Sequencing the 500-kb GC-rich Symbiotic Replicon of *Rhizobium* sp. NGR234 Using Dye Terminators and a Thermostable "Sequenase": a beginning

Freiberg, Christoph; Perret, Xavier; Broughton, William John; Rosenthal, André

RESEARCH

Sequencing the 500-kb GC-rich Symbiotic Replicon of *Rhizobium* sp. NGR234 Using Dye Terminators and a Thermostable "Sequenase": A Beginning

Christoph Freiberg,[1] Xavier Perret,[2] William J. Broughton,[2] and André Rosenthal[1,3]

[1]Institut für Molekulare Biotechnologie, 07745 Jena, Germany; [2]Laboratoire de Biologie Moléculaire des Plantes Supérieures, Université de Genève, Genève, Switzerland

Genomes of the soil-borne nitrogen-fixing symbionts of legumes [*Azo(Brady)Rhizobium* species] typically have GC contents of 59–65 mol%. As a consequence, compressions (up to 400 per cosmid) are common using automated dye primer shotgun sequencing methods. To overcome this difficulty, we have exclusively applied dye terminators in combination with a thermostable "sequenase" for shotgun sequencing GC-rich cosmids from pNGR234*a*, the 500-kbp symbiotic replicon of *Rhizobium* sp. NGR234. A thermostable sequenase incorporates dye terminators into DNA more efficiently than *Taq* DNA polymerase, thus reducing the concentrations needed (20- to 250-fold). Unincorporated dye terminators can simply be removed by ethanol precipitation. Here, we present data of pXB296, one of 23 overlapping cosmids representing pNGR234*a*. We demonstrate that the greatly reduced number of compressions results in a much faster assembly of cosmid sequence data by comparing assembly of the shotgun data from pXB296 and the data from another pNGR234*a* cosmid (pXB110) sequenced using dye primer methods. Within the 34,010-bp sequence from pXB296, 28 coding regions were predicted. All of them showed significant homologies to known proteins, including oligopeptide permeases, an essential cluster for nitrogen fixation, and the C_4-dicarboxylate transporter DctA.

Soil bacteria of the genera *Azorhizobium, Bradyrhizobium,* and *Rhizobium* establish symbiotic associations with leguminous plants leading to the formation of nitrogen-fixing nodules. Plant exudates (particularly flavonoid compounds) modulate the coordinated expression of many bacterial genes leading to the production of mitogenic Nod factors, which provoke nodule formation (Fellay et al. 1995a). In *Azorhizobium* and *Bradyrhizobium* species, symbiotic loci are carried on the chromosome, whereas in *Rhizobium* species most of them are plasmid-borne (Martinez et al. 1990; Fischer 1994; van Rhijn and Vanderleyden 1995). Complete understanding of legume–*Rhizobium* interactions thus requires a catalog of symbiotic genes and exhaustive analysis of their function.

The broad host-range *Rhizobium* sp. NGR234 carries a plasmid of 500 kbp (pNGR234*a*), which confers on *Agrobacterium tumefaciens* recipients the ability to nodulate certain legumes (Broughton et al. 1984). Identification of symbiotic pNGR234*a* genes has been performed by host-range extension (complementation) and mutagenesis (Broughton et al. 1986; Lewin et al. 1990), techniques that are not only time-consuming but that fail to identify subtle phenotypes. To facilitate analysis of this replicon, a canonical ordered cosmid library of pNGR234*a* was constructed (Perret et al. 1991). Flavonoid-inducible loci were mapped to discrete *Xho*I fragments on these cosmids by competitive RNA hybridization (Fellay et al. 1995b). Other symbiotic loci were identified by subtractive DNA hybridization against the genome of the closely related strain *Rhizobium fredii* USDA257 (Perret et al. 1994). Nevertheless, a number of symbiotic genes remain to be identified, and for this reason, we wish to establish the complete nucleotide sequence of pNGR234*a*.

[3]Corresponding author.
E-MAIL arosenth@imb-jena.de; FAX 49-3641-656255.

6:590–600 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

Automated fluorescent methods have been used to sequence cosmids from eukaryotic organisms, including *Saccharomyces cerevisiae* (Levy 1994), *Caenorhabditis elegans* (Sulston et al. 1992), *Drosophila melanogaster* (Hartl and Palazzolo 1993), and *Homo sapiens* (Bodmer 1994), as well as chromosomes from the prokaryotes *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995). In most large-scale sequencing centers, including ours, this technology is based mainly on the shotgun approach. After random fragmentation of DNA [e.g., cosmids, bacterial artificial chromosomes (BACs), entire chromosomes] using sonication or mechanical forces, size-selected fragments are subcloned into M13 phages, phagemids, or plasmids and sequenced by cycle sequencing using dye primers (Craxton 1993). A disadvantage of this method is that DNA regions with elevated GC contents produce large numbers of compressions (unresolvable foci in sequence gels) in the dye primer sequences leading to several hundred compressions per assembled cosmid sequence. It is known that the use of dye terminators—fluorescently labeled dideoxynucleosidetriphosphates—instead of dye primers reduces the number of compressions (Rosenthal and Charnock-Jones 1993). Therefore, dye terminators are frequently being used for gap closure and proofreading after assembly of the shotgun data.

To sequence GC-rich cosmids with the highest accuracy, we have investigated the effectiveness of shotgun sequencing with dye terminators in comparison to dye primer sequencing. To improve the incorporation of dye terminators into DNA, we have used a modified *Taq* DNA polymerase carrying a single mutation (Tabor and Richardson 1995). This enzyme has properties similar to a thermostable "sequenase" and is commercially available as Thermo Sequenase (Amersham) or AmpliTaq FS (Perkin-Elmer). Concentrations of dye terminators needed in the cycle sequencing reactions can be reduced by 20–250 times. We have found that dye terminator shotgun sequencing leads to compression-free raw data that can be assembled much faster than shotgun data mainly obtained by dye primer sequencing. This strategy thus allows a severalfold increase in speed to sequence individual cosmids. We demonstrate this by comparing assembly of the sequence data of two cosmids from pNGR234*a* generated by different chemistries: Cosmid pXB296 was sequenced with dye terminators, whereas data for pXB110 were obtained using the common dye primer method. Furthermore, we present the analysis of the entire pXB296 sequence.

RESULTS

Comparison of Fluorescent Traces Created by Different Cycle Sequencing Methods

When using a thermostable sequenase [Thermo Sequenase (Amersham)], we were able to reduce the concentrations of dye terminators (Perkin-Elmer) by 20- to 250-fold in comparison to the concentrations needed for *Taq* DNA polymerase without compromising the quality of the sequencing results (Table 1).

To compare the dye terminator and dye primer cycle sequencing procedures, representative templates derived from the pXB296 library were sequenced by both methods, each performed with Thermo Sequenase and *Taq* DNA polymerase (Fig. 1). In general, dye terminator traces do not contain the many compressions (on average, one compression every 50 bases in single reads) that are common with dye primers if mixes do not contain nucleotide analogs like deoxyino-

Table 1. Concentrations (in μM) of Dye Terminators in Each Cycle Sequencing Reaction with Two Different Thermostable DNA Polymerases

Dye terminator	AmpliTaq DNA polymerase	Thermo Sequenase DNA polymerase	Dilution factor for dye terminators[a]
A *Taq*	0.751	0.017	40
C *Taq*	22.500	0.137	160
G *Taq*	0.200	0.009	20
T *Taq*	45.000	0.183	250

[a]Thermo Sequenase vs. AmpliTaq.

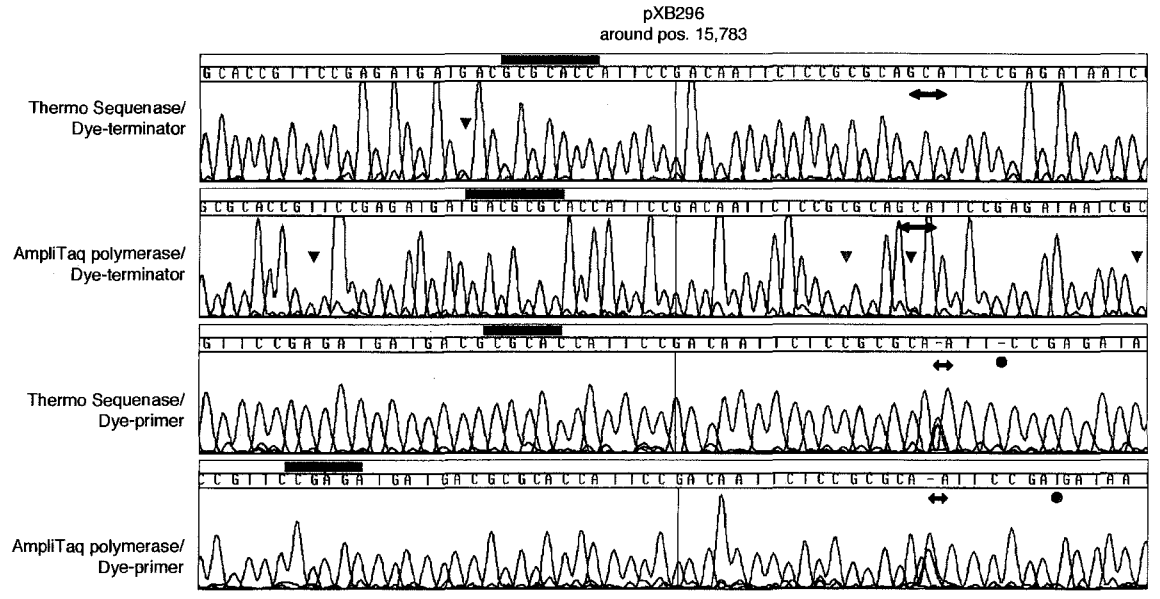

Figure 1 Comparison of sequences from pXB296 created by different cycle sequencing methods. The graphic outputs (program XGAP) of four electropherograms (traces) with the corresponding sequences generated by automatic base calling are shown around the pXB296 sequence position 15,783 (vertical line in the middle of each sequence). The readings cover a part of the minus strand of pXB296. (▼) Extremely low signals produced with dye terminators; (↔) the sequence GCA, which is compressed in dye primer scans; (●) automatic base-calling inaccuracies.

sine or 7-deaza-deoxyguanosine triphosphates or if sequencers are used without active heating systems. In addition, dye terminator traces obtained with Thermo Sequenase show more uniform signal intensities over those obtained with *Taq* DNA polymerase, thus resulting in a reduced number of weak and missing peaks (e.g., a weak G-signal following an A-signal in Thermo Sequenase traces or a weak C-signal following a G-signal in *Taq* DNA polymerase traces). Using Applied Biosystems (ABI) 373A sequencers, errors in automatic base-calling of Thermo Sequenase/dye terminator scans only arise after 300–350 bases. The average number of resolved bases in dye primer gels (378) is 46 bases longer than in those produced with dye terminators (332 bases). Furthermore, in Thermo Sequenase/dye primer sequences the peaks are very regular and the number of stops and missing bases decreases in comparison to *Taq* DNA polymerase/dye primer electropherograms. The number of compressions, however, is not significantly reduced.

Shotgun Sequencing of Entire Cosmids Using Dye Terminators or Dye Primers

To compare the efficiency of both methods, cosmid pXB296 of pNGR234*a* was shotgun sequenced using a combination of dye terminators and thermostable sequenase (Thermo Sequenase), whereas another cosmid, pXB110, was sequenced using a combination of dye primers and *Taq* DNA polymerase (Table 2). Over 93% (736 clones) of 786 dye terminator reads of pXB296 were accepted by XGAP with a maximal alignment mismatch of 4%. By increasing this level to 25%, so that most of the remaining data could be included in the assembly, 775 reads led to three 6- to 10-kbp stretches of contiguous sequence (contigs), two of which were joined after editing. To close the last gap and to complete single-stranded regions with data derived from the opposite strand, only 32 additional dye terminator reads using custom-made primers were obtained. It took <1 week to assemble and finalize the 34,010-bp DNA sequence of pXB296 (EMBL accession no. Z68203; eightfold redundancy; GC content, 58.5 mol%).

In contrast, only 308 (34%) of 899 shotgun reads obtained by dye primer/*Taq* DNA polymerase cycle sequencing of pXB110 were included in the first assembly (4% alignment mismatch). At the 25% alignment mismatch level, 879 reads were assembled, leading to 25 short contigs (1–2

kbp). These ("contigs") had to be edited extensively in order to join most of them. "Primer walks" covering gaps and complementing single-stranded regions, were not sufficient to clarify all of the remaining ambiguities in the assembled sequence. Every 100–150 bp, a compression in one strand could not be resolved by sequence data from the complementary strand. Therefore, it was necessary to resequence clones using dye terminators and universal primer. In total, 191 additional dye terminator reads had to be created. As a result, assembling and finalizing the 34,573-bp sequence of pXB110 (10.5-fold redundancy; GC content, 58.3 mol%) took much more time than pXB296 did.

Analysis of Cosmid pXB296

Putative genes were located on the 34,010-bp sequence of pXB296 using the programs TESTCODE (Fickett 1982) and CODONPREFERENCE (Gribskov et al. 1984), the latter in combination with a codon frequency table based on previously sequenced genes of *Rhizobium* sp. NGR234 (as well as the closely related *R. fredii*). All 28 open reading frames (ORFs) and their deduced amino acid sequences exhibited significant homologies to known genes and/or proteins. The positions of the ORFs along pXB296, as well as the best homologs, are displayed in Table 3 and Figure 2. Ribosomal binding site-like sequences (Shine and Dalgarno 1974) precede each putative gene except for ORF9 (position 11,124–12,455). If one disregards the homology to known glutamate dehydrogenases in the first 32 amino acids deduced from this ORF, a downstream alternative start codon (position 11,220) preceded by a Shine–Dalgarno sequence can be identified. Most of the ORFs are organized in five clusters (ORFs with only short intergenic spaces or overlaps between them). Cluster I containing ORF1 to ORF5 encodes proteins homologous to *trans*-membrane and membrane-associated oligopeptide permease proteins and to a *Bacillus anthracis* encapsulation protein. Cluster II includes ORF6 and ORF7 homologous to aminotransferase and (semi)aldehyde dehydrogenase genes. Homologies to transposase genes [ORF8; cluster III (ORF10 and ORF11)] and to various *nif* and *fix* genes [cluster IV (ORF12 to ORF20); ORF23, part of cluster V] are also reported.

Presumed promoter and stem–loop sequences that might represent ρ-independent terminator-like structures (Platt 1986) are shown in Figure 2. Significant σ^{54}-dependent promoter consensus sequences (5′-TGGCACG-N_4-TTGC-3′; Morett and Buck 1989), as well as *nifA* upstream activator sequences (5′-TGT-N_{10}-ACA-3′;

Table 2. Comparison of the Assembly of the Sequence Data from Cosmids pXB296 (Dye Terminator Shotgun Reads) and pXB110 (Dye Primer Shotgun Reads)

Data assembly	pXB296	pXB110
Average length of the shotgun reads (bases)	332	378
No. of shotgun reads used for assembly	786	899
No. of shotgun reads assembled with 4% mismatch[a]	736	308
No. of shotgun reads assembled with 25% mismatch[a]	775	879
No. of contigs[b] longer than 1 kbp	3	25
No. of contigs left after editing[c]	2	4
No. of additional reads (gap closure and proofreading)[d]	32	191
Total length of cosmid insert (bp)	34,010	34,573
Sequencing redundancy (per bp)	8.0	10.5

[a]Assembling program: XGAP; principal autoassembling conditions: normal shotgun assembly, joins permitted, minimum initial match = 15, maximum no. of pads per reading during the alignment procedure = 8, maximum no. of pads per reading in contig to align any new reading = 8, alignment mismatches 4% and 25%, respectively.
[b]Contiguous parts of sequence created by overlapping reads.
[c]Lengths of contigs: 6–10 kbp (pXB296); 2–12 kbp (pXB110).
[d]Reads necessary for closing gaps and making single-stranded regions double-stranded by primer walking on selected templates and, in case of pXB110, for solving ambiguities (compressions) by the resequencing of clones with universal primer and dye terminators.

Morett and Buck 1988), are found upstream of the *nifB* homolog ORF15, the *fixA* homolog ORF20, and ORF21, ORF22, and ORF23. ORF23 is part of cluster V in pXB296, which includes the *dctA* gene of *Rhizobium* sp. NGR234 (van Slooten et al. 1992). Surprisingly, the published *dctA* sequence shows important discrepancies. Therefore, a fragment encompassing this locus was amplified by PCR using NGR234 genomic DNA as template. By sequencing this fragment, our cosmid sequence was confirmed.

DISCUSSION

Advantages of the Dye Terminator/Thermostable Sequenase Shotgun Strategy

We have examined whether GC-rich cosmids can be sequenced much more efficiently using dye

Table 3. Putative Genes of pXB296 and Homologies of the Deduced Amino Acid Sequences to Known Proteins

ORF[a]	st.[b]	position on cosmid (base no.)[c]	ribosomal binding site: SD-sequence - distance from start codon (bases)- start codon[d]	no. of deduced amino acids	homologous amino acids (position)	homologous protein				identity (%)[g]	similarity (%)[g]
			SD-sequence: 5´-TAAGGAGGTGA-3´			name	length (aa)[e]	function[f]	accession no.		
ORF1[h]	+	00001-00625		>207	1-207	OppB	306	oligopeptide	X05491	45	68
ORF2	+	00628-01503	GTATCCGGT-7-ATG	291	2-289	OppC	305	permease	X56347	37	63
ORF3	+	01505-02512	AGCGGAGG-7-ATG	335	8-327	OppD	336	proteins	X56347	49	69
ORF4	+	02509-03570	TGAAGTGGT-6-ATG	353	2-323	OppF	334		X05491	51	69
ORF5	+	03606-04991	CAAGGA-6-ATG	461	1-458	CapA	411	encapsulation protein	M24150	25	48
ORF6	+	05460-06863	CCGAGAGG-8-ATG	467	1-464	BioA	455	aminotransferase	M29292	29	55
ORF7	+	06888-08426	GCCTTCGG-5-GTG	512	97-509 34-510	ORF[i] GapD	417 482	unknown succinic semialdehyde dehydrogenase	D37877 M38417	36 33	58 57
ORF8	-	09781-10860	GAACGTGG-8-ATG	359	72-299	ORF[i]	414	transposase homologue, minicircle DNA	X15942	30	48
ORF9	+	11124-12455	?-7-ATG	443	2-443	GLUD1	558	glutamate dehydrogenase	M37154	41	60
ORF10	-	13370-14116	AAAGGA-6-ATG	248	1-245	ORF2[i]	231	transposase	X79443	45	64
ORF11	-	14128-15672	CATGGAG-7-TTG	514	1-513	ORF1[i]	558	homologues, IS1162	X79443	41	62
ORF12	-	16712-16942	GAAGGA-8-ATG	76	1-70	FixU	70	unknown	P42710	63	80
ORF13	-	16939-17265	ACAAGAGG-7-ATG	109	1-79 15-107	ORF2[i] NifZ	>78 159	unknown involved in FeMo-cofactor synthesis	X07567 M20568	53 39	81 56
ORF14	-	17349-17543	CCAGGAG-9-ATG	64	1-64	FdxN	64	ferredoxin-like	M21841	80	87
ORF15	-	17585-19066	AGTGGAG-7-ATG	493	1-493	NifB	490	involved in FeMo-cofactor synthesis	M15544	73	84
ORF16	-	19292-20962	ATTGG-12-ATG	556	9-556	NifA	541	transcriptional regulator	X02615	59	72
ORF17	-	21129-21422	AGGGGAG-7-ATG	97	1-97	FixX	98	required for	M15546	84	87
ORF18	-	21437-22744	AACTGAGGT-7-ATG	435	1-435	FixC	435	nitrogen	M15546	83	90
ORF19	-	22755-23864	ATAGGAG-6-ATG	369	18-369	FixB	353	fixation	M15546	79	89
ORF20	-	23874-24731	TAAAGAG-5-ATG	285	1-285	FixA	292		M15546	74	85
ORF21	-	25148-25468	CCAGGAG-10-ATG	106	1-106	ORF118[i]	108	unknown	X13691	55	71
ORF22	-	26145-26711	GAAGGAG-9-ATG	188	9-199 1-173	- -	241 166	hypothetical protein peroxisomal protein	U32739 U11244	47 32	64 57

terminators throughout the shotgun phase instead of dye primers. As a test case, cosmid pXB296 with a GC content of 58 mol% from pNGR234*a*, the symbiotic plasmid of *Rhizobium* sp. NGR234, was exclusively sequenced using dye terminators in combination with a thermostable sequenase [Thermo Sequenase (Amersham)]. Another rhizobial cosmid with identical GC content, pXB110, was sequenced using traditional dye primer chemistry and *Taq* DNA polymerase.

Using the dye terminator/thermostable sequenase shotgun strategy, we have shown that most, if not all, compressions could be resolved and reads were produced with the highest fidelity among all sequencing chemistries tested. As a result, we obtained a much faster assembly of cosmid pXB296 in comparison to pXB110. The shotgun data could be assembled to a high-quality sequence without extensive editing and proofreading. By measuring the error rate in overlapping regions between individual cosmids from pNGR234*a*, as well as in the cosmid vector sequence itself (data not shown), we estimate that the accuracy of the pXB296 sequence is higher than 99.98%. Using other thermostable sequenases such as AmpliTaq FS (Perkin-Elmer), we would expect similar results because thermostable sequenases have similar properties.

We also examined dye primer chemistry in combination with Thermo Sequenase. Although the peak uniformity of signals was much improved over dye primer/*Taq* DNA polymerase data, the number of compressions in GC-rich shotgun reads was not reduced significantly. Compressions in shotgun raw data enormously increase the overall effort of editing, proofreading, and finishing a cosmid as shown for pXB110 (Table 2). We have not investigated whether compressions in dye primer sequencing of rhizobial cosmids could be reduced efficiently using either nucleotide analogs like deoxyinosine or 7-deaza-deoxyguanosine triphosphates or el-

Table 3. *(Continued)*

ORF23	+	27169-27861	GAAGGA-7-ATG	230	1-167	NifQ	167	probably involved in Mo-processing	X13303	37	57
ORF24	+	27920-29434	CTGGGAGG-18-ATG	504	1-454 8-454	DctA1 DctA2	456 449	C_4-dicarboxylate transporter	S38912 S38912	97 97	98 98
ORF25	+	29431-30675	TTCGGCGG-12-ATG	414	2-414	CamC	415	cytP450-like	M12546	34	53
ORF26	+	30676-31332	TTGGG-5-TTG	218	30-190	LinA	155	γ-hexachloro-cyclohexan-dechlorinase	D90355	27	51
ORF27	+	31329-33035	AGTGGAG-10-ATG	568	28-270 294-534	FabG	244	reductase	M84991	38 30	57 57
ORF28[k]	+	33173-34010	CAAGGAG-5-ATG	>279	1-279	LuxA	355	luciferase α-subunit	M10961	23	49

[a](ORF) Open reading frame.
[b](st.) Plus or minus strand.
[c]Position on cosmid: from the first base of the start codon to the last base of the stop codon; alternative start points are 6912/6927/7017 (ORF7), 10665/10656 (ORF8), 11220 (ORF9), 15699/15651 (ORF11), 17322/17271 (ORF13), 20995/21076 (ORF16), 26744 (ORF22), 27229/27304 (ORF23), 27941 (ORF24), and 30751/30754 (ORF26).
[d](SD sequence) Shine–Dalgarno sequence (Shine and Dalgarno 1974). Bases underlined are identical with the Shine–Dalgarno sequence. The following possible start codons were considered: ATG, GTG, or TTG.
[e](aa) Amino acids.
[f]Organisms: *Salmonella typhimurium, Bacillus subtilis* (OppBCDF), *Bacillus anthracis* (CapA), *Bacillus sphaericus* (BioA), *Streptomyces hygroscopicus* (ORF7 homolog), *Escherichia coli* (GapD), *Streptomyces coelicolor* (ORF8 homolog), *Homo sapiens* (GLUD1), *Pseudomonas fluorescens* (ORF10, ORF11 homologs), *Rhizobium leguminosarum* (FixU), *Rhodobacter capsulatus* (ORF13 homolog), *Azotobacter vinelandii* (NifZ), *Rhizobium meliloti* (FdxN, NifBA, FixXCBA), *Bradyrhizobium japonicum* (ORF118), *Haemophilus influenzae* (hypothetical protein), *Lipomyces kononenkoae* (peroxisomal protein), *Klebsiella pneumoniae* (NifQ), *Rhizobium* sp. NGR234 (DctA), *Pseudomonas putida* (CamC), *Pseudomonas paucimobilis* (LinA), *Escherichia coli* (FabG), *Vibrio harveyi* (LuxA).
[g]Identity and similarity were calculated using the program BESTFIT (Smith and Waterman 1981).
[h](ORF1) 3′ end.
[i]Translated ORF.
[k](ORF28) 5′ end.

evated gel running temperatures. Because of their longer reading potential, dye primer reads are definitely helpful for gap closure. However, using ABI 373A sequencers, dye primer reads are, on average, only ~50 bases longer than dye terminator reads.

We are aware that from the methodological point of view our comparison is not quite parallel because two different cosmids were compared using two different enzymes. However, under our experimental conditions, shotgun sequencing with dye terminators and a thermostable sequenase is superior because for GC-rich cosmid templates it removes most of the compressions and thus leads to a severalfold improvement in assembling and finishing of cosmid-sized projects. Although dye terminators are slightly more expensive than dye primers, the overall saving in time for finishing projects has, in our experience, a much greater effect on general costs.

We believe that our strategy is effective for high-throughput shotgun sequencing of GC-rich templates. We have therefore used this strategy to sequence the remaining 21 overlapping cosmids of the symbiotic replicon of *Rhizobium* sp. NGR234.

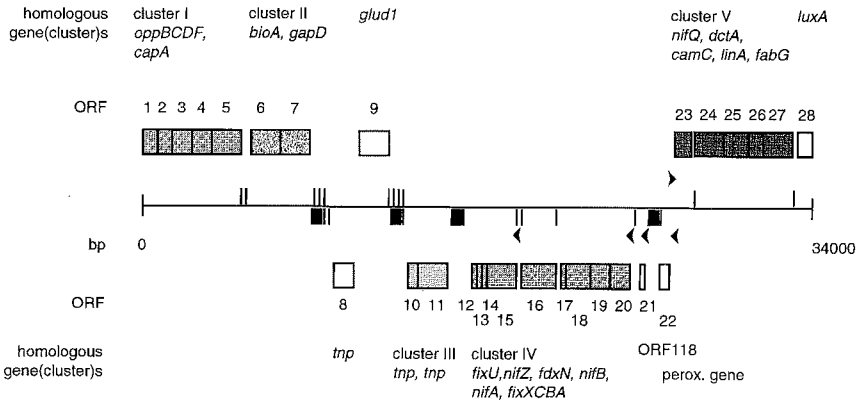

Figure 2 Organization of the predicted genes in pXB296 from *Rhizobium* sp. NGR234. Significant stem–loops (I)/stem–loop clusters (■), which might function as ρ-independent transcription terminators, are represented. Sequence motifs (open arrowhead), similar to σ^{54}-dependent promoter consensus sequences (T**GG**CACG-N_4-TT**GC**) + *nifA* upstream activator sequences (TGT-N_{10}-ACA), are located at the following positions on the cosmid: 19,107–19,120 + 19,195–19,210; 24,787–24,800 + 24,878–24,893; 25,508–25,521 + 25,567–25,582; 26,815–26,828 + 26,941–26,956 (all minus strand); 27,074–27,131 (several possibilities) + 26,969–26,984 (plus strand).

Genetic Organization of pXB296

All 28 predicted ORFs in pXB296 (Fig. 2) show significant homologies to data base entries. The first putative gene cluster (cluster I) containing ORF1 to ORF5 corresponds to various oligopeptide permease operons (Hiles et al. 1987; Perego et al. 1991). Only ORF5 shows homology to a gene from a different bacterium, *B. anthracis* (Makino et al. 1989). Each homolog encodes membrane-bound or membrane-associated proteins suggesting that all five genes are involved in oligopeptide permeation.

Organization of the predicted gene cluster IV, including the *nifA* homolog ORF16 (*fixABCX, nifA, nifB, fdxN,* ORF, *fixU* homologs, position 16,746–24,731), as well as the predicted locations of the σ^{54}-dependent promoters and the *nifA* upstream activator sequences (Fig. 2), corresponds to the organization found in *Rhizobium meliloti* and *Rhizobium leguminosarum* bv. *trifolii*. (Iismaa et al. 1989; Fischer 1994). NifA is a positive transcriptional activator (Buikema et al. 1985), whereas *nif* and *fix* genes are essential for symbiotic nitrogen fixation. Identification of σ^{54}-dependent promoter sequences, together with the upstream activator motifs upstream of ORF21, ORF22, and ORF23 suggests that these genes may play an important, but still undefined role, in symbiosis.

Inevitably, large-scale sequencing uncovers differences with already published sequences. van Slooten et al. (1992) cloned a 5.8-kb *Eco*RI fragment from *Rhizobium* sp. NGR234 and sequenced 2067 bp by manual radioactive methods (EMBL accession no. S38912). This sequence exhibits 2.4% mismatches with the corresponding sequence in pXB296. It contains the gene *dctA* (encoding a C_4-dicarboxylate permease), which is 144 bases shorter than in pXB296. In this respect, a single nucleotide deletion in position 29,248 of the cosmid sequence close to the 3′ end of the gene causes a frameshift leading to a DctA product extended by 48 residues. van Slooten et al. (1992) also failed to identify the *nifQ* homolog, ORF23 (position 27,169–27,861), presumably be-

cause they overlooked a small *Xho*I fragment located between positions 27,349 and 27,536 on pXB296. Expression studies allowed these investigators to define a putative σ^{54}-dependent promoter in a 1.7-kb *Sma*I fragment (position 27,094–28,818 in pXB296). This fragment stretches from the upstream region of ORF23 to the 5′ part of *dctA*. The 58-bp intergenic region between ORF23 and *dctA* contains a stem–loop structure but no obvious promoter sequence. Possibly the promoter that controls *dctA* is located upstream of ORF23 (e.g., the minimal consensus sequence included in **GG**GGGCACAATT**GC** at position 27,098–27,111). Although clones containing *dctA* complemented mutants of *R. meliloti* and *R. leguminosarum* for growth on dicarboxylates, the growth of the NGR234 *dctA* deletion mutant was not affected (van Slooten et al. 1992). Nevertheless, this mutant was unable to fix nitrogen in nodules. Because *dctA* is now possibly part of a larger transcription unit, the symbiotic phenotype may also result from the inactivation of downstream genes.

Interestingly, the GC content of the predicted pXB296 ORFs ranges from 53.3 mol% to 64.6 mol%, with an overall cosmid GC content of 58.5 mol%. Genomes of *Azorhizobium, Bradyrhizobium,* and *Rhizobium* species have GC contents of 59 mol% to 65 mol% (Padmanabhan et al. 1990), with 62 mol% reported for *Rhizobium* sp. NGR234 (Broughton et al. 1972). Although pXB296 covers <7% of the complete symbiotic plasmid sequence, its lower overall GC value suggests that symbiotic genes might have evolved by lateral transfer from other organisms. In this case, methods of the type applied here will become even more relevant in sequencing the whole genome.

Although functional analyses of selected ORFs in pXB296 still have to be performed, large-scale sequencing gives a global picture of their genomic organization and possible roles. Determination of putative functions of predicted genes by homology searches and identification of sequence motifs (promoters, *nod* boxes, *nifA* activator sequences, and other regulatory elements) will aid in finding new symbiotic genes. High-fidelity sequence data covering long stretches of the genome are a prerequisite for these studies. The dye terminator/thermostable sequenase shotgun approach will allow completion of the entire 500-kb sequence of pNGR234*a* within several months and open up new avenues for genetic analysis of symbiotic function.

METHODS

Bacteria and Plasmids

Escherichia coli was grown on SOC, in TB, or in twofold YT medium (Sambrook et al. 1989). The cosmid clones pXB296 and pXB110 (Perret et al. 1991) were raised in *E. coli* strain 1046 (Cami and Kourilsky 1978). Subclones in M13mp18 vectors (Yanisch-Perron et al. 1985) were grown in *E. coli* strain DH5αF′IQ (Hanahan 1983).

Construction of pXB296 and pXB110 Libraries

Cosmid DNA was prepared by standard alkaline lysis procedures followed by purification in CsCl gradients (Radloff et al. 1967). DNA fragments sheared by sonication of 10 μg of cosmid DNA were treated for 10 min at 30°C with 30 units of mung bean nuclease (New England Biolabs, Beverly, MA), extracted with phenol/chloroform (1:1), and precipitated with ethanol. DNA fragments, ranging in size from 1 to 1.4 kbp, were purified from agarose gels using Geneclean II (Bio101, Vista, CA) and ligated into *Sma*I-digested M13mp18. Electroporation of aliquots of the ligation reaction into competent *E. coli* DH5αF′IQ was performed according to standard protocols (Dower et al. 1988; Sambrook et al. 1989).

M13 Template Preparation

Fresh 1-ml *E. coli* cultures in twofold YT held in 96-deep-well microtiter plates (Beckman Instruments, Fullerton, CA) were infected with recombinant phages from white plaques growing on plates containing X-gal (5-bromo-4-chloro-indoyl-β-D-galactoside) and IPTG (isopropyl-β-thiogalactopyranoside). Rapid preparation of ~0.5 μg of single-stranded M13 template DNA was carried out as follows: 190-μl portions of the phage cultures grown for 6 hr at 37°C were transferred into 96-well microtiter plates. Lysis of the phages was obtained by adding 10 μl of 15% (wt/vol) SDS followed by 5 min incubation at 80°C. Template DNA was trapped using 10 μl (1 mg) of paramagnetic beads (Streptavidin MagneSphere Paramagnetic Particles Plus M13 Oligo, Promega, Madison, WI) and 50 μl of hybridization solution [2.5 M NaCl, 20% (wt/vol) polyethylene glycol (PEG-8000)] during an annealing step of 20 min at 45°C. Beads were pelleted by placing microtiter plates on appropriate magnets and washed three times with 100 μl of 0.1-fold SSC. The DNA was recovered in 20 μl of water by a denaturation step of 3 min at 80°C. When required, larger amounts of single-stranded recombinant DNA (>10 μg) were purified using QIAprep 8 M13 Purification Kits (Qiagen, Hilden, Germany) from 3 ml of supernatant of phage cultures grown for 6 hr at 37°C.

Sequencing

Two sequencing methods were used: dye terminator and dye primer cycle sequencing, each in combination with AmpliTaq DNA polymerase (Perkin-Elmer, Foster City, CA) and Thermo Sequenase (Amersham, Buckinghamshire, UK). All reactions, including ethanol precipitation, were performed in microtiter plates. Reagents were pipetted us-

ing 12-channel pipettes. Where necessary, sequencing reaction mixtures, including enzymes, were pipetted into the plates in advance and held at −20°C until needed.

Dye Terminator Cycle Sequencing

For dye terminator/AmpliTaq DNA polymerase sequencing, 0.5 μg of template DNA, and the PRISM Ready Reaction DyeDeoxy Terminator Cycle Sequencing Kit (Perkin-Elmer) were used. Cycle sequencing was performed in microtiter plates using 25 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 4 min at 60°C). Prior to loading the amplified products on electrophoresis gels, unreacted dye terminators were removed using Sephadex columns scaled down to microtiter plates (Rosenthal and Charnock-Jones 1993).

Dye terminator/Thermo Sequenase sequencing was performed using the same experimental conditions except that the reaction mix contained 16.25 mM Tris-HCl (pH 9.5), 4.0 mM $MgCl_2$, 0.02% (vol/vol) NP-40, 0.02% (vol/vol) Tween 20, 42 μM 2-mercaptoethanol, 100 μM dATP/dCTP/dTTP, 300 μM dITP, 0.017 μM A/ 0.137 μM C/0.009 μM G/0.183 μM T from Taq Dye Terminators (Perkin-Elmer; no. A5F034), 0.67 μM primer, 0.2–0.5 μg of template DNA, and 10 units of Thermo Sequenase (Amersham) in a 30-μl reaction volume. Unincorporated dye terminators were removed from reaction mixtures by precipitation with ethanol.

Dye Primer Cycle Sequencing

Dye primer/AmpliTaq DNA polymerase sequencing reactions were performed according to the instructions accompanying the Taq Dye Primer, 21M13 Kit (Perkin-Elmer). Cycle sequencing was carried out on 0.5 μg of template DNA with 19 PCR cycles (30 sec at 95°C, 30 sec at 5°C, and 90 sec at 72°C) followed by six cycles, each consisting of 95°C for 30 sec and 72°C for 2.5 min. Prior to electrophoresis, the four base-specific reactions were pooled and precipitated with ethanol.

Identical PCR conditions and the Thermo Sequenase Fluorescent Labeled Primer Cycle Sequencing Kit (Amersham) were used for dye primer/Thermo Sequenase sequencing reactions.

Sequence Acquisition and Analysis

Gel electrophoresis and automatic data collection were performed with ABI 373A DNA sequencers (Perkin-Elmer). After removing cosmid vector and M13mp18 sequences from the shotgun sequence data, the data were assembled using the program XGAP (Dear and Staden 1991) and edited against the fluorescent traces. To close remaining gaps, to make single-stranded regions double-stranded, and to clarify ambiguities, additional cycle sequencing reactions with selected shotgun templates were carried out using either custom-made primers (primer-walks) or universal primer.

The complete double-stranded DNA sequence of cosmid pXB296 was analyzed using programs from the Wisconsin Sequence Analysis Package (v. 8, Genetics Computer Group, Madison, WI). Homology searches were performed with BLAST (v. 1.4; Altschul et al. 1990) and FASTA (v. 2.0; Pearson and Lipman 1988). Several nucleotide and protein data bases were screened (GenBank/Genpept, SwissProt, EMBL, and PIR). Identities and similarities between homologous amino acid sequences were calculated with the alignment program BESTFIT (Smith and Waterman 1981).

ACKNOWLEDGMENTS

We thank Matthias Platzer for help and critical reading of the manuscript, Evelyn Michaelis for excellent technical assistance, and Bernd Drescher for managing the computer system. X.P. gratefully acknowledges the receipt of a European Molecular Biology Organization (EMBO) short-term fellowship. Financial support was provided by the Fonds National Suisse de la Rechèrche Scientifique (grant 31-36454.92). The sequence data described in this paper have been submitted to the EMBL data library under accession no. Z68203.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., G. Warren, W. Miller, E.M. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bodmer, W.F. 1994. The Human Genome Project. *Rev. Invest. Clin.* (Suppl.) 3–5.

Broughton, W.J., M.J. Dilworth, and I.K. Passmore. 1972. Base ratio determination using unpurified DNA. *Anal. Biochem.* **46:** 164–172.

Broughton, W.J., N. Heycke, H. Meyer z.A., and C.E. Pankhurst. 1984. Plasmid-linked *nif* and *"nod"* genes in fast growing rhizobia that nodulate *Glycine max, Psophocarpus tetragonolobus,* and *Vigna unguiculata. Proc. Natl. Acad. Sci.* **81:** 3093–3097.

Broughton, W.J., C.-H. Wong, A. Lewin, U. Samrey, H. Myint, H. Meyer z.A., D.N. Dowling, and R. Simon. 1986. Identification of *Rhizobium* plasmid sequences involved in recognition of *Psophocarpus, Vigna,* and other legumes. *J. Cell Biol.* **102:** 1173–1182.

Buikema, W.J., W.W. Szeto, P.V. Lemley, W.H. Orme-Johnson, and F.M. Ausubel. 1985. Nitrogen fixation specific regulatory genes of *Klebsiella pneumoniae* and *Rhizobium meliloti* share homology with the general nitrogen regulatory gene *ntr*C of *K. pneumoniae. Nucleic Acids Res.* **13:** 4539–4555.

Cami, B. and P. Kourilsky. 1978. Screening of cloned recombinant DNA in bacteria by in situ colony hybridization. *Nucleic Acids Res.* **5:** 2381–2390.

Craxton, M. 1993. Cosmid sequencing. *Methods Mol. Biol.* **23:** 149–167.

Dear, S. and R. Staden. 1991. A sequence assembly and

editing for efficient management of large projects. *Nucleic Acids Res.* **19:**3907–3911.

Dower, W.J., J.F. Miller, and C.W. Ragsdale. 1988. High efficiency transformation of *E. coli* by high voltage electroporation. *Nucleic Acids Res.* **16:** 6127–6145.

Fellay, R., P. Rochepeau, B. Relić, and W.J. Broughton. 1995a. Signals to and emanating from *Rhizobium* largely control symbiotic specificity. In *Pathogenesis and host specificity in plant diseases. Histopathological, biochemical, genetic, and molecular bases* (ed. U.S. Singh, R.P. Singh, and K. Kohmoto), Vol. I, pp. 199–220. Pergamon/Elsevier Science Ltd., Oxford, UK.

Fellay, R., X. Perret, V. Viprey, W.J. Broughton, and S. Brenner. 1995b. Organization of host-inducible transcripts on the symbiotic plasmid of *Rhizobium* sp. NGR234. *Mol. Microbiol.* **16:** 657–667.

Fickett, J.W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10:** 5303–5318.

Fischer, H.-M. 1994. Genetic regulation of nitrogen fixation in Rhizobia. *Microbiol. Rev.* **58:** 352–386.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, et al. 1995. The minimal gene complement of *Mycoplasma genitalium. Science* **270:** 397–403.

Gribskov, M., J. Devereux, and R.R. Burgess. 1984. The codonpreference plot: Graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* **12:** 539–549.

Hanahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166:** 557–580.

Hartl, D.L. and M.J. Palazzolo. 1993. Drosophila as a model organism in genome analysis. In *Genome research in molecular medicine and virology* (ed. K.W. Adolf), pp. 115–129. Academic Press, Orlando, FL.

Hiles, I.D., M.P. Gallagher, D.J. Jamieson, and C.F. Higgins. 1987. Molecular characterization of the oligopeptide permease of *Salmonella typhimurium. J. Mol. Biol.* **195:** 125–142.

Iismaa, S.E., P.M. Ealing, K.F. Scott, and J.M. Watson. 1989. Molecular linkage of the *nif/fix* and *nod* gene regions in *Rhizobium leguminosarum* biovar *trifolii. Mol. Microbiol.* **3:** 1753–1764.

Levy, J. 1994. Sequencing the yeast genome: An international achievement. *Yeast* **10:** 1689–1706.

Lewin, A., E. Cervantes, C.-H. Wong, and W.J. Broughton. 1990. *nodSU,* two new *nod* genes of the broad host range *Rhizobium* strain NGR234 encode host-specific nodulation of the tropical tree *Leucaena leucocephala. Mol. Plant Microbe Interact.* **3:** 317–326.

Makino, S.-I., I. Uchida, N. Terakado, C. Sasakawa, and M. Yoshikawa. 1989. Molecular characterization and protein analysis of the cap region, which is essential for encapsulation in *Bacillus anthracis. J. Bacteriol.* **171:** 722–730.

Martinez, E., D. Romero, and R. Palacios. 1990. The Rhizobium genome. *Crit. Rev. Plant Sci.* **9:** 59–93.

Morett, E. and M. Buck. 1988. NifA-dependent *in vivo* protection demonstrates that the upstream activator sequence of *nif* promoters is a protein binding site. *Proc. Natl. Acad. Sci.* **85:** 9401–9405.

———. 1989. *In vivo* studies on the interaction of RNA polymerase-σ^{54} with the *Klebsiella pneumoniae* and *Rhizobium meliloti nifH* promoters: The role of *nifA* in the formation of an open promoter complex. *J. Mol. Biol.* **210:** 65–77.

Padmanabhan, S., R.-D. Hirtz, and W.J. Broughton. 1990. Rhizobia in tropical legumes: Cultural characteristics of *Bradyrhizobium* and *Rhizobium* sp. *Soil Biol. Biochem.* **22:** 23–28.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Perego, M., C.F. Higgins, S.R. Pearce, M.P. Gallagher, and J.A. Hoch. 1991. The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. *Mol. Microbiol.* **5:** 173–185.

Perret, X., W.J. Broughton, and S. Brenner. 1991. Canonical ordered cosmid library of the symbiotic plasmid of *Rhizobium* species NGR234. *Proc. Natl. Acad. Sci.* **88:** 1923–1927.

Perret, X., R. Fellay, A.J. Bjourson, J.E. Cooper, S. Brenner, and W.J. Broughton. 1994. Subtraction hybridization and shotgun sequencing: A new approach to identify symbiotic loci. *Nucleic Acids Res.* **22:** 1335–1341.

Platt, T. 1986. Transcription termination and regulation of gene expression. *Annu. Rev. Biochem.* **55:** 339–372.

Radloff, R., W. Bauer, and J. Vinograd. 1967. A dye-buoyont-density method for the detection and isolation of closed circular duplex DNA: The closed circular DNA in HELA cells. *Proc. Natl. Acad. Sci.* **57:** 1514–1521.

Rosenthal, A. and D.S. Charnock-Jones. 1993. Linear amplification sequencing with dye terminators. *Methods Mol. Biol.* **23:** 281–296.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual,* 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Shine, J. and L. Dalgarno. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementary to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci.* **71:** 1342–1346.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sulston, J., Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, et al. 1992. The *C. elegans* genome sequencing project: A beginning. *Nature* **356:** 37–41.

Tabor, S. and C.C. Richardson. 1995. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci.* **92:** 6339–6343.

van Rhijn, P. and J. Vanderleyden. 1995. The *Rhizobium*-plant symbiosis. *Microbiol. Rev.* **59:** 124–142.

van Slooten, J.C., T.V. Bhuvanasvari, S. Bardin, and J. Stanley. 1992. Two C_4-dicarboxylate transport systems in *Rhizobium* sp. NGR234: Rhizobial dicarboxylate transport is essential for nitrogen fixation in tropical legume symbioses. *Mol. Plant Microbe Interact.* **5:** 179–186.

Yanisch-Perron, C., J. Ira, and J. Messing. 1985. Improved M13 phage cloning vectors and host strains: Nucleotide sequences of M13mp18 and pUC19 vectors. *Gene* **33:** 103–119.

Received February 22, 1996; accepted in revised form May 23, 1996.