Cite as: M. Xue *et al.*, *Science* 10.1126/science.aax2685 (2019).

Structure elucidation of colibactin and its DNA cross-links

Mengzhao Xue[1]*, Chung Sub Kim[1,2]*, Alan R. Healy[1,2]†, Kevin M. Wernke[1], Zhixun Wang[1]‡, Madeline C. Frischling[1], Emilee E. Shine[2,3], Weiwei Wang[4,5], Seth B. Herzon[1,6]§, Jason M. Crawford[1,2,3]§

[1]Department of Chemistry, Yale University, New Haven, CN 06520, USA. [2]Chemical Biology Institute, Yale University, West Haven, CN 06516, USA. [3]Department of Microbial Pathogenesis, Yale School of Medicine, New Haven, CN 06536, USA. [4]Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, P.O. Box 208114, New Haven, CN 06520, USA. [5]W. M. Keck Biotechnology Resource Laboratory, Yale University School of Medicine, 300 George Street, New Haven, CN 06510, USA. [6]Department of Pharmacology, Yale School of Medicine, New Haven, CN 06520, USA.

***These authors contributed equally to this work.**

†Present address: New York University Abu Dhabi, Post Office Box 129188, Abu Dhabi, United Arab Emirates.

‡Present address: Department of Process Research and Development, Merck, Rahway, NJ 07065, USA.

§Corresponding author. Email: jason.crawford@yale.edu, seth.herzon@yale.edu

Colibactin is a complex secondary metabolite produced by some genotoxic gut *Escherichia coli* strains. The presence of colibactin-producing bacteria correlates with the frequency and severity of colorectal cancer in humans. However, because colibactin has not been isolated or structurally characterized, it has been difficult to study physiological effects of colibactin-producing bacteria in the human gut. We use a combination of genetics, isotope labeling, tandem mass spectrometry, and chemical synthesis to deduce the structure of colibactin. Our structural assignment accounts for all known biosynthetic and cell biology data and suggests roles for the final unaccounted enzymes in the colibactin gene cluster.

Research on the human microbiota has now established a large number of correlative relationships between bacterial species and host physiology or disease. However, deriving causal relationships from correlations or associations remains challenging (*1*). Evidence suggests that molecular-level approaches may ultimately be required to unveil many causal relationships in the microbiome; success here will illuminate therapeutic strategies to treat disease and improve human health (*2*). Toward this end, a large amount of research has been devoted to studying certain strains of Enterobacteriaceae that contain a 54 kb biosynthetic gene cluster (BGC) termed *clb* (also referred to as *pks*). The *clb* gene cluster encodes the biosynthesis of a non-proteogenic metabolite known as colibactin. *Clb*$^+$ *Escherichia coli* are commonly found in the human colon (*3*, *4*), induce DNA damage in eukaryotic cells (*5*, *6*), promote tumor formation in mouse models of colorectal cancer (CRC) (*7*–*9*), and are more prevalent in CRC patients than healthy subjects (*7*, *10*). These findings have been attributed to colibactin, but experiments designed to test this hypothesis have been impossible to conduct because colibactin does not appear to be isolable, and its structure has remained incompletely defined (*11*–*15*). Understanding the full structure of colibactin will lay the foundation to probe for a causal relationship between one of the most well studied human microbiota phenotypes and its associated disease with atomic resolution.

Because colibactin has been recalcitrant to isolation, knowledge of its structure and bioactivity derives from diverse interdisciplinary findings. Enzymology, bioinformatic analysis of the *clb* BGC, stable isotope feeding experiments, characterization of biosynthetic intermediates, and gene deletion and editing studies have given insights into many elements of colibactin's biosynthesis, bioactivity, and cellular trafficking (*11*–*15*). Consistent with the determination that *clb*$^+$ *E. coli* are genotoxic (*5*, *6*), a *clb* metabolite isolated from a mutant strain was shown to damage DNA in cell-free experiments (*16*). Subsequently, chemical synthesis was used to access other *clb* metabolites and putative biosynthetic intermediates, and further a mechanistic model to explain colibactin's genotoxic properties (*17*).

Merging this data forms a picture, albeit incomplete, of colibactin's biosynthesis, structure, and mode of genotoxicity. Colibactin is assembled in a linear prodrug form referred to as precolibactin (Fig. 1) (*1*). Key structural elements of precolibactins include a terminal *N*-myristoyl-D-Asn amide (blue in **1**) (*18*–*20*) and an aminocyclopropane residue (green in **1**) (*16*, *21*, *22*). The terminal amide is cleaved in the periplasm by a pathway-dedicated serine protease known as colibactin peptidase (ClbP) (*23*, *24*). The resulting amine **2** undergoes a series of cyclization reactions to generate spirocyclopropyldihydro-2-pyrrolones resembling **3** (*17*, *25*). These cyclizations place the cyclopropane in conjugation with both an imine and amide, rendering the cyclopropane electrophilic and capable of alkylating DNA (*14*, *17*) (For a detailed mechanism of cyclization and DNA alkylation, see fig. S1). The mono(adenine) adduct **4** was identified in the digestion mixture of

Downloaded from http://science.sciencemag.org/ on August 8, 2019

linearized pUC19 DNA exposed to *clb*$^+$ *E. coli* (*26*) and in colonic epithelial cells of mice infected with *clb*$^+$ *E. coli* (*27*). However, a recent study established that *clb*$^+$ *E. coli* cross-link DNA (*28*), suggesting that colibactin contains a second DNA-reactive site that has yet to be elucidated. The full structure of colibactin and the site of the second alkylation have remained undefined.

Mutation of *clbP* has been widely-employed to promote the accumulation of precolibactins and facilitate isolation. Precolibactins A–C (**5**–**7**) and precolibactin 886 (**8a**) exemplify the metabolites produced in Δ*clbP* cultures (*16*, *20*, *22*, *29*–*32*). The persistence of the *N*-myristoyl-D-Asn residue (deriving from mutation of *clbP*) changes the fate of the linear precursor **1** and promotes pyridone formation (for **5**–**7**) (*14*, *17*) or macrocyclization (for **8**) (*33*). Precolibactin 886 (**8a**) is an advanced metabolite that requires every biosynthetic gene in the pathway except polyketide synthase (PKS) *clbO*, type II thioesterase *clbQ,* and amidase *clbL* (*25*). Recently, precolibactin 969 (**8b**, Fig. 1), which bears a terminal oxazole ring, was reported, but this product still does not account for every biosynthetic step encoded in the *clb* gene cluster (vide infra) (*34*). Genetic studies established that deletion of any biosynthetic gene in the *clb* locus abolishes cytopathic effects (*5*), thus the full biosynthetic product is believed to possess additional chemical functionalities not contained in **8a** or **8b**.

Characterization of colibactin–DNA crosslinks and biosynthetic proposal

Because colibactin has proven recalcitrant to isolation, we focused on structural elucidation of the DNA cross-links generated by *clb*$^+$ *E. coli* (*28*). This approach circumvents the challenges in obtaining pure samples of the metabolite from fermentation extracts, and instead relies intensively on MS and tandem MS analysis (rather than conventional NMR analysis). The stereochemical assignments in the structures that follow are based on established intermediates and nonribosomal peptide synthetase (NRPS)-polyketide synthase (PKS) biosynthetic logic.

Tandem MS analysis of the digestion products of linearized pUC19 DNA that had been exposed to *clb*$^+$ *E. coli* was used to elucidate the structure of the colibactin–adenine adduct, **4** (*26*). In that study, wild-type *E. coli* BW25113 and its cysteine and methionine auxotrophs (Δ*cys*E, Δ*met*A) (*35*) containing *clb* on a bacterial artificial chromosome (BAC) were employed. The latter two cultures were supplemented with L-[U-^{13}C]-Cys or L-[U-^{13}C]-Met, which are known precursors to the thiazole (*16*) and aminocyclopropane (*16*, *21*, *36*) residues of colibactin, respectively. This approach allowed for the identification of *clb* metabolite–nucleobase adducts by mining for shifts in the mass spectra between unlabeled wild-type and labeled auxotrophic cultures.

Further analysis of this data revealed a compound of *m/z* = 537.1721 (*z* = 2) (Fig. 2A and supplementary materials), which corresponds to a molecular formula of $C_{47}H_{50}N_{18}O_9S_2^{2+}$ (error = 0.37 ppm). The doubly charged ion (*m/z* = 537.1721) was shifted by 3 or 4 units in cultures containing L-[U-^{13}C]-Cys or L-[U-^{13}C]-Met, respectively, supporting the presence of two thiazole and two cyclopropane residues. To gain further insights into the structure, we analyzed its production in glycine (Δ*gly*A) and serine (Δ*ser*A) auxotrophs. These cultures were supplemented with [U-^{13}C]-Gly, L-[U-^{13}C]-Ser, or L-[U-^{13}C, ^{15}N]-Ser. Glycine serves as the CN extension in the 2-methylamino thiazole of precolibactin A (**5**, highlighted by green spheres in **5**, Fig. 1) (*16*), whereas serine is incorporated into precolibactin 886 (**8a**) via an unusual α-aminomalonate extender unit (*31*, *32*, *37*). The doubly charged ion (*m/z* = 537.1721) was shifted by 1 unit in cultures containing [U-^{13}C]-Gly, indicating incorporation of one glycine building block. However, this ion was shifted by 1.5 units in cultures containing L-[U-^{13}C]-Ser, and by 2 units in cultures containing L-[U-^{13}C, ^{15}N]-Ser indicating that three carbon atoms and one nitrogen atom are derived from serine. This unexpectedly suggests that two α-aminomalonate building blocks are transformed into two distinct fragments which are incorporated into colibactin's structure, rather than only one, or two identical, serine-derived building blocks. We note that these cultures produced a range of higher molecular weight isotopologs owing to amino acid metabolism and incorporation into other building blocks (see below).

When wild-type *clb*$^+$ *E. coli* cultures were grown in medium lacking amino acids and supplemented with D-[U-^{13}C]-glucose, the doubly charged ion (*m/z* = 537.1721) was shifted by 18.5 units, establishing that the colibactin residue contained 37 carbon atoms. Cultivation in minimal medium containing [^{15}N]-ammonium chloride shifted the doubly charged ion by 4 units, indicating the colibactin residue contained eight nitrogen atoms. A double-labeling experiment using D-[U-^{13}C]-glucose and [^{15}N] ammonium chloride resulted in a shift of 22.5 units, confirming the results of the individual labeling experiments. The singly charged (*z* = 1) and triply charged (*z* = 3) ions were also detected in many of these auxotrophs and provided data of comparable quality (supplementary materials). A fragment ion corresponding to protonated adenine was observed in the tandem MS of each of the isotopically labeled and unlabeled adducts. Additionally, the consecutive loss of two adenine bases was observed in all labeling experiments. Finally, the doubly charged ion (*m/z* = 537.1721) was detected when the experiment was conducted with poly(AT) as the substrate. Collectively these data suggest the generation of a bis(adenine) adduct and a molecular formula of $C_{37}H_{38}N_8O_9S_2$ for the colibactin residue contained therein.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

Based on these data, we reconsidered the unaccounted functions of ClbO, ClbQ, and ClbL (Fig. 3). ClbO is a polyketide synthase that accepts an α-aminomalonyl-extender unit in protein biochemical studies (*32*), suggesting a canonical extension step. ClbQ serves as an editing thioesterase and also off-loads intermediary structures with an observed preference for hydrolyzing thioester intermediates toward the middle of the assembly line (*25*, *31*, *38*). While these off-loaded structures enhance the metabolite diversity arising from the pathway, we reasoned that they could also serve as downstream substrates. In this scenario, off-loading of intermediate **A** (Fig. 3) followed by an uncharacterized ClbL-mediated transpeptidase activity could promote a heterodimerization. The resulting structure would accommodate the isotopic labeling studies, including the presence of two aminocyclopropane units derived from methionine (Fig. 3, and see below), and the detection of double nucleobase adducts arising from two-fold alkylation of DNA. While our work was under revision, a recent study supporting ClbL as an amidase was published, which is in agreement with our model (*39*, *40*).

Taking all of these data into consideration, we formulated the structure of the observed parent ion as the bis(adenine) adduct **9** (Fig. 2B). The experimental and theoretical masses for **9** are in agreement (error = 0.37 ppm). The positions of the cysteine, methionine, serine, and glycine isotopic labels depicted in **9**, which were determined by tandem MS analysis, are fully supported by all known elements of colibactin biosynthesis (supplementary materials). The tandem MS fragments **10**–**12** shown in Fig. 2C provide further robust support for the structure **9**. Each of the ions **10**–**12** possessed the expected mass shift in the individual labeling experiments (supplementary materials).

The structure **9** is fully supported by all published data in the field, to our knowledge. The bis(adenine) adduct **9** derives from two-fold alkylation of DNA by cyclopropane ring-opening, which is in agreement with the discovery that colibactin derivatives containing one cyclopropane residue alkylate DNA by a parallel pathway (*17*). Additionally, two-fold alkylation of DNA to form **9** is consistent with the observation that *clb*$^+$ *E. coli* cross-link DNA and activate cross-link repair machinery in human cells (*28*). The proposed ClbL-mediated transacylation appends the second (pro)warhead, and these data explain why *clbL* mutants alkylate but do not crosslink exogenous DNA (*41*). The aminal functional groups in **9** derive from aerobic oxidation of the nucleotide ring-opened products, as previously established in studies of simpler colibactin derivatives (*26*, *42*). Finally, in agreement with the well-established propensity of α-diketones to hydrate under aqueous conditions (the K_d for dissociation of the monohydrate of butane-2,3-dione = 0.30) (*43*), the product of hydration of C37 (**S1**) was also detected (supplementary materials). Tandem MS and isotopic labeling data for **S1** fully support the structure of the hydrate and are in agreement with the diketone form **9** (supplementary materials).

Additional nucleobase adducts were detected at discrete retention times (Fig. 2D). The methylaminoketone **13** and its corresponding hydrate (**S15**) are of special significance (supplementary materials): These are likely formed by hydrolytic off-loading of the biosynthetic product **E** (Fig. 3); its enzyme precursor serves as the acceptor in the ClbL transacylation step we proposed. The known adduct **4** (Fig. 1A) (*26*, *27*) and the right-hand fragment **14** (Fig. 2D) were also detected (supplementary materials). Fragment **14** is significant because we have demonstrated that that the C36–C37 bond in advanced colibactins [see (*9*) for numbering] is susceptible to oxidative cleavage in the presence of weak nucleophiles, such as water or methanol (*33*). Hydrolytic degradation of **9** at this bond accounts for isolation of the earlier monoadenine adduct **4** (*26*, *27*) and now, the right-hand fragment **14**. We also detected the hydrate and diketone of a full-length mono(adenine) adduct (**S2** and **S7**) (supplementary materials).

The cross-linking, digestion, and MS experiments performed above served to reveal the presence of the bis(adenine) adduct **9** and its corresponding hydrate **S1**. Although the relevance of these bis(nucleobase) adducts to colibactin genotoxicity remains to be determined, we sought to probe for their production in human tissue culture. Accordingly, HCT-116 colon cells were infected with *clb*$^+$ *E. coli* BW25113. Following a 2 hours infection, the human cells were separated, and their genomic DNA was isolated, digested, and subjected to MS analysis. We were able to detect trace levels of the bis(adenine) adduct **9** (error = 1.49 ppm) and its corresponding hydrate **S1** (error = 0.37 ppm). The retention time of these materials were identical to the material derived from pUC19 DNA exposed to the *clb*$^+$ *E. coli* BW25113 (supplementary materials).

Identification of colibactin (17)

We then searched *clb*$^+$ *E. coli* cultures for the structures of the α-ketoamine **16a**, the corresponding α-ketoimine **16b**, and the α-dicarbonyl **17** (Fig. 3), which were anticipated based on the structure of the bis(adenine) adduct **9** and established biosynthetic logic. Although our data do not allow us to exclude **16a** or **16b** as active *clb* genotoxic contributors (oxidation of **16a** and hydrolysis of **16b** lead to the observed α-dicarbonyl under our experimental conditions) our prior studies established that α-ketoamines and α-ketoimines structurally-related to **16a** and **16b** rapidly transform to the corresponding α-dicarbonyl under mild conditions (*33*). Moreover, we were unable to detect **16a** or **16b** in freshly prepared *E. coli* extracts. However, the proton and sodium adducts of colibactin (**17**) were observed in *E. coli* DH10B harboring the *clb* BAC (supplementary materials). Colibactin

Downloaded from http://science.sciencemag.org/ on August 8, 2019

(**17**) was not detectable in a *clbO* deletion mutant and a *clbL* active site point mutant (S179A) (Fig. 4A). Because colibactin (**17**) was detected at low abundance, we also confirmed production in the wild-type probiotic *E. coli* Nissle 1917. Deletion of the *clb* genomic island (*20*) in Nissle 1917 or a *clb*$^-$ BAC control strain abolished production, as expected. We hypothesized that the titer of colibactin (**17**) might be higher in a *clbS* Nissle 1917 mutant (*42*, *44*), as we previous established that ClbS is a self-resistance enzyme that catalyzes hydrolytic ring-opening of the cyclopropane ring (*42*). While deletion of *clbS* leads to a fitness defect and activates a *clb*-dependent bacterial SOS DNA damage response (*44*), this genetic modification resulted in an 8.5-fold improvement in the signal intensity.

We then individually supplemented *E. coli* Nissle 1917 Δ*clbS* cultures with labeled amino acids. Colibactin (**17**) (Fig. 4, B and C) incorporated two equivalents of Cys, Met, and Ala, as expected based on its proposed structure. The cultures labeled with L-[U-^{13}C]-Ser and [U-^{13}C]-Gly produced a range of isotopologs owing to their metabolism and incorporation into other building blocks (supplementary materials). To account for this variation, we used Ser-derived enterobactin, an iron-scavenging siderophore in *E. coli*, as an internal control for comparison (fig. S118). We also repeated Gly labeling in *clb*$^+$ DH10B for confirmation of dominant mono-labeling of Gly in colibactin (**17**). The key tandem MS ions **19** and **20** were observed and provide further support for colibactin's structure (Fig. 4D). We note that because of the nearly-C_2 symmetric structure of colibactin (**17**), the two structures of **20** shown are equally plausible based on the available data. Similar to the colibactin–DNA adducts, we observed the C37 hydrate of colibactin (**17**) (**S34**) (supplementary materials).

Characterization of precolibactin 1489 (18)

Every biosynthetic enzyme encoded in the *clb* gene cluster is necessary to observe the genotoxic phenotype associated with *clb*$^+$ *E. coli* (*5*). While truncated precolibactins such as precolibactin 886 (**8a**, Fig. 1B) can be detected as macrocyclization products in non-genotoxic *clbP* peptidase mutants (*31*), precolibactin 886 (**8a**) is still produced in mutants of *clbL*, *clbO*, and *clbQ* in a *clbP*-deficient genetic background (ClbP S95A active site point mutant) (*25*). Additionally, the recently characterized metabolite precolibactin 969 (**8b**), isolated from a *clbP/ clbQ*/*clbS* triple mutant (*34*), does not account for *clbL* and *clbQ* and was undetectable in freshly prepared organic extracts of a *clbP*-deficient strain under our experimental conditions. Given the structure of colibactin (**17**) and the requirement of every biosynthetic enzyme for cytopathic effects (*5*), we reasoned that more complex precolibactins existed.

Accordingly, we searched for the precolibactin that could account for colibactin (**17**) in *clb*$^+$ DH10B (ClbP S95A) (*25*). While we were not able to detect the expected unstable linear precursor precolibactin 1491 (**15**, Fig. 3) or its oxidation products, we detected both the proton and sodium ion adducts of a metabolite predicted to be the macrocycle precolibactin 1489 (**18**) (supplementary materials). We employed genome editing to individually inactivate the catalytic domains from ClbH to ClbL in the biosynthetic pathway (*25*). Precolibactin 1489 (**18**) was genetically dependent on all of the enzymatic steps in the pathway (Fig. 5A). Production was only detected in an acyltransferase (AT) domain mutant of ClbI; metabolites dependent on this single domain can be complemented *in trans* by other ATs in the cell (*25*). Thus, precolibactin 1489 (**18**) represents the first reported product derived from the complete *clb* biosynthetic pathway. A similar analysis confirmed that precolibactin 886 (**8a**) was still produced in *clbL, clbO,* or *clbQ* mutants.

The structure of precolibactin 1489 (**18**) is supported by extensive ^{13}C-isotopic amino acid labeling and tandem MS analysis (Figs. 5, B to D, and fig. S125). Labeled Met, Gly, Ala, Cys, and Ser precursors incorporated into precolibactin 1489 (**18**) in a manner fully consistent with its biosynthesis and proposed structure. Additionally, two units of L-[U-^{13}C]-Asn were incorporated, supporting the presence of two *N*-myristoyl-D-Asn residues, as expected. Tandem MS analysis of precolibactin 1489 (**18**) produced the ions **21**–**23** and **S35**–**S37**, which are also consistent with the proposed structure (Fig. 5D and fig. S125C). Based on the recent determination that ClbP-deacylation of precolibactin 886 (**8a**) produces a non-genotoxic pyridone (*33*), it seems likely that precolibactin 1489 (**18**) is simply a stable product arising from oxidation and macrocyclization of the putative linear precursor precolibactin 1491 (**15**, Fig. 3). Regardless, these studies support a two-fold *N*-acyl-D-Asn prodrug activation mechanism, in which ClbP peptidase sequentially initiates the formation of two electrophilic architectures.

Confirmation of the structure of colibactin (17)

The structure of colibactin (**17**) was confirmed by chemical synthesis. The presence of two electrophilic spirocyclopropyldihydro-2-pyrrolones (*17*) and the hydrolytically-labile C36–C37 α-dicarbonyl (*33*) necessitated a careful analysis of potential synthetic pathways. Figure 6A outlines the essential elements of our strategy. While in earlier studies (*17*) monomeric colibactins were assembled by a linear approach [stepwise formation of bonds a, b, and c, in that order (Fig. 6A) (*24*)], we recognized that colibactin (**17**) could be assembled by a two-fold coupling (a, a´ bond formation) of the diamine **26** with the novel β-ketothioester **25**. In addition to increased convergence, this approach masks the reactive (*17*) spirocyclopropyldihydro-2-pyrrolones as identical stable vinylogous imides. We have established that *N*-deacylation followed by mild neutralization is sufficient to induce

Downloaded from http://science.sciencemag.org/ on August 8, 2019

cyclization and formation of the spirocyclopropyldihydro-2-pyrrolones residues (*17*, *41*). Based on our observations that C36–C37 α-aminoketones undergo spontaneous oxidation (*33*), we targeted an α-hydroxyketone in place of the α-dicarbonyl in colibactin (**17**). This was projected to allow for the assembly of **26** by benzoin addition, followed by late-stage oxidation to generate the sensitive α-dicarbonyl. In our synthesis of precolibactin 886 (**8a**) (*33*), the initial ketone was generated at C36. However, in exploratory experiments, we found that intermediates with a C37 ketone were more stable and pursued these, as outlined below.

The synthesis of the β-ketothioester **25** is shown in Fig. 6B. Silver trifluoroacetate-mediated coupling of the known β-ketothioester **27** (*17*) with ethyl 1-aminocyclopropyl-1-carboxylate (**28**) generated a β-ketoamide (not shown) that was cyclized to the vinylogous imide **29**. Addition of the lithium enolate of *tert*-butyl thioacetate to **29** then provided the β-ketothioester **25.** The diamine **26** was synthesized by the route shown in Fig. 6C. Selenium dioxide oxidation of the commercial reagent ethyl 2-methylthiazole-4-carboxylate (**30**) generated the aldehyde **31**. Reduction of the aldehyde, followed by saponification of the ester provided the hydroxy acid **32**. Treatment of the hydroxy acid **32** with excess 1,1´-carbonyldiimidazole (CDI) resulted in acylation of the primary alcohol and activation of the carboxylic acid as the expected acyl imidazole (LC-MS analysis). Addition of sodium nitromethanide, followed by in situ hydrolysis of the acylated alcohol, formed the β-nitroketone ***33***. Hydrogenolysis of the nitro group, followed by protection of the resulting primary amine and oxidation of the primary alcohol (2-iodoxybenzoic acid, IBX) provided the aldehyde **34**. Silyl cyanohydrin formation, deprotonation, and addition of the aldehyde **36** (*33*) generated the α-silyloxy ketone ***37***. The carbamate protecting groups were removed under acidic conditions to furnish the diammonium salt **26**.

Silver-mediated coupling of the diamine **26** with an excess of the β-ketothioester **25** provided the expected two-fold coupling product (LC-MS). However, all attempts to purify this product resulted in extensive decomposition deriving from cleavage of the C36–C37 bond (LC-MS analysis). To circumvent this, we developed conditions to protect this residue in situ. Thus, immediately following the fragment coupling, the enedislylether **24** was formed by silylation of the product mixture. The stereochemistry of the central alkene was determined to be (*E*), as shown, by 2D-ROESY analysis. The yield of this two-fold coupling–protection sequence was 17% (based on ^{1}H NMR analysis of the unpurified product mixture using an internal standard), and **24** was isolated in 11.5% yield following reverse phase HPLC purification. By this approach, 5–7 mg batches of **24** were readily-prepared.

Conversion of the protected intermediate **24** to colibactin (**17**) proved to be challenging since we found that introduction of the C36–C37 α-dicarbonyl rendered the intermediates exceedingly unstable. This is consistent with an earlier model study (*33*) demonstrating rupture of the C36–C37 bond under slightly basic conditions. Ultimately, we found that treatment with concentrated hydrochloric acid in ethanol resulted in instantaneous cleavage of the carbamate protecting groups and one silyl ether; this was followed by slower and sequential cleavage of the remaining silyl ether, and aerobic oxidation to the α-dicarbonyl **38**. The α-dicarbonyl **38** was accompanied by variable amounts of the diketone hydrate (LC-MS analysis) (fig. S127 and table S70) as observed for the bis(adenine) adduct **9** and colibactin (**17**, vide supra).

On dissolving **38** in rigorously deoxygenated aqueous citric acid buffer (pH = 5.0), we observed double cyclodehydration to form colibactin (**17**) (Fig. 7A). This mild cyclization is consistent with earlier studies establishing synthetic iminium ions resembling **38** cyclize to spirocyclopropyldihydro-2-pyrrolone genotoxins instantaneously under aqueous conditions (*17*, *41*), and genetic studies supporting the off-loading of linear biosynthetic intermediates, followed by spontaneous transformation to the unsaturated imine electrophile (*25*). Although we were unable to separate small amounts of side products deriving from hydrolytic ring-opening of the vinylogous urea of **38,** synthetic colibactin (**17**) obtained in this way was indistinguishable from natural material by LC-MS coinjection and tandem MS analysis using a range of collision energies (20–50 eV) (Fig. 7, B and C, and fig. S127).

Although we could enhance mass spectral detection of natural colibactin (**17**) in the *clbS* mutant of Nissle 1917, the titers remained too low to facilitate isolation. Consequently, we turned to functional analysis of synthetic colibactin (**17**) in the DNA cross-linking assay to further confirm the structural assignment. We observed dose-dependent cross-linking of DNA (Fig. 7D) by forming colibactin (**17**) in situ from the iminium diion **38** at pH 5 in the presence of DNA. Additionally, the DNA cross-links induced by synthetic colibactin (**17**) were indistinguishable from those produced by *clb*$^+$ *E. coli* (figs. S129 to S132) under basic denaturing gel conditions. Cross-linking was strongest at pH 5.0 and diminished as the pH was increased, an observation consistent with the known instability of the α-diketone under basic conditions (*33*). This was also consistent with the stability of the cross-links derived from *clb*$^+$ bacteria (supplementary materials). The DNA cross-links derived from **38** were isolated, digested, and subjected to tandem MS using the same parameters employed to analyze the natural colibactin–bis(adeinine) adduct, which confirmed the assignment (Fig. 7E). All of the ions detected from crosslinking products derived from *clb*$^+$ *E. coli* BW25113 were detected using synthetic **38** (supplementary materials). Collectively, the abundance of genetics data as well as these synthetic efforts confirm the structure of the major colibactin as **17**.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

Conclusion

Elucidating the complete structure of colibactin (**17**) puts to rest the decade long debate over the structure of the metastable metabolite. Correlative relationships abound in the microbiome field, but causative relationships are far more rare, primarily owing to a lack of detailed, molecular-level structure–function analysis. The development of a chemical synthesis of colibactin (**17**) enables researchers to probe for a causative relationship between the metabolite and CRC formation. The interdisciplinary approach we developed to determine and confirm colibactin's structure may be extensible to other low abundance bioactive metabolites from complex backgrounds such as the human microbiome.

Materials and Methods

NMR spectroscopy

Proton nuclear magnetic resonance spectra (^{1}H NMR) were recorded at 400, 500 or 600 MHz at 24°C, unless otherwise noted. Chemical shifts are expressed in parts per million (ppm, δ scale) downfield from tetramethylsilane and are referenced to residual protium in the NMR solvent ($CDCl_3$, δ 7.26; CD_2HOD, δ 3.31; $CDHCl_2$, δ 5.33; C_2D_5HSO, δ 2.50). Data are represented as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet and/or multiple resonances, br = broad, app = apparent), coupling constant in Hertz, integration, and assignment. Proton-decoupled carbon nuclear magnetic resonance spectra (^{13}C NMR) were recorded at 100, 125 or 150 MHz at 24°C, unless otherwise noted. Chemical shifts are expressed in parts per million (ppm, δ scale) downfield from tetramethylsilane and are referenced to the carbon resonances of the solvent ($CDCl_3$, δ 77.17; CD_3OD, δ 49.0; CD_2Cl_2, δ 54.0; C_2D_6SO, δ 39.5). Signals of protons and carbons were assigned, as far as possible, by using the following two-dimensional NMR spectroscopy techniques: [^{1}H, ^{1}H] COSY (Correlation Spectroscopy), [^{1}H, ^{13}C] HSQC (Heteronuclear Single Quantum Coherence) and long range [^{1}H, ^{13}C] HMBC (Heteronuclear Multiple Bond Connectivity).

Infrared spectroscopy

Attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectra were obtained using a Thermo Electron Corporation Nicolet 6700 FTIR spectrometer referenced to a polystyrene standard. Data are represented as follows: frequency of absorption (cm^{-1}), intensity of absorption (s = strong, m = medium, w = weak, br = broad).

Analytical LC-MS for synthetic chemistry

Analytical ultra high-performance liquid chromatography-mass spectrometry (UPLC-MS) was performed on a Waters UPLC-MS instrument equipped with a reverse-phase C_{18} column (1.7 μm particle size, 2.1 × 50 mm), dual atmospheric pressure chemical ionization (API)/electrospray (ESI) mass spectrometry detector, and photodiode array detector. Samples were eluted with a linear gradient of 5% acetonitrile–water containing 0.1% formic acid→100% acetonitrile containing 0.1% formic acid over 0.75 min, followed by 100% acetonitrile containing 0.1% formic acid for 0.75 min, at a flow rate of 800 μL/min.

HRMS for synthetic intermediates

High-resolution mass spectrometry (HRMS) spectra were obtained on either a Waters UPLC-HRMS instrument equipped with a dual API/ESI high-resolution mass spectrometry detector and photodiode array detector eluting over a reverse-phase C_{18} column (1.7 μm particle size, 2.1 × 50 mm) with a linear gradient of 5% acetonitrile–water containing 0.1% formic acid→95% acetonitrile–water containing 0.1% formic acid for 1 min, at a flow rate of 600 μL/min or an Agilent 6550A QTOF Hi Res LC-MS equipped with a 1290 dual spray API source eluting over an Agilent Eclipse Plus C_{18} column (1.7μm particle size, 4.5 × 50 mm) with a linear gradient of 5% acetonitrile–water containing 0.1% formic acid→95% acetonitrile–water containing 0.1% formic acid for 6 min, at a flow rate of 500 μL/min.

HRMS for natural (pre)colibactins

HRMS and tandem MS data were acquired by an Agilent iFunnel 6550 quadrupole time-of-flight (QTOF) mass spectrometer coupled to an Agilent Infinity 1290 HPLC, scanning from *m/z* 25–1700 and a Phenomenex Kinetex 1.7 μ C18 100 Å column (100 × 2.1 mm, flow rate 0.3 mL/min, a water–acetonitrile gradient solvent system containing 0.1% formic acid (FA): 0–2 min, 5% acetonitrile; 2–26 min, 5 to 98% acetonitrile; hold for 10 min, 98% acetonitrile). The domain-targeted metabolomics result for precolibactin 1489 (**18**) was obtained by reanalyzing data from our previous study (*25*).

HPLC enrichment for natural colibactin-nucleobase adducts detected from the genomic DNA

For colibactin-monoadenine adduct

The digested mixture was dissolved in 100 μL of water and injected onto a semipreparative reverse phase HPLC system equipped with a Phenomenex Luna C8 (2) 100 Å column [250 × 10 mm, flow rate 4.0 mL/min, a gradient elution from 5 to 100% aqueous acetonitrile with 0.01% trifluoroacetic acid over 30 min (0–5 min, 5%; 5-30 min, 5–100%)] using a 1 min fraction collection window. Fractions 11–20 were combined, dried, and dissolved in 20 μL of methanol for further LC-MS anlalysis.

For colibactin-(bis)adenine adduct

The digested mixture was dissolved in 10 mL of water and injected onto a preparative reverse phase HPLC system

Downloaded from http://science.sciencemag.org/ on August 8, 2019

equipped with Agilent Polaris C18-A 5 μm column [21.2 × 250 mm, flow rate 8.0 mL/min, a gradient elution from 5 to 100% aqueous acetonitrile with 0.01% trifluoroacetic acid over 30 min (0–5 min, 5%; 5-30 min, 5–100%)] using a 1 min fraction collection window. Fractions 21–30 were combined, dried, and dissolved in 20 μL of methanol for further LC-MS anlalysis.

HRMS for natural colibactin–nucleobase adducts

HRMS and tandem MS data were obtained at the Mass Spectrometry and Proteomics Resource of the W.M. Keck Foundation Biotechnology Resource Laboratory at Yale University (New Haven, CT). All HRMS/MS samples were prepared in 1-mL screw neck total recovery vials (Waters, Milford, MA). The concentration of the digested nucleosides was adjusted to 50 ng/μL before injection. 5 μL of sample was injected at 4°C. UPLC analysis was performed on an AcQuity M-Class Peptide BEH C18 column (130 Å pore size, 1.7 μm particle size, 75 μm × 250 mm) equipped with an M-Class Symmetry C18 trap column (100 Å pore size, 5 μm particle size, 180 μm × 20 mm) at 37°C. Trapping was initiated at 5 μL/min at 99.5% of aqueous mobile phase (0.1% formic acid in water) for 3 min, and the gradient for separation began at 3% organic mobile phase (0.1% formic acid in acetonitrile), and increased to 5% over 1 min, 25% over 32 min, 50% over 5 min, 90% over 5 min and then maintained at 90% for 5 min and then 3% over 2 min, and equilibrated for an additional 20 min. Mass spectrometry was acquired on an Orbitrap Elite FTMS (Thermo Scientific) or on an Orbitrap Fusion FTMS (Thermo Scientific). The Orbitrap Elite FTMS (Thermo Scientific) was set at full scan from m/z = 150–1800 at a resolution ranging from 30,000 to 60,000, and the data-dependent MS^2 scans were collected with CID (collision-induced dissociation) at collision energies ranging from 35 eV to 40 eV. The Orbitrap Fusion FTMS (Thermo Scientific) was set to scan from m/z = 150–1100 with a resolution of 60,000, and the data-dependent MS^2 scans were collected with HCD (higher-energy collisional dissociation) at 32 eV collision energy using quadrupole isolation. Data was analyzed using the Thermo Xcalibur Qual Browser software (Version 2.2).

Cell lines

E. coli strains include the *E.coli* K-12 BW25113 parent strain and its single gene knock-out strains: cysteine auxotroph JW3582-2 (Δ*cysE720::kan*), methionine auxotroph JW3973-1 (Δ*metA780::kan*), serine auxotroph JW2880-1 (Δ*SerA764::kan*), and glycine auxotroph JW2535-2 (Δ*glyA725::kan*). The isolated BAC DNA (pBAC *clb*$^+$ and *clb*$^-$) were separately transformed into these BW25113-derived strains.

Mutant strains

E. coli Nissle 1917 Δ*clbS* was constructed as previously described (*20*). Briefly, the FRT-flanked spectinomycin resistance cassette of pIJ778 was amplified using primers with short sequence extensions homologous to the flanking regions of *clbS*. Purified PCR products were desalted and transformed into *E. coli* Nissle 1917 carrying the lambda red recombinase system on plasmid pKD46. Transformants were selected by plating on streptomycin (50 μg/mL). Colonies were analyzed with overspanning PCR and the resulting product was sequenced to confirm the replacement of gene *clbS* with the spectinomycin resistance gene.

The DH10B Δ*clbO* strain was generated in a wildtype *clb*$^+$ BAC background (containing a functional copy of the colibactin peptidase, ClbP), as previously described (*25*). This full gene-deletion was generated in the same manner as above using the lambda red recombinase system, but with apramycin as the selection marker. To avoid potential polar effects on the pathway, recombineering plasmid pKD46 was cured and plasmid pCP20 encoding the FLP recombinase was introduced in order to flip out the apramycin gene cassette. Successful deletion was confirmed by overspanning PCR. The DH10B Δ*clbL*-S179A strain was generated in a wildtype background (functional copy of the colibactin peptidase, ClbP), as previously described (*25*). Briefly, multiplex automated genome engineering (MAGE) was used to insert a single codon mutation into an active site serine residue of *clbL*, as determined by homology alignments to characterized amidase domains. Multiplex allele-specific colony PCR (MASC-PCR) was used to screen for mutations introduced and verified through overspanning PCR of the gene of interest and subsequent sequencing.

DNA and nucleic acids

The 2686 bp plasmid pUC19 was purchased from New England Biolabs and linearized with the endonuclease EcoRI (New England Biolabs, 5 U/μg DNA). The linearized plasmid was purified using the Monarch® PCR and DNA Cleanup Kit (New England Biolabs) and eluted with 10 mM Tris–1 mM EDTA pH 8.0 buffer.

Preparation of media

Isotopically-labeled reagents were purchased from Cambridge Isotope Laboratories, including L-[U-^{13}C]-asparagine: H_2O ([$^{13}C_4$]-Ala, 99% ^{13}C), L-[U-^{13}C]-alanine ([$^{13}C_3$]-Ala, 99% ^{13}C), L-[U-^{13}C]-cysteine ([$^{13}C_3$]-Cys, 99% ^{13}C), L-[U-^{13}C]-methionine ([$^{13}C_5$]-Met, 99% ^{13}C), L-[U-^{13}C]-serine ([$^{13}C_3$]-Ser, 99%^{13}C), L-[U-^{13}C, ^{15}N]-serine ([$^{13}C_3$, ^{15}N]-Ser, 99% ^{13}C, 99%^{15}N), [U-^{13}C]-glycine ([$^{13}C_2$]-Gly, 99% ^{13}C), D-[U-^{13}C]-glucose ([$^{13}C_6$]-glucose, 99% ^{13}C), and [^{15}N]-ammonium chloride ($^{15}NH_4Cl$, 99% ^{15}N). To prepare the media for isolating partially labeled colibactin–nucleobase adducts, the labeled amino acids were

Downloaded from http://science.sciencemag.org/ on August 8, 2019

separately incorporated into modified M9-CA medium for culturing the corresponding auxotrophs including JW3582-2 (cysteine), JW3973-1 (methionine), JW2880-1 (serine), and JW2535-2 (glycine). Natural abundance cysteine, methionine, serine, and glycine were incorporated into modified M9-CA medium for culturing the BW25113 parent strain as a control. To prepare the modified M9-CA medium, the M9 minimal medium (Sigma) was supplemented with 0.4% glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, chloramphenicol (12.5 μg/mL), and the following L-amino acid mass composition (5 g/L total): 3.5% Arg, 20.0% Glu, 2.5% His, 5.0% Ile, 8.0% Leu, 7.0% Lys, 4.5% Phe, 9.5% Pro, 4.0% Thr, 1.0% Trp, 6.0% Tyr, 5% Val, 4% Asn, 4% Ala, 4% Met, 4% Gly, 4% Cys, and 4% Ser. To prepare the media for isolating universally-labeled colibactin–nucleobase adducts, [$^{13}C_6$]-glucose ([U-^{13}C]), $^{15}NH_4Cl$ ([U-^{15}N]), and a combination of [$^{13}C_6$]-glucose and $^{15}NH_4Cl$ ([U-^{13}C, U-^{15}N]) were separately incorporated into modified M9-glucose medium for culturing the BW25113 parent strain. Natural abundance glucose and ammonium chloride salt were incorporated into the modified M9-glucose medium for culturing the BW25113 parent strain as a control. The modified M9-glucose medium contained 6.78 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, and 0.5 g/L NaCl, and was supplemented with 0.4% glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, and chloramphenicol (12.5 μg/mL). All amino acids were excluded from this medium. For detection of colibactin (**17**) and the hydrate **S34** from *E. coli* Nissle 1917 Δ*clbS* strain, the modified M9-CA medium was prepared with Difco M9 minimal medium powder (10.5 g/L), 0.4% glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, spectinomycin (100 μg/mL), and with the following L-amino acid mass composition (5 g/L total): 3.5% Arg, 20.0% Glu, 2.5% His, 5.0% Ile, 8.0% Leu, 7.0% Lys, 4.5% Phe, 9.5% Pro, 4.0% Thr, 1.0% Trp, 6.0% Tyr, 5% Val, 4% Asn, 4% Ala, 4% Met, 4% Gly, 4% Cys, and 4% Ser. D-[U-^{13}C]-amino acids were supplemented instead of normal amino acids for isotopic labeling experiments. For detection of precolibactin 1489 (**18**) from the *E. coli* DH10B Δ*clbP* S95A strain, the same media compositions were used as for colibactin (**17**) and **S34** described above with a different antibiotic, chloramphenicol (12.5 μg/mL).

Preparation of DNA cross-links from natural colibactin

For each DNA cross-link preparation derived from the BW25113 parent strain, the JW3582-2 cysteine auxotroph, and the JW3973-1 methionine auxotroph, 3200 ng of linearized plasmid DNA was added to 800 μL of modified M9-CA media (containing the appropriate isotopically-labeled amino acid for each auxotroph) and then inoculated with 2.4×10^7 bacteria growing in exponential phase. The DNA–bacteria mixture was incubated for 4.5 hours at 37°C before isolation of the DNA. For each DNA cross-link preparation derived from the JW2880-1 serine auxotroph, 1000 ng of linearized plasmid DNA was added to 250 μL of modified M9-CA media containing either L-[U-^{13}C]-serine or L-[U-^{13}C, ^{15}N]-serine inoculated with 9.0×10^6 bacteria growing in exponential phase. The DNA–bacteria mixture was incubated for a total of 4.5 hours at 37°C before isolation of the DNA. During the incubation, 0.1 μg of appropriately labeled serine was added to the growing culture separately 1 hour and 3 hours after the initial inoculation. Each preparation was repeated in triplicate to accumulate sufficient DNA sample for analysis. For each DNA cross-link derived from the JW2535-2 glycine auxotroph, 1000 ng of linearized plasmid DNA was added to 250 μL of modified M9 minimal medium containing [U-^{13}C]-glycine inoculated with 3.2×10^7 bacteria growing in exponential phase. The final O.D. was adjusted to 0.2. The DNA–bacteria mixture was incubated for a total of 5 hours at 37°C before isolation of the DNA. For each universally labeled DNA cross-link, 1000 ng of linearized plasmid DNA was added to 250 μL of modified M9-glucose media containing D-[U-^{13}C]-glucose, or ^{15}N-ammonia chloride, or a combination of D-[U-^{13}C]-glucose and [^{15}N]-ammonium chloride. Each mixture was separately inoculated with 2.5×10^7 *clb*$^+$ BW25113 parent strain bacteria growing in exponential phase. The DNA–bacteria mixture was incubated for a total of 7 hours at 37°C before isolation of the DNA. To isolate the DNA from the cultures, the bacteria were pelleted by centrifugation. The DNA was isolated from the supernatant using the Monarch® PCR and DNA Cleanup Kit (New England Biolabs) and eluted using ultra purified water (Invitrogen). The isolated DNA was stored at −20°C until further use. To verify the presence of a DNA cross-link, a small quantity of DNA was analyzed by denaturing electrophoresis. To prepare the positive control for cross-linked DNA, 200 ng of linearized pUC19 DNA was treated with 100 μM of cisplatin (Biovision) in 10 mM sodium citrate pH 5 buffer with 5% final DMSO concentration. Cross-linking with cisplatin (generates both intrastrand and interstrand crosslinks) was conducted for 3 hours at 37°C.

Denaturing gel electrophoresis

The concentration of each DNA sample was adjusted to 10 ng/μL using water. 5 μL (50 ng) of the DNA sample was removed and mixed with 15 μL of 0.4% denaturing buffer (0.53% sodium hydroxide, 10% glycerol, 0.013% bromophenol blue) or 1% denaturing buffer (1.33% sodium hydroxide, 10% glycerol, 0.013% bromophenol blue). The DNA was denatured for 10 min at 4°C and then immediately loaded onto a 1% agarose Tris Borate EDTA (TBE) gel. The samples were run in TBE buffer for 1.5 hours at 90 V. The DNA was visualized by staining with Sybr® Gold (Thermo Fisher) for 2 hours.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

Digestion of clb⁺ cross-linked DNA

Following gel verification of the DNA cross-link, 2000 ng of the remaining DNA was digested using the Nucleoside Digestion Mix (New England Biolabs) for 1 hour at 37°C. The digested DNA was stored at −80°C prior to MS analysis.

Preparation of* E. coli *for HCT116 cell infection

The *clb*⁺ *E. coli* BW25113 was inoculated in the modified M9-CA medium and grown at 37°C for 8 hours to reach stationary phase, and then 10 ml of the *E. coli* culture was pelleted by centrifugation. The spent supernatant was removed via aspiration, and the *E. coli* pellet was resuspended into 12 ml of DMEM/F12 medium supplemented with 15 mM HEPES, 10% FBS, and 12.5 μg/ml chloramphenicol. The resuspended cells were pre-warmed at 37°C prior to use.

HCT116 cell infection experiment

The HCT116 cells were grown in T75 flasks to >80% confluence. The cultivation medium was aspirated, followed by a 1× PBS wash (2 × 10 mL). Then the HCT116 cells were infected with 12 ml of pre-warmed *clb*⁺ *E. coli* BW25113 cells for 2 hours at 37°C. After the infection was completed, the HCT116 cells were washed with 1× PBS (2 × 10 mL), trypsinized, and centrifuged at 300 × g for 4 min at room temperature. The supernatant was removed via aspiration, and the remaining HCT116 cell pellet was washed twice with 1.5 mL of 1× PBS with cell recovery at 250 × g for 4 min at room temperature. The cell pellets were then resuspended in 1× PBS for genomic DNA isolation.

Genomic DNA isolation and reprecipitation

The genomic DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. After the DNA was eluted, DNA was reprecipitated to remove the remaining detergent residue from the kit. To reprecipitate the genomic DNA, 90 μL of 1 M sodium chloride was added into 360 μL of eluted genomic DNA, followed by addition of 1050 μL of 100% ethanol. The mixture was briefly vortexed and then incubated at −20°C for 2 hours. The resulting DNA precipitant was pelleted via centrifugation (14,000 × g, 5 min, 4°C). The DNA pellet was further washed using 70% ethanol (1.5 mL × 2) and pelleted via centrifugation (14,000 × g, 5 min, 4°C). The supernatant was removed via aspiration. The post-washed DNA pellet was air dried at room temperature for 30 min and resuspended in water prior to digestion.

DNA digestion

DNA was digested using the Nucleoside Digestion Mix (New England Biolabs) in 1 hour at 37°C. Alternatively, DNA was digested in the step-wise method using NEBuffer 1.1 (10 mM Bis-Tris-Propane-HCl, 10 mM magnesium chlorids, 100 μg/ml BSA, pH 7 New England Biolabs) supplemented with 0.5 mM calcium chloride and 0.5 mM zinc chloride. First 2 units/μg DNA of DNase I (New England Biolabs) was added to the genomic DNA, and the digestion occurred at 37°C for 1 hour. Then 10 units/μg of Nuclease P1 (New England Biolabs) was added to the digestion mix, and the second step digestion lasted at 37°C for 1 hour. Finally, 1 unit/μg DNA of Quick Dephosphorylation Kit (New England Biolabs) was added to the digestion mix, and the third step digestion lasted at 37°C for 30 min.

Sample preparation for colibactin (17) and S34

Single colonies of *E. coli* DH10B *clb*⁻, *E. coli* DH10B *clb*⁺, *E. coli* DH10B Δ*clbO*, and *E. coli* DH10B Δ*clbL*-S179A were individually used to inoculate of 5 mL of LB with chloramphenicol (12.5 μg/mL). After incubation at 37°C with 250 rpm for 20 hours, 25 μL of each seed culture was used to inoculate 5 mL of 3 replicates of 5 mL of production media described above. The cultures were grown at 37°C with 250 rpm to an OD_{600} of 0.4–0.6 and cooled on ice for 10 min before inducing with isopropyl β-D-1-galactopyranoside (IPTG) at a final concentration of 0.2 mM. After cultures were incubated at 25°C with 250 rpm for 42 hours 6 mL of ethyl acetate was added to each culture. The cultures were vortexed for 20 s and separated by centrifugation (3000 rpm × 10 min). The 5 mL of ethyl acetate was transferred and removed in vacuo. The dried extracts were dissolved in 100 μL of methanol for LC-HRMS analysis. Similar sample preparation method was performed from *E. coli* Nissle 1917, *E. coli* Nissle 1917 Δ*clb*, and *E. coli* Nissle 1917 Δ*clbS* strains with some modification. Overnight cultures were prepared with a different antibiotic, spectinomycin (100 μg/mL), for *E. coli* Nissle 1917 Δ*clbS*, or without antibiotic for *E. coli* Nissle 1917 and *E. coli* Nissle 1917 Δ*clb*. 25 μL of each seed culture was used to inoculate 5 mL of 3 replicates of 5 mL of production media. The cultures were grown at 37°C with 250 rpm for 48 hours before LC-HRMS samples were prepared as described above. Samples for isotopic labeling analysis were prepared from *E. coli* Nissle 1917 Δ*clbS* with L-[U-^{13}C]-Ala, Met, Gly, or Cys, and *E. coli* DH10B *clb*⁺ with L-[U-^{13}C]-Gly.

Sample preparation for precolibactin 1489 (18)

The same method of colibactin (**17**) and **S34** was used for sample preparation of precolibactin 1489 (**18**) with a different strain, *E. coli* DH10B Δ*clbP*-S95A.

Dose-dependent cross-linking assay using the synthetic intermediate 38

A sample of **38** was diluted in DMSO such that each reaction consisted of a fixed 5% DMSO final concentration. 200 ng (15.4 μM in base pairs) of linearized pUC19 DNA as prepared above was added into every reaction with a total

Downloaded from http://science.sciencemag.org/ on August 8, 2019

volume of 20 μL. The final concentration of **38** was adjusted to 200 μM, 100 μM, 10 μM, 1 μM, and 100 nM (absolute concentrations of **38** were approximate). 100 μM cisplatin was used as the positive control, and 5% DMSO was used as the negative control. Pure cisplatin (Biovision) stock solutions were diluted into DMSO immediately before use. All reactions were carried out in 10 mM sodium citrate pH 5.0 buffer and incubated for 3 hours at 37°C. The DNA was immediately analyzed by gel electrophoresis after incubation.

pH-dependent cross-linking assay using the synthetic intermediate 38

A sample of **38** was diluted in DMSO such that each reaction consisted of a fixed 5% DMSO final concentration. 200 ng (15.4 μM in base pairs) of linearized pUC19 DNA was added into every reaction with a total volume of 20 μL. The final concentration of **38** was adjusted to 100 μM (absolute concentrations of **38** were approximate). Reactions were conducted using the following buffer conditions with pH ranging from 5.0 to 7.4: 10 mM sodium citrate (pH 5.0), 10 mM sodium acetate (pH 5.5), 10 mM sodium citrate (pH 6.0), 10 mM sodium citrate (pH 6.5), 10 mM sodium phosphate (pH 7.0), and 10 mM sodium phosphate (pH 7.4). 100 μM cisplatin was used as the positive control, and 5% DMSO was used as the negative control. Both of these control reactions were carried out in 10 mM Tris–1 mM EDTA (pH 8.0) buffer. Pure cisplatin (Biovision) stock solutions were diluted into DMSO immediately prior to use. All of the reactions were incubated for 3 hours at 37°C. The DNA was immediately analyzed by gel electrophoresis after incubation.

Preparation of cross-linked DNA using the synthetic intermediate 38

A sample of **38** was diluted in DMSO and water such that the reaction consisted of a final concentration of 0.28% DMSO. 2400 ng (15.4 μM in base pairs) of linearized pUC19 DNA was added into every reaction with a total volume of 240 μL. The final concentration of **38** was adjusted to 25 μM (absolute concentrations of **38** were approximate). The reaction was conducted in 10 mM sodium citrate (pH 5.0) buffer, and incubated for 4 hours at 37°C. The DNA was re-purified from the reaction mix using the Monarch® PCR and DNA Cleanup Kit (New England Biolabs) and eluted using ultra purified water (Invitrogen). The isolated DNA was stored at −20°C, until further use. To prepare the positive control for cross-linked DNA, 200 ng of linearized pUC19 DNA was treated with 100 μM of cisplatin (Biovision) in 10 mM sodium citrate (pH 5.0) buffer with 0.28% final DMSO concentration. Cross-linking with cisplatin was conducted for 4 hours at 37°C. To verify the presence of a DNA cross-link, a small quantity of DNA was analyzed by denaturing electrophoresis.

DNA gel electrophoresis for cross-linking assays

For each DNA sample, the concentration was pre-adjusted to 10 ng/μL. For non-denatured native gels, 4 μL (40 ng) of DNA was taken out and mixed with 1.5 μL of 6× purple gel loading dye, no SDS (NEB). The mixed DNA samples were immediately loaded onto 1% agarose Tris Borate EDTA (TBE) gels, and the gel was run for 1.5 hours at 90 V. The gel was post stained with SybrGold (Thermo Fisher) for 2 hours. For denaturing gels, 5 μL (50 ng) of DNA was taken out each time and separately mixed with 15 μL of 0.2% denaturing buffer (0.27% sodium hydroxide, 10% glycerol, 0.013% bromophenol blue), 0.4% denaturing buffer (0.53% sodium hydroxide, 10% glycerol, 0.013% bromophenol blue), or 1% denaturing buffer (1.33% sodium hydroxide, 10% glycerol, 0.013% bromophenol blue) at 0°C. The mixed DNA samples were denatured at 4°C for 10 min and immediately loaded onto 1% agarose Tris Borate EDTA (TBE) gels. The gel was run for 1.5 hours at 90 V. The gel was post stained with SybrGold (Thermo Fisher) for 2 hours.

Digestion of DNA cross-linked by 38

Following gel verification of the DNA cross-link, 2000 ng of the remaining DNA was digested using the Nucleoside Digestion Mix (New England Biolabs) for 1 hour at 37°C. The digested DNA was stored at −80°C prior to MS analysis.

General synthetic procedures

All reactions were performed in single-neck, flame-dried, round-bottomed flasks fitted with rubber septa under a positive pressure of nitrogen unless otherwise noted. Air- and moisture-sensitive liquids were transferred via syringe or stainless steel cannula, or were handled in a nitrogen-filled drybox (working oxygen level <10 ppm). Organic solutions were concentrated by rotary evaporation at 28–32°C. Flash-column chromatography was performed as described by Still *et al.* (*45*), employing silica gel (60 Å, 40–63 μm particle size) purchased from Sorbent Technologies (Atlanta, GA). Analytical thin-layered chromatography (TLC) was performed using glass plates pre-coated with silica gel (0.25 mm, 60 Å pore size) impregnated with a fluorescent indicator (254 nm). TLC plates were visualized by exposure to ultraviolet light (UV).

Materials

Commercial solvents and reagents were used as received with the following exceptions. Dichloromethane, ether, and *N*,*N*-dimethylformamide were purified according to the method of Pangborn *et al.* (*46*) Triethylamine was distilled from calcium hydride under an atmosphere of argon immediately before use. Di-*iso*-propylamine was distilled from calcium hydride and was stored under nitrogen. Methanol was distilled from magnesium turnings under an atmosphere of nitrogen immediately before use. Tetrahydrofuran was

Downloaded from http://science.sciencemag.org/ on August 8, 2019

distilled from sodium–benzophenone under an atmosphere of nitrogen immediately before use. Commercial solutions of lithium di-*iso*-propyl amide in tetrahydrofuran–heptane–ethylbenzene were titrated by a variation of the procedure of Ireland and Meissner (*47*) using menthol and 1,10-phenanthroline. The β-ketothioester **27** was prepared according to a published procedure (*17*).

REFERENCES AND NOTES

1. N. K. Surana, D. L. Kasper, Moving beyond microbiome-wide associations to causal microbe identification. *Nature* **552**, 244–247 (2017). doi:10.1038/nature25019 Medline
2. M. A. Fischbach, Microbiome: Focus on causation and mechanism. *Cell* **174**, 785–790 (2018). doi:10.1016/j.cell.2018.07.038 Medline
3. J. R. Johnson, B. Johnston, M. A. Kuskowski, J. P. Nougayrede, E. Oswald, Molecular epidemiology and phylogenetic distribution of the Escherichia coli pks genomic island. *J. Clin. Microbiol.* **46**, 3906–3911 (2008). doi:10.1128/JCM.00949-08 Medline
4. J. Putze, C. Hennequin, J.-P. Nougayrède, W. Zhang, S. Homburg, H. Karch, M.-A. Bringer, C. Fayolle, E. Carniel, W. Rabsch, T. A. Oelschlaeger, E. Oswald, C. Forestier, J. Hacker, U. Dobrindt, Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infect. Immun.* **77**, 4696–4703 (2009). doi:10.1128/IAI.00522-09 Medline
5. J.-P. Nougayrède, S. Homburg, F. Taieb, M. Boury, E. Brzuszkiewicz, G. Gottschalk, C. Buchrieser, J. Hacker, U. Dobrindt, E. Oswald, *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848–851 (2006). doi:10.1126/science.1127059 Medline
6. G. Cuevas-Ramos, C. R. Petit, I. Marcq, M. Boury, E. Oswald, J.-P. Nougayrède, *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11537–11542 (2010). doi:10.1073/pnas.1001261107 Medline
7. J. C. Arthur, E. Perez-Chanona, M. Mühlbauer, S. Tomkovich, J. M. Uronis, T.-J. Fan, B. J. Campbell, T. Abujamel, B. Dogan, A. B. Rogers, J. M. Rhodes, A. Stintzi, K. W. Simpson, J. J. Hansen, T. O. Keku, A. A. Fodor, C. Jobin, Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012). doi:10.1126/science.1224820 Medline
8. A. Cougnoux, G. Dalmasso, R. Martinez, E. Buc, J. Delmas, L. Gibold, P. Sauvanet, C. Darcha, P. Déchelotte, M. Bonnet, D. Pezet, H. Wodrich, A. Darfeuille-Michaud, R. Bonnet, Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932–1942 (2014). doi:10.1136/gutjnl-2013-305257 Medline
9. S. Tomkovich, Y. Yang, K. Winglee, J. Gauthier, M. Mühlbauer, X. Sun, M. Mohamadzadeh, X. Liu, P. Martin, G. P. Wang, E. Oswald, A. A. Fodor, C. Jobin, Locoregional effects of microbiota in a preclinical model of colon carcinogenesis. *Cancer Res.* **77**, 2620–2632 (2017). doi:10.1158/0008-5472.CAN-16-3472 Medline
10. E. Buc, D. Dubois, P. Sauvanet, J. Raisch, J. Delmas, A. Darfeuille-Michaud, D. Pezet, R. Bonnet, High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. *PLOS ONE* **8**, e56964 (2013). doi:10.1371/journal.pone.0056964 Medline
11. E. P. Trautman, J. M. Crawford, Linking biosynthetic gene clusters to their metabolites via pathway-targeted molecular networking. *Curr. Top. Med. Chem.* **16**, 1705–1716 (2016). Medline
12. E. P. Balskus, Colibactin: Understanding an elusive gut bacterial genotoxin. *Nat. Prod. Rep.* **32**, 1534–1540 (2015). doi:10.1039/C5NP00091B Medline
13. F. Taieb, C. Petit, J. P. Nougayrède, E. Oswald, The enterobacterial genotoxins: Cytolethal distending toxin and colibactin. *Ecosal Plus* **7**, (2016). doi:10.1128/ecosalplus.ESP-0008-2016 Medline
14. A. R. Healy, S. B. Herzon, Molecular basis of gut microbiome-associated colorectal cancer: A synthetic perspective. *J. Am. Chem. Soc.* **139**, 14817–14824 (2017). doi:10.1021/jacs.7b07807 Medline
15. T. Faïs, J. Delmas, N. Barnich, R. Bonnet, G. Dalmasso, Colibactin: More than a new bacterial toxin. *Toxins (Basel)* **10**, 151 (2018). doi:10.3390/toxins10040151 Medline
16. M. I. Vizcaino, J. M. Crawford, The colibactin warhead crosslinks DNA. *Nat. Chem.* **7**, 411–417 (2015). doi:10.1038/nchem.2221 Medline
17. A. R. Healy, H. Nikolayevskiy, J. R. Patel, J. M. Crawford, S. B. Herzon, A mechanistic model for colibactin-induced genotoxicity. *J. Am. Chem. Soc.* **138**, 15563–15570 (2016). doi:10.1021/jacs.6b10354 Medline
18. C. A. Brotherton, E. P. Balskus, A prodrug resistance mechanism is involved in colibactin biosynthesis and cytotoxicity. *J. Am. Chem. Soc.* **135**, 3359–3362 (2013). doi:10.1021/ja312154m Medline
19. X. Bian, J. Fu, A. Plaza, J. Herrmann, D. Pistorius, A. F. Stewart, Y. Zhang, R. Müller, In vivo evidence for a prodrug activation mechanism during colibactin maturation. *ChemBioChem* **14**, 1194–1197 (2013). doi:10.1002/cbic.201300208 Medline
20. M. I. Vizcaino, P. Engel, E. Trautman, J. M. Crawford, Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *J. Am. Chem. Soc.* **136**, 9244–9247 (2014). doi:10.1021/ja503450q Medline
21. X. Bian, A. Plaza, Y. Zhang, R. Müller, Two more pieces of the colibactin genotoxin puzzle from *Escherichia coli* show incorporation of an unusual 1-aminocyclopropanecarboxylic acid moiety. *Chem. Sci.* **6**, 3154–3160 (2015). doi:10.1039/C5SC00101C Medline
22. C. A. Brotherton, M. Wilson, G. Byrd, E. P. Balskus, Isolation of a metabolite from the *pks* island provides insights into colibactin biosynthesis and activity. *Org. Lett.* **17**, 1545–1548 (2015). doi:10.1021/acs.orglett.5b00432 Medline
23. D. Dubois, O. Baron, A. Cougnoux, J. Delmas, N. Pradel, M. Boury, B. Bouchon, M.-A. Bringer, J.-P. Nougayrède, E. Oswald, R. Bonnet, ClbP is a prototype of a peptidase subgroup involved in biosynthesis of nonribosomal peptides. *J. Biol. Chem.* **286**, 35562–35570 (2011). doi:10.1074/jbc.M111.221960 Medline
24. A. Cougnoux, L. Gibold, F. Robin, D. Dubois, N. Pradel, A. Darfeuille-Michaud, G. Dalmasso, J. Delmas, R. Bonnet, Analysis of structure-function relationships in the colibactin-maturating enzyme ClbP. *J. Mol. Biol.* **424**, 203–214 (2012). doi:10.1016/j.jmb.2012.09.017 Medline
25. E. P. Trautman, A. R. Healy, E. E. Shine, S. B. Herzon, J. M. Crawford, Domain-targeted metabolomics delineates the heterocycle assembly steps of colibactin biosynthesis. *J. Am. Chem. Soc.* **139**, 4195–4201 (2017). doi:10.1021/jacs.7b00659 Medline
26. M. Xue, E. Shine, W. Wang, J. M. Crawford, S. B. Herzon, Characterization of natural colibactin-nucleobase adducts by tandem mass spectrometry and isotopic labeling. Support for DNA alkylation by cyclopropane ring opening. *Biochemistry* **57**, 6391–6394 (2018). doi:10.1021/acs.biochem.8b01023 Medline
27. M. R. Wilson, Y. Jiang, P. W. Villalta, A. Stornetta, P. D. Boudreau, A. Carrá, C. A. Brennan, E. Chun, L. Ngo, L. D. Samson, B. P. Engelward, W. S. Garrett, S. Balbo, E. P. Balskus, The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019). doi:10.1126/science.aar7785 Medline
28. N. Bossuet-Greif, J. Vignard, F. Taieb, G. Mirey, D. Dubois, C. Petit, E. Oswald, J.-P. Nougayrède, The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* **9**, e02393-17 (2018). doi:10.1128/mBio.02393-17 Medline
29. Z. R. Li, Y. Li, J. Y. H. Lai, J. Tang, B. Wang, L. Lu, G. Zhu, X. Wu, Y. Xu, P.-Y. Qian, Critical intermediates reveal new biosynthetic events in the enigmatic colibactin pathway. *ChemBioChem* **16**, 1715–1719 (2015). doi:10.1002/cbic.201500239 Medline
30. A. R. Healy, M. I. Vizcaino, J. M. Crawford, S. B. Herzon, Convergent and modular synthesis of candidate precolibactins. Structural revision of precolibactin A. *J. Am. Chem. Soc.* **138**, 5426–5432 (2016). doi:10.1021/jacs.6b02276 Medline
31. Z. R. Li, J. Li, J.-P. Gu, J. Y. H. Lai, B. M. Duggan, W.-P. Zhang, Z.-L. Li, Y.-X. Li, R.-B. Tong, Y. Xu, D.-H. Lin, B. S. Moore, P.-Y. Qian, Divergent biosynthesis yields a cytotoxic aminomalonate-containing precolibactin. *Nat. Chem. Biol.* **12**, 773–775 (2016). doi:10.1038/nchembio.2157 Medline
32. L. Zha, M. R. Wilson, C. A. Brotherton, E. P. Balskus, Characterization of polyketide synthase machinery from the *pks* island facilitates isolation of a candidate precolibactin. *ACS Chem. Biol.* **11**, 1287–1295 (2016). doi:10.1021/acschembio.6b00014 Medline
33. A. R. Healy *et al.*, Synthesis and reactivity of precolibactin 886. chemRxiv (2019); https://chemrxiv.org/articles/Synthesis_and_Reactivity_of_Precolibactin_886/7849151.
34. Z.-R. Li *et al.*, Macrocyclic colibactin induces DNA double-strand breaks via copper-mediated oxidative cleavage. bioRxiv 530204 (2019).

Downloaded from http://science.sciencemag.org/ on August 8, 2019

35. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, H. Mori, Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* **2**, 0008 (2006). doi:10.1038/msb4100050 Medline
36. L. Zha, Y. Jiang, M. T. Henke, M. R. Wilson, J. X. Wang, N. L. Kelleher, E. P. Balskus, Colibactin assembly line enzymes use *S*-adenosylmethionine to build a cyclopropane ring. *Nat. Chem. Biol.* **13**, 1063–1065 (2017). doi:10.1038/nchembio.2448 Medline
37. A. O. Brachmann, C. Garcie, V. Wu, P. Martin, R. Ueoka, E. Oswald, J. Piel, Colibactin biosynthesis and biological activity depend on the rare aminomalonyl polyketide precursor. *Chem. Commun.* **51**, 13138–13141 (2015). doi:10.1039/C5CC02718G Medline
38. N. S. Guntaka, A. R. Healy, J. M. Crawford, S. B. Herzon, S. D. Bruner, Structure and functional analysis of ClbQ, an unusual intermediate-releasing thioesterase from the colibactin biosynthetic pathway. *ACS Chem. Biol.* **12**, 2598–2608 (2017). doi:10.1021/acschembio.7b00479 Medline
39. Y. Jiang, A. Stornetta, P. W. Villalta, M. R. Wilson, P. D. Boudreau, L. Zha, S. Balbo, E. P. Balskus, The reactivity of an unusual amidase may explain colibactin's DNA cross-linking activity. *J. Chem. Soc.* **141**, 11489–11496 (2019). doi:10.1021/jacs.9b02453 Medline
40. Y. Jiang, A. Stornetta, P. W. Villalta, M. R. Wilson, P. D. Boudreau, L. Zha, S. Balbo, E. P. Balskus, Reactivity of an unusual amidase may explain colibactin's DNA cross-linking activity. *J. Am. Chem. Soc.* **141**, 11489–11496 (2019). doi:10.1021/jacs.9b02453 Medline
41. E. E. Shine, M. Xue, J. R. Patel, A. R. Healy, Y. V. Surovtseva, S. B. Herzon, J. M. Crawford, Model colibactins exhibit human cell genotoxicity in the absence of host bacteria. *ACS Chem. Biol.* **13**, 3286–3293 (2018). doi:10.1021/acschembio.8b00714 Medline
42. P. Tripathi, E. E. Shine, A. R. Healy, C. S. Kim, S. B. Herzon, S. D. Bruner, J. M. Crawford, ClbS is a cyclopropane hydrolase that confers colibactin resistance. *J. Am. Chem. Soc.* **139**, 17719–17722 (2017). doi:10.1021/jacs.7b09971 Medline
43. R. P. Bell, in *Advances in Physical Organic Chemistry*, V. Gold, Ed. (Academic Press, 1966), vol. 4, pp. 1–29.
44. N. Bossuet-Greif, D. Dubois, C. Petit, S. Tronnet, P. Martin, R. Bonnet, E. Oswald, J.-P. Nougayrède, *Escherichia coli* ClbS is a colibactin resistance protein. *Mol. Microbiol.* **99**, 897–908 (2016). doi:10.1111/mmi.13272 Medline
45. W. C. Still, M. Kahn, A. Mitra, Rapid chromatographic technique for preparative separations with moderate resolutions. *J. Org. Chem.* **43**, 2923–2925 (1978). doi:10.1021/jo00408a041
46. A. B. Pangborn, M. A. Giardello, R. H. Grubbs, R. K. Rosen, F. J. Timmers, Safe and convenient procedure for solvent purification. *Organometallics* **15**, 1518–1520 (1996). doi:10.1021/om9503712
47. R. E. Ireland, R. S. Meissner, Convenient method for the titration of amide base solutions. *J. Org. Chem.* **56**, 4566–4568 (1991). doi:10.1021/jo00014a050

ACKNOWLEDGMENTS

Funding: Financial support from the National Institutes of Health (R01GM110506 to S.B.H., 1DP2-CA186575 to J.M.C., R01CA215553 to S.B.H. and J.M.C.), the Chemistry Biology Interface Training Program (T32GM067543 to K.M.W.), the Charles H. Revson foundation (postdoctoral fellowship to A.R.H.), the NSF graduate research fellowship program (E.E.S), and Yale University is gratefully acknowledged. The structure of colibactin (**17**) was first disclosed by S.B.H. on 12/3/18 at the Merck Lecture, Department of Chemistry, The University of Illinois, Urbana–Champaign, by J.M.C. on 1/8/19 at a Departmental Symposium, Department of Chemistry, Massachusetts Institute of Technology, and by M.Z. on 3/11/19 during a lecture entitled: Characterization of Natural Colibactin Nucleobase Adducts by Tandem MS and Isotopic Labeling, at the Keystone Symposia meeting "Microbiome: Chemical Mechanisms and Biological Consequences"; **Author contributions:** M.X. discovered and characterized the natural colibactin–diadenine adduct **9**, conducted tandem MS analysis of synthetic colibactin–DNA adducts, carried out bacterial infection studies, and identified the colibactin–adenine adducts **9**, **S1**, **4**, and **14** in genomic DNA; C.S.K. characterized natural colibactin (**17**) and precolibactin 1489 (**18**) in bacterial extracts; A.R.H. contributed to the conception of the synthesis, conducted preliminary synthetic studies, and suggested protection of the fragment coupling product **24** as its enoxysilane; K.M.W. conceived the two-fold coupling approach to colibactin, developed a synthesis of the β-ketothioeseter **25**, and optimized the synthetic route; Z.W. optimized the synthetic route and completed the synthesis of colibactin; M.C.F. developed a scalable synthetic route to the α-nitroketone **33**; E.E.S. generated new strains, contributed to the bacterial infection studies, and developed the *clbS* mutant strategy to enhance detection of natural colibactin; W.W. assisted with tandem MS analysis of DNA-colibactin adducts. S.B.H. and J.M.C. conceived the study, oversaw experiments, and wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data availability:** All data to support the conclusions of this manuscript are included in the main text or Supplementary Materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/cgi/content/full/science.aax2685/DC1
Supplementary Text
Figs. S1 to S156
Tables S1 to S78
NMR spectra

7 March 2019; accepted 24 July 2019
Published online 8 August 2019
10.1126/science.aax2685

Downloaded from http://science.sciencemag.org/ on August 8, 2019

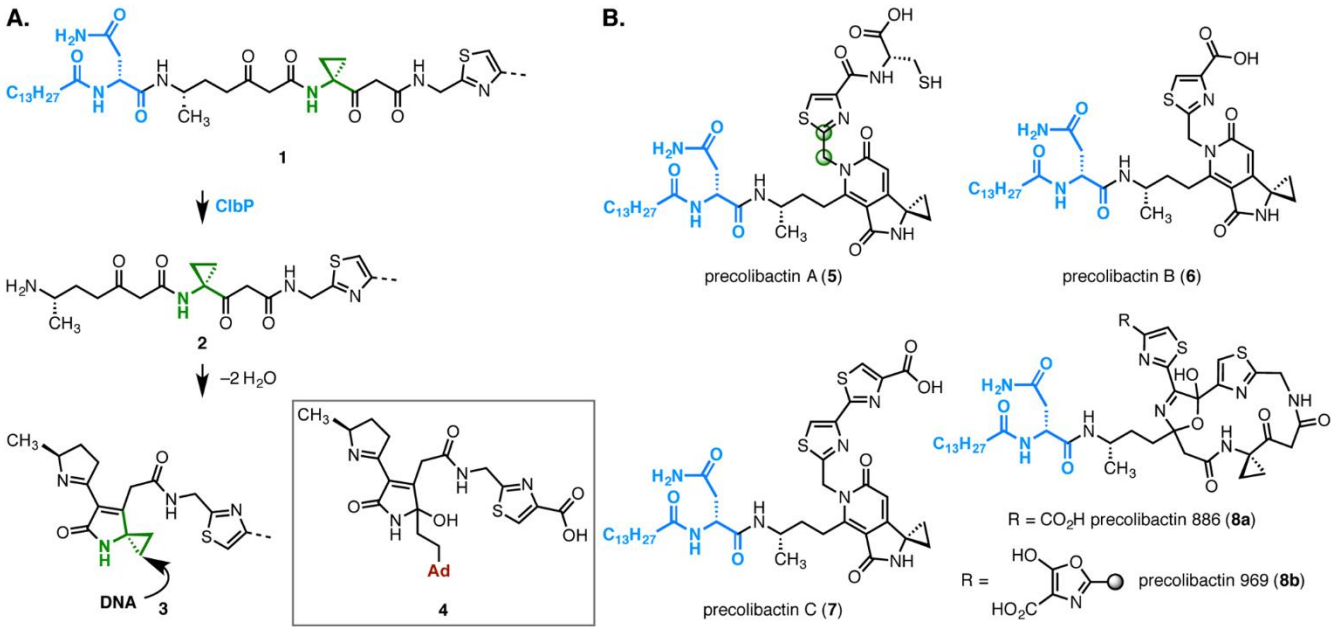

Fig. 1. Structures and reaction pathways of selected *clb* biosynthetic products. (A) Established mechanism of DNA mono-alkylation by *clb* metabolites formed in wild-type cultures. (B) Structures of *clb* metabolites formed in Δ*clbP clb*$^{+}$ *E. coli* cultures. The green spheres in structure **5** denote the carbon atoms derived from glycine.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

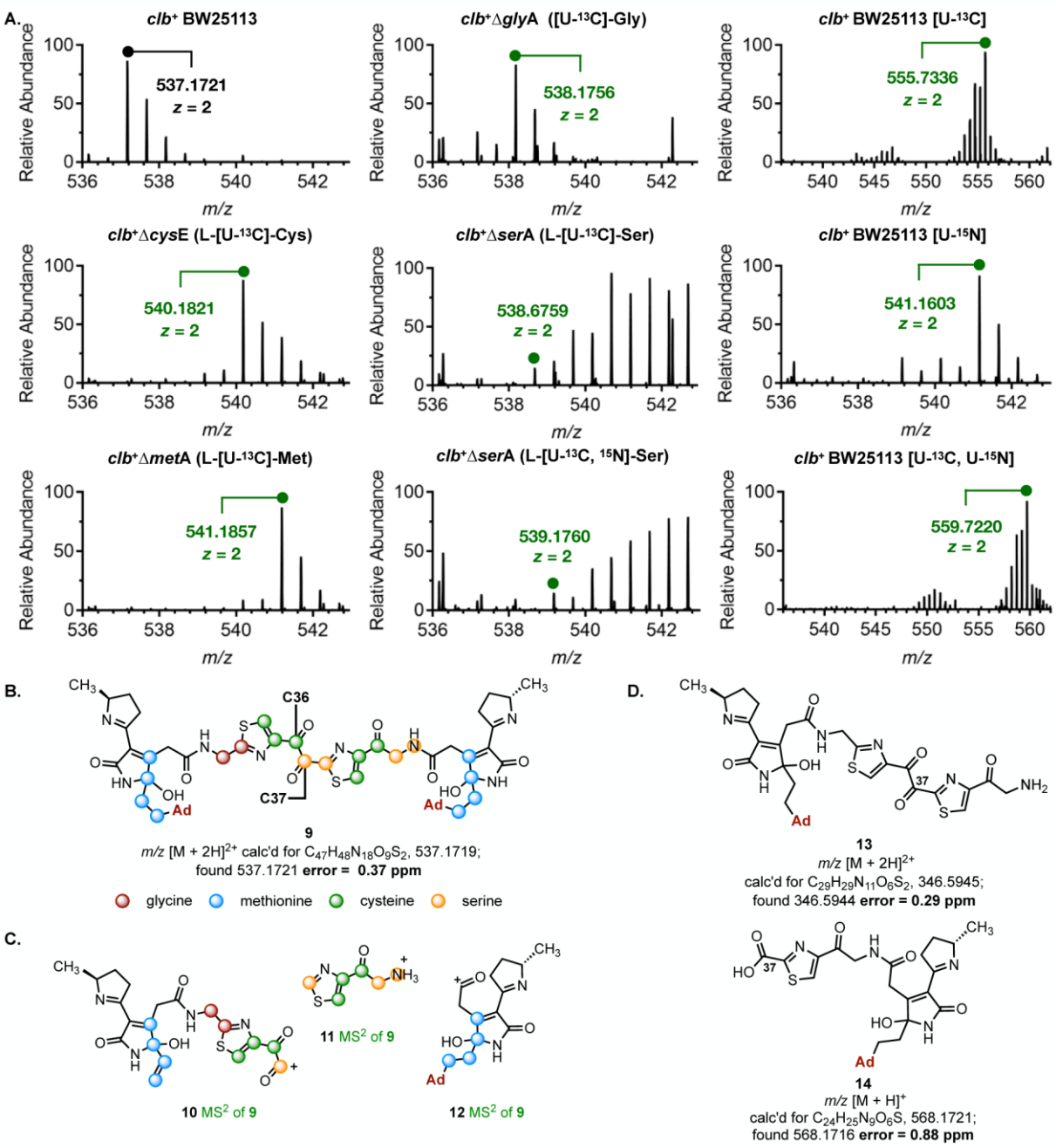

Fig. 2. Selected HRMS signals deriving from treatment of linearized pUC19 DNA with *clb*$^+$ *E. coli*, followed by digestion. (A) Natural abundance and stable isotope derivatives of **9**. The highest intensity labeled peaks (green) were selected for analysis except for Ser, which was extensively metabolized. All selected ions were confirmed by tandem MS. $[M+2H]^{2+}$ ions are marked. (B) Structure of the colibactin–bis(adenine) adduct **9**. (C) Structures of the daughter ions **10**–**12**. D. The novel DNA adducts **13** and **14**.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

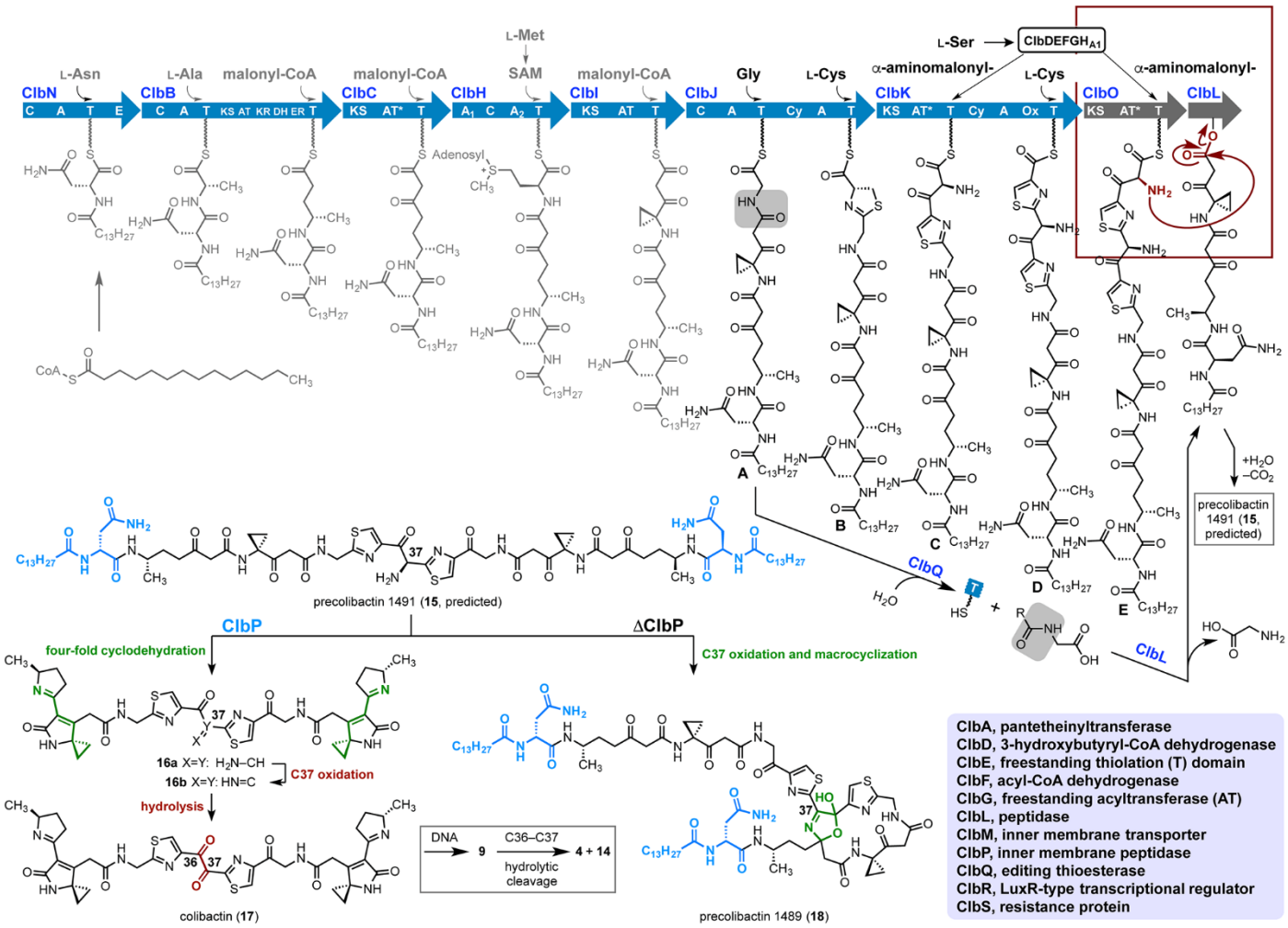

Fig. 3. Proposed biosynthesis of (pre)colibactin. The early stages in the biosynthetic pathway are greyed for clarity. The heterodimerization is highlighted in the red box (upper right). Intermediates B–E are also possible substrates for thioesterase ClbQ, although promiscuous ClbQ has a known preference for hydrolyzing intermediates toward the middle of the assembly line. Amino acids are depicted at their sites of pathway entry. Domain abbreviations: C, condensation; A, adenylation; E, epimerization; KS, ketosynthase; KR, ketoreductase; DH, dehydratase; ER, enoylreductase; AT*, inactivated acyltransferase (AT); Cy, dual condensation/cyclization; Ox, oxidase.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

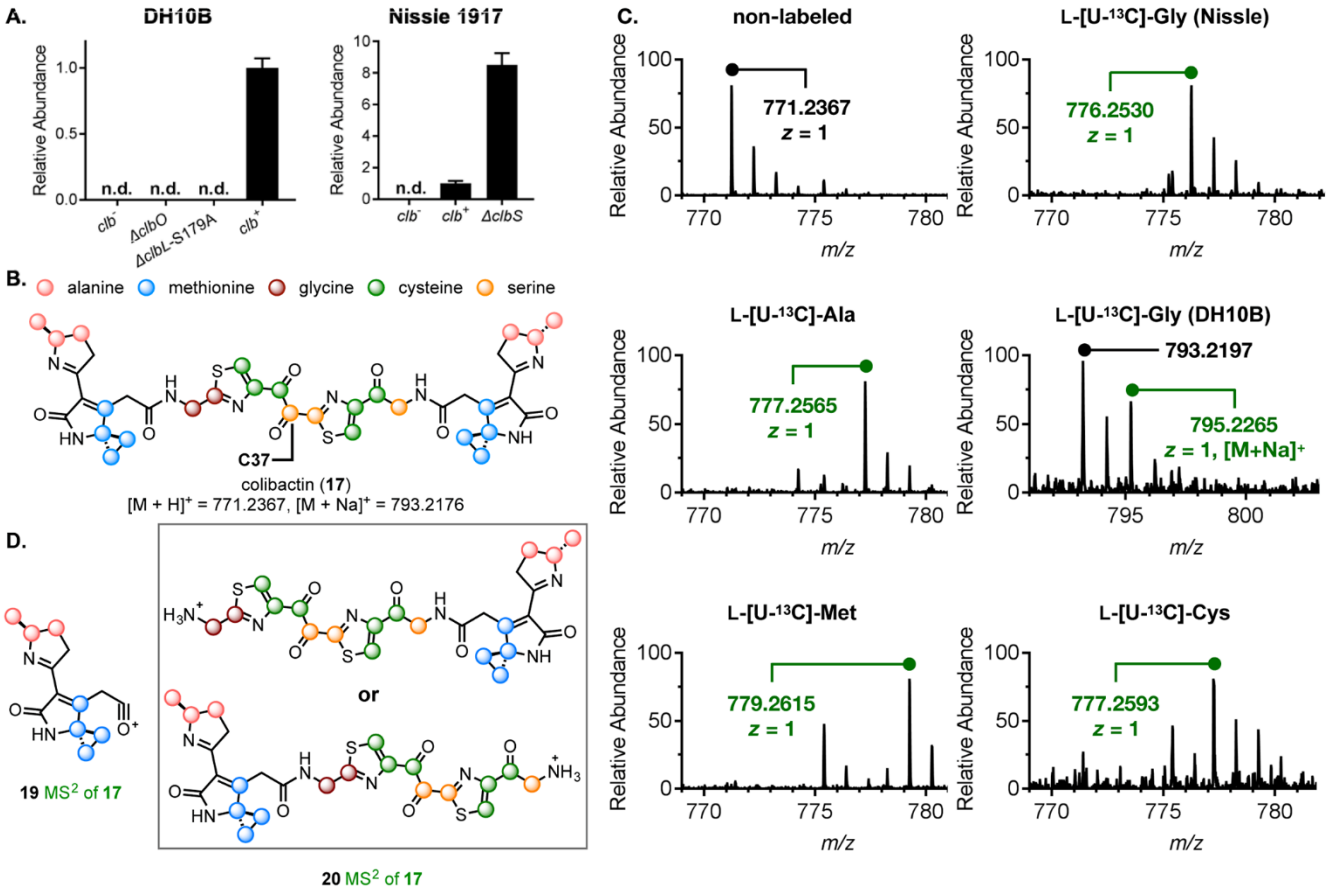

Fig. 4. Stimulation, genetic dependence, and isotopic labeling of natural colibactin (17). (A) Genetic dependence of colibactin (17) production in *clb*⁺ DH10B and Nissle 1917. n = 3 biological replicates; error represents standard deviation. n. d., not detected. (B) Isotopic labeling pattern of colibactin (17). (C) Results of isotopic labeling studies of colibactin (17) in Nissle 1917 (Δ*clbS*). [U-^{13}C]-Gly labeling was conducted in both Nissle 1917 (Δ*clbS*) and *clb*⁺ DH10B. The highest intensity labeled peaks (green) were selected for analysis. $[M+H]^+$ ions are marked unless otherwise noted. (D) Ions observed in the tandem MS of colibactin (17). The two structures of ion 20 are equally plausible based on the MS data.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

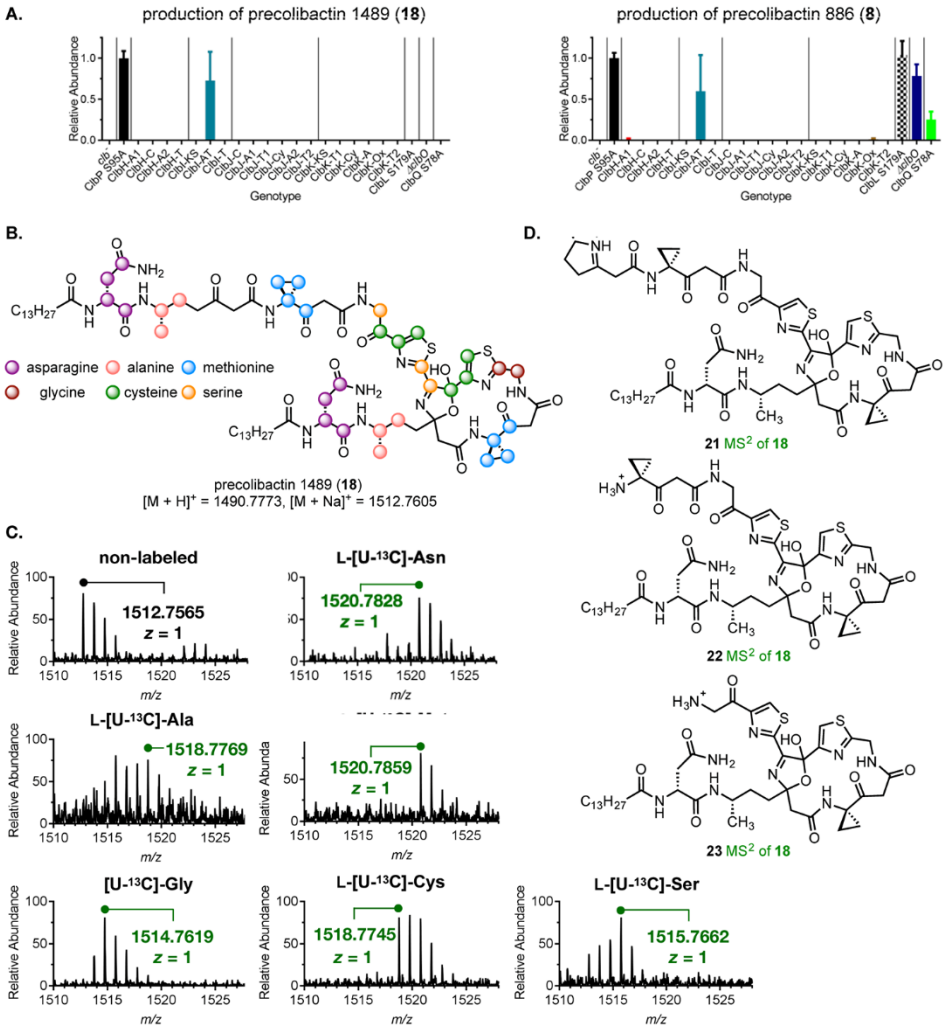

Fig. 5. Genetic, tandem MS, and isotopic labeling support for precolibactin 1489 (18). (A) Precolibactin 1489 (18) biosynthesis requires all biosynthetic enzymes in the *clb* gene cluster. Precolibactin 886 (8a) is still produced in *clbL*, *clbO*, or *clbQ* mutants. n = 5 biological replicates; error represents standard deviation. (B) Position of isotopic labels in precolibactin 1489 (18), as established by tandem MS analysis. (C) Isotopic labeling studies of precolibactin 1489 (18) in a *clb*$^+$ DH10B strain deficient in ClbP catalytic activity. (D) Proposed structures of ions 21–23 (fig. S125C) derived from the tandem MS of precolibactin 1489 (18).

Downloaded from http://science.sciencemag.org/ on August 8, 2019

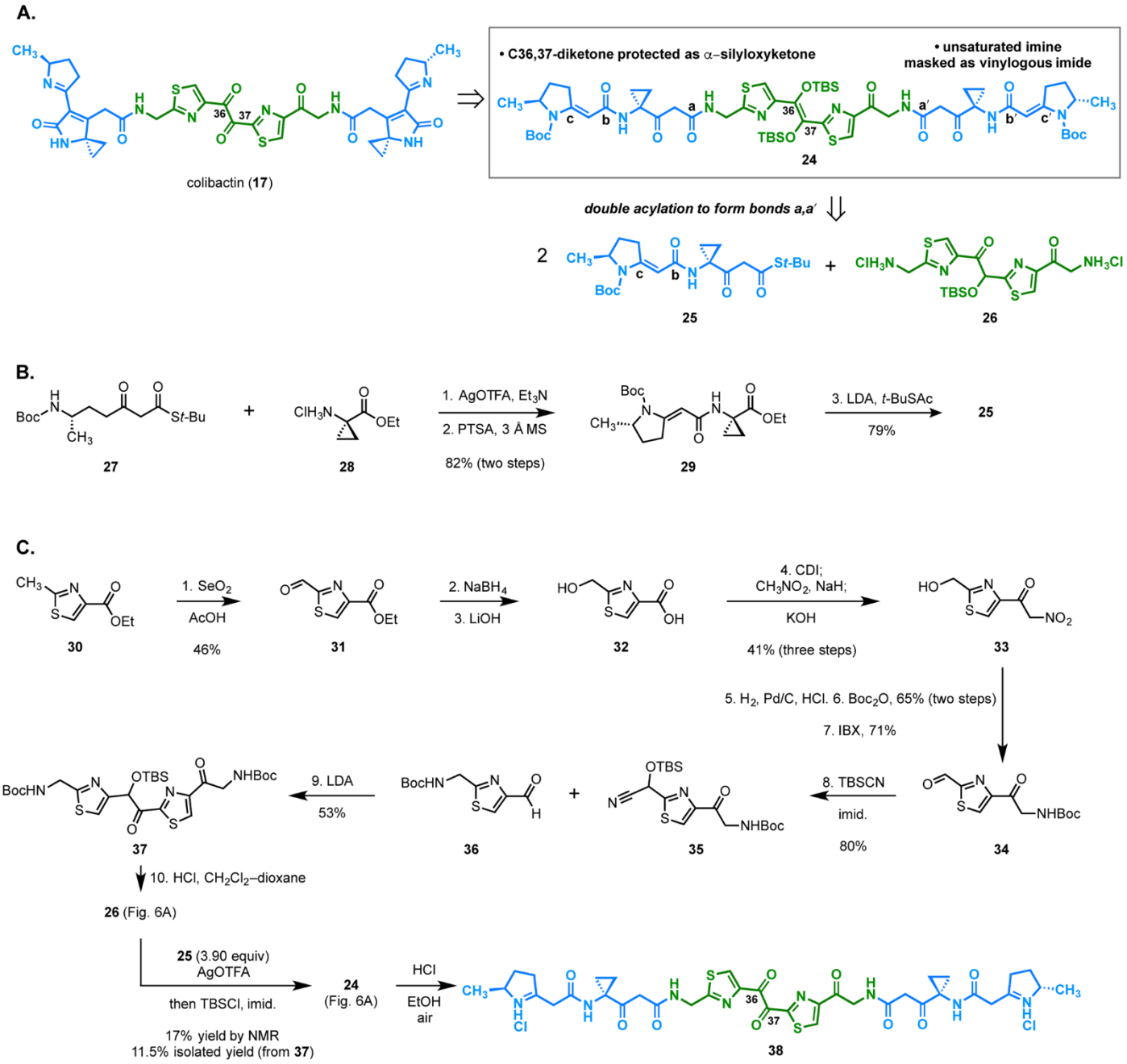

Fig. 6. Synthesis of the colibactin precursor 38.

Downloaded from http://science.sciencemag.org/ on August 8, 2019

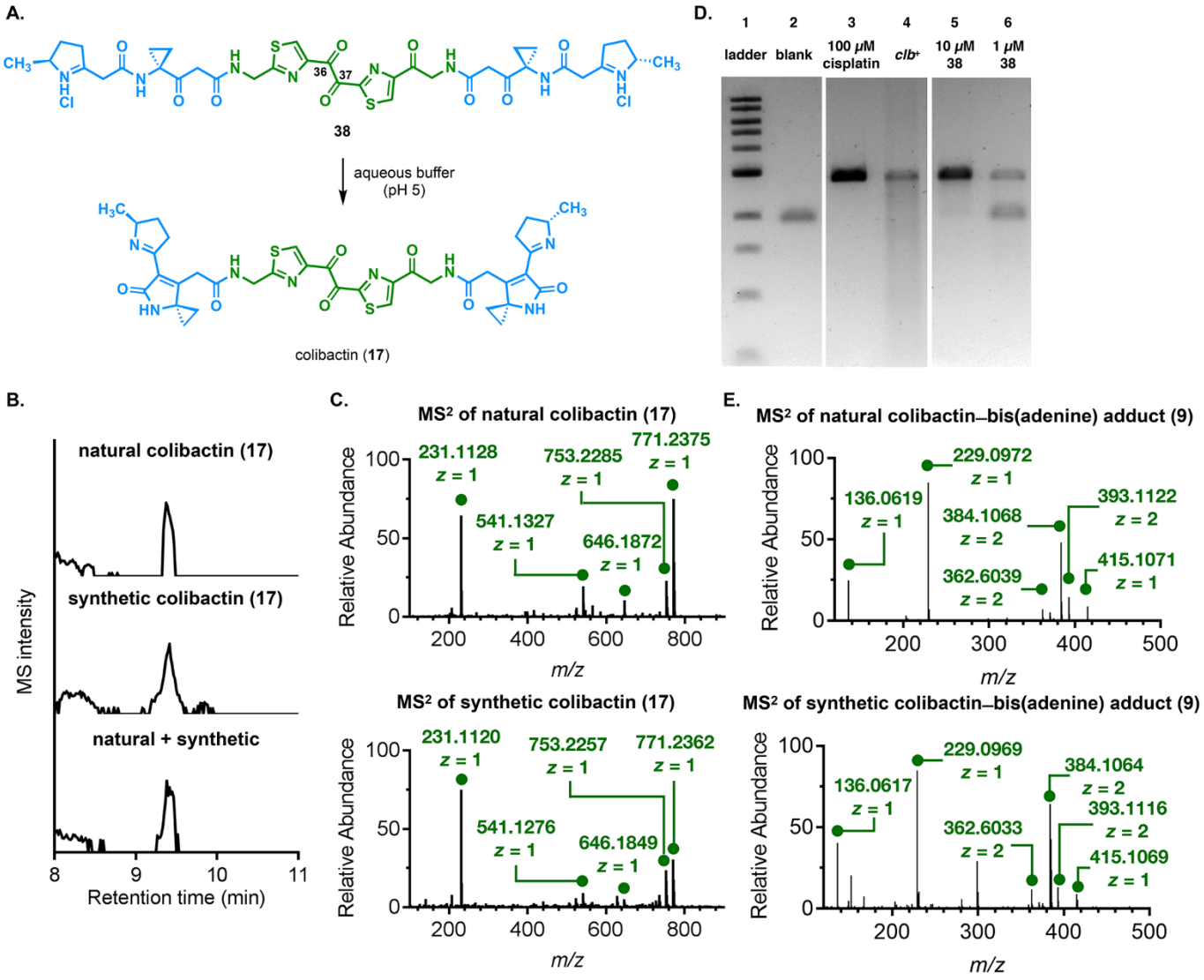

Fig. 7. Confirmation of the predicted structure of colibactin (17). (A) Cyclization of intermediate 38 to colibactin (17). (B) LC-MS coinjection analysis of colibactin (17): natural (top), synthetic (middle), and coinjection (bottom). (C) Tandem MS data of natural colibactin (17, top) and synthetic colibactin (17, bottom). Collision energy = 30 eV. For additional data see fig. S127. (D) DNA crosslinking assay employing linearized pUC19 DNA and synthetic intermediate 38. (E) Tandem MS data of the bis(adenine) adduct 9 derived from natural and synthetic colibactin (17).

Downloaded from http://science.sciencemag.org/ on August 8, 2019

Science

Structure elucidation of colibactin and its DNA cross-links

Mengzhao Xue, Chung Sub Kim, Alan R. Healy, Kevin M. Wernke, Zhixun Wang, Madeline C. Frischling, Emilee E. Shine, Weiwei Wang, Seth B. Herzon and Jason M. Crawford

published online August 8, 2019originally published online August 8, 2019

ARTICLE TOOLS http://science.sciencemag.org/content/early/2019/08/08/science.aax2685

SUPPLEMENTARY MATERIALS http://science.sciencemag.org/content/suppl/2019/08/07/science.aax2685.DC1

REFERENCES This article cites 44 articles, 10 of which you can access for free
http://science.sciencemag.org/content/early/2019/08/08/science.aax2685#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science* is a registered trademark of AAAS.

Downloaded from http://science.sciencemag.org/ on August 8, 2019