ACCOUNTS of chemical research

pubs.acs.org/accounts Article

Leveraging Microbial Genomes and Genomic Context for Chemical Discovery

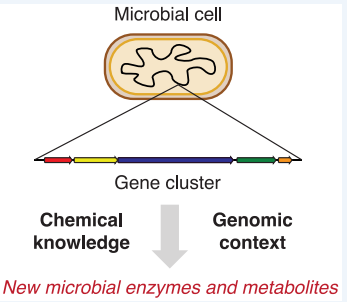
Published as part of the Accounts of Chemical Research special issue "Bacterial Multi-Drug Resistance".

Duncan J. Kountz and Emily P. Balskus*

Cite This: *Acc. Chem. Res.* 2021, 54, 2788−2797 Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: The genomic era has dramatically changed how we discover and investigate microbial biochemistry. In particular, the exponential expansion in the number of sequenced microbial genomes provides investigators with a vast wealth of sequence data to exploit for the discovery of biochemical functions and mechanisms, as well as novel enzymes and metabolites. In contrast to early biochemical work, which was largely characterized by "forward" approaches that proceed from biomass to enzyme to gene, the availability of genome sequences enables the discovery of new microbial metabolic activities, enzymes, and metabolites by "reverse" approaches that originate with genetic information or by approaches that incorporate features of both forward and reverse methodologies. In the genomic era, the canonical organization of microbial genomes into gene clusters presents a singular opportunity for the utilization of genomic data. Specifically, genomic context (information gleaned from the genes surrounding a gene of interest in the chromosome) is a powerful tool for chemical discovery in microbial systems because of the functional and/or physiological relationship that usually exists between genes found within a gene cluster. This means that the investigator can use this inferred link to generate hypotheses about the functions of individual genes in the cluster or even the function of the entire cluster itself. Here, we discuss how analysis of genomic context in combination with a mechanistic understanding of enzymes can facilitate numerous facets of microbial biochemical research including the identification of biosynthetic gene clusters, the discovery of important and novel enzymes, the elucidation of natural product structures, and the identification of new metabolic pathways. We highlight work from our laboratory using genomic context to discover and study biosynthetic pathways that produce natural products, including the cylindrocyclophanes, nitrogen−nitrogen bond-containing metabolites, and the gut microbial genotoxin colibactin. Although use of genomic context is most commonly associated with studies of natural product biosynthesis, we also show that it can be applied to the study of primary metabolism. We illustrate this with examples from our work studying the members of the glycyl radical enzyme superfamily involved in choline and 4-hydroxyproline degradation in the human gut. Looking forward, we envision increased opportunities to use such information, with the combination of biochemical knowledge and computational tools poised to fuel a new revolution in our ability to connect genes and their biochemical functions. In particular, we note a need for methods that computationally formalize the functional association between genes when such associations are not obvious from manual gene annotations. Such tools will drastically augment the feasibility and scope of gene cluster analysis and accelerate the discovery of new microbial enzymes, metabolites, and metabolic processes.



KEY REFERENCES

• Ng, T. L.; Rohac, R.; Mitchell, A. J.; Boal, A. K.; Balskus, E. P. An *N*-Nitrosating Metalloenzyme Constructs the Pharmacophore of Streptozotocin. *Nature* **2019**, *566*, 94−99. 10.1038/s41586-019-0894-z.[1] *This paper describes the identification of the biosynthetic gene cluster for the clinically utilized antitumor compound streptozotocin using resistance-guided genome mining.*

• Nakamura, H.; Schultz, E. E.; Balskus, E. P. A New Strategy for Aromatic Ring Alkylation in Cylindrocyclophane Biosynthesis. *Nat. Chem. Biol.* **2017**, *13*, 916−921. 10.1038/nchembio.2421.[2] *This paper describes the identification of the novel halogenase CylC and alkylating enzyme CylK using comparative genomics.*

• Brotherton, C. A.; Balskus, E. P. A Prodrug Resistance Mechanism Is Involved in Colibactin Biosynthesis and Cytotoxicity. *J. Am. Chem. Soc.* **2013**, *135*, 3359−3362. 10.1021/ja312154m.[3] *This paper describes the discovery of the prodrug-resistance mechanism in colibactin biosynthesis based on comparative genomic analysis. It led to the discovery of the first*

Received: February 7, 2021
Published: June 4, 2021

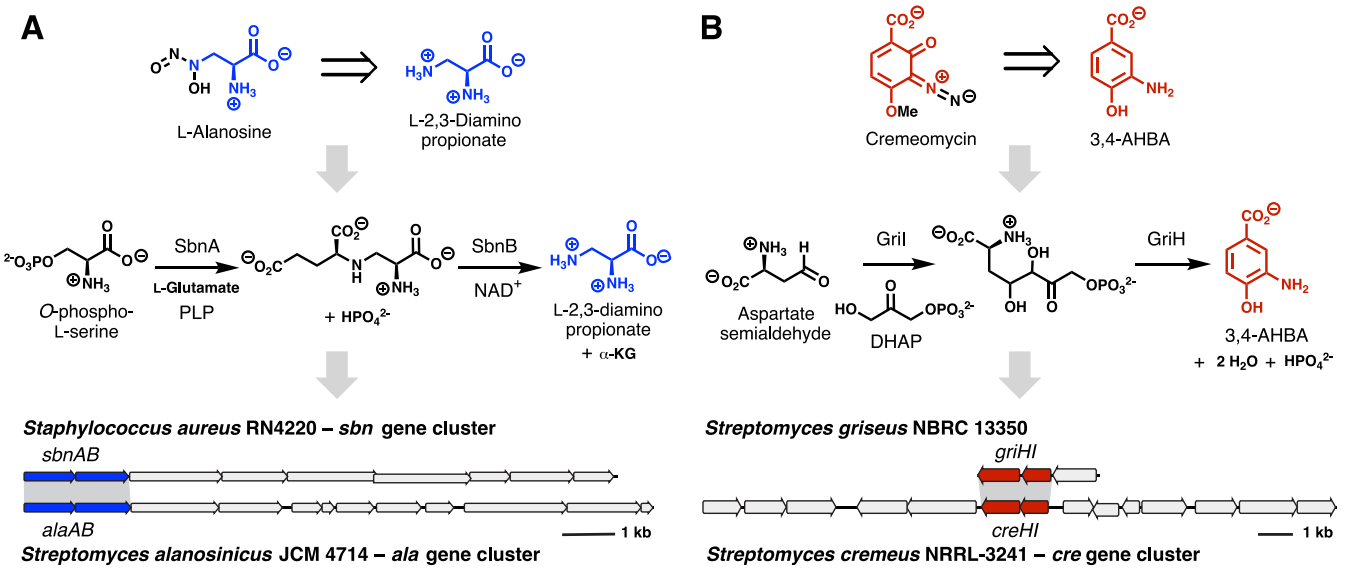


Figure 1. Structure-guided approaches for biosynthetic gene cluster (BGC) identification. (A) Discovery of the alanosine BGC. (B) Discovery of the cremeomycin BGC. PLP = pyridoxal phosphate; NAD^+ = nicotinamide adenine dinucleotide; 3,4-AHBA = 3-amino-4-hydroxybenzoic acid; DHAP = dihydroxyacetone phosphate.

precolibactins and paved the way for the elucidation of colibactin's molecular structure.

- Craciun, S.; Balskus, E. P. Microbial Conversion of Choline to Trimethylamine Requires a Glycyl Radical Enzyme. *Proc. Natl. Acad. Sci.* **2012**, *109*, 21307−21312. 10.1073/pnas.1215689109.[4] *This paper describes the discovery of the gut microbial glycyl radical enzyme choline-TMA lyase (CutC) fueled by analogies with ethanolamine metabolism in E. coli.*

INTRODUCTION

Microbial biochemistry has been studied for well over 100 years, and this field has revealed microbes, with their incredibly varied lifestyles, habitats, and metabolisms, to be a rich source of new chemistry. The discovery and characterization of microbial enzymes, metabolic pathways, and metabolites have influenced many branches of chemistry, informing the study of chemical reactivity,[5] providing targets for total and semisynthesis,[6] fueling drug development and medicinal chemistry,[7] and revealing new tools for biocatalysis and biotechnology.[8]

Though well-established, microbial biochemistry has been transforming in the genomic era. This period (which begins roughly in the mid 2000s and proceeds through the present) encompasses a time during which genomic sequencing has been widely available and, in recent years, ubiquitous. As of November 2020, over 229,000 permanent draft genomes and 20,100 complete genomes have been deposited in NCBI and are listed in the Genomes Online Database (GOLD). The majority of these (62%) are from bacteria and archaea.[9] This bounty is accompanied by a formidable gap between the number of genes sequenced and the number for which we can confidently assign a function. Even in *E. coli*, there is experimental evidence for the function for only about two-thirds of the predicted genes.[10] In less well characterized organisms, that number is much lower. For instance, as part of the expanded Human Microbiome Project, researchers could only functionally annotate 35−45% of the genes identified.[11] This challenge extends to genes that are critical for life. Of the 473 apparently essential genes in a "minimal bacterial genome" derived from the genome of *Mycoplasma mycoides*, 31.5% had no known function at the time of publication.[12] Clearly, the genomic era has provided vast amounts of data that have the potential to transform our understanding of microbes and their chemistry. The major challenge we face in our current research era is how to best harness this information for discovery.

Early biochemical discoveries stemmed from a common investigative approach, regardless of whether the investigator studied microbes or mammals. One usually found a compound or enzyme of interest by obtaining biomass, lysing the cells (or collecting the spent media), and then purifying the active species of interest. When genetic mapping and manipulation became available, the investigator could use purified protein to track down the encoding gene and perhaps disrupt or delete it. Overall, the process was a "forward" one that began with biomass and ended with a gene.

In the genomic era, we can more frequently use "reverse" approaches that start with genes and end with biochemical functions. In these endeavors, microbial biochemists can also exploit a powerful genomic trait. Prokaryotes (bacteria and archaea), and sometimes fungi, organize their genomic information into gene clusters (defined below). Researchers can exploit this feature for biochemical discovery through the analysis of genomic context, providing opportunities that are largely unavailable for the study of macroscopic organisms, which typically do not utilize gene clusters.

This Account discusses how the analysis of gene clusters, operons, and genomic context can enable many facets of microbial biochemical research. In particular, we highlight how work in our laboratory has combined a chemical understanding of enzymes and metabolism with genomic context to solve difficult problems and simplify scientific challenges. The continued development and refinement of computational tools that exploit this feature will greatly accelerate efforts to connect genes with new biochemical functions.

What Is a Gene Cluster and Why Do Microorganisms Use Them?

A gene cluster consists of a syntenic set of genes, their intervening noncoding sequences, and adjacent regulatory elements. These genes are typically functionally related

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797

(involved in the same pathway or process) and may be positively or negatively coregulated. Prokaryotes are known to cotranscribe sets of gene that are oriented in the same direction, generating a single molecule of mRNA containing multiple open reading frames (ORFs). An operon is a stretch of genomic DNA that serves as a transcriptional template for a multi-ORF mRNA, and the simplest gene cluster consists of a single operon. Here, all of the genes in the operon are physiologically tied together by their cotranscription. However, gene clusters are often larger and more complex than single operons. For example, the choline utilization (*cut*) gene cluster, which allows microbes to use choline as a source of carbon and energy under anaerobic conditions, contains multiple apparent operons, all of which are transcribed in the same direction.[4] It is also quite common to observe gene clusters consisting of two operons that are arranged "head-to-head", divergently transcribed from the same DNA segment. An example of this type of gene cluster is the *D. desulfuricans* DSM 642 isethionate metabolism gene cluster.[13] Regulatory proteins are also often found oriented for divergent transcription just upstream of an operon.[14] Finally, some gene clusters are chaotically organized, containing functionally related genes in multiple operons that are divergently transcribed with no apparent pattern. One example is the *cre* gene cluster that biosynthesizes cremeomycin in *Streptomyces cremeus* (Figure 1B).[15] Such gene clusters are particularly difficult to identify. It is therefore important to keep in mind the many structural varieties of gene clusters.

Why do microbes organize their genes into clusters? Several plausible explanations have been proposed. First, this organization mode provides the opportunity to easily coregulate gene expression.[16] Such coregulation would enable coordinated changes in metabolism and other important phenotypes. This is particularly clear in the cases of operons, where an entire biochemical pathway is coexpressed as a single transcriptional unit. Less obviously, the colocalization of genes may permit the coregulation of separate operons using common promotor and regulatory sequences.[16] Finally, organization into gene clusters may permit facile horizontal gene transfer of an entire pathway between different microbes, as compared to gene-by-gene transfer in separate events.[17] Therefore, it is likely that multiple evolutionary driving forces are responsible for this mode of genomic organization in prokaryotes. Why eukaryotes do not consistently use the same strategy is an open question.

Identifying Biosynthetic Gene Clusters

A biosynthetic gene cluster (BGC) is a gene cluster that encodes the enzymatic machinery responsible for the biosynthesis of a natural product or secondary metabolite. Microbial natural products are an important source of drugs and other bioactive compounds.[7] While many organisms do not harbor BGCs, they can be quite numerous in certain microbes. *Streptomyces* species, for instance, typically encode dozens of BGCs.[18] The vast majority of BGCs are "cryptic", meaning that the gene cluster is not associated with a known natural product.

Extensive investigations into the genetics, enzymology, and biosynthetic logic of microbial natural product assembly, as well as the availability of powerful bioinformatic tools,[19,20] make BGCs among the easiest gene clusters to locate. Nonetheless, linking a natural product to its corresponding BGC can still be nontrivial. Strategies that leverage natural product structural information to guide BGC identification are known as structure-guided discovery approaches. As the examples below illustrate, analysis of genomic context plays a key role in structure-guided BGC identification.

L-Alanosine is a natural product produced by *Streptomyces alanosinicus* ATCC 15710 (Figure 1A).[21] It has been investigated as an antitumor compound but was of interest to us because of its unusual nitrogen−nitrogen (N−N) bond-containing *N*-hydroxynitroso (diazeniumdiolate) functional group. We recently identified the L-alanosine BGC using a structure-guided approach. Inspection of L-alanosine's structure suggested it might originate from L-diaminopropionic acid, a metabolite known to be made in the staphloferrin B biosynthetic pathway by two enzymes, SbnA and SbnB.[22,23] SbnA uses a pyridoxal phosphate (PLP) cofactor to convert *O*-phospho-L-serine into *N*-(1-amin-1-carboxyl-2-ethyl)-glutamic acid, which SbnB oxidatively cleaves to L-diaminopropionic acid.[22] Thus, we searched the *S. alanosinicus* genome for genes encoding homologues of SbnA and SbnB. We identified one candidate (the *ala* gene cluster) that encodes the requisite L-diaminopropionic acid biosynthetic enzymes, along with nonribosomal peptide synthase (NRPS) machinery predicted to activate and tether the L-diaminopropionate during N−N bond formation (Figure 1A). Deletion of components of the *ala* gene cluster confirmed its role in L-alanosine biosynthesis.[24]

Structure-guided approaches can uncover BGCs that might not otherwise be readily recognized. Cremeomycin is an intriguing *o*-diazoquinone natural product (Figure 1B) that is produced by *Streptomyces cremeus* NRRL 3241.[25] Based on its structure, we hypothesized that cremeomycin was derived from 3-amino-4-hydroxybenzoic acid (3,4-AHBA). Enzymes involved in assembling 3,4-AHBA, aldolase GriI, and cyclase GriH had been characterized in the context of grixazone biosynthesis.[26] We identified a gene cluster in *Streptomyces cremeus* (the *cre* gene cluster) that encoded GriI and GriH homologues, along with genes later found to be responsible for diazo formation (*creDEM*). Heterologous expression of the *cre* gene cluster verified its connection to cremeomycin biosynthesis.[15]

Applying structure-guided approaches may be challenging if natural products contain structural features that are of unknown origin or are installed by enzyme classes that participate in many cellular pathways and processes. A potential alternative strategy leverages information about the natural product's biological activity. For example, it is common for BGCs to encode self-resistance proteins that protect the producing organism from the natural product.[27] In some cases, the self-resistance protein is an extra, resistant copy of the natural product's molecular target (e.g., a BGC that produces an antibiotic targeting topoisomerase might encode a resistant topoisomerase[28]). This phenomenon presents the opportunity for resistance-guided genomic mining strategies. In this approach, knowledge of a natural product's target or mechanism of action is used to identify gene clusters that contain potential resistance genes.

We used a resistance-guided genome mining approach to identify the streptozotocin BGC (Figure 2). Streptozotocin is a *N*-nitrososourea-containing natural product that is used as a chemotherapeutic for treatment of pancreatic and gastrointestinal cancers.[29] Because streptozotocin is a DNA methylating agent, we searched the genome of the producer *Streptomyces achromogenes* NRRL 2697 for BGCs that encoded

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797

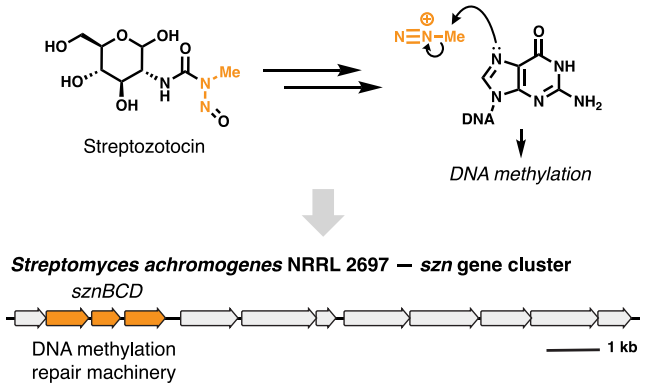


Figure 2. Resistance-guided approach identified the BGC that produces the DNA alkylating agent streptozotocin.

DNA repair enzymes. This analysis identified the *szn* gene cluster, which encodes homologues of the DNA alkylation repair enzymes AlkB and AGT.[1] Characterizing the encoded streptozotocin biosynthetic enzymes led to the discovery of SznF, an iron-dependent oxygenase that catalyzes an unusual N−N bond forming oxidative rearrangement to generate an *N*-nitrososourea-containing biosynthetic intermediate. This is the first example of enzymatic *N*-nitrosation, a transformation previously considered to be derived only from nonenzymatic reactions in living systems.

Identifying Unusual Enzymes in Natural Product Biosynthesis

The discovery of unusual enzymes has long been a motor of progress in biochemical research. In addition to expanding the scope of known enzymatic chemistry, novel enzymes inspire the development of biomimetic catalysts and synthetic strategies, provide new tools for biocatalysis and bioengineering, and highlight new targets for therapeutic development.[30] Identifying an unprecedented enzyme activity in natural product biosynthesis can also facilitate the discovery of new microbial natural products by searching for additional BGCs encoding homologues of the enzyme of interest.

Genomic context can guide the discovery of unusual enzymes. As highlighted above, a deep understanding of natural product biosynthesis informs structure-guided gene cluster identification. Once a new gene cluster is identified, however, one typically finds "gaps" in the proposed biosynthetic pathway—steps that can be predicted by chemical logic but cannot be accounted for by known enzymatic chemistry. Perhaps unsurprisingly, these missing reactions are often the most interesting transformations. In this situation, the investigator can inspect other genes in the gene cluster that have no obvious, alternative role to identify the enzymes responsible for the missing steps. This approach can be used to resolve biosynthetic pathways and uncover intriguing new classes of enzymes.

We employed this strategy in studies of the cylindrocyclophanes, cyanobacterial polyketide natural products with a structurally unusual, all-carbon [7.7]paracyclophane core.[31] Using a structure-guided strategy, we identified the *cyl* gene cluster, which produces cylindrocyclophanes A−F (Figure 3).[2] Annotation of the gene cluster revealed no obvious candidate enzyme(s) responsible for paracyclophane assembly, but genomic context analysis offered a path forward. Comparison of *cyl* gene clusters from the paracyclophane producing cyanobacteria *Cylindrospermum licheniforme* ATCC 29412 and *C. stagnale* PCC 7417 revealed four conserved genes (*cylC*, *cylJ*, *cylK*, and *cylP*) that could not be assigned an obvious biosynthetic role, narrowing the candidates for macrocycle assembly (Figure 3A). Analyses of additional BGCs progressively simplified the outlook. We identified CylC homologues encoded in several BGCs that produce chlorinated natural products but lack homologues of characterized halogenases.[32] This suggested that that CylC was a new type of halogenase that activated a monomeric precursor for

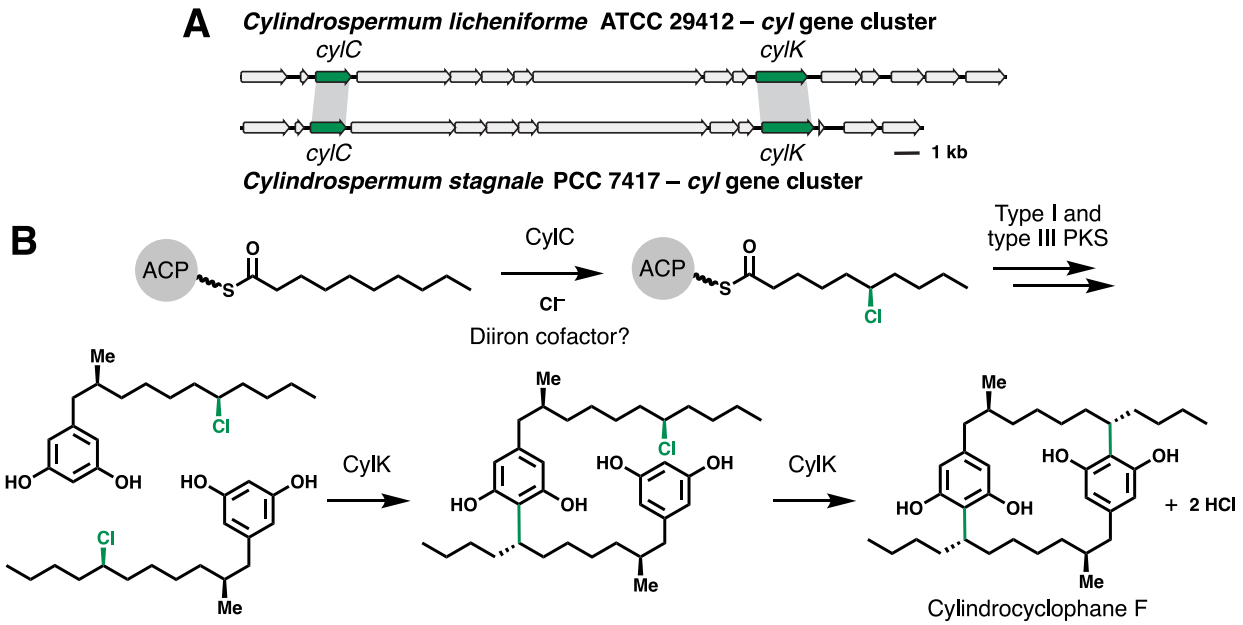


Figure 3. Discovery of unusual enzymes in cylindrocyclophane biosynthesis. (A) Unusual halogenase (CylC) and Friedel−Crafts alkylating (CylK) enzymes were identified using comparative analyses. (B) The roles of CylC and CylK in assembling cylindrocyclophane F. ACP = acyl carrier protein.

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797

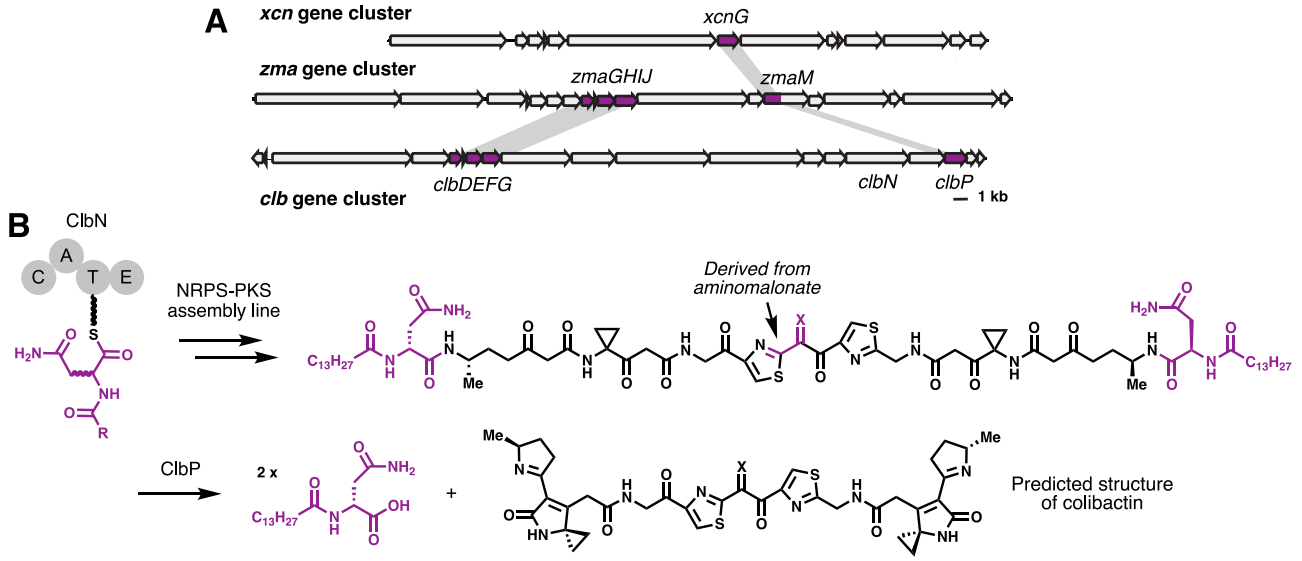


Figure 4. Genomic context analysis informs colibactin structure elucidation. (A) Comparing the zwittermicin, xenocoumacin, and colibactin BGCs revealed conserved enzymes involved in a prodrug resistance mechanism and synthesis of an aminomalonate extender unit. (B) Roles of NRPA ClbN and peptidase ClbP in biosynthesis of the proposed colibactin structure. C = condensation domain; A = adenylation domain; PCP = peptidyl carrier protein domain; E = epimerization domain.

subsequent alkylation. Indeed, recombinant CylC chlorinates an early biosynthetic intermediate that is elaborated to a chlorinated alkylresorcinol precursor (Figure 3B).[32]

The question remained as to which enzyme catalyzed paracyclophane formation, a reaction evocative of the classic Friedel–Crafts alkylation, which typically requires harsh reaction conditions. The potential use of a secondary alkyl chloride substrate, which is typically a poor electrophile, was particularly striking as there was no biological precedent for aromatic ring alkylation with alkyl halides. Of the remaining uncharacterized enzymes, CylP was found to be a benzylic hydroxylase and putative *O*-methyltransferase CylJ was deemed an unlikely candidate. This left CylK, a predicted hemolysin-type calcium-binding protein. *In vitro* assays with recombinant CylK showed that it catalyzes sequential, stereospecific Friedel–Crafts alkylation reactions using the two activated chloroalkylresorcinol monomers to build the paracyclophane macrocycle (Figure 3).[2] CylK is a relatively promiscuous enzyme, accepting a variety of alkyl chlorides and resorcinols.[33] This reaction is biologically unprecedented and potentially of biocatalytic utility.

Natural Product Structural Elucidation and Discovery

It is common to discover natural products using bioactivity-guided isolation followed by structural characterization. However, traditional isolation approaches are sometimes unsuccessful due to metabolite instability, low titers, or variable production. In such cases, the availability of genome sequences and microbial genetics, together with analysis of genomic context, can provide alternative strategies for natural product characterization. This is illustrated by our efforts to characterize the structure of the elusive gut bacterial genotoxin colibactin.

In 2006, Oswald and co-workers discovered that strains of *E. coli* belonging to the phylogenetic group B2 induce DNA double-strand breaks and cell cycle arrest in mammalian cells.[34] Transposon mutagenesis linked this genotoxic activity to a BGC (the *pks* island or *clb* gene cluster) that encodes NRPS and polyketide synthase (PKS) machinery. Based on the presence of these assembly line enzymes, the *clb* gene cluster was inferred to biosynthesize a genotoxic natural product named colibactin.[35] Attempts to isolate colibactin using traditional, activity-guided approaches failed, presumably due to the molecule's instability.[36] Despite the initial lack of structural information, evidence supporting a potential connection between colibactin's genotoxic activity and colorectal cancer has accumulated.[37,38]

Our group revitalized the quest to characterize colibactin by studying the biosynthetic enzymes encoded in the *clb* gene cluster prior to isolation, reversing the typical paradigm for biosynthetic investigations. If the activities of key enzymes could be assigned through *in vitro* biochemical characterization, then perhaps this information could help inform proposals for the final natural product structure, even if we were unable to isolate that product. We envisioned analyzing the *clb* gene cluster and applying a wealth of knowledge regarding NRPS-PKS assembly lines to generate testable biosynthetic hypotheses.

Immediately, we noticed that the *clb* gene cluster encoded homologues of enzymes from the xenocoumacin and zwittermicin biosynthetic pathways: an initiating NRPS module and a periplasmic peptidase. In their studies of zwittermicin biosynthesis, Thomas and co-workers proposed that the initiating NRPS module (ZmaO) initially generated a *N*-acyl-D-asparagine motif which would be later hydrolyzed by the periplasmic peptidase (ZmaM), but this hypothesis was not tested experimentally.[39] Later, genetic studies of xenocoumacin biosynthesis by Bode and co-workers revealed that deleting the gene encoding the periplasmic peptidase XcnG resulted in accumulation of a larger, inactive *N*-acyl-D-asparagine-containing precursor (prexenocoumacin). They proposed that processing of this "prodrug" by XcnG in the periplasm functioned as a self-resistance mechanism.[40]

The *clb* gene cluster encodes a peptidase (ClbP) that is a close homologue of both XcnG and ZmaM,[33] as well as a homologous initiating NRPS module (ClbN) predicted to assemble the *N*-acyl-D-asparagine prodrug motif. This led us to

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788–2797

hypothesize that related biosynthetic logic might be used in colibactin assembly. We began by confirming *in vitro* that ClbN generated an *N*-acyl-D-asparagine motif, identifying myristoyl-CoA as a preferred acyl donor. Next, we found that ClbB, the NRPS that we predicted would act next in the pathway, accepts and elongates the *N*-acyl-D-asparagine intermediate using either L-alanine or L-valine as a building block. Finally, we tested ClbP's activity toward synthetic substrate analogs and found it selectively hydrolyzes the N-acyl-D-asparagine motif both *in vitro* and *in vivo*.[3,41] Together, these results suggested that colibactin is initially synthesized as an inactive precursor (precolibactin) and that removal of the prodrug motif by ClbP generates the active genotoxin. Indeed, genetic analyses showed that the catalytic activity of ClbP is required for genotoxicity.

Identifying the colibactin prodrug resistance mechanism had an immediate impact on further characterization efforts. Deletion of *clbP* disrupts the enzymatic assembly line, resulting in the accumulation of more stable "candidate precolibactins", biosynthetic intermediates, and shunt products which could be isolated and structurally characterized.[35] Using this strategy, we and other groups were able to infer structural motifs present in colibactin and link these features to specific genes through genetic deletion and *in vitro* enzyme characterization. Genomic context analysis was also important in this endeavor. For instance, we noticed additional genes in the *clb* gene cluster (where it is *clbDEFG*, see Figure 4A) resembled elements of the zwittermicin A (*zwa*) gene cluster (*zwaGHIJ*) that generate the unusual PKS extender unit aminomalonate from L-serine. Hypothesizing the corresponding colibactin biosynthetic enzymes would share this activity, we and others verified the roles of ClbDEFG and identified the specific components of the assembly line that accept aminomalonate.[42−44] This knowledge informed subsequent structural elucidation efforts of candidate precolibactins and, ultimately, proposals for colibactin's structure.[5,45−48]

Studies of candidate precolibactins also provided information about colibactin's mode of action. Specifically, the unexpected discovery of cyclopropane rings in metabolites isolated from *clbP* mutants suggested that colibactin was a DNA alkylating agent. This hypothesis guided efforts to identify and characterize the products of colibactin-mediated DNA damage. The initial characterization of two colibactin-derived DNA adducts revealed alkylation of adenine by an activated, cyclopropane electrophilic "warhead" generated by removal of the *N*-acyl-D-asparagine residue by ClbP. This DNA alkylating activity was potentially consistent with reports that colibactin producing *E. coli* form DNA interstrand cross-links. However, this activity could not be rationalized from the structural information available at the time. Our investigation of the final uncharacterized gene in the *clb* gene cluster (*clbL*), together with knowledge of the biosynthetic pathway, helped to resolve this issue. ClbL was annotated as an amidase, but *in vitro* assays with substrate mimics suggested that ClbL links two biosynthetic intermediates together via amide bond formation to generate a pseudodimeric precolibactin containing two *N*-acyl-D-asparagine motifs. Cleavage of this intermediate by ClbP would generate a proposed colibactin structure containing two electrophilic warheads, explaining its interstrand cross-linking activity and the formation of known DNA adducts.[1,47,48]

This work shows that natural product isolation need not be a prerequisite for structure elucidation. Knowledge gleaned from genomic context is a crucial part of genome-guided characterization efforts, informing biochemical hypotheses for subsequent testing. The predicted structures of natural products encoded by cryptic BGCs may also constitute an exciting new class of targets for synthetic chemists.

Discovery of New Primary Metabolic Enzymes

Microbial proteins that participate in primary metabolic pathways are also commonly encoded in gene clusters. As for natural product biosynthesis, exploitation of genomic context provides an additional option for identifying primary metabolic enzymes, complementing more traditional options including activity-guided isolation, genetic deletion, and transcriptomic/proteomic analyses. Thus, if an investigator is interested in discovering a primary metabolic enzyme, knowledge of transformations elsewhere in the pathway may be helpful.

We leveraged genomic context in our discovery of primary metabolic enzymes involved in anaerobic microbial choline metabolism (Figure 5A). Choline is a dietary nutrient that is metabolized to trimethylamine (TMA) by anaerobic bacteria, including inhabitants of the human gut.[9,48] In the past decade TMA production by the gut microbiota has been linked to a variety of diseases including nonalcoholic fatty liver disease, cardiovascular disease, chronic kidney disease, type 2 diabetes, and atherosclerosis.[50] In 2012, our laboratory identified the microbial enzyme responsible for TMA production from choline.[4] We started with the previous observation that acetaldehyde was an intermediate in choline catabolism. This suggested a redox-neutral lyase activity that eliminated TMA. A similar reaction occurs during the catabolism of ethanolamine by *E. coli* and other organisms. In this case, an adenosylcobalamin-dependent enzyme, ethanolamine ammonia-lyase, catalyzes the elimination of ammonia from ethanolamine through a radical mechanism, thereby generating acetaldehyde. This volatile, reactive intermediate is generated and sequestered within a protein-based organelle called a bacterial microcompartment.[51]

Taking note of this analogous biochemical logic, we hypothesized that the enzyme responsible for liberating TMA from choline would be encoded together with homologues of proteins involved in ethanolamine metabolism, including microcompartment shell proteins (Figure 5A). Thus, we looked for homologues of microcompartment proteins encoded in the genome of *D. desulfuricans*, a known choline catabolizer, and identified the choline utilization (*cut*) gene cluster.[6] Intriguingly, the *cut* gene cluster lacked a gene encoding a homologue of ethanolamine ammonia-lyase. Instead, it encoded a glycyl radical enzyme (CutC). Because of its genomic context and powerful catalytic capabilities, we proposed that CutC was a choline TMA-lyase. Subsequent biochemical studies demonstrated that it converts choline to TMA and acetaldehyde by a radical mechanism.[4]

The discovery of CutC aided researchers seeking to understand the metabolism of methylated amines in anoxic environments, particularly the human intestinal tract. It enabled the identification of gut organisms capable of producing TMA from choline,[52] as well as the development of inhibitors that could disrupt choline metabolism in the gut.[49,53] Both these advances have potential diagnostic and therapeutic value.

It is also possible to discover new primary metabolic enzymes starting from uncharacterized microbial genes rather

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797

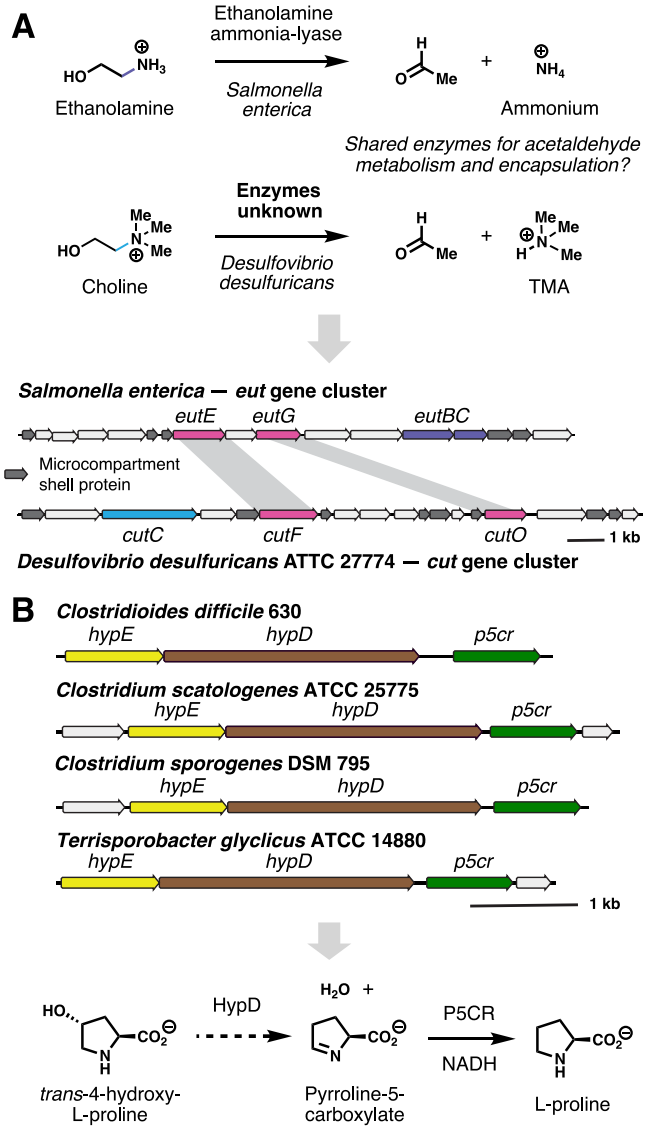


Figure 5. Genomic context analysis enables the discovery of microbial primary metabolic enzymes. (A) Recognizing parallels between the biochemical logic of anaerobic choline metabolism and ethanolamine utilization guided identification of the choline utilization (*cut*) gene cluster. (B) The genomic context of a gene encoding an uncharacterized glycyl radical enzyme (HypD) led to the discovery of its role in *trans*-4-hydroxy-L-proline metabolism. P5CR = pyrroline-5-carboxylate reductase.

than metabolic activities. Most members of large protein superfamilies, including the glycyl radical enzymes, cannot currently be confidently assigned a biochemical function. Examining the genomic contexts of genes encoding uncharacterized superfamily members may guide the generation of hypotheses for experimental testing.

We applied this approach to the glycyl radical enzymes, which are involved in many primary anaerobic metabolic pathways and are among the most abundant protein families in the human gut microbiota. We noticed that many human gut Clostridia, including the intestinal pathogen *Clostridioides* (formerly *Clostridium*) *difficile*, encoded an uncharacterized glycyl radical enzyme next to a gene encoding a homologue of pyrroline 5-carboxylate reductase (P5CR) (Figure 5B). P5CR reduces pyrroline 5-carboxylate to proline, and this context suggested a plausible role for the glycyl radical enzyme (HypD) in *trans*-4-hydroxy-L-proline metabolism. *trans*-4-Hydroxy-L-proline (Hyp) has long been known as a substrate for anaerobic bacterial catabolism,[54] where it serves as an electron acceptor during fermentation, undergoing reduction to 5-aminovalerate. This metabolism was known to occur in *Clostridium sporogenes* and *C. difficile*.[54,55] Previous evidence suggested that proline was an intermediate in Hyp metabolism,[55] but the enzyme(s) responsible for this conversion were unknown. We envisioned that pyrroline 5-carboxylate could be a plausible intermediate between Hyp and proline, implying the existence of an enzyme that could dehydrate Hyp to pyrroline 5-carboxylate.

Chemical reasoning suggested that HypD might be the missing enzyme. The protons vicinal to the Hyp hydroxyl group are not acidic, making a radical mechanism, potentially facilitated by a glycyl radical enzyme, plausible. Indeed, purification and biochemical assays of HypD showed that the enzyme is a Hyp dehydratase (Figure 5B).[56] Discovery of HypD identified a plausible new drug target for *C. difficile* and expanded the range of chemistry known to be catalyzed by glycyl radical enzymes. Metagenomic profiling also revealed that *hypD* is abundant in the healthy human gut microbiota, though the biological roles of HypD in gut commensals are poorly understood.[56,57]

CONCLUSION AND OUTLOOK

The ability to integrate a molecular understanding of enzymes and metabolism with analysis of genomic context is an invaluable tool in the microbial chemical biologist's toolkit, one that can rapidly accelerate the discovery of new enzymes and metabolic pathways. Where does the community go from here? One particularly challenging problem is elucidating the functions of numerous totally uncharacterized gene clusters present in sequenced microbial genomes. Cryptic natural product BGCs are often easier to approach, both because of their presumed role in natural product assembly and because related BGCs have often been studied. Outside of natural product biosynthesis, uncovering the functions of uncharacterized gene clusters is a formidable challenge. Here, we discuss some of the difficulties in characterizing such gene clusters and highlight potential strategies to overcome this problem.

First, the investigator usually requires a "foot in the door"—some pre-existing data that links an uncharacterized gene cluster to a metabolic or biological function. These data can take a variety of forms, including the phenotype of an overexpression or deletion mutant, native expression patterns, and homology of genes in the gene cluster to genes of known function. For most gene clusters, these data do not exist or are difficult to interpret. For instance, one might find conditions under which a gene cluster is expressed but be unable to link that gene cluster to a phenotypic effect. Second, with so many gene clusters available to investigate, one must prioritize specific gene clusters for characterization.

Analysis of genomic context can address both of these problems. Analyzing a gene cluster can reveal whether surrounding genes fall into the same (or related) COGs (Clusters of Orthologous Genes) in similar gene clusters. If the gene cluster is found in more than one organism, this could indicate a shared or closely related function. Thus, the investigator may use what we term "pan-genomic context", the sum of the genomic contexts for each instance of a gene cluster. Pan-genomic context can provide important clues

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797

about function. For instance, in some organisms a gene of interest might have no suggestive genomic context, but in others, it might be colocalized with genes involved in a characterized metabolic process. This could suggest a plausible role in that process in both organisms. Genomic context can also help to prioritize of uncharacterized gene clusters for further study, for example, by allowing researchers to focus on genomic regions containing uncharacterized members of an enzyme family of interest.

While the analysis of genomic context is indubitably a key component of future work in microbial chemistry, significant challenges remain to be addressed. In particular, one commonly finds that a gene of interest is surrounded by genes with only a family or "domain of unknown function" (DUF) assignment. At first glance, such annotations are of little help. But these nonspecific or even ambiguous annotations represent a new protein family to discover, and linking such gene clusters to microbial activities can provide a path forward for characterization.

A lack of formalization is another difficulty in exploiting genomic context. Outside of natural product BGC analysis programs, genomic context is typically assessed using low-throughput methods. The Enzyme Function Initiative Gene Neighborhood Tool (EFI-GNT) has been a notable advance, allowing the research to compare side-by-side the genomic context of a gene and its placement in a protein sequence similarity network.[58] Still, much of the analysis of genomic context relies on intuition—"eureka" moments when the investigator makes the mental connection between the genomic placement of a gene and its plausible role. To exploit the full potential of genomic context, we should integrate computational approaches that identify colocalized genes with improved automated functional inferences to permit high-throughput hypothesis generation. It is quite possible that appropriate machine-learning algorithms may play a substantial role in future advancements.

It is clear that advances in bioinformatics, genetic manipulation, and metabolomics will make the functional delineation of uncharacterized microbial gene clusters substantially easier in the coming decades. To further accelerate this process, it is critical to improve our ability to leverage the information inherent in genomic context, with biochemists and chemical biologists playing a central role in prioritizing exciting targets for discovery and incorporating biochemical knowledge into computational approaches. We envision chemical knowledge and approaches being critical in transitioning from a genomic to a functional era in microbial biology.

AUTHOR INFORMATION

Corresponding Author

Emily P. Balskus − *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States;* orcid.org/0000-0001-5985-5714; Email: balskus@chemistry.harvard.edu

Author

Duncan J. Kountz − *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.1c00100

Notes

The authors declare no competing financial interest.

Biographies

Duncan J. Kountz was born in Wilmington, DE, and grew up in Charlotte, NC, and West Chester, PA. He attended The Ohio State University, graduating with a B.S. in Biochemistry and Philosophy in 2017, performing research under Prof. Joseph Krzycki. He joined the Harvard University Chemical Biology Ph.D. program in 2017.

Emily P. Balskus is originally from Cincinnati, OH. She received her B.A. in Chemistry from Williams College in 2002 and her Ph.D. in Chemistry from Harvard University in 2008 working with Prof. Eric N. Jacobsen. She then pursued postdoctoral research at Harvard Medical School with Prof. Christopher T. Walsh. She joined the faculty of Harvard University in 2011 and is currently a Professor of Chemistry and Chemical Biology.

ACKNOWLEDGMENTS

We are grateful to all our past and current co-workers who have contributed to the research efforts described here. E.P.B. acknowledges support from a Howard Hughes Medical Institute-Gates Faculty Scholar Award (OPP1158186) and D.J.K. acknowledges support from National Institutes of Health Training Grant #5T32GM095450.

REFERENCES

(1) Ng, T. L.; Rohac, R.; Mitchell, A. J.; Boal, A. K.; Balskus, E. P. An N-Nitrosating Metalloenzyme Constructs the Pharmacophore of Streptozotocin. *Nature* **2019**, *566*, 94−99.

(2) Nakamura, H.; Schultz, E. E.; Balskus, E. P. A New Strategy for Aromatic Ring Alkylation in Cylindrocyclophane Biosynthesis. *Nat. Chem. Biol.* **2017**, *13*, 916−921.

(3) Brotherton, C. A.; Balskus, E. P. A Prodrug Resistance Mechanism Is Involved in Colibactin Biosynthesis and Cytotoxicity. *J. Am. Chem. Soc.* **2013**, *135*, 3359−3362.

(4) Craciun, S.; Balskus, E. P. Microbial Conversion of Choline to Trimethylamine Requires a Glycyl Radical Enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 21307−21312.

(5) Jiang, Y.; Stornetta, A.; Villalta, P. W.; Wilson, M. R.; Boudreau, P. D.; Zha, L.; Balbo, S.; Balskus, E. P. Reactivity of an Unusual Amidase May Explain Colibactin's DNA Cross-Linking Activity. *J. Am. Chem. Soc.* **2019**, *141*, 11489−11496.

(6) Liu, W.; Hong, B.; Wang, J.; Lei, X. New Strategies in the Efficient Total Syntheses of Polycyclic Natural Products. *Acc. Chem. Res.* **2020**, *53*, 2569−2586.

(7) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770−803.

(8) Schmid, A.; Dordick, J. S.; Hauer, B.; Kiener, A.; Wubbolts, M.; Witholt, B. Industrial Biocatalysis Today and Tomorrow. *Nature* **2001**, *409*, 258−268.

(9) *Genomes Online Database*; JGI: Walnut Creek, CA, https://gold.jgi.doe.gov (accessed 2020-11-04).

(10) Ghatak, S.; King, Z. A.; Sastry, A.; Palsson, B. O. The Y-Ome Defines the 35% of Escherichia Coli Genes That Lack Experimental Evidence of Function. *Nucleic Acids Res.* **2019**, *47*, 2446−2454.

(11) Lloyd-Price, J.; Mahurkar, A.; Rahnavard, G.; Crabtree, J.; Orvis, J.; Hall, A. B.; Brady, A.; Creasy, H. H.; McCracken, C.; Giglio, M. G.; McDonald, D.; Franzosa, E. A.; Knight, R.; White, O.; Huttenhower, C. Strains, Functions and Dynamics in the Expanded Human Microbiome Project. *Nature* **2017**, *550*, 61−66.

(12) Hutchison, C. A.; Chuang, R.-Y.; Noskov, V. N.; Assad-Garcia, N.; Deerinck, T. J.; Ellisman, M. H.; Gill, J.; Kannan, K.; Karas, B. J.; Ma, L.; Pelletier, J. F.; Qi, Z.-Q.; Richter, R. A.; Strychalski, E. A.; Sun, L.; Suzuki, Y.; Tsvetanova, B.; Wise, K. S.; Smith, H. O.; Glass, J. I.;

Merryman, C.; Gibson, D. G.; Venter, J. C. Design and Synthesis of a Minimal Bacterial Genome. *Science* **2016**, *351*, aad6253−aad6253.
(13) Peck, S. C.; Denger, K.; Burrichter, A.; Irwin, S. M.; Balskus, E. P.; Schleheck, D. A Glycyl Radical Enzyme Enables Hydrogen Sulfide Production by the Human Intestinal Bacterium Bilophila Wadsworthia. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3171−3176.
(14) Wallenstein, A.; Rehm, N.; Brinkmann, M.; Selle, M.; Bossuet-Greif, N.; Sauer, D.; Bunk, B.; Spröer, C.; Wami, H. T.; Homburg, S.; von Bünau, R.; König, S.; Nougayrède, J.-P.; Overmann, J.; Oswald, E.; Müller, R.; Dobrindt, U. ClbR Is the Key Transcriptional Activator of Colibactin Gene Expression in Escherichia Coli. *mSphere* 2020, *5*, e00591-20. .
(15) Waldman, A. J.; Pechersky, Y.; Wang, P.; Wang, J. X.; Balskus, E. P. The Cremeomycin Biosynthetic Gene Cluster Encodes a Pathway for Diazo Formation. *ChemBioChem* **2015**, *16*, 2172−2175.
(16) Osbourn, A. E.; Field, B. Operons. *Cell. Mol. Life Sci.* **2009**, *66*, 3755−3775.
(17) Price, M. N.; Arkin, A. P.; Alm, E. J. The Life-Cycle of Operons. *PLoS Genet.* **2006**, *2*, e96.
(18) Rigali, S.; Anderssen, S.; Naômé, A.; van Wezel, G. P. Cracking the Regulatory Code of Biosynthetic Gene Clusters as a Strategy for Natural Product Discovery. *Biochem. Pharmacol.* **2018**, *153*, 24−34.
(19) Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S. Y.; Medema, M. H.; Weber, T. AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline. *Nucleic Acids Res.* **2019**, *47*, W81−W87.
(20) Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Muñoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Suarez Duran, H. G.; Pascal Andreu, V.; Selem-Mojica, N.; Alanjary, M.; Robinson, S. L.; Lund, G.; Epstein, S. C.; Sisto, A. C.; Charkoudian, L. K.; Collemare, J.; Linington, R. G.; Weber, T.; Medema, M. H. MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function. *Nucleic Acids Res.* **2019**, gkz882.
(21) Murthy, Y. K. S.; Thiemann, J. E.; Coronelli, C.; Sensi, P. Alanosine, a New Antiviral and Antitumour Agent Isolated from a Streptomyces. *Nature* **1966**, *211*, 1198−1199.
(22) Kobylarz, M. J.; Grigg, J. C.; Liu, Y.; Lee, M. S. F.; Heinrichs, D. E.; Murphy, M. E. P. Deciphering the Substrate Specificity of SbnA, the Enzyme Catalyzing the First Step in Staphyloferrin B Biosynthesis *Biochemistry* **2016**, *55*,927.
(23) Kobylarz, M. J.; Grigg, J. C.; Takayama, S. J.; Rai, D. K.; Heinrichs, D. E.; Murphy, M. E. P. Synthesis of L-2,3-Diaminopropionic Acid, a Siderophore and Antibiotic Precursor. *Chem. Biol.* **2014**, *21*, 379−388.
(24) Ng, T. L.; McCallum, M. E.; Zheng, C. R.; Wang, J. X.; Wu, K. J. Y.; Balskus, E. P. The L -Alanosine Gene Cluster Encodes a Pathway for Diazeniumdiolate Biosynthesis. *ChemBioChem* **2020**, *21*, 1155−1160.
(25) McGuire, J. N.; Wilson, S. R.; Rinehart, K. L. Cremeomycin, a Novel Cytotoxic Antibiotic from Streptomyces Cremeus. Structure Elucidation and Biological Activity. *J. Antibiot.* **1995**, *48*, 516−519.
(26) Suzuki, H.; Ohnishi, Y.; Furusho, Y.; Sakuda, S.; Horinouchi, S. Novel Benzene Ring Biosynthesis from C_3 and C_4 Primary Metabolites by Two Enzymes. *J. Biol. Chem.* **2006**, *281*, 36944−36951.
(27) Foulston, L. Genome Mining and Prospects for Antibiotic Discovery. *Curr. Opin. Microbiol.* **2019**, *51*, 1−8.
(28) Baumann, S.; Herrmann, J.; Raju, R.; Steinmetz, H.; Mohr, K. I.; Hüttel, S.; Harmrolfs, K.; Stadler, M.; Müller, R. Cystobactamids: Myxobacterial Topoisomerase Inhibitors Exhibiting Potent Antibacterial Activity. *Angew. Chem., Int. Ed.* **2014**, *53*, 14605−14609.
(29) Krug, S.; Boch, M.; Daniel, H.; Nimphius, W.; Müller, D.; Michl, P.; Rinke, A.; Gress, T. M. Streptozocin-Based Chemotherapy in Patients with Advanced Neuroendocrine Neoplasms − Predictive and Prognostic Markers for Treatment Stratification. *PLoS One* **2015**, *10*, e0143822. DOI: 10.1371/journal.pone.0143822.
(30) Bollenbach, M.; Ortega, M.; Orman, M.; Drennan, C. L.; Balskus, E. P. Discovery of a Cyclic Choline Analog That Inhibits Anaerobic Choline Metabolism by Human Gut Bacteria. *ACS Med. Chem. Lett.* **2020**, *11*, 1980−1985. DOI: 10.1021/acsmedchemlett.0c00005.
(31) Moore, B. S.; Chen, J. L.; Patterson, G. M. L.; Moore, R. E.; Brinen, L. S.; Kato, Y.; Clardy, J. [7.7]Paracyclophanes from Blue-Green Algae. *J. Am. Chem. Soc.* **1990**, *112*, 4061−4063.
(32) Nakamura, H.; Hamer, H. A.; Sirasani, G.; Balskus, E. P. Cylindrocyclophane Biosynthesis Involves Functionalization of an Unactivated Carbon Center. *J. Am. Chem. Soc.* **2012**, *134*, 18518−18521.
(33) Schultz, E. E.; Braffman, N. R.; Luescher, M. U.; Hager, H. H.; Balskus, E. P. Biocatalytic Friedel−Crafts Alkylation Using a Promiscuous Biosynthetic Enzyme. *Angew. Chem., Int. Ed.* **2019**, *58*, 3151−3155.
(34) Nougayrede, J.-P.; Homburg, S.; Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J.; Dobrindt, U.; Oswald, E. Escherichia Coli Induces DNA Double-Strand Breaks in Eukaryotic Cells. *Science* **2006**, *313*, 848−851.
(35) Putze, J.; Hennequin, C.; Nougayrède, J.-P.; Zhang, W.; Homburg, S.; Karch, H.; Bringer, M.-A.; Fayolle, C.; Carniel, E.; Rabsch, W.; Oelschlaeger, T. A.; Oswald, E.; Forestier, C.; Hacker, J.; Dobrindt, U. Genetic Structure and Distribution of the Colibactin Genomic Island among Members of the Family Enterobacteriaceae. *Infect. Immun.* **2009**, *77*, 4696−4703.
(36) Balskus, E. P. Colibactin: Understanding an Elusive Gut Bacterial Genotoxin. *Nat. Prod. Rep.* **2015**, *32*, 1534−1540.
(37) Arthur, J. C.; Perez-Chanona, E.; Mühlbauer, M.; Tomkovich, S.; Uronis, J. M.; Fan, T.-J.; Campbell, B. J.; Abujamel, T.; Dogan, B.; Rogers, A. B.; Rhodes, J. M.; Stintzi, A.; Simpson, K. W.; Hansen, J. J.; Keku, T. O.; Fodor, A. A.; Jobin, C. Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota. *Science* **2012**, *338*, 120−123.
(38) Cougnoux, A.; Dalmasso, G.; Martinez, R.; Buc, E.; Delmas, J.; Gibold, L.; Sauvanet, P.; Darcha, C.; Déchelotte, P.; Bonnet, M.; Pezet, D.; Wodrich, H.; Darfeuille-Michaud, A.; Bonnet, R. Bacterial Genotoxin Colibactin Promotes Colon Tumour Growth by Inducing a Senescence-Associated Secretory Phenotype. *Gut* **2014**, *63*, 1932−1942.
(39) Kevany, B. M.; Rasko, D. A.; Thomas, M. G. Characterization of the Complete Zwittermicin A Biosynthesis Gene Cluster from Bacillus Cereus. *Appl. Environ. Microbiol.* **2009**, *75*, 1144−1155.
(40) Reimer, D.; Pos, K. M.; Thines, M.; Grün, P.; Bode, H. B. A Natural Prodrug Activation Mechanism in Nonribosomal Peptide Synthesis. *Nat. Chem. Biol.* **2011**, *7*, 888−890.
(41) Volpe, M. R.; Wilson, M. R.; Brotherton, C. A.; Winter, E. S.; Johnson, S. E.; Balskus, E. P. In Vitro Characterization of the Colibactin-Activating Peptidase ClbP Enables Development of a Fluorogenic Activity Probe. *ACS Chem. Biol.* **2019**, *14*, 1097−1101.
(42) Zha, L.; Wilson, M. R.; Brotherton, C. A.; Balskus, E. P. Characterization of Polyketide Synthase Machinery from the Pks Island Facilitates Isolation of a Candidate Precolibactin. *ACS Chem. Biol.* **2016**, *11*, 1287−1295.
(43) Brachmann, A. O.; Garcie, C.; Wu, V.; Martin, P.; Ueoka, R.; Oswald, E.; Piel, J. Colibactin Biosynthesis and Biological Activity Depend on the Rare Aminomalonyl Polyketide Precursor. *Chem. Commun.* **2015**, *51*, 13138−13141.
(44) Bian, X.; Plaza, A.; Zhang, Y.; Müller, R. Two More Pieces of the Colibactin Genotoxin Puzzle from Escherichia Coli Show Incorporation of an Unusual 1-Aminocyclopropanecarboxylic Acid Moiety. *Chem. Sci.* **2015**, *6*, 3154−3160.
(45) Wilson, M. R.; Jiang, Y.; Villalta, P. W.; Stornetta, A.; Boudreau, P. D.; Carrá, A.; Brennan, C. A.; Chun, E.; Ngo, L.; Samson, L. D.; Engelward, B. P.; Garrett, W. S.; Balbo, S.; Balskus, E. P. The Human Gut Bacterial Genotoxin Colibactin Alkylates DNA. *Science* **2019**, *363*, eaar7785.
(46) Xue, M.; Kim, C. S.; Healy, A. R.; Wernke, K. M.; Wang, Z.; Frischling, M. C.; Shine, E. E.; Wang, W.; Herzon, S. B.; Crawford, J. M. Structure Elucidation of Colibactin and Its DNA Cross-Links. *Science* **2019**, *365*, eaax2685.

(47) Brotherton, C. A.; Wilson, M.; Byrd, G.; Balskus, E. P. Isolation of a Metabolite from the Pks Island Provides Insights into Colibactin Biosynthesis and Activity. *Org. Lett.* **2015**, *17*, 1545−1548.
(48) Li, Z.-R.; Li, J.; Gu, J.-P.; Lai, J. Y. H.; Duggan, B. M.; Zhang, W.-P.; Li, Z.-L.; Li, Y.-X.; Tong, R.-B.; Xu, Y.; Lin, D.-H.; Moore, B. S.; Qian, P.-Y. Divergent Biosynthesis Yields a Cytotoxic Aminomalonate-Containing Precolibactin. *Nat. Chem. Biol.* **2016**, *12*, 773−775.
(49) King, G. M. Metabolism of Trimethylamine, Choline, and Glycine Betaine by Sulfate-Reducing and Methanogenic Bacteria in Marine Sedimentst. *Appl. Environ. Microbiol.* **1984**, *48*, 719..
(50) Orman, M.; Bodea, S.; Funk, M. A.; Campo, A. M.; Bollenbach, M.; Drennan, C. L.; Balskus, E. P. Structure-Guided Identification of a Small Molecule That Inhibits Anaerobic Choline Metabolism by Human Gut Bacteria. *J. Am. Chem. Soc.* **2019**, *141*, 33−37.
(51) Kerfeld, C. A.; Aussignargues, C.; Zarzycki, J.; Cai, F.; Sutter, M. Bacterial Microcompartments. *Nat. Rev. Microbiol.* **2018**, *16*, 277−290.
(52) Rath, S.; Heidrich, B.; Pieper, D. H.; Vital, M. Uncovering the Trimethylamine-Producing Bacteria of the Human Gut Microbiota. *Microbiome* **2017**, *5*, 54.
(53) Roberts, A. B.; Gu, X.; Buffa, J. A.; Hurd, A. G.; Wang, Z.; Zhu, W.; Gupta, N.; Skye, S. M.; Cody, D. B.; Levison, B. S.; Barrington, W. T.; Russell, M. W.; Reed, J. M.; Duzan, A.; Lang, J. M.; Fu, X.; Li, L.; Myers, A. J.; Rachakonda, S.; DiDonato, J. A.; Brown, J. M.; Gogonea, V.; Lusis, A. J.; Garcia-Garcia, J. C.; Hazen, S. L. Development of a Gut Microbe−Targeted Nonlethal Therapeutic to Inhibit Thrombosis Potential. *Nat. Med.* **2018**, *24*, 1407−1417.
(54) Stickland, L. H. XXXV. STUDIES IN THE METABOLISM OF THE STRICT ANAEROBES (GENUS CLOSTRIDIUM). *Biochem. J.* **1935**, *29*, 288−290.
(55) Jackson, S.; Calos, M.; Myers, A.; Self, W. T. Analysis of Proline Reduction in the Nosocomial Pathogen Clostridium Difficile. *J. Bacteriol.* **2006**, *188*, 8487−8495.
(56) Levin, B. J.; Huang, Y. Y.; Peck, S. C.; Wei, Y.; Martínez-del Campo, A.; Marks, J. A.; Franzosa, E. A.; Huttenhower, C.; Balskus, E. P. A Prominent Glycyl Radical Enzyme in Human Gut Microbiomes Metabolizes Trans-4-Hydroxy-l-Proline. *Science* **2017**, *355*, eaai8386.
(57) Huang, Y. Y.; Martínez-del Campo, A.; Balskus, E. P. Anaerobic 4-Hydroxyproline Utilization: Discovery of a New Glycyl Radical Enzyme in the Human Gut Microbiome Uncovers a Widespread Microbial Metabolic Activity. *Gut Microbes* **2018**, 1−16.
(58) Zallot, R.; Oberg, N.; Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, *58*, 4169−4182.

https://doi.org/10.1021/acs.accounts.1c00100
Acc. Chem. Res. 2021, 54, 2788−2797