence of a Pseudomonas polyyne BGC has been alluded to as a homologous gene cluster of the collimoinin (8) and caryoyncin (6) BGCs during the discovery of these polynes. As such, we sought to investigate the production of an uncharacterized polyyne in Pseudomonas (Fig. 5a), focusing on Pseudomonas protegens (formerly P. fluorescens) strains PF-5 and CHA0 as model systems (see Table S1 in the supplemental material). High-performance liquid chromatography (HPLC) analysis of these two strains revealed a small chromatographic peak with a characteristic UV absorbance spectrum as observed for other polynes (6, 8). Comparative negative-ion-mode high-resolution electrospray ionization quadrupole time of flight mass spectrometry (HR-ESI-Q-TOF MS) analysis of the wild-type P. protegens PF-5 and CHA0 strains and mutants with in-frame deletions in the fatty acyl-AMP ligase gene (PF-5 ΔpgnD and CHA0 ΔpgnD, respectively) identified a compound, which we named protegencin, with the molecular formula C_{18}H_{18}O_{2} (Calculated for C_{18}H_{18}O_{2}: 265.1234. Found: 265.1239) as the product of the polyyne BGC (Fig. 5b and c; see Fig. S3a and b in the supplemental material).

**NMR spectroscopy confirms protegencin is a novel Pseudomonas polyyne.**

Polynes are notorious for being unstable and difficult to isolate, with recent studies requiring derivatization by click chemistry prior to spectroscopic analysis (6). The isolation of protegencin required careful optimization to enable spectroscopic characterization of the compound without derivatization. Purified fractions of protegencin were dried under vacuum for 2 to 3 h, with the addition of small volumes of MeCN to promote the removal of water from the sample. Freeze-drying of protegencin-containing fractions resulted in a polymerized brown oil. Using this procedure, protegencin was isolated as a brownish, amorphous powder. Its 1H, 13C, correlation spectroscopy (COSY), heteronuclear single quantum coherence (HSQC), and heteronuclear multiple-bond correlation (HMBC) spectra were acquired in deuterated dimethyl sulfoxide (DMSO-d_6) (see Table S2 and Fig. S3c to g in the supplemental material). The 1H NMR spectroscopic data displayed two olefinic protons (δ_H 6.65, 1H, dt, J = 16.0, 6.5, H-9; δ_H 5.79, 1H, d, J = 16.0 H-10), a methine proton (δ_H 4.06, 1H, H-18), and seven pairs of methylene protons. The 13C NMR and HSQC spectroscopic data (Table S2) indicated 18 carbons, including three methine carbons (δ_C 155.4, 107.3, and 74.7), seven methylene carbons (δ_C 34.1, 33.4, 28.9 × 2, 28.8, 28.0, and 24.9), one carbonyl carbon (δ_C 175.0.), and seven quaternary carbons. The above data suggested a similar polyyne structure to carroyncin (5, 6), but lacking a pair of olefinic protons and an oxymethine proton. The structure was further established by COSY and HMBC spectroscopic data analysis (Fig. S3f and g). The HMBC correlations of H-9 with C-11, C-8, and C-7, along with the couplings of H-10 to C-9, C-11, C-12, C-8, and C-13, confirmed a double bond was located at C-9/C-10 next to the polyyne scaffold, as observed in caryoyncin. The double bond at C-7/C-8 and hydroxyl group at C-6 in caryoyncin were missing in protegencin, as evidenced by HMBC correlations from a methylene (H\_2\_8) to two methine carbons (C-9 and C-10) and two methylene carbons (C-6 and C-7), and from a methylene (H\_2\_4) to two methylene carbons (C-6 and C-5), as well as COSY couplings of H\_2\_8 to H\_9 and H\_7. The other COSY correlations of H\_2\_3 to H\_2\_4 and H\_1\_2, and of H\_2\_4 to H\_2\_5, together with HMBC correlations of H\_2\_3 to C-1, C-3, and C-4, and of H\_3 with C-1, C-2, C-4, and C-5, confirmed the structure of the saturated region of this metabolite. Therefore, the structure of protegencin was elucidated as a novel polyyne natural product, as shown in Fig. 5c.

**Distribution of protegencin (pgn) BGC within Pseudomonas.**

Following the discovery of the previously uncharacterized polyyne metabolite protegencin, we sought to fully understand the species distribution of the pgn locus. The Pseudomonas
FIG 5 Organization and distribution of the protegencin (pgn) BGC and analysis of protegencin production. (a) Organization and putative function of genes within the pgn BGC. (b) Extracted-ion chromatograms at \( m/z = 265.12 \pm 0.02 \), corresponding to \([M-2H]^2-\) for protegencin, from LC-MS analyses of crude extracts made from agar-grown cultures of \( P. \) protegens CHA0 (top) and the \( P. \) protegens CHA0 \( \Delta pgnD \) mutant (bottom). (c) Structure of protegencin, determined by a combination of high-resolution mass spectrometry and NMR spectroscopy (see Table S2 and Fig. S3a to g). (d) Core gene-based phylogeny, using 1,487 genes, of 67 \( Pseudomonas \) genomes carrying the pgn BGC. The main nodes that demarcate the \( Pseudomonas \) species are highlighted, and all possess bootstrap values of 100. Representative strains and genome assembly accession numbers are included for each defined species. The scale bar represents the number of substitutions per site. The \( P. asturiensis \) branch was shortened (indicated by a break), and as such, the scale bar does not apply. (e) HPLC chromatograms (220 to 400 nm) of \( P. \) protegens Pf-5 wild-type and in-frame insertional mutant cultures. Only in the presence of all three desaturase genes (\( pgnE, pgnF, \) and \( pgnH \) ) is protegencin produced. No polyyne precursors can be detected in the mutant strains.
branches of the *Gammaproteobacteria* clade represented 67 *Pseudomonas* genomes. Subsequent average nucleotide identity analysis (ANI) of these genomes indicated the presence of multiple species. Based on the established 95% species delineation threshold for ANI (17, 18), six species were identified: these included two named species, *Pseudomonas protegens* (P. *fluorescens* group) and *Pseudomonas asturiensis* (P. *syringae* group) (19), and four unnamed species. The relatedness of these two species to one another is highlighted in the core-gene-based phylogeny (Fig. 5d). *P. protegens* was the dominant species possessing the *pgn* BGC, representing approximately 75% of genomes. A wider search for genome representatives of these six species in the European Nucleotide Archive (ENA) revealed that all genomes available of these species possess the protegencin (*pgn*) BGC, except for *P. asturiensis*. Of the two available *P. asturiensis* genomes, only the type strain LMG 26898T contained the *pgn* BGC. It was absent from *Pseudomonas* sp. strain 286 (98.9% ANI to LMG 26898). The *pgn* locus is present in five out of six *Pseudomonas* species examined in this study.

**A conserved desaturase triad is essential for polyyne formation.** The high conservation of the three desaturase genes and the thioesterase gene across all orthologous polyyne BGCs is notable (Fig. 3). To elucidate their roles, we performed targeted gene replacements. Specifically, we individually replaced the desaturase and thioesterase genes with a kanamycin and apramycin resistance cassette in the *P. protegens* *pgn* and *T. caryophylli* *cay* BGCs, respectively (Fig. 5e; see Fig. S4 in the supplemental material). Sequence analyses indicated that pairs of desaturase genes (*pgnEcayB* and *pgnFcayC*) would have similar functions. The deduced product of *pgnH* is a didomain enzyme with putative desaturase and thioesterase functions that corresponds to *cayE* and *cayF*, respectively. The metabolic profiles of the mutant strains were compared by HPLC (220 to 400 nm) with those of the wild-type strains, with or without the empty pGL42a or pJET1.2/blunt vector used for mutagenesis (Fig. 5e; Fig. S4). Whereas *P. protegens* PF-5 (with or without the empty vector) produces protegencin, in the *ΔpgnE Kanr, ΔpgnF Kanr,* and *ΔpgnH Kanr* mutant strains, no polyyne precursor could be identified (Fig. 5e). Deletions of the desaturase genes *cayB, cayC,* and *cayE* and the thioesterase gene *cayF* in *T. caryophylli* abolished the production of caryoynencin. The wild type (with or without an empty vector) generates the 7/E/Z-isomers of caryoynencin, but the mutant strains (*ΔcayB Aprr, ΔcayC Aprr, ΔcayE Aprr,* and *ΔcayF Aprr*) produce neither polyynes nor pathway intermediates (Fig. S4). These data indicate that the three desaturases and the thioesterase synergize in the production of polyynes. Interestingly, the same multienzyme system that gives rise to a tetrayne in the protegencin and caryoynencin BGCs appears to form a triyne in the collimomin pathway and a diynyl allene in the cepacin pathway (Fig. 1).

**DISCUSSION**

**Highly transmissible alkyne and polyyne cassettes.** Our results identify evidence of a single point of evolution of polyyne biosynthesis within bacteria and demarcate its evolution from alkyne biosynthesis (Fig. 2). The basal positioning of *Proteobacteria* within the polyyne phylogeny hints at a potential origin of this biosynthetic ability (Fig. 4), followed by horizontal gene transfer into *Actinobacteria* and other *Proteobacteria* classes. Additionally, the occurrence of alkyne biosynthetic genes across diverse bacterial lineages was also indicative of multiple horizontal gene transfer events. Few other fatty acid synthase-based biosynthetic capabilities appear to occur across a spectrum of bacterial lineages.

While examples of polyyne biosynthesis exist across plants, fungi, and insects, they appear to have different biosynthetic origins compared to bacteria (2). In contrast to the biosynthetic mechanism for multiple carbon-carbon triple bond formation defined in this study, there is no evidence of other biosynthetic pathways evolving from an alkyne precursor biosynthetic gene cassette. Within bacteria, a separate, evolutionarily independent, mechanism exists for the biosynthesis of multiple carbon-carbon triple bonds in the form of enediyynes (20). In contrast to the seven-gene cassette required for polyyne biosynthesis, a minimal five-gene cassette was defined by comparing 10 biosynthetic pathways associated with production of enediyne-containing natural products (20). Mining of bacterial genomes revealed comparably fewer examples of
the enediyne gene cassette (20, 21); however, there is evidence of horizontal gene transfer across several phyla (20) similar to the alkyne and polyyne gene cassettes.

**Phylogeny-driven metabolite discovery.** Mapping the diversity of polyyne biosynthetic gene clusters through functional gene and protein phylogenies permitted the discovery of an uncharacterized *Pseudomonas* polyyne BGC, *pgn*, and metabolite, protegencin. Hotter et al. (22) have recently demonstrated that this *P. protegens* polyyne, protegencin, acts as an algicidal toxin of the green alga *Chlamydomonas reinhardtii*. In parallel to these studies characterizing protegencin, Murata et al. (23) identified the same polyyne biosynthetic gene cluster in the biocontrol strain *P. protegens* Cab57, designating the molecules produced as protegenins.

Function-based phylogenies have been exploited previously to gain insight into natural product diversity. For example, ketosynthase (KS) and condensation (C) domains have been used to identify polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) BGCs, respectively (24). Mining for genes known to encode enzymes that biosynthesize specific structural moieties also enables discovery and comparison to other structurally related metabolites. A novel glutarimide, gladiostatin, was recently discovered in *Burkholderia gladioli* by identifying a BGC possessing genes similar to those associated with the biosynthesis of glutarimide antibiotics in *Streptomyces* species (25, 26).

The deep branching observed within the *Actinobacteria* clade of the polyyne phylogeny represents evidence of sequence divergence and may translate into structural diversity of the resulting polyyne natural products. No *Actinobacteria* polyyne has been associated with a biosynthetic gene cluster to date, and the only published *Actinobacteria* polyyne, Sch 31828, originated from a strain that lacks a genome sequence and has not been characterized at the species level, *Microbispora* sp. strain SCC 1438 (11). In *Cyanobacteria*, many alkyne-containing natural products have been characterized (1); in contrast, only two polyynes have been discovered to date (12, 13). The lack of a genome sequence for the *Fischerella muscicola* strains that produce fischerellins also impedes our mapping of their phylogenetic relationship to other polyyne biosynthetic gene clusters, and they potentially represent an uncharacterized *Cyanobacteria* clade.

**Evidence for an uncharacterized polyyne in *P. protegens***. We identified and characterized a novel *Pseudomonas* polyyne metabolite produced by the widely studied *P. protegens* strains Pf-5 and Cha0 (Table S1). Both strains have an extensive history of bioprostecidal properties (27, 28), indicative of the array of potent antimicrobial natural products biosynthesized by this species, such as the antifungal metabolites 2,4-diacetylphloroglucinol and pyoluteorin (27, 28). Previous sequence comparisons had highlighted the existence of a polyyne BGC in *P. protegens* with similarities to the caryoyncin (6) and collimolin (8) BGCs. However, homology to only the core biosynthetic region was defined in these studies (6, 8) (Fig. 2), and the metabolic product was not identified. Additionally, a transcriptomics analysis of the Gac global regulatory system highlighted a locus possessing similarities to those in *Burkholderia* (29), with a gene organization and putative gene functions like those found in the cepacin BGC (4).

Overall, we sought to understand the evolution and diversity of polyyne biosynthesis following emergence from the alkyne biosynthetic gene cassette. This study exploited functional gene phylogenetics alongside evolutionary analyses to explore polyyne biosynthetic diversity. Bioinformatics analyses supported by molecular biology and analytical chemistry led to the discovery of a *Pseudomonas*-derived polyyne BGC, *pgn*, and its metabolic product, protegencin. The conserved multienzyme system was proven to be essential for polyyne formation in both protegencin and caryoyncin biosynthesis (Fig. 5e).

**MATERIALS AND METHODS**

**Detection of alkyne and polyyne biosynthetic gene clusters.** A BLASTp (30) search of NCBI genomes, excluding *Burkholderia* (taxid: 32008) and a local database of *Burkholderia* assemblies (3,002 downloaded genomes and 4,434 genomes assembled from publicly available Illumina read data) was performed with the cepacin homologue (CcnK) (4) of the desaturase JamB as the query. *Burkholderia* genomic assemblies were downloaded from the European Nucleotide Archive (ENA) using a script from...
Preparative HPLC purification and structure elucidation by NMR spectroscopy. *P. protegens* Pf-5 metabolite production was scaled up by growth on 53 PEM agar plates (1.5 liters of medium in total). After growth at 22°C for 3 days, the medium was processed as described for the LC-MS analysis. The purification was performed on an Agilent 1200 series HPLC instrument equipped with a diode array detector and an Agilent Zorbax C_{18} column (100 by 2.1 mm, 5 µm), and the crude EtOAc extract was separated with an MeCN-H_{2}O gradient (0 min, 5% MeCN; 5 min, 30% MeCN; 50 min, 30% MeCN; 80 min, 100% MeCN; 90 min, 100% MeCN) at a flow rate of 9 ml/min and monitoring absorbance at 260 nm. This resulted in the isolation of a putative polyyne metabolite (1.5 mg, t_{R} = 76.8 min). The structure of this compound was elucidated using NMR spectroscopy. The sample was dissolved in 0.6 ml of deuterated DMSO in a Norell standard series 5-mm NMR tube, and 1D/2D spectra (\textsuperscript{1}H, \textsuperscript{13}C, COSY, HSQC, and HMBC) were obtained at 500 MHz for \textsuperscript{1}H NMR and 125 MHz for \textsuperscript{13}C NMR on a Bruker Avance III HD 500-MHz spectrometer. Chemical shifts (\delta) are given in ppm, and coupling constants (J) are given in hertz (Hz). Additional HPLC methods are described in the supplemental material.

Data availability. All bacterial genome assemblies and Illumina reads analyzed during this study were downloaded from the National Center for Biotechnology Information (NCBI) or European Nucleotide Archive public databases.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, PDF file, 0.1 MB.
**ACKNOWLEDGMENTS**

A.J.M., G.W., J.Z., E.M., G.L.C., J.A.H.M., and T.R.C. acknowledge funding from the Biotechnology and Biological Sciences Research Council (BBSRC) grant references BB/S007652/1 and BB/S008020/1. A.J.M. acknowledges previous funding by the BBSRC South West Doctoral Training Partnership (BB/M009122/1). T.R.C. acknowledges funding from Medical Research Council award MR/L015080/1, which funded the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) used for data analysis. The Dionex 3000RS/Bruker MaXis Impact instrument used in this work was purchased with a grant to G.L.C. from the BBSRC Discovery Fellowship (BB/R012121/1). C.H. and H.J.K. acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—SFB 1127/2, ChemBioSys—239748522 and the Pakt für Forschung und Innovation.

We thank Gail Preston (Department of Plant Sciences, Oxford) for supplying the strain *P. protegens* PF-5 and George O’Toole (Dartmouth College, Hanover, NH, USA) for providing the pMQ30 mutagenesis strains and constructs. We thank the School of Biosciences Genomics Research Hub at Cardiff University for genome sequencing services and project student George Mears for laboratory technical assistance during his B.Sc. (Honors) final year project.

G.L.C. is a codirector of Erebage, Ltd. The other authors declare no competing interests.

Conceptualization, A.J.M., G.W., E.M., H.J.K., and C.H.; Data curation, A.J.M., J.Z., M.J., and H.J.K.; Formal analysis, A.J.M., G.W., Y.D.P., J.Z., M.J., and H.J.K.; Funding acquisition, E.M., G.L.C., J.A.H.M., and C.H.; Investigation, A.J.M., G.W., Y.D.P., J.Z., C.E.R., M.J., and H.J.K.; Methodology, A.J.M., Y.D.P., J.Z., M.J., E.M., and H.J.K.; Project administration, A.J.M., J.Z., G.L.C., and E.M.; Resources, E.M., G.L.C., J.A.H.M., and C.H.; Software, A.J.M.; Supervision, E.M., G.L.C., and C.H.; Validation, A.J.M., Y.D.P., J.Z., M.J., and H.J.K.; Visualization, A.J.M., J.Z., M.J., and H.J.K.; Writing—original draft, A.J.M., G.W., Y.D.P., J.Z., M.J., E.M., and H.J.K.; Writing—review & editing, A.J.M., E.M., G.W., J.Z., T.R.C., M.J., G.L.C., H.J.K., C.H.

**REFERENCES**

1. Chai QY, Yang Z, Lin HW, Han BN. 2016. Alkynyl-containing peptides of marine origin: a review. Mar Drugs 14:216. https://doi.org/10.3390/md14110216.

2. Shi Shun ALK, Tykwinski RR. 2006. Synthesis of naturally occurring polyynes. Angew Chem Int Ed Engl 45:1034–1057. https://doi.org/10.1002/anie.200502071.

3. Parker WR, Rathnum ML, Seiner V, Trejo WH, Principe PA, Sykes RB. 1984. Cepacin A and cepacin B, two new antibiotics produced by *Pseudomonas cepacia*. J Antimicrob Chemother (Tokyo) 37:431–440. https://doi.org/10.7164/antibiotics.37.431.

4. Mullins AJ, Murray JAH, Bull MJ, Jenner M, Jones C, Webster G, Green AE, Neill DR, Connor TR, Parkhill J, Chaliss GL, Mahenthiralingam E. 2019. Genome mining identifies cepacin as a plant-protective metabolite of the pesticidal bacterium *Burkholderia ambifaria*. Nat Microbiol 4:996–1005. https://doi.org/10.1038/s41564-019-0383-z.

5. Kusumi T, Ohtani I, Nishiyama K, Kakisawa H. 1987. Caryoynencins, potent antibacterials from a plant pathogen *Pseudomonas carboxylytica*. Tetrahedron Lett 28:3981–3984. https://doi.org/10.1016/S0040-4099(00)96437-2.

6. Ross C, Scherlach K, Kloss F, Hertweck C. 2014. The molecular basis of conjugated polyyne biosynthesis in phytopathogenic bacteria. Angew Chem Int Ed Engl 53:7794–7798. https://doi.org/10.1002/anie.201403344.

7. Flórez LV, Scherlach K, Gaube P, Ross C, Sitte E, Hermes C, Rodrigues A, Hertweck C, Kaltenpoth M. 2017. Antibiotic-producing symbionts dynamically transition between plant pathogenicity and insect-defensive mutualism. Nat Commun 8:15172. https://doi.org/10.1038/ncomms15172.

8. Fritsche K, van den Berg M, de Boer W, van Beek TA, Raaijmakers JM, van Veen JA, Leveau JH. 2014. Biosynthetic genes and activity spectrum of antifungal polyynes from *Collimonas fungivorans* Ter331. Environ Microbiol 16:1334–1345. https://doi.org/10.1111/1462-2920.12440.

9. Kai K, Sogame M, Sakurai F, Nasu N, Fujita M. 2018. Collimonins A–D, unstable polyynes with antifungal or pigmentation activities from the fungus-feeding bacterium *Collimonas fungivorans* Ter331. Org Lett 20:3536–3540. https://doi.org/10.1021/acs.orglett.8b01311.

10. Ueoka R, Bhushan A, Probst SI, Bray WM, Lokey RS, Linington RG, Piel J. 2018. Genome-based identification of a plant-associated marine bacterium as a rich natural product source. Angew Chem Int Ed Engl 57:14519–14523. https://doi.org/10.1002/anie.201805673.

11. Patel M, Conover M, Horan A, Loebenberg D, Marquez J, Mierzwa R, Puar MS, Yarborough R, Waizt JA. 1988. Sch 31828, a novel antibiotic from a *Microbispora* sp. Taxonomy, fermentation, isolation and biological properties. J Antimicrob Chemother 21:794–797. https://doi.org/10.1093/jac/21.4.794.
12. Hagmann L, Jüttner F. 1996. Fischerellin A, a novel photosystem-II-inhibiting allelochemical of the cyanobacterium Fischerella musculica with anti-fungal and herbicidal activity. Tetrahedron Lett 37:6539–6542. https://doi.org/10.1016/0040-4099(96)01445-1.

13. Papke U, Gross EM, Francke W. 1997. Isolation, identification and determination of the absolute configuration of fischerellin B. A new algicide from the freshwater cyanobacterium Fischerella musculica (Thuret). Tetrahedron Lett 38:379–382. https://doi.org/10.1016/S0040-4099(96)02284-8.

14. Zhu X, Su M, Manickam K, Zhang W. 2015. Bacterial genome mining of newzymatic tools for alkyne biosynthesis. ACS Chem Biol 10:2785–2793. https://doi.org/10.1021/acschembio.5b00641.

15. Ziemert N, Alanyar M, Weber T. 2016. The evolution of genome mining in microbes—a review. Nat Prod Rep 33:988–1005. https://doi.org/10.1039/cbn00025h.

16. Zerikly M, Challis GL. 2009. Strategies for the discovery of new natural products by genome mining. ChemBioChem 10:625–633. https://doi.org/10.1002/cbic.200800389.

17. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.

18. Goris J, Klappenbach JA, Vandamme P, Coenye T, Konstantinidis KT, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:811–91. https://doi.org/10.1099/ijs.0.64483-0.

19. Peix A, Ramírez-Bahena MH, Velázquez E. 2018. The current status on the taxonomy of Pseudomonas revisited: an update. Infect Genet Evol 57:106–116. https://doi.org/10.1016/j.meegid.2017.01.026.

20. Shen B, Hindra, Yan X, Huang T, Ge H, Yang D, Rudolf JD, Lohman JR. 2015. Enediyne: exploration of microbial genomics to discover new anticancer drug leads. Bioorganic Med Chem Lett 46. https://doi.org/10.1002/ chin.201506306.

21. Yang X, Ge H, Huang T, Hindra, Yang D, Teng Q, Cnopovici I, Li X, Rudolf JD, Lohman JR, Gansemans Y, Zhu X, Huang Y, Zhao LX, Jiang Y, van Nieuwerburgh F, Rader C, Duan Y, Shen B. 2016. Strain prioritization and genome mining for enediyne natural products. mBio 7:e02104-16. https://doi.org/10.1128/mBio.02104-16.

22. Hotter V, Zopf D, Kim HJ, Silge A, Schmitt M, Aiyar P, Fleck J, Matthäus C, Hotter V, Zopf D, Kim HJ, Silge A, Schmitt M, Aiyar P, Fleck J, Matthäus C, Aiyar P, Fleck J, Matthäus C, Fleck J, Matthäus C. 2020. Insect-associated bacteria assemble the antifungal butenolidine gladiofungin by non-canonical polyketide chain termination. Angew Chem Int Ed Engl 59:2312–2316. https://doi.org/10.1002/anie.202005711.

23. Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers G, Gavrodhi DV, DeBoy RT, Seshadri R, Ren Q, Madupu R, Dodson RJ, Durkin AS, Brinkac LM, Daugherty SC, Sullivan SA, Rosovitz MJ, Gwinn ML, Zhou L, Schneider DJ, Cattinhuor SW, Nelson WC, Weidman J, Watts R, Tran Khan, Khouri H, Piron EA, Pionter LS, Thomashow LS, Loper JE. 2005. Complete genome sequence of the plant commensal Pseudomonas fluorescens PF-5. Nat Biotechnol 23:873–878. https://doi.org/10.1038/nbt1110.

24. Haas D, Défago G. 2005. Biological control of soil-borne pathogens by fluorescent pseudomonads. Nat Rev Microbiol 3:307–319. https://doi.org/10.1038/nrmicro1129.

25. Hassan KA, Johnson A, Shaffer BT, Ren Q, Kidarsa TA, Elbourne LDH, Hartney S, Duboy G, Goebel NC, Zabriskie TM, Paulsen IT, Loper JE. 2010. Reactivation of the GacA response regulator in Pseudomonas fluorescens PF-5 has far-reaching transcriptomic consequences. Environ Microbiol 12:809–915. https://doi.org/10.1111/j.1462-2920.2009.01214.x.

26. Morquis A, Coulouris G, Raytelsi Y, Madden TL, Agarwala R, Schaffer AA, Madden TL, Ainscough R, Alexander F, et al. 2008. BLAST+: architecture and applications. Bioinformatics 24:1757–1764. https://doi.org/10.1093/bioinformatics/btn232.

27. Jones P, Binnis D, Chang H-Y, Fraser M, Li W, McNamul CA, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesset S, Quinn AM, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

28. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software ver 7: improvements in performance and usability. Mol Biol Evol 30:770–780. https://doi.org/10.1093/molbev/mst101.

29. Page AJ, Cummings CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC.

30. Morgulis A, Coulouris G, Raytelsi Y, Madden TL, Agarwala R, Schaffer AA, Madden TL, Ainscough R, Alexander F, et al. 2008. BLAST+: architecture and applications. Bioinformatics 24:1757–1764. https://doi.org/10.1093/bioinformatics/btn232.

31. Jones P, Binnis D, Chang H-Y, Fraser M, Li W, McNamul CA, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesset S, Quinn AM, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

32. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software ver 7: improvements in performance and usability. Mol Biol Evol 30:770–780. https://doi.org/10.1093/molbev/mst101.

33. Page AJ, Cummings CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC.

34. Morgulis A, Coulouris G, Raytelsi Y, Madden TL, Agarwala R, Schaffer AA, Madden TL, Ainscough R, Alexander F, et al. 2008. BLAST+: architecture and applications. Bioinformatics 24:1757–1764. https://doi.org/10.1093/bioinformatics/btn232.

35. Jones P, Binnis D, Chang H-Y, Fraser M, Li W, McNamul CA, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesset S, Quinn AM, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

36. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software ver 7: improvements in performance and usability. Mol Biol Evol 30:770–780. https://doi.org/10.1093/molbev/mst101.

37. Page AJ, Cummings CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes DR, Pipe SC.

38. Sullivan MJ, Petty NK, Beatson SA. 2011. Easy protein function classification using Pfam and Pfam-A. Bioinformatics 27:1009–1017. https://doi.org/10.1093/bioinformatics/btr039.

39. Sullivan MJ, Petty NK, Beatson SA. 2011. Easy protein function classification using Pfam and Pfam-A. Bioinformatics 27:1009–1017. https://doi.org/10.1093/bioinformatics/btr039.