APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Jan. 2003, p. 49–55 Vol. 69, No. 1
0099-2240/03/$08.00+0 DOI: 10.1128/AEM.69.1.49–55.2003

Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products

Sophie Courtois,[1,2]† Carmela M. Cappellano,[2] Maria Ball,[3] Francois-Xavier Francou,[3] Philippe Normand,[1] Gérard Helynck,[2] Asuncion Martinez,[4] Steven J. Kolvek,[4] Joern Hopke,[4] Marcia S. Osburne,[4]* Paul R. August,[4] Renaud Nalin,[1]‡ Michel Guérineau,[3] Pascale Jeannin,[2] Pascal Simonet,[1] and Jean-Luc Pernodet[3]

Laboratoire d'Ecologie Microbienne du Sol, UMR CNRS 5557, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex,[1] Aventis Pharma, Centre de Recherche de Vitry-Alfortville, 94403 Vitry sur Seine Cedex,[2] and Institut de Génétique et Microbiologie, UMR CNRS 8621, Université Paris Sud XI, 91405 Orsay Cedex, France,[3] and Aventis Pharmaceuticals Inc., Cambridge Genomics Center, Cambridge, Massachusetts 02139[4]

Received 10 July 2002/Accepted 1 October 2002

To further explore possible avenues for accessing microbial biodiversity for drug discovery from natural products, we constructed and screened a 5,000-clone "shotgun" environmental DNA library by using an *Escherichia coli-Streptomyces lividans* shuttle cosmid vector and DNA inserts from microbes derived directly (without cultivation) from soil. The library was analyzed by several means to assess diversity, genetic content, and expression of heterologous genes in both expression hosts. We found that the phylogenetic content of the DNA library was extremely diverse, representing mostly microorganisms that have not been described previously. The library was screened by PCR for sequences similar to parts of type I polyketide synthase genes and tested for the expression of new molecules by screening of live colonies and cell extracts. The results revealed new polyketide synthase genes in at least eight clones. In addition, at least five additional clones were confirmed by high-pressure liquid chromatography analysis and/or biological activity to produce heterologous molecules. These data reinforce the idea that exploiting previously unknown or uncultivated microorganisms for the discovery of novel natural products has potential value and, most importantly, suggest a strategy for developing this technology into a realistic and effective drug discovery tool.

Intensive screening of microbial isolates over the last 50 years has resulted in the commercialization of numerous biomolecules, the products of microbial secondary metabolism. However, recent progress in molecular microbial ecology has shown that microbial diversity in nature is far greater than that reflected in laboratory strain collections, since only a very small fraction of the total bacterial community can be cultured under standard laboratory conditions (2). The accepted phylogenetic definitions of major microbial groups, including archaea, have been dramatically altered by cultivation-independent analyses of microbial rRNA sequences; many newly detected but uncultivated microbial groups may represent major components of indigenous microbial communities. The vast majority of microorganisms in environmental samples remain unexplored and unknown (13, 18, 20), since access to this enormous reservoir of secondary metabolite producers has been hindered by the difficulty of culturing most of them.

A new technology developed to overcome the difficulties of culturing microorganisms involves extracting DNA directly from natural bacterial environments and cloning and expressing that DNA in surrogate expression hosts (16, 19, 20). This approach allows access to the DNA of whole bacterial communities, with modifications of standardized cloning techniques being used to form gene libraries. Environmental DNA libraries that recover functional genes from uncultivated bacteria provide a promising drug discovery tool that requires validation and optimization. The recovery of bacterial DNA from complex environments has been achieved either by direct extraction of DNA from soil (in situ lysis of bacteria) or from preisolated (but uncultivated) bacteria. An assessment of possible cloning bias resulting from these two techniques has shown that although there are some variations in the amount and quality of the DNA recovered, no major phylogenetic bias is observed when these methods are compared (7).

Recent studies have begun to confirm the enormous potential of this technology for discovering new enzymes and small molecules (16, 19, 22, 23, 28). In the study reported here, we constructed an environmental DNA library containing large DNA inserts, with an average size of 50 kb, in order to enhance the chances of isolating gene clusters encoding biosynthetic pathways for secondary metabolites. Our library comprised 5,000 environmental DNA clones in a cosmid vector. We assessed the diversity of the cloned DNA and screened the clones in various ways to analyze their genetic content and to increase our understanding of the expression capabilities of the surrogate host strains. We found that our library encoded several new polyketide synthase (PKS) genes and expressed several

* Corresponding author. Mailing address: Aventis Cambridge Genomics Center, 26 Landsdowne St., Cambridge, MA 02139. Phone: (617) 768-4101. Fax: (617) 374-8811. E-mail: drothstein@rcn.com.

† Present address: ONDEO Services, Centre International de Recherche sur l'Eau et l'Environnement, 78 230 Le Pecq, France.

‡ Present address: LIBRAGEN, 69622 Villeurbanne Cedex 2, France.

new compounds. These results provide valuable information regarding the feasibility of this type of approach for accessing microbial diversity and help lead to an understanding of how best to convert this approach into a realistic drug discovery effort.

MATERIALS AND METHODS

Bacterial strains and DNA manipulations. *Escherichia coli* strains DH10B, DH5α, and TOP10 were obtained from Invitrogen Life Technologies, Carlsbad, Calif. *Streptomyces lividans* TK24 was obtained from The John Innes Centre Collection, Norwich, United Kingdom.

Unless otherwise indicated, all DNA manipulations were performed according to standard protocols (21).

Construction of *E. coli-S. lividans* shuttle cosmid pOS700I. pOS700I was constructed from cosmid pWED1, a pWE15 derivative (11, 27) which encodes an ampicillin resistance gene for selection in *E. coli*, the ColE1 origin of replication for maintenance in *E. coli*, and the *cos* sequence for packaging in λ phage particles. pOS700I is integrative in *S. lividans* via the *attP* site and the *int* gene from the *Streptomyces* integrative element pSAM2, allowing site-specific integration of pOS700I in many *Streptomyces* species (24). pOS700I also carries the *Whyg* gene cassette (5), which confers hygromycin resistance in both *S. lividans* and *E. coli*.

Transformation of *S. lividans*. Cosmid DNA from *E. coli* clones was extracted by using a semiautomated procedure. After manual lysis, the final extraction step was carried out by means of a Qiagen BioRobot according to the manufacturer's recommendations. Standard protocols (15) were followed for the transformation of *S. lividans*, except that we inoculated 25 ml of YEME medium (12) with 1.8×10^9 *S. lividans* spores. Under these conditions, the transformation efficiency was up to 10^4 CFU/μg of DNA with the insert-free cosmid vector pOS700I.

Extraction of soil DNA and construction of a cosmid soil DNA library. Soil samples were obtained from the upper 5 to 10 cm of an arable field in La Cote Saint Andre (Isere, France). Soil from this site was described and used in a previous diversity study (7). The soil was a sandy loam, pH 5.6, with an organic matter content of 40.6 g/kg of dry soil. After all visible roots were removed, the soil was sieved through 2-mm mesh and stored at 4°C. Cells and soil particles were separated by high-speed centrifugation on a Nycodenz density gradient (Nycomed Pharma AS, Oslo, Norway) as follows. Five grams of soil was suspended in 50 ml of 0.9% NaCl and homogenized in a Waring blender by three 1-min pulses at full speed, with cooling on ice every minute. Twenty milliliters of the soil suspension was applied to a Nycodenz density gradient as described previously (7). The cell pellet was resuspended in 10 mM Tris-500 mM EDTA (pH 8.0). Cells were lysed in a lysozyme-achromopeptidase solution (5 and 0.5 mg/ml, respectively) for 1 h at 37°C. Lauryl sarcosyl (final concentration, 1%) was added, and the solution was incubated at 60°C for 30 min. The DNA solution was then purified on a cesium chloride-ethidium bromide density gradient (35,000 rpm in a 65.13 Kontron rotor for 36 h at room temperature).

To avoid the need for digesting environmental DNA before cloning (which could reduce the insert size and possibly introduce bias based on G+C content), an alternative strategy was adopted in which terminal transferase was used to add polynucleotide tails to the 3′ ends of the insert and vector DNAs. Five micrograms of purified, uncut soil DNA was incubated with 35 U of terminal deoxynucleotidyl transferase (Amersham Phamacia Biotech) and 1.5 mM dTTP according to the manufacterer's directions. Similarly, 7.5 μg of shuttle cosmid pOS700I, linearized with *Hin*dIII, was incubated with 25 U of terminal deoxynucleotidyl transferase and 5 mM dATP. Two microliters of treated vector was mixed with 10 μl of treated soil DNA, and the mixture was incubated for 15 min at 65°C and then for 2 h at 57°C. DNA was then packaged into λ phage particles, which were used to infect *E. coli* cells.

We determined the insert-vector junction sequences of 17 recombinant cosmids by using primer 5′-CCGCGAATTCTCATGTTGACCG, which is complementary to the vector sequence between the *Bam*HI and *Hin*dIII sites. We thus determined that the homopolymeric tails were 12 to 60 nucleotides long.

PCR amplification, cloning, and sequencing of 16S rRNA genes. Cosmids extracted from pools of library clones were used as templates for the amplification of 16S rRNA genes with universal primers 63f (5′-CAGGCCTAACACATGCAAGTC-3′) and 1387r (5′GGGCGGWGTGTACAAGGC-3′) (17). Amplification products (each approximately 1.3 kb) were cloned and sequenced, and sequences were analyzed for diversity by comparison with known sequences by using Blast 2.0 (1) and other standard software.

PCR amplification, cloning, and sequencing of PKS I gene sequences. Cosmid DNA was extracted from library pools of 96 clones by using a Qiagen plasmid mini kit. For PCR amplification, DNA (100 to 500 ng) from each pool was used as a template. Two primer sets complementary to highly conserved regions of type I PKS (PKS I) genes from actinomycetes and flanking the active site of the enzyme were used to screen the library for the presence of homologous PKS I genes: set 1—sense, 5′-CCSCAGSAGCGCSTSTTSCTSGA-3′, and antisense, 5′-GTSCCSGTSCCGTGSGTSTCSA-3′; set 2—sense, 5′-CCSCAGSAGCGCSTSCTSCTSGA-3′, and antisense, 5′-GTSCCSGTSCCGTGSGCCTCSA-3′. The specificities of the two primer sets were confirmed by testing with a collection of polyketide-producing strains (*Streptomyces coelicolor* ATCC 101478, *Streptomyces ambofaciens* NRRL2420, *Streptomyces lactamandurans* ATCC 27382, and *Streptomyces rimosus* ATCC 109610) and nonhomologous strains (*Bacillus subtilis* ATCC 6633 and *Bacillus licheniformis* from an in-house strain collection). PCR mixtures for amplification of the soil cosmid DNA library contained 200 μM deoxynucleoside triphosphates, 2.5 mM $MgCl_2$, 7% dimethyl sulfoxide, Qiagen buffer, 0.4 μM each primer, and 2.5 U of Hot-Start *Taq* polymerase (Qiagen). The amplification reactions were performed with a Robocycler (Biometra) by using 15 min of initial denaturation at 95°C, 30 cycles of 1 min of denaturation at 95°C and 1 min of hybridization at 65°C for the first cycle and 62°C for the remaining cycles, 1 min of elongation at 72°C, and 10 min of extension at 72°C. The resulting PCR products (about 700 bp) were purified on agarose gels (Qiagen gel extraction kit) and then subcloned and transformed into *E. coli* TOP10. Plasmid DNA encoding the subclones was then extracted by alkaline lysis with a BioRobot and dialyzed for 2 h by using a 0.025-μm-pore-size VS membrane (Millipore). Samples were sequenced with M13 primers by using an ABI model 377 automated sequencer (Perkin-Elmer). The resulting sequences were compared with the nonredundant sequence database at the National Center for Biotechnology Information (NCBI) by using BLAST.

Sequencing of cosmid insert DNA. Cosmid inserts were sequenced by using either a transposon-mediated or a "shotgun" subcloning method. For the former, an apramycin-resistant version of plasmid pGPS (New England BioLabs, Beverly, Mass.) was used essentially according to the manufacturer's instructions. Transformants were grown with 100 μg of apramycin/ml. For the latter, 1.5-kb DNA fragments were generated by sonicating recombinant cosmids. These fragments were then subcloned into vector pBluescript KSII(+) *Sma*I after treatment with T4 DNA polymerase and the Klenow fragment to obtain blunt ends. Cosmid DNA was isolated by using a BioRobot (model 9600; Qiagen) according to the manufacturer's instructions, and sequencing reactions were performed with purified plasmid DNA by using ABI Big Dye at one-quarter strength and an ABI model 3700 DNA sequencer. The DNA sequence was determined with the UNIX program Phred. Sequence data were assembled by using the program Phrap and edited by using the program Consed (8, 9, 10). Open reading frames (ORFs) were identified and translated by using the program ORF Finder at NCBI. BLAST and PSI-BLAST analyses were also performed at NCBI.

Colony screening. For antibacterial assays, clones were grown on Luria-Bertani agar plates with ampicillin (50 μg/ml) or on ATCC medium 765 agar plates with ampicillin (25 μg/ml) for 6 days at 30°C. Plates were then overlaid with top agar containing exponentially growing *B. subtilis* strain BR151(pPL608) (Bacillus Genetic Stock Center, Columbus, Ohio) and incubated overnight at 37°C and then at room temperature for several more days. Clones producing antibacterial activities were identified by a zone of inhibition in the lawn surrounding the clone. For kanamycin resistance assays, *E. coli* colonies were plated on Luria-Bertani agar containing kanamycin (50 μg/ml) and incubated at 30°C for up to 6 days.

Preparation of culture extracts for high-pressure liquid chromatography (HPLC) screening. For *E. coli*, 1 ml of cosmid-containing cells grown in 3 ml of TB medium (25) with ampicillin (50 μg/ml) for 7 h at 37°C was used to inoculate 20 ml each of TB medium and ATCC medium 765 in 100-ml Erlenmeyer flasks. Cells were grown for 36 h at 37°C, frozen at −30°C, and lyophilized. For *S. lividans* cosmid transformants, colonies isolated on R2YE medium (26) plus hygromycin (200 μg/ml) were inoculated into liquid growth medium (5 g of peptone/liter, 5 g of yeast extract/liter, 5 g of meat extract/liter, 15 g of glucose/liter, 3 g of $CaCO_3$/liter, 5 g of NaCl/liter, 1 g of agar/liter, 50 μg of hygromycin/ml). After 3 days of incubation at 28°C (220 rpm), 2 ml of this preculture was inoculated into 50 ml of the same medium in a 250-ml Erlenmeyer flask and incubated as described above. Cultures were then prepared on three agar media known to support the production of secondary metabolites in actinomycetes: medium 1—10 g of corn steep (Roquette)/liter, 50 g of starch (Glucidexp; Roquette)/liter, 23 g of corn germ oil/liter, 225 g of KH_2PO_4 (Prolabo)/liter, 10 g of NaCl/liter; 5 g of $(NH_4)_2SO_4$/liter, 5.4 g of $CaCO_3$/liter, 6.8 g of dry yeast/liter, 20 g of agar (Difco)/liter; medium 2—33.3 g of soya flour/liter, 10.4 g of starch (Fluitex)/liter, 18.9 g of glucose/liter, 0.5 g of $ZnSO_4 \cdot 7H_2O$/liter, 7.8 g of $CaCO_3$/liter, 4 g of dry yeast/liter, 20 g of agar (Difco)/liter; medium 3—20 g of glucose/liter, 5 g of yeast extract (Difco)/liter, 4 g of $CaCO_3$/liter, 3 g of

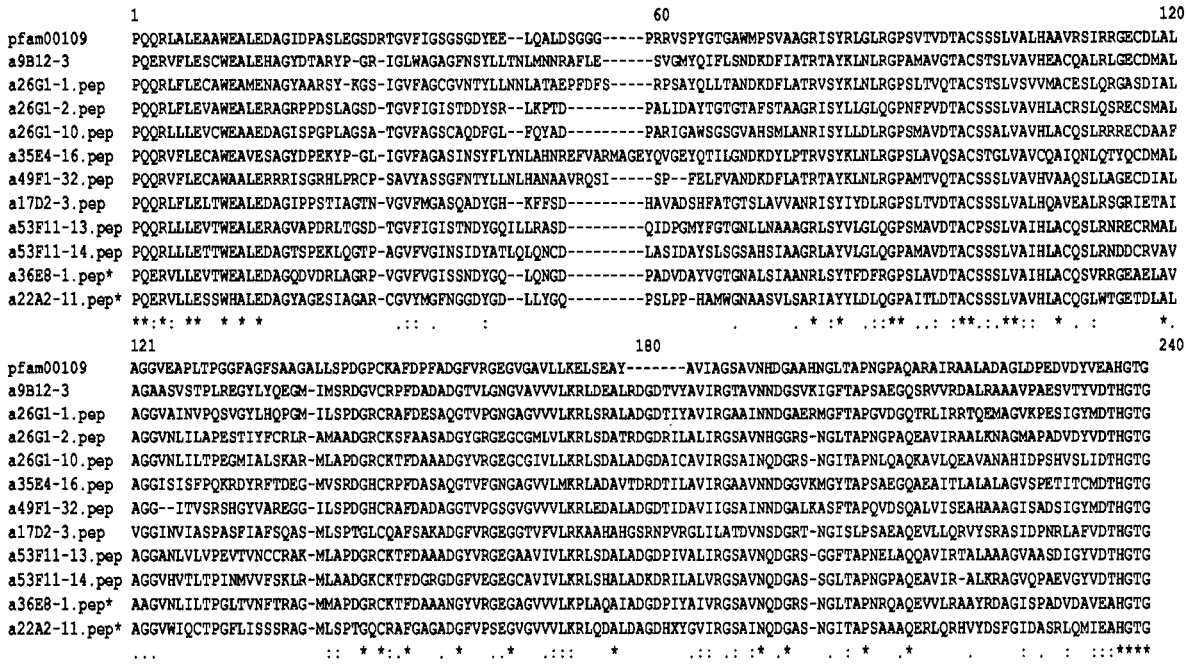

FIG. 1. ClustalX alignment of predicted amino acid sequences of soil PKS I genes and gene fragments. The PKS I consensus sequence pfam00109 is indicated. Soil genes and identities were as follows: a9B12-3, 54% identity to *Nostoc* sp. strain GSV224 NosB gene (227-amino-acid [aa] alignment; GenBank accession number AAF15892.2); a26G1-1.pep, 56% identity to *Microcystis aeruginosa* McyG gene (239-aa alignment; GenBank accession number AAF00957.1); a26G1-2.pep, 60% identity to *S. aurantiaca* MtaE gene (222-aa alignment; GenBank accession number AAF19813.1); a26G1-10.pep, 61% identity to *Mycobacterium tuberculosis* PpsA gene (247-aa alignment; GenBank accession number spQ10977); a35E4-16.pep, 59% identity to *S. aurantiaca* MtaD gene (234-aa alignment; GenBank accession number AAF19812.1); a49F1-32.pep, 55% identity to *Nostoc* sp. strain GSV224 NosB gene (228-aa alignment; GenBank accession number AAF15892.2); a17D2-3.pep, 46% identity to *Mycobacterium leprae* PKS gene (224-aa alignment; GenBank accession number embCAC29609.1); a53F11-13.pep, 59% identity to *S. aurantiaca* MtaB gene (249-aa alignment; GenBank accession number AAF19810.1); a53F11-14.pep, 58% identity to *S. aurantiaca* MtaE gene (244-aa alignment; GenBank accession number AAF19813.1); a36E8-1.pep*, 60% identity to *S. aurantiaca* MtaB gene (225-aa alignment; GenBank accession number AAF19810.1); and a22A2-11.pep*, 50% identity to *Saccharopolyspora spinosa* PKS gene (GenBank accession number AAG23263.1). An asterisk in the designations denotes sequences derived from primer set 2. All other sequences were derived from primer set 1. *, identity; :, strong similarity; ., weak similarity.

$(NH_4)_2SO_4$/liter, 15 g of Pharmamedia (cotton flour)/liter, 3 mg of $ZnSO_4 \cdot 7H_2O$/liter, 20 g of agar (Difco)/liter. Each medium was dispensed into 90-ml petri dishes; inoculated by overlaying with 1 ml of the liquid culture; incubated at 28°C for 3, 5, and 10 days before freezing at −30°C; and lyophilized.

Lyophilized *E. coli* and *S. lividans* cultures were extracted with 30 ml of methanol to eliminate major macromolecules that may interfere with screening. An aliquot of the methanol solution, concentrated 10-fold for *S. lividans* strains, was analyzed by HPLC with a photodiode-array detector. The remaining extract was stored at −30°C for further analysis.

High-throughput HPLC analysis. The HPLC screen detects recombinant compounds by comparing UV spectra obtained from library clone extracts with those for the host strains, *E. coli* DH10B and *S. lividans* TK24, each carrying the empty cosmid vector pOS700I. Methanol extracts (30 μl) were loaded onto a reverse-phase (RP) C_{18} silica gel column (5 by 250 mm, 5 μm, 100 Å) at a flow rate of 1 ml/min. A linear elution gradient was performed from 100% water–0.05% (vol/vol) trifluoroacetic acid (TFA) to 100% acetonitrile– 0.05%(vol/vol) TFA over 30 min, followed by a 10-min elution with 100% acetonitrile–0.05% (vol/vol) TFA. A reequilibration step (10-min wash at 1 ml/min with 100% water– 0.05% [vol/vol] TFA) was performed prior to the next injection. A Waters 717+ injector, a 616 HPLC pump, and a 2690 separation module were used.

UV spectra were stored by using a Waters 996 photodiode-array detector. The technical setup for the library was as follows: optical resolution of 1.2 nm, sampling rate of one spectrum per second, and scan range of 200 to 600 nm. For spectrum comparison, the match angle threshold was set at 10°. All chromatographic systems and data were controlled and calculated by using Waters Millennium 32 software. We generated two host UV libraries, containing 129 and 314 spectra from *E. coli* and *S. lividans*, respectively. Comparisons of the UV spectra were carried out with the software. New candidate compounds were then characterized by further comparisons of the spectral features with available natural product structural databases.

RESULTS AND DISCUSSION

Initial characterization of the soil environmental DNA library. A 5,000-clone ampicillin-resistant environmental DNA library was constructed in *E. coli* by using *E. coli-S. lividans* shuttle cosmid pOS700I (see Materials and Methods). Microbial DNA used to construct the library was obtained from cells isolated with a Nycodenz density gradient, which separates microbial cells from the soil matrix. A prior study indicated that DNA obtained by this method shows no major phylogenetic bias compared with DNA obtained by the direct DNA extraction method (7) and that the method recovers only bacteria, without DNA contamination from other organisms (3, 7).

The library clones were manually ordered in 96-well plates for subsequent ease of handling. Sequence analysis beyond the vector-insert junction regions of 17 soil DNA inserts revealed that, interestingly, the G+C content of the soil DNA was 53 to 70%. This result showed that the *E. coli* host strain, with an average DNA G+C content of 51%, exhibited no serious bias against DNA with a higher G+C content. Similar results have been reported by others (6).

Forty-seven rRNA gene sequences were amplified directly from the environmental DNA library (see Materials and Methods), representing about 1% of the total library clones. Anal-

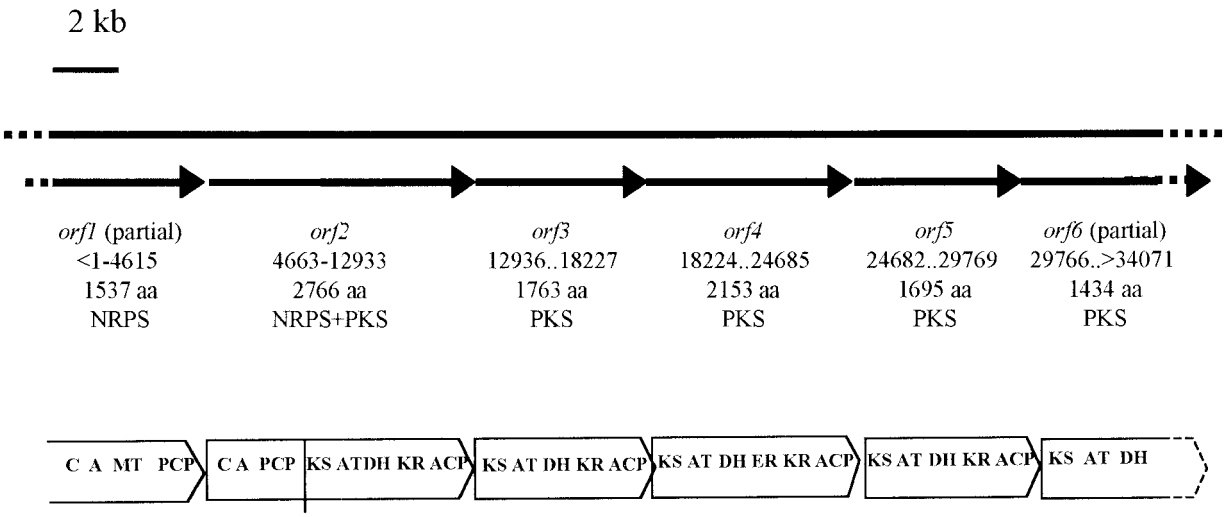

FIG. 2. ORF map of cosmid a26G1 insert DNA. Domains were as follows: C, condensation; A, adenylation; MT, methylation; PCP, peptidyl carrier protein; KS, ketoacyl synthetase; AT, acyl transferase; DH, dehydratase; KR, keto reductase; ACP, acyl carrier protein; ER, enoyl redutase. aa, amino acids.

ysis of these sequences confirmed that all 47 were unique and that the library appears to have been derived from phylogeneticaly diverse microorganisms, many of which have never been isolated or screened. The majority of the sequences analyzed belong to the *Proteobacteria* (data not shown). The results are consistent with previous work documenting the diversity found in DNA extracted from various soils (7, 16, 19, 20). However, most importantly, our current diversity analysis was carried out directly on the library clones, rather than on the soil used for constructing the library, thus extending previous results to now show that diversity was carried over to large pieces of DNA that were extracted from the soil and cloned into vectors to form an environmental DNA library.

Molecular screening of library clones for PKS I DNA sequences. In order to link the phylogenetic diversity analysis of this environmental library to an assessment of its potential for encompassing genes of functional relevance, we devised a method to amplify specific biosynthetic gene sequences from recombinant cosmid DNA preparations. As one class of natural products of potential interest, we targeted genes encoding enzymes involved in the biosynthesis of polyketides, a vast group of structurally diverse natural compounds produced by a large variety of soil microorganisms. The existence of highly conserved regions of actinomycete PKS I, flanking the active site of the ketoacyl synthetase domain, provided two sets of primer sequences (Materials and Methods) that could be used for the PCR amplification of homologous genes. The specificity of these primer sets was first validated by showing that they specifically amplified DNA from PKS I-producing strains (data not shown). Next, recombinant cosmid DNA prepared from 96-clone pools was analyzed by PCR with both primer sets as described in Materials and Methods. Following dereplication, 10 and 15 positive clones were detected with primer sets 1 and 2, respectively. Some of the PCR products were purified from agarose gels, cloned, and sequenced. Twelve unique nucleotide sequences were thus obtained. The alignment of the predicted protein sequences revealed that 11 of the nucleotide sequences encoded the highly conserved region corresponding to the active site of the β-ketoacyl synthetase consensus region of PKS I genes. Furthermore, a comparison of the cloned soil DNA sequences with the GenBank database showed that all of the cloned soil PKS I sequences were novel and highly similar to the sequences of PKS I genes from known microorganisms. The results of these analyses are summarized in Fig. 1. In particular, the highest similarity values were observed with PKS sequences from myxobacteria (*Stigmatella aurantiaca*), cyanobacteria (*Microcystis* and *Nostoc*), and *Mycobacterium* species. Moreover, the identity between the PKS I sequences from the soil clones and from the polyketide erythromycin cluster was about 53%.

Three of the 11 different PCR products described above were derived from one cosmid, a26G1, suggesting that genes encoding at least three different PKS I modules were encoded on that cosmid insert. We therefore determined the complete insert sequence (Fig. 2). Analysis of the sequence with FramePlot (14) revealed six large ORFs, all in the same orientation, with the upstream ORF overlapping the start codon of the following one in three instances. The first and last ORFs were truncated. The product of the first ORF resembled nonribosomal peptide synthetase (NRPS), and the product of the second ORF resembled a protein with one module of NRPS and one module of PKS. The products of ORFs 3, 4, 5, and 6 all resembled PKS, with each ORF (or partial ORF, in the case of ORF 6) encoding one module only. The predicted products of these ORFs are most similar to myxobacterial PKS and NRPS modules involved in the biosynthesis of myxothiazole in *S. aurantica* or in the biosynthesis of epothilone in *Sorangium cellulosum*. The G+C content (64%) of this insert was comparable to that of myxobacterial DNA.

We studied the prevalence of PKS I genes as a representative example for assessing the abundance of potentially interesting natural products encoded by our environmental DNA library. We were encouraged to find 11 PKS I gene sequences, a number much higher than expected in a random and rela-

```
                         1                                                 50                                                100
8E12.AAT        Unk      --------MTSRVATRSSLADDLSAIGLADGDAVLVHAALRQVGKIVGGP DAIIDALRDVIGPAGTILGYCDWQLEDELRDD------P-SMRPHIAAFD
AAA25683.1      Psa      MFSRWSKPLVLAAVTRASLAADLAALGLAAGDAVMVHAAVSKVGRLLDGP DTIIAALSDAGRPAGTILAYADWEARYEDLVDED-GRVPQEWREHIPPFD
AAA25682.1      Psa      ----------------------------------MVHAAVSRVGRLLDGP DTIIAALRDTVGPGGTVLAYADWEARYEDLVDDA-GRVPPEWREHVPPFD
AAA88552.1      Str      MDELALLKRSDGPVTRTRLARDLTALGLGDGDTVMFHTRMSAVGYVAGGP ETVIGALRDVVGERGTLMVTCGWNDAPPYDFTDWPQTWQDARRAEHPAYD

                         101                                               150                                               200
8E12.AAT        Unk      PERSRSTRDNGYWPEALRTTPGALRSGSPGASMAALGGEAEWFTADHALD YGYGPQSPLGKLVEAEGKVLMLGAPLDTMTLLHHAEHLADFPNKRILRYE
AAA25683.1      Psa      PRRSRAIRDNGVLPEFLRTTPGALRSGNPGASMVGLGARAEWFTADHPLD YGYGEGSPLARLVEAGGKVLMLGAPLDTLTLLHHAEHLADIPGKRIRRIE
AAA25682.1      Psa      PQRSRAIRDNGVLPEFLRTTPGTLRSGNPGASLVALGAKAEWFTADHPLD YGYGEGSPLAKLVEAGGKVLMLGAPLDTLTLLHHAEHLADIPGKRIKRIE
AAA88552.1      Str      PVLSEADHNNGRLPEALRRRPGAVRSRHPDASFAALGAAATALTADHPWD DPHGPDSPLARLVAMGGRVLLLGAPLEALTLLHHAEALADAPGKRFVDYE

                         201                                               250                                   288
8E12.AAT        Unk      APILVDGETVWRWFEEFDTSEPP-DG-----LPEDYFATIVEAFLATGRG KRGEVGEASSVLVPAAAMVAFGVDWLERWGKTL-----
AAA25683.1      Psa      VPLATPTGTQWRMIEEFDTGDPIVEG-----LAEDYFAEIVTAFLAGGRG RQGLIGTAPSVLVDAAAITAFGVAWLESRFGSPSS---
AAA25682.1      Psa      VPFATPTGTQWRMIEEFDTGDPIVAG-----LAEDYFAGIVTEFLASGQG RQGLIGAAPSVLVDAAAITAFGVTWLEKRFGTPSP---
AAA88552.1      Str      QPILVDGERVWRRFHDIDSEDGAFDYSALVPEGTEAFEIIGRDMRAAGIG RRGTVGAADSHLFEARDVVDFGVAWMEEKLGRERGPGG
```

FIG. 3. Similarity of predicted aminoglycoside acetyltransferase sequence to sequences of known proteins. Sequences were as follows: 8E12.AAT, putative aminoglycoside acetyltransferase in cosmid a8E12 (nucleotides 30829 to 31617; GenBank accession number AF486581); AAA25683.1, aminoglycoside 3′-*N*-acetyltransferase of *Pseudomonas aeruginosa*; AAA25682.1, AAC(3)-IIIB of *P. aeruginosa*; AAA88552.1, aminocyclitol 3-*N*-acetyltransferase, type VII, of *S. rimosus*. Unk, unknown.

tively small (<250-Mb) DNA sample. The partial NRPS/PKS pathway encoded on a26G1 strongly suggests that one could reasonably expect to find complete clusters of polyketide or other biosynthetic genes in a library containing more clones and larger inserts.

Colony screening of library clones for biological activity. As another method for exploring the expression of heterologous DNA in the soil cosmid library, we screened colonies of library clones for various biological activities. The library in the *E. coli* host was arrayed on agar plates, grown for several days at 30°C, and overlaid with *B. subtilis* to detect antibacterial activity (see Materials and Methods). In addition, to detect genes encoding kanamycin resistance, which might lie adjacent to a biosynthetic gene cluster, colonies were plated on medium containing kanamycin. We detected one antibacterial activity (clone a10B12) and one kanamycin resistance activity (clone a8E12) expressed in *E. coli*. Although the antibacterial activity appeared to comprise a small molecule encoded on the cosmid

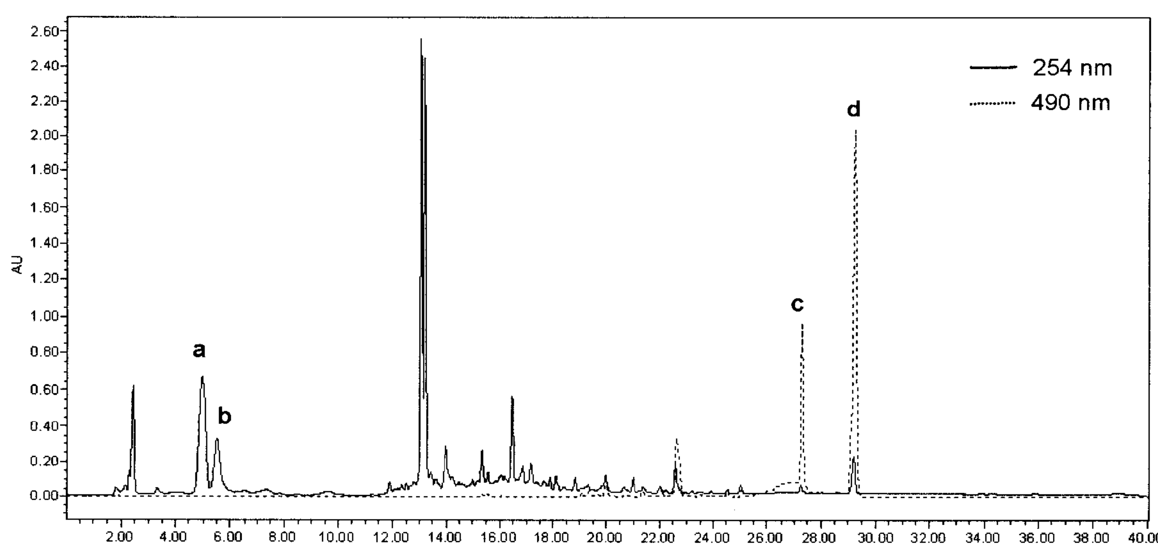

FIG. 4. RP HPLC elution profile of an extract of *S. lividans* TK24 containing library cosmid a22G9. Modified R5 agar plates (as described in reference 14, but omitting sucrose) were cut into pieces of approximately 0.5 cm^3, transferred to 50-ml tubes, lyophilized for 48 h (Labconco Freezone 4.5), ground to a fine powder, extracted with methanol, filtered through Whatman Autovial PTFE filters (0.2-mm-pore size), placed in Waters SepPak Plus C_{18} cartridges, concentrated to 1 ml (Savant Speedvac SC210A), and filtered again (Whatman 4-mm-diameter, 0.2-mm-pore-size PTFE syringe filters) prior to HPLC analysis. An Inertsil ODS-3 column (5 μm, 150 [length] by 4.6 [diameter] mm; GL Sciences) was used for analytical RP HPLC on a Waters 600 system with a Waters 996 photodiode-array detector (210 to 560 nm, 1.2-nm resolution; Millennium 4.0 software). The mobile phases were 0.08% TFA in water (A) and 0.08% TFA in acetonitrile (B). Elution was started with 100% A for 2 min, and a linear gradient was run from 0 to 100% B over 20 min with a 10-min hold at 100% B. The flow rate was 1 ml/min, and the injection volume was 10 ml. Identification of known compounds (undecylprodigiosin and actinorhodin) was based on their λ_{max} values. Traces for 254 and 490 nm are shown. Peaks a and b are new compounds that were not present in the control extract. Peaks c and d correspond to undecylprodigiosin and actinorhodin, respectively. AU, arbitrary units.

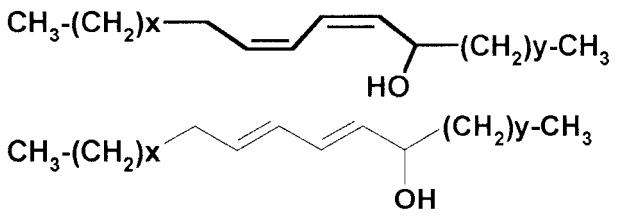

FIG. 5. Structures of two fatty dienic alcohol isomers. For the isomers with a relative atomic mass of 294, $x + y = 12$.

insert and present in *E. coli* extracts, the activity was lost before we could determine its structure, possibly due to strong negative selection in *E. coli* resulting from the expression of this molecule. The kanamycin resistance activity was stably expressed in *E. coli*, and the DNA sequence of the cosmid insert revealed that although the ORF likely to be responsible did not appear to be part of a biosynthetic gene cluster encoding an antibiotic, it did encode a putative protein with high similarity to aminoglycoside acetyltransferase proteins of several species, including *Pseudomonas* and *Streptomyces* (Fig. 3). Cosmid a8E12 was transformed into *S. lividans* but did not express the kanamycin resistance activity in that host strain, underscoring the importance of using multiple expression systems to capture a wider spectrum of possible activities.

Some additional cosmid clones (including a10B12 and clones encoding PKS I homologs) were transformed into *S. lividans* TK24 and into an *S. lividans* derivative with a deletion of endogenous pigment genes (A. Martinez, unpublished results). Although no antibacterial activities were found in the strain with a deletion of endogenous pigment genes, clone a22G9, one of 30 clones tested, caused *S. lividans* TK24 to overproduce blue (actinorhodin) pigment, which can be an indication of the production of heterologous molecules (Martinez, unpublished). HPLC analysis of an extract of strain TK24 bearing this cosmid revealed two new peaks not present in the host strain bearing the cosmid backbone alone (Fig. 4).

HPLC screening of library clone extracts for heterologous molecules. To further investigate the expression of heterologous DNA in the soil cosmid library, we carried out chemical screening by HPLC, aimed at detecting and characterizing new metabolites produced by cultured library clones. Library clones that were found positive in the PKS I prescreening in addition to other library clones selected randomly were transformed into *S. lividans* TK24. Cells were grown in various media, and a total of 2,500 extracts, 1,700 from *E. coli* and 800 from *S. lividans* (corresponding to 480 and 40 *E. coli* and *Streptomyces* clones, respectively), were analyzed by HPLC with a photodiode-array detector.

Out of 12,000 peaks, more than 100 peaks could not be matched to the UV library (i.e., were not present in the host strains). The majority of these peaks fell below a threshold of reliable purity (purity angle, $>10°$) and were omitted from further analysis. However, two recombinant strains, *S. lividans* clones a24H2 and a24A3, yielded the same chromatographic profile, containing peaks that revealed the presence of homologous compounds not detected in the host strains.

Having defined the best isocratic conditions for analyzing them, we separated a series of six closely related compounds which exhibited nearly the same UV spectra. A liquid chromatography-mass spectrometry analysis performed under these conditions yielded a relative atomic mass of 294 for four of them. A liquid chromatography-nuclear magnetic resonance spectrometry study showed that two of these compounds were a mixture of *E,E* and *Z,Z* dienes with the structures shown in Fig. 5. These two fatty dienic alcohol isomers have not been described in the literature (4; Chemical Abstracts databases through 1999; American Chemical Society, Washington, D.C.).

Summary. The work presented here provides encouraging evidence that accessing microbial biodiversity for drug discovery from natural products by this type of technology is a very promising approach. The small library (5,000 clones) described in this work was found to contain a number of interesting genes and activities and was quite phylogenetically diverse. The data obtained in this study suggest a strategy for developing the technology further; i.e., we confirmed that genes encoding natural products can be readily captured by using this strategy, but larger libraries with larger inserts, expression in multiple host systems, and the strategic use of prescreening should greatly enhance the ability to detect novel and useful secondary metabolites.

ACKNOWLEDGMENTS

Sophie Courtois and Carmela M. Cappellano contributed equally to the work presented here.

We thank Françoise Le Gall for excellent technical assistance and Nathalie Bamas-Jacques for supplying the PKS I primer sets.

This work was supported by Aventis Pharmaceuticals, Centre National de la Recherche Scientifique (CNRS), Université Paris Sud XI (UMR 8621), and Université Claude Bernard Lyon 1 (UMR 5557).

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Amann, R. I., W. Ludwig, and K. H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143–169.
3. **Bakken. L. R., and V. Lindahl.** 1995. Recovery of bacterial cells from soil, p. 9–27. *In* J. D. van Elsas and J. T. Trevors (ed.), Nucleic acids in the environment: methods and applications. Springer-Verlag, Berlin, Germany.
4. **Berdy, J.** 1980. Handbook of antibiotic compounds. CRC Press, Inc., Boca Raton, Fla.
5. **Blondelet-Rouault, M. H., J. Weiser, A. Lebrihi, P. Branny, and J. L. Pernodet.** 1997. Antibiotic resistance gene cassettes derived from the omega interposon for use in *E. coli* and *Streptomyces*. Gene **190:**315–317.
6. **Chatzinotas, A., R. A. Sandaa, W. Schonhuber, R. Amann, F. L. Daae, V. Torsvik, J. Zeyer, and D. Hahn.** 1998. Analysis of broad-scale differences in microbial community composition of two pristine forest soils. Syst. Appl. Microbiol. **21:**579–587.
7. **Courtois, S., A. Frostegard, P. Goransson, G. Depret, P. Jeannin, and P. Simonet.** 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. **3:**431–439.
8. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. **8:**186–194.
9. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. **8:**175–185.
10. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
11. **Gourmelen, A., M. H. Blondelet-Rouault, and J. L. Pernodet.** 1998. Characterization of a glycosyl transferase-inactivating macrolide, encoded by *gimA* from *Streptomyces ambofaciens*. Antimicrob. Agents Chemother. **42:**2612–2619.
12. **Hopwood, D. A., M. J. Bibb, K. F. Chater, T. Kieser, C. J. Bruton, H. M. Kieser, D. J. Lydiate, C. P. Smith, J. M. Smith, J. M. Ward, and H. S. Schrempf.** 1985. Genetic manipulation of streptomycetes: a laboratory manual. John Innes Institute, Norwich, United Kingdom.
13. **Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace.** 1998. Novel

division-level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180:**366–376.

14. **Ishikawa, J., and K. Hotta.** 1999. FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. FEMS Microbiol. Lett. **174:**251–253.
15. **Kieser, T., M. J. Bibb, M. J. Buttner, K. F. Chater, and D. A. Hopwood.** 2000. Practical *Streptomyces* genetics. The John Innes Foundation, Norwich, England.
16. **MacNeil, I. A., C. L. Tiong, C. Minor, P. R. August, T. H. Grossman, K. A. Loiacono, B. A. Lynch, T. Phillips, S. Narula, R. Sundaramoorthi, A. Tyler, T. Aldredge, H. Long, M. Gilman, D. Holt, and M. S. Osburne.** 2001. Expression and isolation of antimicrobial small molecules from soil DNA libraries. J. Mol. Microbiol. Biotechnol. **3:**301–308.
17. **Marchesi, J. R., T. Sato, A. J. Weightman, T. A. Martin, J. C. Fry, S. J. Hiom, D. Dymock, and W. G. Wade.** 1998. Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. Appl. Environ. Microbiol. **64:**795–799.
18. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.
19. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman.** 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66:**2541–2547.
20. **Rondon, M. R., R. M. Goodman, and J. Handelsman.** 1999. The Earth's bounty: assessing and accessing soil microbial diversity. Trends Biotechnol. **17:**403–409.
21. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
22. **Seow, K. T., G. Meurer, M. Gerlitz, E. Wendt-Pienkowski, C. R. Hutchinson, and J. Davies.** 1997. A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms. J. Bacteriol. **179:**7360–7368.
23. **Short, J. M.** 1997. Recombinant approaches for accessing biodiversity. Nat. Biotechnol. **15:**1322–1323.
24. **Smokvina, T., P. Mazodier, F. Boccard, C. J. Thompson, and M. Guerineau.** 1990. Construction of a series of pSAM2-based integrative vectors for use in actinomycetes. Gene **94:**53–59.
25. **Tartoff, K. D., and C. A. Hobbs.** 1987. Improved media for growing plasmid and cosmid clones. Bethesda Research Labs Focus **9:**12.
26. **Thompson, C. J., J. M. Ward, and D. A. Hopwood.** 1980. DNA cloning in *Streptomyces*: resistance genes from antibiotic-producing species. Nature **286:**525–527.
27. **Wahl, G. M., K. A. Lewis, J. C. Ruiz, B. Rothenberg, J. Zhao, and G. A. Evans.** 1987. Cosmid vectors for rapid genomic walking, restriction mapping, and gene transfer. Proc. Natl. Acad. Sci. USA **84:**2160–2164.
28. **Wang, G. Y., E. Graziani, B. Waters, W. Pan, X. Li, J. McDermott, G. Meurer, G. Saxena, R. J. Andersen, and J. Davies.** 2000. Novel natural products from soil DNA libraries in a streptomycete host. Org. Lett. **2:**2401–2404.