Published in final edited form as:
Biopolymers. 2010 September ; 93(9): 833–844. doi:10.1002/bip.21450.

Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR

Jeffrey H. Kim, **Zhiyang Feng**, **John D. Bauer**, **Dimitris Kallifidas**, **Paula Y. Calle**, and **Sean F. Brady***
Howard Hughes Medical Institute Laboratory of Genetically Encoded Small Molecules, The Rockefeller University, 1230 York Avenue, New York, NY 10065

Abstract

A single gram of soil can contain thousands of unique bacterial species, of which only a small fraction is regularly cultured in the laboratory. Although the fermentation of cultured microorganisms has provided access to numerous bioactive secondary metabolites, with these same methods it is not possible to characterize the natural products encoded by the uncultured majority. The heterologous expression of biosynthetic gene clusters cloned from DNA extracted directly from environmental samples (eDNA) has the potential to provide access to the chemical diversity encoded in the genomes of uncultured bacteria. One of the challenges facing this approach has been that many natural product biosynthetic gene clusters are too large to be readily captured on a single fragment of cloned eDNA. The reassembly of large eDNA-derived natural product gene clusters from collections of smaller overlapping clones represents one potential solution to this problem. Unfortunately, traditional methods for the assembly of large DNA sequences from multiple overlapping clones can be technically challenging. Here we present a general experimental framework that permits the recovery of large natural product biosynthetic gene clusters on overlapping soil-derived eDNA cosmid clones and the reassembly of these large gene clusters using transformation-associated recombination (TAR) in *Saccharomyces cerevisiae*. The development of practical methods for the rapid assembly of biosynthetic gene clusters from collections of overlapping eDNA clones is an important step towards being able to functionally study larger natural product gene clusters from uncultured bacteria.

Introduction

Cultured soil bacteria have been a productive source of both biologically active and structurally diverse natural products.[1,2] Molecular phylogenetic analyses of soil microbiomes now indicate that a single gram of soil can contain thousands of unique bacterial species, only a small fraction of which is regularly cultured in the laboratory.[3–6] Uncultured bacteria represent one of the largest pools of genetic diversity that has not been examined for the production of natural products. Culture-independent analyses of microbial communities using DNA extracted directly from environmental samples, which is commonly defined as metagenomics, has the potential to provide access to the biosynthetic capacity of uncultured bacteria.[7]

All of the genes required for the biosynthesis of a natural product, including genes that encode biosynthetic, regulatory, and self-immunity enzymes, are typically clustered on

Corresponding Author: Sean F. Brady, Address: Laboratory of Genetically Encoded Small Molecules, The Rockefeller University, 1230 York Avenue, New York, NY 10065, Phone: 212-327-8280, Fax: 212-327-8281, sbrady@rockefeller.edu.

bacterial chromosomes. Natural product gene clusters can range in size from a few kilobases to over 100 kilobases. The heterologous expression of natural product biosynthetic gene clusters captured on individual eDNA clones has begun to provide access to some of the natural products encoded in the genomes of uncultured bacteria (Figure 1). However, a major limitation of this strategy has been the inability to routinely construct very large eDNA libraries with inserts big enough to capture large biosynthetic pathways on individual clones. Figure 1 shows a collection of metabolites that have been isolated from the culture broths of soil-derived eDNA clones. In each case, a single cosmid/fosmid eDNA clone was responsible for the production of the metabolites by a heterologous host. Successful functional metagenomic natural product discovery studies carried out on marine samples and other microbiomes have also largely been restricted to single clones.[8,9]

While the construction of 30–40 kb insert cosmid libraries from environmental samples is now routine, the construction of larger insert libraries that can be used to capture large natural product gene clusters has been challenging. Bacterial artificial chromosome (BAC)-derived libraries are capable of capturing larger inserts but generally yield metagenomic libraries that are two to three orders of magnitude smaller than those constructed using cosmid-based cloning strategies.[10] Theoretically, all gene clusters that are too large to be captured on a single cosmid-sized clone can be reassembled from collections of overlapping eDNA cosmid clones (Figure 3a, 4c). Existing gene cluster assembly strategies depend on either unique restriction sites or lambda-mediated recombination to reassemble large DNA fragments. Both of these strategies are technically challenging when working with very large DNA fragments or with sequences that span more than two overlapping clones.[11,12,13–17] Transformation-associated recombination (TAR) in *Saccharomyces cerevisiae* relies on homologous recombination to selectively capture a known sequence from a mixture of genomic DNA.[18,19] In TAR cloning protocols, genomic DNA and a “capture” vector with short homology arms corresponding to sequences flanking the region of interest are co-transformed into *S. cerevisiae.* The capture vector arms and homologous target DNA undergo recombination to yield a stable plasmid containing the targeted genomic region. TAR was originally developed to facilitate cloning large genomic fragments without having to construct and screen genomic DNA libraries. Recent studies extended the scope of this methodology by showing that it could be used to assemble 25 co-transformed overlapping DNA fragments into a complete 592 kb synthetic genome and that multiple PCR products could be assembled into small biochemical pathways.[20–22] These studies led us to believe that TAR could also be used to assemble large natural product gene clusters from multiple overlapping eDNA clones.

In this report, we show that TAR in *S. cerevisiae* can be used to rapidly reassemble large natural product biosynthetic gene clusters from overlapping eDNA cosmid clones. The rich microbial diversity present in soils makes them attractive, but challenging, starting points for the culture-independent discovery of new natural product biosynthetic gene clusters. Much of the difficulty in working with soil-derived eDNA libraries stems from their inherent complexity, which necessitates the construction of very large clone libraries in order to ensure that large biosynthetic pathways can be recovered in their entirety. Using two of the largest soil eDNA cosmid libraries reported to date as examples, we have also empirically investigated the minimum size eDNA libraries will likely need to be to recover complete large natural product gene clusters on overlapping cosmid clones. Taken together, these studies provide an experimental framework for gaining access to large, intact natural product biosynthetic gene clusters from soil microbiomes.

Materials and Methods

Library construction and formatting

Top soil collected in Utah and California was used to construct cosmid-based eDNA libraries following methods previously described.[23] Briefly, the soil was incubated at 70°C in lysis buffer (2% sodium dodecyl sulfate (w/v), 100 m*M* Tris-HCl, 100 m*M* ethylenediaminetetraacetic acid (EDTA), 1.5 *M* NaCl, 1% cetyl trimethylammonium bromide (w/v)) for two hours. Large particulates were then removed by centrifugation (4,000 × g, 30 min). DNA was precipitated from the resulting supernatant with the addition of 0.6 volumes of isopropyl alcohol, pelleted by centrifugation (4,000 × g, 30 min), washed with 70% ethanol and resuspended in a minimum volume of TE (10 m*M* Tris, 1 m*M* EDTA, pH 8). High molecular weight DNA that was purified from the crude extract by gel electrophoresis (1% agarose, 0.5x Tris/Borate/EDTA, 16 hours, 20 V) was blunt-ended (End-It, Epicentre Biotechnologies), ligated into precut pWEB or pWEB-TNC (Epicentre Biotechnologies), packaged into lambda phage and transduced into *Escherichia coli* (EC100, Epicentre Biotechnologies). Individual library aliquots equivalent to approximately 4,000–5,000 colony forming units were either plated on agar plates or inoculated into 5 mL of liquid LB and then allowed to incubate overnight at 37°C with the appropriate selection. Once colonies formed, the plate-grown aliquots were resuspended in 5 mL of LB. Matching glycerol stocks (15% glycerol) and DNA miniprep pairs were created from each unique library aliquot. The minipreps were arrayed in 8 × 8 grids corresponding to 250,000–320,000 total cosmids and DNA from the rows and columns of each grid was pooled. To facilitate library screening, pooled rows and columns were further combined to yield master aliquots, each representing a single 8 × 8 grid of minipreps. Each unique *E. coli* transduction yielded three master aliquots (~750,000 clones) of the Utah library and one master aliquot (~320,000 clones) of the California library. In total, the Utah soil library contains ~10 million unique cosmid clones and the California soil library contains ~15 million unique cosmid clones.

Library size analysis

DNA from each unique *E. coli* transduction reaction was used as a template in PCR reactions with degenerate primers designed to amplify β-Ketoacyl synthase gene sequences (dp:KS_β, 5′-TTCGGSGGNTTCCAGWSNGCSATG-3′ and dp:ACP, 5′-TCSAKSAGSGCSANSGASTCGTANCC-3′).[24,25] Each 25 μL PCR reaction contained 50 ng eDNA template, 2.5 μ*M* of each primer, 2 m*M* dNTPs, 1× ThermoPol Reaction Buffer (New England Biolabs), 0.5 U *Taq* DNA polymerase (New England Biolabs) and 5% dimethyl sulfoxide. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 2 min), then 8 touchdown cycles (95°C, 45 sec; 65°C (dt −1°C/cycle), 1 min; 72°C, 2 min), 35 standard cycles (95°C, 45 sec; 58°C, 1 min; 72°C, 2 min) and a final extension step (72°C, 2 min).[24,25] Amplicons of the correct predicted size (~1.5 kb) were identified by gel electrophoresis, gel purified and directly sequenced. In total, DNA from 7 unique *E. coli* transductions of the Utah library and 20 unique *E. coli* transductions of the California library was examined.

Identification of gene clusters of interest

PCR reactions with degenerate primers designed to amplify β-Ketoacyl synthase gene sequences were used to detect Type II polyketide synthase (PKS) sequences.[24,25] Degenerate primers designed to detect flavin-dependent halogenases (TyrohalF3: 5′-CGGCTGGTTCTGGTACATCCC-3′, TyrohalR2: 5′-GAACTCGTAGAASACSCCGTACTC-3′) were used to identify the nonribosmal peptide synthetase (NRPS) gene cluster. The FRI gene cluster was identified using primers that recognize conserved sequences in acyl-CoA ligases found in lipopeptide antibiotic gene

clusters (DpFrEFWD1: 5′-TSMTSCAGTACACSTCSGG-3 ′ and DpFrEREV1: 5′-WDGTCGTASGCGAAGTCSG-3′). Type II PKS sequences were amplified using the same PCR conditions outlined for the library size analysis. Flavin-dependent halogenases were amplified using the following PCR conditions: Each 20 μL reaction contained primer added to a final concentration of 2.5 μ*M*, 0.5 μL of eDNA template (~100 ng), 1× FailSafe Buffer G (Epicentre Biotechnologies), and 1 U of *Taq* DNA polymerase. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 2 min); 9 touchdown cycles (95°C, 30 sec; 70°C (dt −1°C/cycle), 30 sec; 72°C, 30 sec), 30 standard cycles (95°C, 30 sec; 60°C, 30 sec; 72°C, 30 sec) and a final extension step (72°C, 5 min). The acyl-CoA ligase homologues were identified using the following reaction conditions: 25 μL reactions contained primer added to a final concentration of 2.5 μ*M*, 0.5 μL of eDNA template (~100 ng), 1× ThermoPol Buffer, 2 m*M* dNTPs, and 0.5 U of *Taq* DNA polymerase. Reactions were cycled using the following touchdown protocol: initial denaturation (95°C, 2 min); 6 touchdown cycles (95°C, 30 sec; 65°C (dt −1°C/cycle), 30 sec; 72°C, 30 sec), 30 standard cycles (95°C, 30 sec; 58°C, 30 sec; 72°C, 30 sec) and a final extension step (72°C, 2 min). Amplicons of the correct predicted size were gel purified and directly sequenced.

General procedure for clone recovery

Individual clones were recovered from a 4,000–5,000-membered sub-library by plating a 10^{-5} or 10^{-6} dilution of the corresponding glycerol stock into 96-well microtiter plates and screening the diluted cultures by whole-cell PCR with primers designed to recognize amplicons detected in the initial screen. PCR positive wells were then either subjected to a second round of dilution plating or plated directly on LB agar with ampicillin (50 μg/mL) to yield distinct colonies that were screened by whole-cell PCR to identify individual clones of interest. Each recovered cosmid was end sequenced using vector-specific (pWEB, pWEB-TNC) universal primers (M13(−40) and the T7 promoter). All clones were fully sequenced using 454 GLX FLX pyrosequencing, assembled using Newbler (Roche), and annotated using Genemark and BLASTX.[26–28] Gene cluster images were generated using MacVector. The amino acid substrate specificity for each adenylation domain found in the cryptic NRPS gene cluster was predicted using NRPSpredictor.[29]

pTARa vector construction

The yeast *ARSH4* (autonomous replicating sequence), *CEN6* (plasmid maintenance element), and *URA3* markers were obtained from pLLX13 by digestion with EcoRI and HindIII.[30] After gel purification, the fragment was ligated into similarly digested pCC1-BAC (Epicentre Biotechnologies). The resulting vector was digested with HpaI and ligated to a DraI fragment from pOJ436 containing an origin of transfer (OriT), integrase and apramycin resistance gene.[31] Transformation into EPI300 *E. coli* (Epicentre Biotechnologies) and selection on chloramphenicol (12.5 μg/mL) and apramycin (50 μg/mL) yielded the capture vector pTARa (TAR-ready BAC with the *Streptomyces* attP integration system, GenBank accession number: GQ452294).

TAR cloning

TAR cloning was initially developed to selectively isolate regions of genomes without the need to construct and screen a genomic library.[18,19,30,32,33] The procedures outlined below describe our adaptation of these methods for the isolation of sequenced natural product gene clusters and the assembly of large natural product biosynthetic gene clusters captured on multiple overlapping eDNA clones.

Pathway-specific capture vector construction

The cycloheximide counter selection cassette (CYH2/*bla*) was PCR amplified using pLLX8 as a template following reported protocols.[30] The cassette was amplified using primers pLLX8/fw/: 5′-TTTTCTAGAACGCGTTTAATTAAAATCTAAAGTATATATGAGTAAAC-3′ and pLLX8/rv/: 5′-CCCTCTAGAGTTAACGTTTAAACAAAAAACGGTGAAAATGGGTGATAG-3′. Each 50 μL reaction contained 1× FailSafe Buffer B (Epicentre Biotechnologies), 2.5 μ*M* of each primer, 100 ng of pLLX8 template, and 1 U of *Taq* DNA polymerase. Reactions were cycled using the following protocol: initial denaturation (95°C, 2 min), 35 standard cycles of (95°C, 30 sec; 65°C, 30 sec; 72°C, 3 min) and a final extension step (72°C, 7 min). The 2.95 kb PCR product was gel purified prior to capture vector assembly (MinElute™, Qiagen). eDNA clone assembly homology arms were PCR amplified in 25 μL reactions containing 100 ng of template cosmid, 2.5 μ*M* of each primer, 1× FailSafe Buffer D (Epicentre Biotechnologies), and 0.5 U *Taq* DNA polymerase. Reactions were cycled using the following protocol: initial denaturation (95°C, 2 min), 35 standard cycles (95°C, 1 min; 60°C, 1 min; 72°C, 1 min) and a final extension step (72°C, 5 min). PCR primers for homology arms were designed to contain 40 bp of homology to the pTARa vector and 40 bp of homology to the counter selection cassette.[30] These homology regions were incorporated to allow pathway-specific capture vector construction using recombination in *S. cerevisiae*.[30] Upstream homology arm amplification primers contained a sense primer extension: 5′-ATATTACCCTGTTATCCCTAGCGTAACTATCGATCTCGAG-3′, and an antisense primer extension: 5′-CATATATACTTTAGATTTTAATTAAACGCGTTCTAGAAAA-3′, which add 40 bp of homology to pTARa and the counter selection cassette, respectively. The downstream targeting sequence sense primer extension is: 5′-CATTTTCACCGTTTTTTGTTTAAACGTTAACTCTAGAGGG-3′, which provides homology to the counter selection cassette and the antisense primer extension is: 5′-TAACAGGGTAATATAGAGATCTGGTACCCTGCAGGAGCTC-3′, which provides homology to pTARa. Each primer pair was designed to yield a 600–900 bp amplicon that acts as a homology arm in a pathway-specific capture vector used for a TAR reassembly reaction.[30] Cosmids X16 and V48 were used as templates to generate upstream and downstream homology arms for the PKS gene cluster. Cosmids ZA41 and J2 were used as templates to generate upstream and downstream homology arms for the NRPS gene cluster. Cosmids 1679 and 201 were used as templates to generate upstream and downstream homology arms for the FRI gene cluster. 300 ng of purified *Citrobacter Koseri* genomic DNA (MasterPure™ Complete DNA Purification Kit, Epicentre Biotechnologies) was used as a template to generate upstream and downstream homology arms for the colibactin gene cluster (GenBank accession number: AM229678). Each PCR amplified component was gel purified prior to its use in the assembly of a pathway-specific capture vector.

For the assembly of a pathway-specific capture vector, 200 ng of pTARa was linearized with NheI and added to 200 μg of heat denatured single stranded carrier DNA (heated to 95°C for 10 min then kept on ice), 600 ng of CYH2/*bla* counter selection cassette amplicon[30] and 200 ng of an upstream and downstream homology arm amplicon pair prepared as described above. All components were added to lithium acetate prepared chemically competent CRY1–2 (uracil deficient, *ura*$^{-}$) yeast, plated on synthetic complete (SC) uracil dropout agar (Invitrogen) and incubated at 30°C.[34] Colonies typically began to appear within 24–48 hours. Assembled capture vectors were isolated in bulk by resuspending yeast colonies from a 100 mm SC dropout agar plate in 5 mL of 1 × phosphate buffered saline. Plasmid DNA was isolated from 1 mL of resuspended cells (ChargeSwitch™ Yeast Plasmid Isolation Kit, Invitrogen). 100 ng of the purified DNA was transformed into electrocompetent EPI300 *E. coli* and plated on LB agar containing ampicillin (100 μg/mL), chloramphenicol (12.5 μg/

mL), and apramycin (50 μg/mL) to yield a pathway-specific capture vector containing a counter-selection cassette.

TAR cloning and pathway assembly

Direct TAR cloning of the colibactin gene cluster from genomic DNA was carried out using reported protocols.[32,33] For eDNA pathway assembly, each cosmid to be used in an assembly reaction was initially linearized by digestion with DraI and the capture vector was linearized by digestion with PmeI. 200 ng of each linearized cosmid and an equimolar amount (~100 ng) of a linearized pathway-specific capture vector were added to 200 μL of *S. cerevisiae* spheroplasts prepared as previously reported.[33] The transformed spheroplasts were added to 7 mL of top agar equilibrated to 50°C (1 *M* sorbitol, 1.92 g/L SC uracil dropout supplement (Invitrogen), 6.7 g/L yeast nitrogen base (Invitrogen), 2% glucose, 2.5% agar). The top agar containing transformed spheroplasts was overlaid onto SC dropout agar containing 2.5 μg/mL cycloheximide. The plates were incubated at 30°C and spheroplast growth was typically seen within 72 hours. The resulting recombinants were patched onto SC uracil dropout agar with cycloheximide (2.5 μg/mL) for overnight growth at 30°C.

For initial PCR detection of reassembled pathways, a small portion of each yeast patch was resuspended in 10 μL of 20 m*M* NaOH and heated at 95°C for 10 minutes. 1.5 μL of the cell lysate was then used as a template in a 50 μL multiplex PCR reaction following the manufacturers directions (Multiplex PCR Kit, Q solution™, Qiagen). The primer sets used in this analysis were designed to recognize unique regions from each overlapping cosmid clone that was used in an assembly reaction. In the colibactin TAR experiment, PCR primer pairs were designed to detect the previously reported boundaries of the biosynthetic gene cluster.[35]

Analysis of TAR recombined clones

Yeast recombinants that produced PCR amplicons of correct size for all portions of a pathway were grown overnight (30°C, 225 RPM) in 2 mL of SC uracil dropout media (or on SC uracil dropout agar) with 2.5 μg/mL cycloheximide and TAR assembled pathways were isolated from these cultures (ChargeSwitch™, Invitrogen). 5 μL of ChargeSwitch™ prepared DNA (1/10 elution volume) was transformed into electrocompetent EPI300 *E. coli* which were outgrown at 30°C for 2 hours (225 RPM) and then plated on LB agar with 12.5 μg/mL chloramphenicol. Whole-cell PCR was used to identify *E. coli* colonies containing correctly reassembled gene clusters. DNA was then isolated from 5 mL cultures of PCR positive *E. coli* transformants using alkaline lysis and isopropanol precipitation (CopyControl™ pCC1-BAC Induction Protocol, Epicentre Biotechnologies). *E. coli* transformants containing the colibactin gene cluster were identified using 8 sets of previously reported PCR primers designed to detect different ORF's in the pathway (data not shown).[35] Detailed restriction mapping was carried out on each reassembled pathway using an enzyme (PKS, EcoRI; NRPS, EcoRI; FRI, BglII; Colibactin, HindIII) that was predicted to yield restriction fragments that could be easily resolved using agarose gel electrophoresis (1% agarose, 0.5x Tris/Borate/EDTA, 30 V, overnight). The lambda HindIII and 50 bp molecular weight makers were obtained from New England Biolabs. Full pathway sequencing for each gene cluster was deposited with GenBank under the following accession numbers: NRPS: GQ475282, FRI: GQ475284, and PKS: GQ475283.

Conjugation and preparation for heterologous expression

Assembled pathways were transformed into S17-1 *E. coli* for conjugation into *Streptomyces* using published protocols.[31] All three reassembled eDNA gene clusters were successfully conjugated and chromosomally integrated into a number of *Streptomyces* including *Streptomyces lividans, Streptomyces albus* and *Streptomyces toyocaensis.*[36]

Results and Discussion

Library size analysis

The genes responsible for the biosynthesis of a natural product are typically clustered on a bacterial chromosome, and therefore theoretically can be cloned on a single continuous fragment of eDNA. While the heterologous expression of biosynthetic gene clusters captured on eDNA-derived clones has begun to yield novel natural products, many natural product gene clusters are too large to be routinely captured on individual eDNA cosmid clones (Figure 1). With metagenomic libraries of sufficient size and sequence coverage, large gene clusters that cannot be captured on a single cosmid clone could be accessed by recovering collections of overlapping eDNA clones. Soil microbiomes are among the most genetically diverse environments characterized to date and are therefore attractive starting points for the discovery of natural products using a metagenomic approach.[3] However, because of this complexity, it is difficult to predict the size a soil-based eDNA library must be to permit the recovery of overlapping clones from a diverse collection of large natural product gene clusters. We set out to empirically investigate this problem using eDNA libraries constructed from two different soil samples. For this study, DNA isolated from soil collected in Utah was used to construct a series of independent 750,000-membered eDNA cosmid libraries (~10,000,000 clones in total) and DNA isolated from a soil sample collected in California was used to construct a series of independent 320,000-membered eDNA cosmid libraries (~15,000,000 clones in total).

The reassembly of large natural product gene clusters from multiple overlapping eDNA fragments begins with the detection of specific sequence(s) of interest located on two or more unique library clones. We therefore wanted to determine the point at which redundant sequences of interest began to regularly appear in unique eDNA library aliquots constructed from the same soil sample. Culture-based studies suggest that type II (aromatic, iterative) PKS biosynthetic systems are common in bacteria and the PKS genes found in these systems are highly conserved. We therefore chose type II PKS pathways as a model system for studying large (>30 kb) gene clusters present in soil-derived eDNA libraries. Both the California and Utah libraries were screened for the presence of β-ketoacyl synthase (KS_β) gene sequences using degenerate PCR primers designed to recognize type II PKS systems.[24,25] In total, 19 distinct KS_β gene sequences were amplified from the Utah library and 73 distinct KS_β gene sequences were amplified from the California library (Figure 2). In the Utah library, redundant KS_β sequences began to regularly appear once ~3×10^6 clones had been examined, while in the California library redundant KS_β sequences began to regularly appear once ~2.25×10^6 clones had been examined. Additional screens using primers designed to recognize other conserved natural product biosynthetic gene sequences have shown similar results. In these studies, redundant sequences begin to regularly appear once libraries exceed 1–3×10^6 clones in size (data not shown).[37] The libraries used in our efforts to recover natural product gene clusters to be used in TAR assembly experiments were therefore expanded until they contained at least 1–1.5×10^7 unique clones, which corresponds to 5–10 times the number of clones needed to identify the first redundant type II PKS sequences. While even an eDNA library of 1–1.5×10^7 clones is unlikely to permit the recovery of rare gene clusters, our analysis suggests that it will likely contain collections of clones encompassing complete PKS gene clusters and, by extension, overlapping clones from many other types of biosynthetic gene clusters found in the genomes of uncultured bacteria.

Natural product gene cluster identification and recovery

In excess of 35,000 unique microbial natural products have been characterized using culture-based methods.[38,39] This amazing assortment of natural products is biosynthesized using a

much smaller number of conserved enzyme families. The structural diversity seen in natural products appears to arise in large part from the natural combinatorial shuffling of these conserved biosynthetic enzyme families.[40] Degenerate primers designed to recognize conserved natural product biosynthetic gene sequences should therefore be useful for identifying eDNA derived gene clusters that encode the biosynthesis of a diverse collection of small molecules. In this study, three different sets of degenerate primers were used to recover three large natural product biosynthetic gene clusters from the Utah and California soil eDNA libraries. A cryptic type II PKS gene cluster was identified using the type II PKS-specific degenerate primers we used in our initial library size analysis.[24,25] A cryptic NRPS gene cluster was identified using degenerate primers designed to amplify flavin-dependent halogenases known to tailor aromatic amino acids found in halogenated nonribosomal peptides. Degenerate primers designed to recognize acyl-CoA ligases found in lipopeptide antibiotic gene clusters were used to identify a gene cluster that is predicted to encode the known metabolite friulimicin. These three eDNA-derived gene clusters are referred to as the PKS, NRPS and FRI gene clusters, respectively. The PKS and NRPS gene clusters were found in an eDNA library derived from topsoil collected in Utah while the FRI gene cluster was found in an eDNA library derived from desert soil collected in California.

Individual cosmid clones containing genes recognized by the degenerate primers used in initial library screens were recovered from the appropriate library and then end sequenced (Figure 3). PCR primers designed against the end sequences were subsequently used to identify and recover overlapping clones from the same library. The process of clone recovery and end sequencing was iteratively repeated until genes predicted to be involved in primary metabolism were found on the distal ends of a recovered cosmid (Figure 3). This initial end-sequencing analysis suggested that the NRPS and FRI gene clusters were recovered on three cosmids each (NRPS: clones ZA41, Q87, J2; FRI: clones 1697, 1451, 201). The PKS gene cluster appeared to be present on two overlapping cosmids (PKS: clones X16, V48).

Each clone that was predicted to be part of a gene cluster was fully sequenced and annotated (Figure 6). The eDNA-derived FRI gene cluster and the friulimicin gene cluster from *A. friuliensis* have the same gene organization and are 89% identical over the 68 kb region that is predicted to comprise the functional biosynthetic pathway.[41] A comparison of these two gene clusters suggests that the entire FRI gene cluster was likely captured on the three overlapping eDNA cosmids that were recovered. While the eDNA-derived PKS and NRPS gene clusters do not closely resemble any known gene clusters, the appearance of primary metabolic enzymes in the sequence surrounding the conserved natural product biosynthetic genes found on these clones suggests they were also likely recovered in their entirety. Sequencing of a fourth overlapping clone that extends 20 kb beyond the NRPS gene cluster found no enzymes associated with secondary metabolism. As suggested by our initial eDNA library size analyses, cosmid libraries containing in excess of 10 million clones appear to provide sufficient coverage of soil metagenomes to allow access to a diverse range of complete natural product biosynthetic gene clusters.

TAR vector design and construction

To facilitate TAR reassembly of large natural product gene clusters as well as subsequent heterologous expression studies with reassembled pathways, we created pTARa, a BAC-based *S. cerevisiae/E. coli/Streptomyces* shuttle capture vector (Figure 4a). This vector contains elements that allow pathways to be assembled in *S. cerevisiae*, characterized and maintained in *E. coli*, and conjugatively transferred into a wide range of *Streptomycetes* for heterologous expression studies.[31] We included these elements to facilitate *Streptomyces*-based heterologous expression studies, but any number of species-specific genetic elements can be incorporated into pTARa to allow the transfer of pathways into a wide variety of

bacterial hosts.[30] As a demonstration of the utility of pTARa as a shuttle vector, we propagated the vector in *S. cerevisiae* (CRY1–2), transformed and isolated the vector from *E. coli* and successfully conjugated into a number of different *Streptomycetes* including *S. toyocaensis, S. lividans,* and *S. albus.*

Capturing natural product gene clusters from sequenced genomes using pTARa

The cloning of natural product gene clusters from cultured organisms traditionally requires the construction and screening of a genomic DNA library.[42,43] Using TAR cloning, a sequenced biosynthetic gene cluster of any size can be directly cloned without the need to construct or screen a genomic library (Figure 5c).[30] To demonstrate the utility of pTARa for culture-based natural products research, we directly cloned the 56 kb colibactin gene cluster directly from genomic DNA isolated from the cultured bacterium, *C. koseri*.[32,33] Previous studies determined the functional boundaries of the colibactin gene cluster via transposon mutagenesis.[35] In order to TAR clone this gene cluster, we simply designed a pathway-specific capture vector using this information (Figure 4b, 5), and co-transformed the capture vector and *C. koseri* genomic DNA into *S. cerevisiae* spheroplasts.[32,33] We screened yeast spheroplasts using colibactin gene cluster specific PCR primers and were able to quickly identify clones containing intact colibactin gene clusters. Detailed restriction mapping of the TAR cloned pathway confirmed that we had specifically cloned the colibacin gene cluster (pTARa-Colibactin) directly from *C. koseri* genomic DNA (Figure 5b).[35] As demonstrated by this experiment, TAR cloning should provide a general and rapid means to access intact natural product biosynthetic gene clusters from sequenced microorganisms without the need to construct or screen a genomic library (Figure 5c).

TAR assembly of multi-clone gene clusters

For each reassembly experiment, we constructed a unique pathway-specific capture vector with homology arms corresponding to sequences at the proximal and distal ends of the gene cluster to be reassembled (Figure 4c, 6). Homologous recombination in *S. cerevisiae* is stimulated by the presence of double stranded breaks adjacent to recombination sites.[18] The individual cosmids to be used in the reassembly of a gene cluster were therefore linearized by restriction digestion with DraI and then co-transformed with a linearized pathway-specific capture vector into competent CRY1–2 *S. cerevisiae*. DraI, which recognizes the AT rich hexamer, TTTAAA, digests the cosmid backbone, yet rarely cuts in GC rich sequences found in biosynthetic gene clusters thus providing a means to generate linear DNA fragments for TAR reassembly reactions. The concentration of the components used in the co-transformation step was empirically determined and selected to yield, on average, one assembled construct per spheroplast. After 3–5 days of recovery on SC uracil dropout agar, recovered spheroplasts were restruck on new SC uracil dropout agar plates. This step is necessary to reduce the chance of cross contamination caused by DNA from the TAR reaction during the PCR analysis that is used to identify yeast colonies with assembled gene clusters. Yeast colonies were then screened using multiplex PCR with primers specific to each unique cosmid fragment predicted to be present in a re-assembled gene cluster construct. Between 30–70% of the yeast colonies were found to be PCR positive for all fragments predicted to be present in a pathway. Using this approach we were able to rapidly identify yeast colonies that contained intact biosynthetic gene clusters.

Large constructs isolated from PCR positive yeast clones were electroporated into *E. coli* and analyzed by detailed restriction analysis (Figure 6). In each case, the large construct obtained from a TAR reassembly reaction produced a restriction map that was identical to the map predicted to arise from assembling the individual overlapping clones used in the reaction (Figure 6). The 39 kb PKS gene cluster was successfully subcloned from the central region of cosmids X16 and V48, two cosmids that contain 2.1 kb of overlap. The entire 89

kb cryptic NRPS gene cluster was successfully reconstructed in a single *S. cerevisiae* spheroplast transformation reaction from three overlapping eDNA cosmid clones. In a similar fashion, we reassembled the 90 kb eDNA-derived FRI gene cluster using a single *S. cerevisiae* spheroplast transformation reaction and three overlapping eDNA-derived cosmid clones.

While the PKS and NRPS gene clusters were initially assembled from fully sequenced sets of cosmids, reassembly experiments can also be performed in the absence of comprehensive sequencing. The FRI gene cluster was originally reassembled with only end-sequencing data for each cosmid clone predicted to comprise the complete gene cluster (Figure 4c, 6). A capture vector based on the end-sequencing data from the distal ends of the two outermost clones, cosmids 1679 and 201, was used to reassemble the gene cluster (Figure 3). We confirmed the successful reassembly of the fragments using PCR and by comparing restriction maps of the reassembled construct with those produced by the cosmids used in the reassembly experiment (data not shown). Subsequent full sequencing of the clones comprising the FRI gene cluster confirmed the restriction mapping and successful sequencing-independent TAR assembly experiment (Figure 6).

Traditional gene cluster assembly strategies can become technically impractical when working with large naturally derived DNA sequences. Unique and conveniently located restriction sites needed for traditional “cut and paste” strategies are often not available when working with long natural DNA sequences. Recently, lambda-based recombination has been used to reconstruct functional gene clusters, circumventing many of the problems associated with traditional strategies.[12] Lambda-based recombination becomes difficult, however, for large gene clusters captured on multiple overlapping clones because it requires the step-wise recombination of two clones at a time. This step-wise recombination process requires the introduction of a unique selectable marker into each fragment used in an assembly experiment. As demonstrated here, TAR-dependent assembly of multi-clone natural product gene clusters can be performed in a single reaction without any of these barriers. The maximum number of DNA fragments that can be simultaneously assembled in TAR experiments has yet to be determined, but even the largest gene clusters are unlikely to require more than 3 or 4 overlapping cosmids which is well within the established limits of TAR.[20,21,44]

Conclusions

Previous studies have demonstrated that metagenomic strategies can be used to uncover metabolites encoded by gene clusters captured on individual soil-derived eDNA clones (Figure 1). Cloning large natural product gene clusters presents a challenge for both culture dependent and culture independent studies. We have shown that T AR can be used to rapidly reassemble overlapping eDNA-derived clones into a single construct containing large eDNA derived natural product gene clusters. We have also shown that TAR can be used to directly and specifically clone natural product gene clusters from sequenced organisms without constructing and screening a genomic library. TAR-dependent assembly of natural product gene clusters from overlapping clones found in eDNA soil-libraries provides an experimental framework for rapidly accessing intact natural product gene clusters that exceed conventional eDNA cloning limits (Figure 1b). In doing so, it eliminates one of the major roadblocks associated with current metagenomic natural product discovery efforts. In this study, this experimental approach provided access to both a new example of what was thought to be a rare gene cluster (FRI) as well as what appear to be new gene clusters (PKS, NPRS). The heterologous expression of large TAR-assembled gene clusters should form a basis for the identification of new natural products from eDNA. The major remaining challenge to the discovery of new natural products from uncultured bacteria, that of

heterologous expression, is not unique to culture-independent studies and will likely need to be addressed using many different gene cluster specific strategies.

Acknowledgments

We thank Stephen Lory (Harvard Medical School) for kindly providing plasmids pLLX8 and pLLX13, Daniel Gibson (J. Craig Venter Institute) for experimental suggestions and the core genomics facility at Memorial Sloan Kettering Cancer Center (MSKCC) for 454 sequencing.

This work was supported by the Howard Hughes Medical Institute, NIH GM077516 and by the Beckman and Searle Foundations.

References

1. Newman DJ, Cragg GM. J Nat Prod. 2004; 67:1216–1238. [PubMed: 15332835]
2. Newman DJ, Cragg GM. J Nat Prod. 2007; 70:461–477. [PubMed: 17309302]
3. Gans J, Wolinsky M, Dunbar J. Science. 2005; 309:1387–1390. [PubMed: 16123304]
4. Rappe MS, Giovannoni SJ. Annu Rev Microbiol. 2003; 57:369–394. [PubMed: 14527284]
5. Torsvik V, Goksoyr J, Daae FL. Appl Environ Microbiol. 1990; 56:782–787. [PubMed: 2317046]
6. Torsvik V, Ovreas L, Thingstad TF. Science. 2002; 296:1064–1066. [PubMed: 12004116]
7. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Chem Biol. 1998; 5:R245–249. [PubMed: 9818143]
8. Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG, Ravel J. Proc Natl Acad Sci U S A. 2005; 102:7315–7320. [PubMed: 15883371]
9. Guan C, Ju J, Borlee BR, Williamson LL, Shen B, Raffa KF, Handelsman J. Appl Environ Microbiol. 2007; 73:3669–3676. [PubMed: 17435000]
10. Liles MR, Williamson LL, Rodbumrer J, Torsvik V, Goodman RM, Handelsman J. Appl Environ Microbiol. 2008; 74:3302–3305. [PubMed: 18359830]
11. Zhang Y, Buchholz F, Muyrers JP, Stewart AF. Nat Genet. 1998; 20:123–128. [PubMed: 9771703]
12. Wenzel SC, Gross F, Zhang Y, Fu J, Stewart AF, Muller R. Chem Biol. 2005; 12:349–356. [PubMed: 15797219]
13. Holt RA, Warren R, Flibotte S, Missirlis PI, Smailus DE. Bioessays. 2007; 29:580–590. [PubMed: 17508395]
14. Sharan SK, Thomason LC, Kuznetsov SG, Court DL. Nat Protoc. 2009; 4:206–223. [PubMed: 19180090]
15. Thomason, L.; Court, DL.; Bubunenko, M.; Costantino, N.; Wilson, H.; Datta, S.; Oppenheim, A. Curr Protoc Mol Biol. Vol. Chapter 1. 2007. p. 16
16. Sawitzke JA, Thomason LC, Costantino N, Bubunenko M, Datta S, Court DL. Methods Enzymol. 2007; 421:171–199. [PubMed: 17352923]
17. Court DL, Sawitzke JA, Thomason LC. Annu Rev Genet. 2002; 36:361–388. [PubMed: 12429697]
18. Larionov V, Kouprina N, Eldarov M, Perkins E, Porter G, Resnick MA. Yeast. 1994; 10:93–104. [PubMed: 8203155]
19. Larionov V, Kouprina N, Graves J, Resnick MA. Proc Natl Acad Sci U S A. 1996; 93:13925–13930. [PubMed: 8943037]
20. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA 3rd, Smith HO. Science. 2008; 319:1215–1220. [PubMed: 18218864]
21. Gibson DG, Benders GA, Axelrod KC, Zaveri J, Algire MA, Moodie M, Montague MG, Venter JC, Smith HO, Hutchison CA 3rd. Proc Natl Acad Sci U S A. 2008; 105:20404–20409. [PubMed: 19073939]
22. Shao Z, Zhao H. Nucleic Acids Res. 2009; 37:e16. [PubMed: 19074487]
23. Brady SF. Nat Protoc. 2007; 2:1297–1305. [PubMed: 17546026]

24. King RW, Bauer JD, Brady SF. Angew Chem Int Ed Engl. 2009; 48:6257–6261. [PubMed: 19621341]
25. Seow KT, Meurer G, Gerlitz M, Wendt-Pienkowski E, Hutchinson CR, Davies J. J Bacteriol. 1997; 179:7360–7368. [PubMed: 9393700]
26. Lukashin AV, Borodovsky M. Nucleic Acids Res. 1998; 26:1107–1115. [PubMed: 9461475]
27. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Nature. 2005; 437:376–380. [PubMed: 16056220]
28. Tatusov RL, Altschul SF, Koonin EV. Proc Natl Acad Sci U S A. 1994; 91:12091–12095. [PubMed: 7991589]
29. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. Nucleic Acids Res. 2005; 33:5799–5808. [PubMed: 16221976]
30. Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E, Olavarietta R, Doud M, Smith RS, Montgomery P, White JR, Godfrey PA, Kodira C, Birren B, Galagan JE, Lory S. Proc Natl Acad Sci U S A. 2008; 105:3100–3105. [PubMed: 18287045]
31. Bierman M, Logan R, O'Brien K, Seno ET, Rao RN, Schoner BE. Gene. 1992; 116:43–49. [PubMed: 1628843]
32. Kouprina N, Larionov V. Nat Protoc. 2008; 3:371–377. [PubMed: 18323808]
33. Kouprina N, Noskov VN, Larionov V. Methods Mol Biol. 2006; 349:85–101. [PubMed: 17071976]
34. Gietz RD, Schiestl RH. Nat Protoc. 2007; 2:31–34. [PubMed: 17401334]
35. Nougayrede JP, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. Science. 2006; 313:848–851. [PubMed: 16902142]
36. Tobias Kieser, MJB.; Buttner, Mark J.; Chater, Keith F.; Hopwood, David A. Practical Streptomyces Genetics. John Innes Centre; Colney, Norwich NR4 7UH, England: 2000.
37. Banik JJ, Brady SF. Proc Natl Acad Sci U S A. 2008; 105:17273–17277. [PubMed: 18987322]
38. Buckingham, J. Dictionary of Natural Products. CRC Press; London: 2007.
39. Laatsch, H.; Laatsch, H., editors. Wiley-VCH; 2009.
40. Dewick, PM. Medicinal Natural Products: A Biosynthetic Approach. John Wiley and Sons Ltd; West Sussex, England: 2002.
41. Muller C, Nolden S, Gebhardt P, Heinzelmann E, Lange C, Puk O, Welzel K, Wohlleben W, Schwartz D. Antimicrob Agents Chemother. 2007; 51:1028–1037. [PubMed: 17220414]
42. Miao V, Coeffet-Legal MF, Brian P, Brost R, Penn J, Whiting A, Martin S, Ford R, Parr I, Bouchard M, Silva CJ, Wrigley SK, Baltz RH. Microbiology. 2005; 151:1507–1523. [PubMed: 15870461]
43. McHenney MA, Hosted TJ, Dehoff BS, Rosteck PR Jr, Baltz RH. J Bacteriol. 1998; 180:143–151. [PubMed: 9422604]
44. Gibson DG. Nucleic Acids Res. 2009; 37:6984–6990. [PubMed: 19745056]
45. Wang GY, Graziani E, Waters B, Pan W, Li X, McDermott J, Meurer G, Saxena G, Andersen RJ, Davies J. Org Lett. 2000; 2:2401–2404. [PubMed: 10956506]
46. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM, Handelsman J. Appl Environ Microbiol. 2002; 68:4301–4306. [PubMed: 12200279]
47. Brady SF, Clardy J. Angew Chem Int Ed Engl. 2005; 44:7063–7065. [PubMed: 16206308]
48. Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G, Martinez A, Kolvek SJ, Hopke J, Osburne MS, August PR, Nalin R, Guerineau M, Jeannin P, Simonet P, Pernodet JL. Appl Environ Microbiol. 2003; 69:49–55. [PubMed: 12513976]

49. Lim HK, Chung EJ, Kim JC, Choi GJ, Jang KS, Chung YR, Cho KY, Lee SW. Appl Environ Microbiol. 2005; 71:7768–7777. [PubMed: 16332749]
50. Brady SF, Chao CJ, Handelsman J, Clardy J. Org Lett. 2001; 3:1981–1984. [PubMed: 11418029]
51. Brady SF, Clardy J. J Nat Prod. 2004; 67:1283–1286. [PubMed: 15332842]
52. Brady SF, Chao CJ, Clardy J. J Am Chem Soc. 2002; 124:9968–9969. [PubMed: 12188643]
53. Brady SF, Clardy J. Org Lett. 2005; 7:3613–3616. [PubMed: 16092832]
54. Brady SF, Chao CJ, Clardy J. Appl Environ Microbiol. 2004; 70:6865–6870. [PubMed: 15528554]
55. Clardy J, Brady SF. J Bacteriol. 2007; 189:6487–6489. [PubMed: 17586635]
56. Thompson, JD.; Gibson, TJ.; Higgins, DG. Curr Protoc Bioinformatics. Vol. Chapter 2. 2002. p. 3

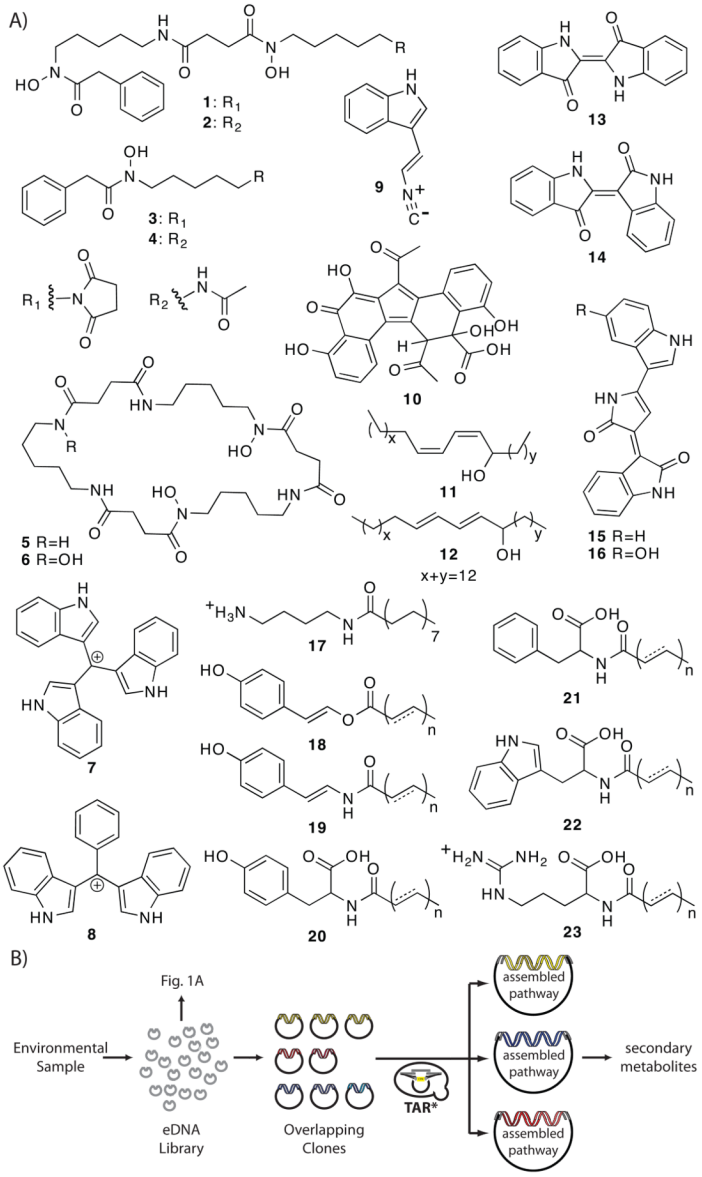

Figure 1.

(A) Natural products that have been isolated and characterized using metagenomic methods have all been derived from single clones. These include terragines A–E (**1–5**)[45], norcardamine (**6**)[45], turbomycin A (**7**) and B (**8**)[46], a C3-isocyanide functionalized indole derivative (**9**)[47], erdacin (**10**)[24], aliphatic dienic alcohol isomers (**11, 12**)[48], indirubin (**13**)[9,49], indigo (**14**)[9,49], deoxyviolacein (**15**)[50], violacein (**16**)[50], palmitoylputrescine (**17**)[51], long chain enol esters (**18**)[52], long chain eneamides (**19**)[52], and various long chain *N*-acyl amino acids (**20–23**).[53–55] (B) TAR-based gene cluster reassembly strategies can provide access to larger natural product gene clusters captured on overlapping eDNA clones.

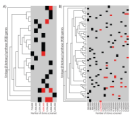

Figure 2.

Degenerate primers targeting minimal type II PKS genes were used to identify KS_{β} sequences present in unique eDNA library aliquots constructed from soil samples collected in Utah (A) and California (B). ClustalW[56] derived phylogenetic trees of the KS_{β} sequences identified in these screens are shown. The aliquots from which sequences were amplified and the point at which they began to reappear in the library (red) are shown as a heatmap.

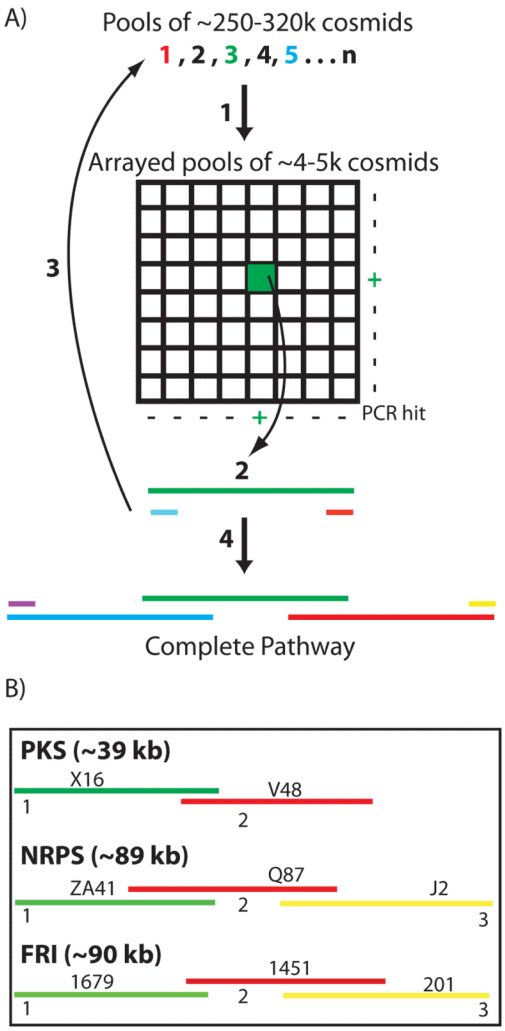

Figure 3.

(A) PCR with degenerate primers was used to identify biosynthetic genes of interest in large library pools (1) and then to subsequently locate these same sequences in arrays of smaller library aliquots (+). Whole cell PCR of serially diluted smaller library aliquots was used to recover individual cosmids of interest (2). Overlapping clones were iteratively recovered (3) until complete biosynthetic pathways were identified (4). (B) The topology of the overlapping clones that are predicted to comprise the eDNA derived PKS, NRPS and FRI gene clusters is shown.

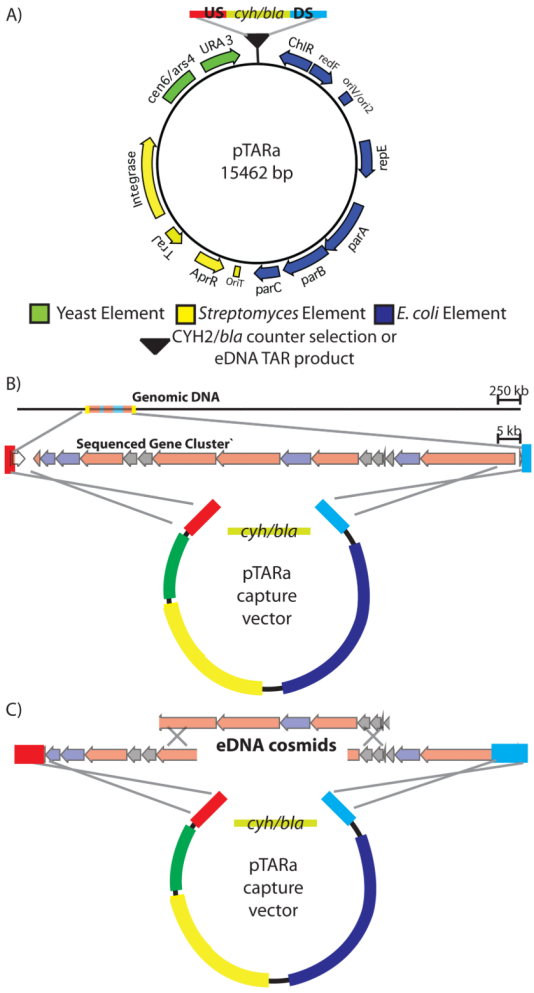

Figure 4.

(A) pTARa contains elements that allow for the rapid assembly and propagation of pathways in *S. cerevisiae (*green), the transformation and analysis of these pathways in *E. coli* (blue) and the integrative conjugation of assembled pathways into *Streptomyces* (yellow). For capture vector construction, pathway-specific upstream (US-blue), and downstream (DS-red) homology arms, as well as a counter selection cassette (*cyh/bla*) are incorporated into the capture vector.[30,32] During recombination, the counter selection cassette is exchanged for a TAR cloned gene cluster (B) or TAR reassembled eDNA pathway (C).

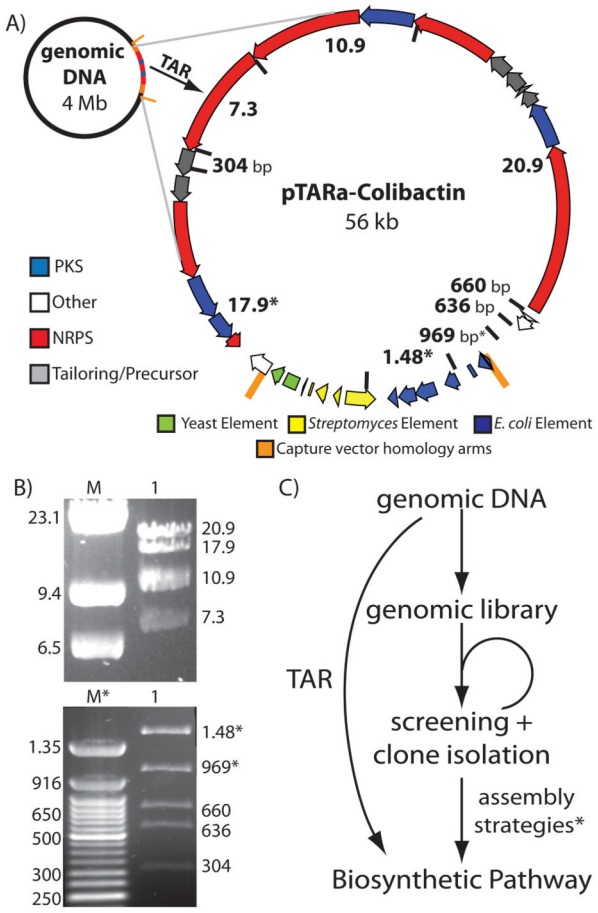

Figure 5.

(A) We used pTARa to directly and specifically clone the colibactin gene cluster from *C. koseri* genomic DNA. Predicted HindIII cut sites and restriction fragment sizes are marked on the map of the pTARa-Colibactin construct. The size of the gene cluster is listed. (B) The experimentally determined HindIII restriction map of pTARa-Colibactin is shown. Two images of the same digest were taken at different points during electrophoresis to highlight fragment sizes more clearly (M=Lambda HindIII digest, M*=50 bp ladder). (C) TAR cloning of gene clusters circumvents the need to construct and screen a genomic library.

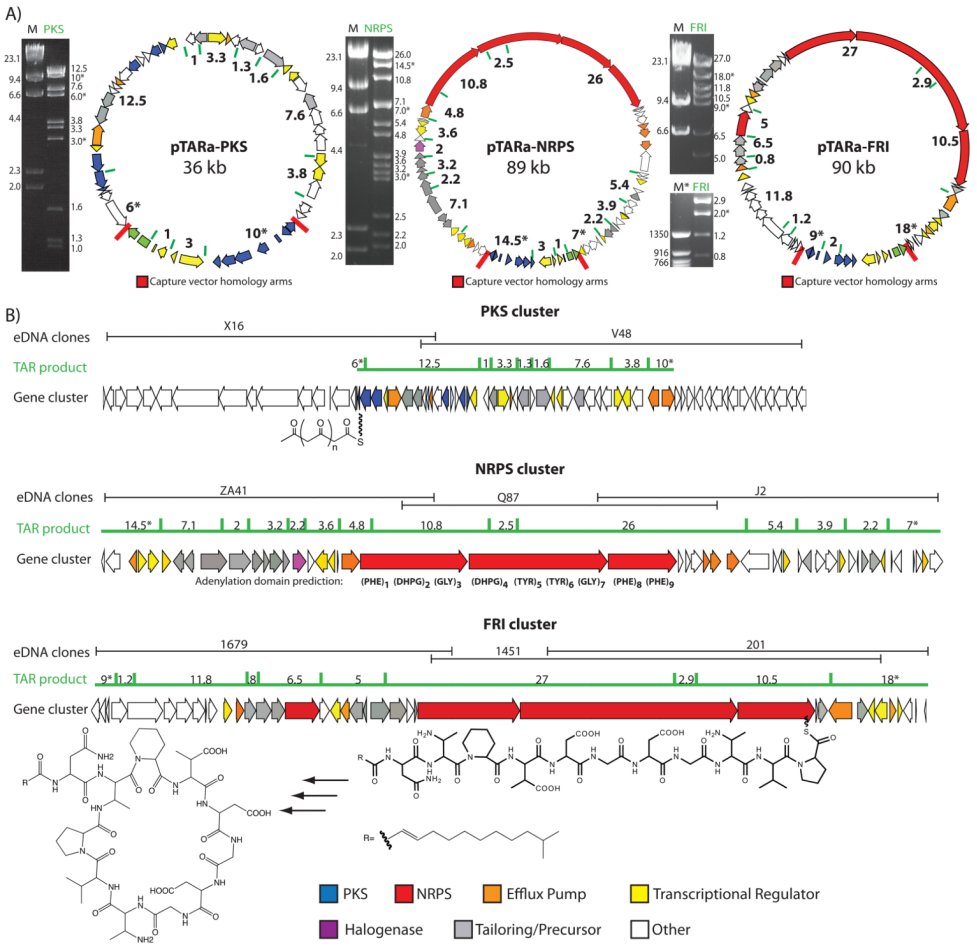

Figure 6.

(A) Experimentally determined restriction maps and predicted restriction enzyme cut sites for each reconstructed gene cluster are shown. The size of each gene cluster is listed for clarity. (M=Lambda HindIII digest, M*=50 bp ladder). B) The overlapping cosmids (black) comprising a complete biosynthetic pathway are shown above the region targeted for TAR assembly (green line). The individual building blocks that are predicted to be used by the conserved modules (PKS and NRPS) found in these biosynthetic pathways appear below each gene cluster (DHPG = dihydroxyphenylglycine).[29]