DNA sequence of *Rhizobium trifolii* nodulation genes reveals a reiterated and potentially regulatory sequence preceding *nodABC* and *nodFE*

Peter R.Schofield¹ and John M.Watson*

CSIRO, Division of Plant Industry, GPO Box 1600, Canberra, ACT 2601, Australia

Received 13 February 1986; Accepted 13 March 1986

ABSTRACT

The Rhizobium trifolii nod genes required for host-specific nodulation of clovers are located on 14 kb of Sym (symbiotic) plasmid DNA. Analysis of the nucleotide sequence of a 3.7 kb portion of this region has revealed open reading frames corresponding to the nodABCDEF genes. A DNA sequencing technique, using primer extension from within Tn 5, has been used to determine the precise locations of Tn 5 mutations within the nod genes and the phenotypes of the corresponding mutants correlate with their mapped locations. The predicted nodA and nodB genes overlap by four nucleotides and the nod F and nodE genes overlap by a single nucleotide, suggesting that translational coupling may ensure the synthesis of equimolar amounts of these gene products. The nodABC and nodFE genes constitute separate transcriptional units and each is preceded by a conserved 76-bp sequence which may be involved in the regulation of expression of these genes.

INTRODUCTION

The soil bacterium <u>Rhizobium</u> is capable of infecting leguminous plants and inducing a highly-differentiated structure, the root nodule, within which atmospheric nitrogen is reduced to ammonia by the bacteria. The symbiotic interaction is characterized by the high degree of specificity observed between temperate legumes and their corresponding fast-growing species of <u>Rhizobium</u> microsymbiont.

Bacterial genes required for nodulation (nod) and nitrogen fixation (nif or fix) are located on large Sym (symbiotic) plasmids in the fast-growing <u>Rhizobium</u> species. In <u>R</u>. trifolii and <u>R</u>. leguminosarum the host-specific nodulation genes are located on 14 kb and 10 kb of Sym plasmid DNA, respectively (1, 2), indicating that there are relatively few bacterial genes involved in nodule induction. In <u>R</u>. meliloti, two separate clusters of nod genes have been identified (3, 4).

We have determined the nucleotide sequence of a 3.7 kb portion of the <u>R</u>. <u>trifolii</u> strain ANU843 nodulation gene region. This region has been shown to contain parts of three nodulation gene transcriptional units by:

i) directed Tn5 mutagenesis (5); ii) directed Mudl1734 mutagenesis (6) and iii) analysis of functions encoded on cloned restriction fragments (7).

The availability of a range of Tn 5-induced, nodulation-deficient <u>R</u>. <u>trifolii</u> mutants enabled us to utilize the Tn 5 insertions as sites from which to directly access <u>nod</u> gene sequences. Previously, Tn 5 insertion sites have been determined by sequencing from the <u>Hpal</u> site, located 185 bp from the end of Tn5 (8, 9). To facilitate sequencing into flanking DNA, we synthesized an oligonucleotide complementary to the ends of Tn5, for use in chain-termination sequencing.

The nodulation genes, identified as open reading frames in the nucleotide sequence, were found to correlate with the locations of Tn 5 insertions known to result in either the loss of plant root hair curling ability or altered plant host range.

Analysis of the DNA sequence revealed the presence of two copies of a conserved 76-bp sequence. Hybridization analysis also indicates the presence of a third copy of this conserved sequence in the 14 kb <u>nod</u> gene region. The two copies, identified within the sequenced region are located at the 5¹ ends of two of the <u>nod</u> gene transcriptional units suggesting that these sequences may be involved in regulating the coordinate expression of these genes.

MATERIALS AND METHODS

Strains and plasmids

The 14 kb fragment of <u>R</u>. <u>trifolii</u> strain ANU843 <u>nod</u> DNA (pRt587), identified previously (9), was used as a source of DNA for sequence analysis. The <u>nod</u>::Tn <u>5</u> mutations used in this study were described by Djordjevic <u>et al</u>., (5): Strains referred to as A22 through O7 are strain ANU845 (pRt032::Tn <u>5</u>) insertions. Strain numbers 245, 246, 249, 252, 258, 274, 276, 277, 297 and 851 are derivatives of strain ANU843 containing Tn <u>5</u> insertions in the Sym plasmid. M13 sequencing vectors mp18 and mp19 were used in <u>E</u>. <u>coli</u> strain JM107 (10).

DNA sequence determination and analysis

Standard molecular biology techniques for electrophoresis, DNA preparation and hybridization were carried out essentially as described previously (11).

The 3.7 kb <u>Eco</u>RI - <u>Bg</u>III restriction fragment was sequenced by the chain-termination method (12), using M13 sequencing vectors. The DNA sequence was determined on both strands using overlapping <u>Sau</u>3A and <u>Taq</u>I

fragments in conjunction with sequence obtained in both directions from the <u>EcoRI</u>, <u>PstI</u>, <u>BamHI</u>, <u>SmaI</u>, <u>SphI</u>, <u>ClaI</u> and <u>BgI</u>II sites indicated in Figure 5. The DNA sequence was compiled and analysed using SEQ and ANALYSEQ (13) programs.

Oligonucleotide synthesis

Both the universal M13 sequencing primer (3'-TGACCGGCAGCAAAATG-5') and the Tn <u>5</u> sequencing primer (see below) (3'-TTCATCGCAGGACTTGC-5') were synthesized in our laboratory by the phosphoramidite method (14)

Tn5-mediated sequence determination

To undertake chain-termination sequencing from <u>Tn5</u> into flanking <u>Rhizobium</u> DNA, a 17-mer oligonucleotide complementary to the sequence from nucleotides 16 to 32 in the arms of Tn5, was synthesized. This sequence was selected as it possessed least homology to other sequences in the arms of Tn5 (15). DNA fragments extending from <u>HindIII</u> sites within Tn5 to <u>EcoRI</u>, <u>BamHI</u> or <u>BgIII</u> sites in flanking <u>Rhizobium</u> DNA were cloned into mp18 and detected by using γ -[³²P]ATP-labelled Tn 5 primer. Tn 5 primer-extended, chain-termination sequencing from a given insertion site allows DNA sequence to be determined on different strands in each direction such that the terminal 10 bp of Tn5 can be seen (see Fig. 2). In each instance, the two sequences can be aligned with respect to each other by virtue of the 9-bp duplication of host DNA generated by Tn5 insertion (15).

This method may be used to determine the nucleotide sequence of regions of DNA that have been genetically probed by Tn5 mutagenesis or for rapidly determining the genetic location of random Tn5 mutants of other bacterial species if homologous genes have already been characterized.

RESULTS AND DISCUSSION

DNA sequence of the R. trifolii nod genes

The DNA sequence of the <u>EcoRI-BgIII</u> nod gene fragment was determined as outlined above and the 3668-bp nucleotide sequence is presented in Figure 1. Several open reading frames were identified, the predicted amino acid sequences of which are homologous to those of <u>nod</u> genes previously characterized in other <u>Rhizobium</u> species (16, 17, 4, 18). The six genes corresponding to these reading frames are designated <u>nodA</u>, <u>B</u>, <u>C</u>, <u>D</u>, <u>E</u> and <u>F</u>. Since we have been unable to detect the corresponding <u>R</u>. <u>trifolii</u> <u>nod</u> gene products (7), we utilized DNA sequence analysis programs that predict probable coding sequences on the basis of unequal

Nucleic Acids Research

codon usage. In all cases, both Staden's (13) and Fickett's (19) programmes indicate that the open reading frames translated in Figure 1 are within coding regions of the DNA.

The predicted <u>R</u>. trifolii nod gene products (nodA, 21.6 kD; nodB,

ABSGATGSCLEGEGATICSGLAGEGACCOCCUCTER TEACCOCCUCATE THE THE CODE OF TEACHAGE ACCUTCE TOTAL ADARGET CONCUCATE THE CONCUTATION CONTRACT TO CONCUTATION C	AGGGATGCCAG	A D	GCGGGG G A	GTATC	CGTT	IGCTG ICGAC	AAGC E	IGCCA ICGGT A V	GGTG V	GGCGT CCGCA A D	CGCG GCGC R	ATTTC TAAAC N F	TGGA ACCT	ACCG TGGC G	TUGTO AGCAG D D	AGCCA TCGGT A V	CATA GTAT Y	AACI TTGA V	ICGC/ AGCG R	AACA TTGT L M	AGGI G	CCGG/ GGCC1 P	ATAAC TATTC Y S	CTCC GAGGI S R	GTGC CACC P	GTTGGC/ CAACCGT Q C	120
To convert the second	PIGP	GCAATC CGTTAG	GGACAG CCTGTC	GACGC	CCCC CCCC	GACT	CTCG GAGC F	AACT	ACGA/ TGCT	AGGCA TCCGT P V	CGAT	GATGT CTACA	CCAC	GCTT CGAA	GCAG	AGAGC TCTCG	TCTC AGAG E	TTCO AAGO E	GACGO CTGCO	- GTGT CACA T N	TTTC 	CCGC GGCG/	TGTT ACAA/ T E	TCCG AGGC E S	ACTI TGAA	IGAGAGT ACTCTC/	240
TCGEMAGESTEMAGENCE TALACTICUATIAGEAGEGAMECAETICUATAMECEGAMECEGEGGEGATE TEGET TECHNOLOW AGE TATE (TECHNOLOW AGE TATE (TECHNOLO			• •		•		÷	* *	•	* *	•				•		-	-			-						
480 CONSIGNATION CONSIGNATION CONSTRUCTION C	TCGGAAAGCGTC AGCCTTTCGCAC R F A H	GAAGGA CTTCCT L V	CCTGTA GGACAT Q V	CACTC GTGAG S	CTAT GATA R Y	AGGA TCCT S	GGAG CCTC S	GAAA CTTT S L	GCAG CGTC	TCCAT AGGTA G Y	AAAG TTTC	CGAGA BCTCT S N	CGGC GCCG	GGCG CCGC A	ATGCT TACGA I S	GGTTG CCAAC T T	TTTC AAAG E	AAGO TTCC	CAAGO STTCO	GTCA CAGT T M	TGG ACC/	TTTC: AAAG/	TCGTO	CGAT CCTA	AA TO	GTTTTAG CAAAATG	360
$ \begin{aligned} \begin{aligned} & TREEDENDED AND EVALUATE LEVEN TO COMPARE THE THE THE COMPARE STATE THE THE COMPARE THE COMP$	* *	*	*	•	*			*		*		*	*		*	* *		*	*	*							
CAACCGTCGTGCTAAAAGACGATTGCAACGTCGGCCGAGACACGCAGGTACCGGGTAACGATAGATTCGGGCGCGGGGGGGG	ACCCUTTCACT	TGCTTC R L	CCTTTT S F *	AGGTG G R	IGGUT ICCGA	A TGC. TACG H	TCAC	GGCC	CTAT	TAGCG	GT TC	TCGCG	GCTT S	TCAC	AAACT. Q	AGTGC D R	GTCA	GGCC G 1	CACTO F L	GGAA K	CGTO CGTO	DAAGO N (GCT/ GCT/	AGAG E	TAGI D	FCCACCO	480
GTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	CAACCGTCGTG	CAAAAG	CACGAT	TGCAC	CTGG	CCGA	GCAG	CAGC	GAGT/	ACCGC	GTCA	ACGAT	AGAA	TCGA	CTCCG	GGACG	GGAC	CAGT	rc tc	TCGG	GTC	TATCO	GACC	AGTG	AACA	GGTCG	600
ASTCCGGCCGTTCATCGATC TTGAAAGGACATCCCGCGTTCGAAGGCCCGTATGTGCGAACATAGCGAACATGGCGAAGAGAGGATGGCTTCGTTTGTTGTTGCCGTTTATCTCT ASTCCGGCCGTAGTTACTAGACTTTCCTGTTAGCGCCAAGGTTCGCGCGCATTAGCACATATGCAACTTGCCGGAAGATGGCGGACGAGAGAGGATGGCTTCGTTTGTTGCGTTCTTCCTTTCCTGTAGCACAAGGAGAGAAGAGAGGAGAGATGGGAGGATAGAAGAGGAG	GTTGGCAGCACO	GTTTTC	GTOCTA V I	ACGTG A G	GACC	GGCT	CGTC A A	GTCG	CTCA	TGGCG √ A	CAGT	TCCTA V I	TCTT S	AGCT D V	GAGGC	CCTGC PR	CCTG S	IGTCA W E	∖GAG/ D R	AGCO P	CAG/ D	ATAGO	CTGG1 S W	TCAC' H	TTG1 V	FCCAGC1	
ABIC DEGRECATION CONCERNMENT C	* * *	* *	* *	* *	*	٠			•	•		• •			•	•		• •	• •	٠	٠	•	• •	٠			
F3 CGTCCATCGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTCCGGCGTTC TCAGGCCGCAAC L G A N	CATCGA GTAGCT/ M S	TCTTGA AGAACT R S	AAGGA TTCCT L V	CATC GTAG D	ICGCG IGCGC A	TTCC AAGG N W	AAGC TTCG A	GCCG GGCC	TATCG ATAGC Y R	TCGT/ AGCA R	ATATO TATAC I H	AACT	ACCGA TGGCA V A	CAACT GTTGA	TGGGG ACCCC P P	ACAC TGTG C	GCAG CGTC A S	SAGAC CTCTC S V	CGAT CTA	GGC1 CCG/ A	FTCG1 NAGCA E N	TTTG1 NAACA NT	FC TCO AGAGO E	CTTI GAAA K	TATCTCT TAGAGA	720
<pre></pre>	• • •	٠		*			*		* •	•	*					• •	*	٠		٠	٠					٠	
R E I R F R I	CGTTCGATCGCC	GAGGCT	GTCACA	80000 20000	TAAG	ATCG	GGGT CCCA	GGGTO CCCA	3 CATCO GTAGO	GTGTG	ATTG	GCGAC	CAAG	TGAC	CTTCT	SCTGC CGACG	AACA TTGT	CGGC	CGGAC	CAAG GTTC	ATCO TAGO	GGGGT	GGTO		2 GACA CTGT	TGCCCA	840
249 CCGAGGACCATCCCGGCGATCCGGCGATCCGGCGATCCGGCAGTCCGGCAGTCCGGCAGTGCGGATCGGCAGTCGGCGCGCGC	REIA	E S	υc	A I	*	*	РН * *	*	* +	і н •	N /	• •	L	н G * *	÷ ·	•	v	кн **	< v	۲ ۲	D	Р F +	1 0	к	۷	нG	
<pre>GCLGASACGAMAAGGIGGC TICGUSSIGNICGGCASIC GGCASIC GGCASIC GGCASIC GGC CGGC CGGC CGGC CGGC CGGC CGGC CG</pre>																						249					
G L V F F T A P V R H E A L V D L I Q G T C H P N P G D D F T L Y I N R D D S G GCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	GGCTCCTGCTT	TTTCCA	CCGACC	ATGCG	GGIG	AAGCI	CGCT	GCAC. CGTG	A ICC/ TAGGI	AGGA I FCC TA	C TGGC GACCO	JCAGT GGTCA	GCAG	ACGO	TAAC	JGACC CCTGG	GICG	AGCT	ITGC/	GTT	GTAC CATC	SATGI	ITACO NATGO	CTAG	GTCG CAGC	CTTCCT	960
CCTCATCGAAGGGCACTTCGCTTATATATAGCGCGCCGCTTCATAGCTCGCATCGCTTCGTTCG	GLVF	FΤ	A P	V R	н	E	A L	۷	DL	- 1	Q (ЗT	с	нP	N	G	D	DF	т	L	Y	1 1	R	D	D	SG	
CGAGAGTAGCTHAGCGATATATTCGCGCGGCGCGAGGGACAAGCAAGCTAGCT	GCCTCATCGAAC	CGGCAC	ттссст	TATAT	AAGC	GCGC	- CGCT	тсат	AGCTO	GCAT	CCGT	TTCGT	TCGA	TCAA	GATC	CAGAA	GGCC	ACTC	стс	ATT	GGTC	стсс	TATO	GGA	N ATA	CCAGGA	
A E D F P V E S I Y A R R K M CGAGCACGTCTTCGGTGCGGTGGAGGGGCATATCCGGCAGCCTCTGGAAGGGTCCACGGTACCGCAACCCCCCGTGACGATATTGGAGATACCGTCTCGGCAAACTCTCCCACGTGGTG GCTCGTGCAGAAGCCACGCCACGCCACGCCATATGCGCGCGGCACCTCTGGAAGGGTCCACCGAACCCCCCGTGACGAATATTGGAGATACCGTCTTGGCAGAGCCGTCTAGAGAGGTGCACCA L V D E T R T S P M D P L A D P L T S R V R V G T V I N S I G D R C F R E V H N TCCGCATAGCATGTCGGACGGCACCGCCACGCGCACGCCCCCGCACGTCCTGCAAGCCCAAGCCCCAGGTCGAGGGATGCCAAGCCCTCTCAGGACGCGCCTCCGGACGGCCCACGCCC R M A H R V T G F A F P V A L E Q L A P A L A R I S H P I G L G E L D P R V G Y ACAAGCCAGCTGGCCGGTCGGACGGCACCTCCTGCAGGCGCAGCTCCTGCGGAGCCCAAGCCCCAGGTCGGCGGCTGCGGGGGGTTCGGGCGGG	CGGAGTAGCTTO	SCCGTG																									1080
CGAGCACGTCTTCGGTGCAGGGGCATATCCGGCAGCGCATCTGGAAGGGTCGAGCGTACCGGACCCCCGTGACGCATATTGGAGATACCGTCCGCGCGAAGCCTTCCACGCGGCGACGCCATAGGCGCGCCGCAGCCCCCGATGGCTGGGGGGCCCGCATATGGCGGCGCGCCCCCCGTATAGGCGGCGCCCCCGGTATAGGCCGCCGCCGCCGCCCCCCGATGGCTGGGGGGCCGCCGCCGCCGCCTCTAGGCGGGCG			NGCGA	ΑΤΑΤΑ	TTCG	CGCG	GCGA	AGTA	TCGAC L E	CGTA	GGCA/ G N	AGCA R	AGCT	AGTT/	CTAG	STCTT S	CCGG PW	TGAG	CGAC	STAA M 1	CCAC P F	SCAGO ≀ G		CCTI P F	TAT		1080
L V D E T R T S P M D P L A D P L T S R V R V G T V T N S T G D R C F R E V H N 1320 TCCGCATAGCATGECGGACGGCACCGCAACGGCACCGCCACGCCCTGCAAAGCGGAGCCAAAGCGGGATGCAAGCGCCAGGCCCACGCCCTCCAGATCGGGCGGCGCGCGC	AEDF	ΡV	E S	ATATA	TTCG A	R R	GCGA R K	AGTA M	TCGAG L E	CGTA C	GGCA/ G N	R R	AGCT. E I	AGTT/ L	S G	S	CCGG PW	TGAG E *	S	M I	CCAC PF	G G	SA TAC	P F	TAT - V		1080
1520 1520 ACAAGCCAGETCGGCAGCGGCAGCGGCAGCGCCAGCTCCTGGCAGCGGAGCCAAGGGGGAGCCAAGGGGGATGGAGGGAG	A E D F CGAGCACGTCTT GCTCGTGCAGAA		E S CGCGTG GCGCAO	ATATA I Y GAGGG CTCCC	A A CATA GTAT	R F R F TCCG	GCGA R K GCAG CGTC	AGTA M # CGCA GCGT/	TCGAC L E TCTGC	C C SAAGG TTCC	GGCA/ G N GTCG/ CAGC1	AGCA R AGCGT	AGCT. E I ACAC TGTG	AGTT/ L GAACO	CTAGI S G CCCCG	STCTT S IGACG	CCGG PW *	TGAG E * TGGA ACCT	S S GATA CTAT	M I	CCAC P F * *	CAGO G G G G G G G G G G G G G G G G G G					1080
R M A H R V T G F A F P V A L E U L A P A L A R I S H P I G L G E L D P R V G Y ACAAGCCAAGTCGGGGACCAGCAAATCGGTATCACCAACTTITATGAATCGACGCAGCAGCACCCCATGTGACTTGCTTTGCATTGCTTAGCATCGTAAGCAAATCGGCGCGGCCCGGCCCCGG TGTTCGGTCAAGCCGCTGGTCGTTAGCCATCGCAACCCAAACCTTATACTTACCTCGCCCCATGGACCCCATGGGCCGGAACCCCCAGCGCTGACGCAAACCGAACCGAACCGAACCGAACCGAACCGAACCGAACGCAAGCGCGCGACC L G L E A V L L D T D G V K I F R R L L G M H S A I G K S D Y A I A R R E P R A 252 	A E D F CGAGCACGTCTT GCTCGTGCAGAA L V D E * * *	PV ICGGTGC AGCCACC ETF	E S CGCGTG GCGCAO R T *	ATATA I Y GAGGG CTCCC S P	A A CATA GTAT	CGCG4 R F TCCG4 AGGC4 D P	GCGA R K GCAG CGTC L	AGTA M * CGCA GCGT/ A	TCGAC L E TCTGC AGACC D P	C C SAAGG C TTCC L	GGCA/ G N GTCG/ CAGCT T S *	AAGCA R AGCGT FCGCA R *	AGCT. E I ACAC TGTG V R * *	AGTT/ L GAACO CTTGO V	ACTAGI S G CCCCG GGGC/ G T *	STCTT S IGACG NCTGC V	CCGG PW * ATAT TATA IN	TGAG E * TGGA ACCT S	GATA GATA CTAT	GTAAG M I (CCG' (GGC) G I *	CCAC PF ** TCTC AGAC DF	CAGG G G G G G G G G G G G G G G G G G G	GATAC I I I I I I I I I I I I I I I I I I I	P F * TCTCT	FTAT FV FCCA AGGT EV		1080
ACAGECAAGTICGECGACACCAGCAATICGETATCACCAACTITITATGAATCGACGCAGCAACCCCATGTGACTTGCTATGCCTTTGAATCGTAAGCAATTGCCCGGGGCTCCGGCCCCGGT TGTICGGTCAAGCCGCTGGTCGTTATGCATCGCAACGATACCTTATGATTGCTCGGGGGGCGCCCAAGCCCAACGCAACCGAACCGAACCGAACCGAACCGAACCGAACCGAACGCCGC	A E D F CGAGCACGTCTT GCTCGTGCAGAA L V D E + + + TCCGCATAGCAT AGGCGTATCGTA	PV TCGGTGG AGCCACG T F TGTCGG/ ACAGCC	E S CGCGTG GCGCAO R T * ACGGTA TGCCAT	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT	A CATA GTAT M CGCG GCGC	CGCGI R F TCCGI AGGCI D P AACGI TTGCI	GCGA R K GCAG CGTC L GCAC CGTG	AGTA M # CGCA GCGT/ A I CGCC/ GCGG	TCGAC L E TCTGC AGACC D P AGCTC TCGAC	C GAAGG TTCC L CTGC GGACG	GGCA/ G N GTCG/ CAGCT T S * AAAGC TTTCC	AGCGT R AGCGT FCGCA R * CCGGA GGCCT	AGCT. E I ACAC TGTG V R * * GCCA CGGT	AGTT/ L GAACO CTTGO V AAGCO TTCGO	CCCCG SGGGC/ G T * SGCGA	STCTT S IGACG NCTGC V ICGAG	CCGG PW ATAT TATAJ IN TGAG	TGAG E * TGGA ACCT S GGAT CCTA	IGATA CTAT CTAT	M I M I GGC/ G I * MAGCO	CCAC PFF TCTC AGAC DFF CCCT GGG4	CAGE G CGGCA CCCGT CCAG	GATAC I I I I I I I I I I I I I I I I I I I	P F * TCTCT KGAG4 R E CGGCC	FTAT F V FCCA AGGT E V CGCA		1080 1200 1320
1G11CG11CG41CG41C41G4CG41TAGC41CG41C41G41G41G4C4AAC13CG1CG1CG1CG4AAC41AGC4AACG4AACG4AACG4AACG	A E D F CGAGCACGTCTT GCTCGTGCAGAA L V D E * * * TCCGCATAGCAT AGGCGTATCGTA R M A F *	PV TCGGTGC AGCCACC ETF F TGTCGG ACAGCCT HR	E S COCOTO COCOTO COCOCAO R T * ACOGTA TOCCAT V T *	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT G F * *	A CATA GTAT M CGCG GCGC A	CGCGI R I TCCGI AGGCO D P AACGO TTGCO F P * *	GCGA R K GCAG CGTC L GCAC CGTG V *	AGTA M * CGCA GCGT/ A CGCC/ GCGG A I	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E	C C C C C C C C C C C C C C C C C C C	GGCA/ G N GTCG/ CAGC1 T S * AAAGC TTTCC L A	AGCGT R AGCGT FCGCA R * CCGGA SGCCT P *	AGCT. E I ACAC TGTG V R * * GCCA. CGGT A L	AGTT/ L GAACO CTTGO V AAGCO TTCGO A	ACTAGI S G CCCCG GGGGC/ G T * CGCGA R I	STCTT S IGACG NCTGC V ICGAG NGCTC S I	CCGG PW ATAT TATA IN TGAG ACTCI HP	TGAG E * TGGA ACCT S GGAT CCTA I *	CGAC S GATA CTAT I GCCA G G *	M I NCCG IGGC/ G I * NAGCO I TCGO L (CCAC P F TCTC AGAC D F CCC1 GGG4 G E	CAGC GGCA CCCGT CCAG	GATAC I GAAT F GATCC TAGG D +	CCCTT PF TCTCT GAGA RE CGGCC CCCGC PF T	TTAT TCCA AGGT E V CGCA SCGT R V		1080 1200 1520
252 H9 CACCCGCCCAACTGCGGCCAGTGTTGGAATGGATGGATGG	A E D F CGAGCACGTCTI GCTCGTGCAGAA L V D E TCCGCATAGCAT AGGCGTATCGTA R M A F ACAAGCCAAGTI	PV TCGGTGG AGCCACG TF TGTCGG ACAGCCT ACAGCCT TCGGCGG	E S CGCGTG GCGCAO R T * ACGGTA TGCCAT V T * * ACCCAGC	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT G F AAATC	A CATA GTAT M CGCG GCGC A GGTA	CGCGI R I TCCGI AGGCI D P AACGI TTGCI F P * *	GCGA R K GCAG CGTC L GCAC CGTG V * CAAC	AGTA M # CGCA GCGT/ A CGCC/ GCGG A I TTTT/	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E ATGAA	C C C C C C C C C C C C C C C C C C C	GGCA/ G N GTCG/ CAGCT T S * AAAGC TTTCC L A * CGCAC	A GCGT R AGCGT FCGCA R * CCGGA GGCCT P *	AGCT. E I ACAC TGTG V R * * GCCA. CGGT A L CCCA	AGTT/ L GAACO CTTGO V AAGCO TTCGO A TGTG/	ACTAGI S G CCCCG GGGC/ G T * CGCGA R I ACTTG	GACG GACG CTGC V TCGAG AGCTC S S TATG	CCGG PW ATAT TATAJ IN TGAGA ACTCO HP CCTT	TGAG E * TGGA ACCT S GGAT CCTA I * TTGA	GATA CTAT CTAT CTAT CGCCA CGGT G ATCG	GTAAC	CCAC P F TCTC AGAC D F CCCT GGGA G E S GCAA	CAGC GGCA CCCGT CCAG GGTC L +	GATAC I GAAT CTTA F GATCC D TAGG D	FCTCT AGAGA R E CGGCC CCCC P F SCGTT	TTAT TCCA AGGT E V CGCA GCGT R V + +		1080 1200 1320 14 4 0
272H9 CACCCCCCCCACTCGCAACGGCAGTCGCAACGGTTAGGATTGGATTGGATTGGACCAATAGGTTGGCGGAGAAAATTCAGACAGTTCCGGATGATCCGACGGTTCGAGACTCATTTTCCC GTGGGGGGGGTTGACGCCGGTCAGAGCTTTCCAAACGGTAACTTAGGCAGCCAGTCGGACCATAGGTTGCCGACGGCTCTTTTAAGTCTGTCAAGGCGGACGCGTCGAGGTCGAGGTCGAGGTCGAGGTCGAGGTCGAGGTCGAGGTCGAGGTCGAGTCGAGACGGTAGTCGAGGCGGAGGCGGTTGGTCGAGGCGGGTCGAGGTCGAGGTCGGAGTCGAGCGGAGCGGTTGGTCGAGCGGCACACGCGGAGCGCGGTCGAGGTCGAGCGGTGGGGCGGTCGGGGTGGGGCGGTTGGTCGAGCGGGTGGGGCGGTGGGGCGGTGGGGCGGTGGTGCGACTGCTCAACCAGGCGGCGGTGGGGCGGGC	A E D F CGAGCACGTCTI CCTCGTCAGAA L V D E * * * TCCGCATAGCAT R M A H AGACCCAATTI TGTTCGGTTCAA L G L E	PV TCGGTGC AGCCACC T F TGTCGGC ACAGCC T R TCGGCG/ AGCCGC T A	E S COCOTO COCOCAO R T * ACCOGTA TOCCAT V T * * ACCCACC TOGTCG V L	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT G F * * AAATC TTTAG L D	TTCG A CATA GTAT. M CCCCG GCCC A GCCA CCAT. T	CGCGI R F TCCGI AGGCI D P AACGI TTGCI F P * * TCACI AGTGI D G	GCGA R K GCAG CGTC L GCAC CGTG V * CAAC GTTG V V	AGTA M CGCA GCGT/ A CGCC/ GCGG A TTTT/ AAAA K	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E * ATGAA TACTI I F	C C C C C C C C C C C C C C C C C C C	GGCA/ G N GTCG/ CAGC1 T S * AAAGC TTTCC L A * CGCAC GCGTC R L	AGCGT R AGCGT FCGCA R * CCGGA SGCCT P * SCAAC	AGCT. E I ACAC TGTG V R * * GCCA CGGT A L CCCA GGGT. G M	AGTT/ L GAACO CTTGO V AAGCO TTCGO A TGTG/ ACAC1 H	ACTAGE S G SCCCCG SGGGC/ G T * SCCGA R I ACTTGC S A	GTCTT S IGACG ACTGC V ICGAG AGCTC S I S TATG GATAC	CCCGG PW H TATAT TGAGG ACTCC HP H CCTT GGAA/ GGAA/ G K	TGAG E * TGGA ACCT S GGAT CCTA I * TTGA AACT S	GGACGAC S GGATA CCTAT I FGCCA CGGT G * ATCCG TAGC D	ACCCG GGC/ GGC/ TCGG LCG ATTCG	CCAC P F TCTC AGAC D F CCCT GGGA G E GCAA CGTT A I	CGGCAGG CGGCA CCCGT CCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCAGCAG CCCAG CCCAG CCCAGCAG CCCAGCAG CCCAGCAG CCGCCA	GATAC I AGAAT TCTTA F GATCC CTAGO D * * CCCGGG GGGCC R	COCTI P F AGAGA R E CGGCCC CGCCGC P F CGCGCI CGCCAA R E	TTAT TTAT TCCA AGGT E V CCCA R V CCCA R V TCCG AGGC E P	IGGTCCT (L) ACGTGGT GCACCA (H) CGCCCGT GCGGCA (G) GCCGCA CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC	1080 1200 1320 1440
GTGGGGGGGTTGACGCCGGTCAGAGCTTTCCAAACCGTAACTTACGCGGTCAGCCTGGTATCCAACGAGCCTTTTTAAGTCTGTCAAGGCTGCCAAGCTCTAGTAAAAGGG G A W S R G T E F P K A N F A G T P G Y T A R F F E S L E A H D S P E L D N E W 246 AGGTTATTTTCCACCGCACTCCAAGCAGACATGCAAAGCTCCAATTCGTTCCTCAGCCCAACAGCCCAAAGCAGCAGAGATGGGTCAGCACAGCTCGAAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA	A E D F CGAGCACGTCTI CCTCGTGCAGAA L V D E * * * TCCGCATAGCAT R M A F ACAAGCCAAGTT TGTTCGGTTCAA L G L E * * *	P V TCGGTGG AGCCACG T F TGTCGG ACAGCC T R 1 TCGGCG AGCCGC T A 1 TCGGCG	E S COCGTG COCGCAD R T * ACCGGTA TOCCAT Y T * * ACCAGCA TGGTCG Y L *	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT G F * * AAATC TTTAG L D * *	TTCG A CATA GTAT M CCCCG GCCC A I GCTA CCAT. T I	R F TCCGG AGGCC D P AACGG TTGCC F P * * TCACC D G	GCGA R K GCAG CGTC L GCAC CGTG V * CAAC STTG V	AGTA M * CCCA GCCC/ A CCCC/ GCCCC/ GCCCC/ GCCCC/ TTTT// AAAAA K K *	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E * ATGAA TACT1 I F * *	C GAAGG CTTCC L CCTGC GGACG Q NTCGAN TAGCT R I	GGCA/ G N GTCG/ CAGC T S * AAAGC TTTCC L A * CGCAC R L *	AGCGT R AGCGT CCCGA R * CCCGGA GGCCT CGTTG L	AGCT. E I ACAC TGTG V R * * GCCA CGGT A L CCCA GGGT. G M *	AGTT/ L GAACC 2TTGC V AAGCC A AGACACA H *	ACTAGE S G SGGGC/ G T * SGGGA R I ACTTGC S A	STCTT S IGACG ACTGC V ICGAG AGCTC S I S TATG S ATACI	CCCGG PW * TATAT IN TGAGG ACTCC HP * CCTTT GGAA/ 3K *	TGGAG E * TGGA ACCT S GGAT CCTA I * TTGA AACT S	GGATA GGATA CTA1 I GGCCA G * ATCG D *	GTAAG M I GGC/ G I TCGC TCCG L G GTAAG CATTC Y /	CCAC P F TCTC AGAC D F CCCT GGGA G E GCAA CGTT A I	CGGCAGG CGGCAG CCCCGT CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCAG CCAG CCAG CCAG CCAG CCAG CCAG CCAG CCAG CCAG CCAG CCGCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCAG	GAATAC I AGAAT TCTTA F GATCC CTAGO D * CCCGG GGGCC R *	COCTT P F * TCTCT KGAGA R E CGGCC CGCCGC P F * * CGCAA R E *	TTAT TCCA AGGT E V CGCA GCGT R V * * TCCG AGGC E P	CGCTCGT GCACCA H N CCCCCGT GCGCCA G Y CGCCTCG CCGGAQC R A	1080 1200 1320 1440
G A W S R G T E F P K A N F A G T P G Y T A R F F E S L E A H D S P E L D N E W 246 AGGTTATTTTCCACCGCACTCCAAGCAGACATGCCAAAGCATCCATTCGTTCCTCAGCCAAAGCAGACAGA	A E D F CGAGCACGTCTI GCTCGTGCAGAU L V D E TCCGCATAGCAT AGGCGTATCGTA R M A F ACAAGCCAAGTT TGTTCGGTTCAA L G L E 252	P V TCGGTGG AGCCACC T F TGTCGGG ACAGCCT H R 1 + 4 TCGGCG AGCCGCT E A 1 + 4 CTGCGGC	E S COCGTG SCOCAO R T * ACCGGTA TGCCAT V T * ACCAGC TGGTCG V L * + * + + + + + + + + + + + + + + + +	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT G F * * AAATC TTTAG L D * *	A CATA GTAT. M CGCG GCGC A GGTA CCAT. T	R F TCCGG AGGCC D P AACGC F P * * TCACC AGTGC D G	GCGA R K GCAG CGTC L GCAC CGTG V * CAAC STTG V CATT(AGTA M T CCCA GCCGT/ A CCCCC/ GCCGC GCCGCG A I TTTT/ AAAA K K K SAAT(TCGAC L E TCTGC AGACCO D P AGCTO TCGAC L E * ATGAA TACTI I F * *	C GAAGG CTTCC L CCTGC GGACG Q ITCGA ITCGA ITCGA	GGCA/ G N GTCG/ CAGCT T S * AAAAGC TTTCC L A * CGCACC R L * *	AGCGT R AGCGT FCGCA R * CCGGA A GGCCT F * CCGGA C GCTTG L	AGCT. E I ACAC TGTG V R * * GCCA GGGT. A L CCCA GGGT. G M *	AGTT/ L GAACCO CTTGO V AAAGCO TTCGO A TGTG/ A CACACT H *	SCCCCG SGCCCCG GGCCCCG GGCCCC R R I CCCCCCT R I CCCCCCT R I CCCCCCG S A	GACGA GACGA V TCGAG S CTATGO CTATGO SATACO	CCCGG PW * ATATATA IN TGAGG ACTCC HP CCTTT GGAA 3K * SACAA	TGAG E * TGGA AACCT S GGAT CCTA I * TTGA AACT S GTTC	GCGAC S GATA CCTAT I GCCCA G * ATCC D *	GTAAG M I GGC/ G I TCGG TTCG GTAAG CATTC Y /	CCAC P F TCTC AGAC D F CCCT GGGA G E SCCAA CGTT A I	CGGCA CGGCA CCCGT CCAG CCAG CCAG CCAG CC	GAATAC I AGAAT TCTTA F GATCC CTAGO D CCTAGO D CCTAGO C CTAGO R *	COCOTTI P F F TCTCTCT AGAGA R E COGOCO COCOGC P F F SCGTTI COCAA R E SCGTTI	TTAT TTAT FCCA AGGT E V CGCA AGGC F F F CCCG F F F F F F F F F F F F F	IGGTCCT L V CGTGGT GCACCA H N CGCCCGT GCGGCA G Y CGCCTCG CGGAGC R A	1080 1200 1520 1440
	A E D F CGAGCACGTCTI GCTCGTGCAGAU L V D E TCCGCATAGCAI AGGCGTATCGTA R M A F ACAAGCCAAGTTI TGTTCGGTTCAA L G L E 252 CACCGGCCCAAC	P V TCGGTGC AGCCACC TGTCGGC ACAGCCT H R N TCGGCGC AGCCGCT E A N CTGCGGC GACGCCCC	E S CCCGTG CCCCAT ACCGTA TCCCAT V T * * ACCCACC TGCTCG V L * * * H9 CCAGTC GGTCAG	ATATA I Y GAGGG CTCCC S P CCGAA GCTTT G F * * AAATCC TTTAG L D * * TCGAA.	TTCG A CATA GTAT M CCCCG A CCAT, T T AGGT A AGGT	CGCGG R F TCCGG AGGCC D P AACGG TTGCC F P * * TCACC AGTGC D G TTGGC	GCGA R K GCAG CGTC L GCAC CGTG V * CAAC STTG V CATTC	AGTA M H CGCA GCGT/ A CGCC/ GCGC/ GCC/ GCC/ GCC/ GCC/ GCC/ GCC/ GCC/ GCC/ GCC/ GC/ G	TCGAC L E TCTGC AGACCO D P AGCTC TCGAC L E * ATGAA TACT1 I F * * GCGCCC	C GAAGG DTTCC L CCTGC GGACG Q NTCGA NTCGA R I CAGTC STCAG	GGCA/ G N GTCG/ CAGCT T S * AAAG0 TTTCC L A * CGCAA GCGTC R L * SGGACC CCTGC	R AGCGT TCGCA R * CCGGA GGCCT P * GCAACC CGTTG L CATAG GTATC	AGCT. E I ACAC TGTG V R * * GCCA CGGT A L CCCA GGGT. G M * *	AGTT/ L GAACCO CTTGC V AAGCC TTCGC A ACACT H * CTCGC GAGCC	SCACAGE SGCCCCG GGGGGC/ G T * SCGGGA GGCT/ R I SCGGGA S A SAAAAA/	GACGA GACGAG CCTQC V CCGAG GACTCC S CTATGO CTATGO CTATGO L CTATGO L CTATGO L CTATGO L CTATGO L CTATGO L CTATGO L CTATGO L CTACGAG GACGAG S C C C C C C C C C C C C C C C C C C	CCCGG PW * ATATT TATA/IN TGAGIA ACTCC HP* CCTTT GGAA/ GGAA/ GACA(CTGTC	TIGAG E * TIGGA ACCT S GGAT CCTA I * TITGA AACT S GTTC CAAG	GGATA GGATA CCTAT I GGCCA GGCCA CCGGT CCGCA CCGCA CCGCA	GTAAG M 1 GGC/ G I TCGC TCGC L C STAAG CATTC Y / *	CCAC P F TCTC AGAC D F CCCT GGGA G E SCAA CGTT A I SCCA AGGC	CGGCA CGGCA CGGCA CCCGT CCAG GGTC L L TTTGC CAACG A SACGG CTGCC	AGAATA I AGAATT CTTAGG D SACCGG GGGCCC R * STTCG SAAGC	COCTI P F * TCTCT AGAGA R E CGGCC CGCCGC CGCCGC CGCCGC CGCCGC CGCCGC	TTAT TTAT TCCA AGGT E V CGCA AGGT R V * * * * * * * * *	GGTCCT CGTGGT CGTGGT GGCGCCA CGCCGT GGCGCCA CGCCGT GGCGCCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCCCG CGGACCA CGCCGCA CA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCCGCA CGCGCA CGCCGCA CGCGCA CGCGCA CGCGCA CGCGCA CGCGCA CGCGCA CGCGCA CGGCA CGGACA CGCGCA CGGACA CGCGCA CGCGCA CGCGCA CGCGCA CGCGCA CGGACA CGCGCA CGGACA CGCGCA CGGACA CGCGCA CGCA CGCGCA CGCGCA CGC	1080 1200 1320 1440 1560
AGETTATTTTC/ACUGA/CICCAGUAGACATGCAAAGCCTCCAATCCAGCCAAACCAAGCCCAAAGCAGCATTGCTTCAGCGCACACCCTGATAGGACAGATGGGACCGCTT 1680 TCCAATAAAAGGTGGCGTGAGGTCGTCTGTACGTTTCGAGGTTAAGCAAGGAGTCGGTTGTTCGGGTTTCGTCGTCTCTAACCAAGTCGCGTGTGCGACTATCCTGTCTACCCTGGCGAA	A E D F CGAGCACGTCTI GCTCGTGCAGAV L V D E TCCGCATAGCAT AGGCGTATCGTA R M A P ACAAGCCAAGTT TGTTCGGTCGATCA L G L E 252 CAACCCGCCCCAAG G A W S	P V TCGGTGC AGCCACC T F TGTCGGC ACAGCCC H R 1 TGCGGCG AGCCGC T A 1 TGCGGCG CTGCGGC CTGCGGC CTGCGGC CTGCGGC CTGCGGC CTGCGGC CTGCGGC C A C TGCGGCC C A C C C A C C A C C C C C C A C C C C C C A C C C C C	E S CGCGTG GCGCAC R T ACCGGTA TGCCAT V T * * ACCAGCA TGGTCG GGTCG GGTCAG GGTCAG GGTCAG G T I	ATATA I Y GAGGG CTCCC S P CCGAA: GGCTTI G F AAATCC TTTAG L D * * TCGAA: AGCTTI E F *	TTCG A CATA GTAT. M CGCG GCGC A CCAT. T I CCAT. T AGGT TCCA. P I *	R F TCCGG AGGCC D P AACGG TTGCC F P * * TCACC AGTGG D G TTGGC AACCC K A *	GCGA R K GCAG CGTC L GCAC CGTC V * CAAC STTG V CATTC STAAC N	AGTA M + CCCCA GCCGT/ A CCCCC/ GCCGC GCCGC GCCGC A I TTTT/ AAAA K K S A CTTT/ K GAAT(CTTA(F /	TCGAC L E TCTGC AGACCC D P AGCTC TCGAC L E * * * * * * * * * * * * * * * * *	C GAAGG L TTCC L CCTGC GGACG Q N TCGA R I R I R I R I R I R I T I	GGCAA G N GTCGA CAGCT T S * AAAAGCC TTTCCA L A * CGCAAC GCCGTCC R L * * GGCACCC CCCTGC P G	R AGCGT TCGCAA R * CCGGAA CGCTTC L CATAG CATAG TATC Y	AGCT. E I ACAC TGTG V R * * GCCA CGGT A L CCCA GGGT. G M * GTTG CCAAC T A	AGTT/ L GAACCO TTGO V AAGCCO TTCGO A TGTG/ H * CTCGO GAGCO R	CCCCCG S G CCCCCG GGGGGC/ G T * CCCCCGA CCCCC/ R I CCCCCGA CCCCC/ R I CCCCCGA CCCCCGA CCCCCGA CCCCCGA CCCCCGA CCCCCGA CCCCGGGCC/ CCCCGGGCC/ G T A GACAGO CCCCCG GGGGCC/ G T A GACAGO CCCCCG GGGGCC/ G T A GCCCCG CCCCG GGGGCC/ G T A GCCCCG CCCCG GGGCC/ G T A GCCCCG CCCCG GGGCC/ G T A GCCCCG CCCG CCCCG CCCG CCCCG CCCCG CCCG CCCCG CCCCG CCCCG CCCCG C	GACGA GACGA V TCGAG GAGCTCC V TCGAG GAGCTCC I TTTTCA I TTTCAI E	CCCGG PW ATATI TATAJIN TGAGGA ACTCC HP * CCTTI GGAAJ GACAG SL	TGAG E * TGGA ACCT S GGAT CCTA A ACT S GTTC CAAG E	GGATA GGATA CTAT I GGCCA CGGT G * AATCO D * CGCCA A CGCCA A	GTAAG M I GGC/ GG I * AAGCC TTCGGG TTCGG TTCGG TTCGG TTCGG TTCGG TTCGG TTCGG T	CCAC P F TCTC AGAC D F CCCT GGGA G E GGCA A CGTT A I TCCC AGGC D S	CGGCA CGGCA CCGCCA CCCAG GGTC L + + + TTTGC AACG A + + CCAG C A CCAG C CCAG C CCAG C CCAG C CCAG C C CCAG C C CCAG C C CCCAG C C C C	AGAAT I AGAAT F SATCC CTAGO D * CCCGG GGGCC R * STTCCG E CAAGC E	P F * TCTCT R E CGGCC CCCGC P F * CGCGT CGCAA R E CGCAA R E CGCAA R E CGCAA CCCGT CCCGAA CCCGT CCCGAA CCCGCAA CCCGAAA CCCGAAA CCCGAA CCCGAAAA CCCGAAA CCCGAAAA CCCGAAAA CCCGAAAAA CCCGAAAAA CCCGAAAAAA CCCGAAAAAA CCCGAAAAAAAAAA	TTAT TTAT TCCA AGGT TCCA TCCG AGGC TCCG AGGC TCCA	GGTCCT GGCACCA H N CGCCCGT GCGCCCG G Y CGCGCCGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGAGC CGGACCA CGCCCG CGCACCA CGCCCG CGCACCA CCACCACCA CGCACCA CCACCACCACCA CGCACCA CCACCACCACCACCACCACCACCACCACCACCAC	1080 1200 1320 1440 1560
TIKWRVGASM	A E D F CGAGCACGTCTI GCTCGTGCAGAV L V D E TCCGCATAGCAT AGGCGTATCGTA R M A F ACAAGCCAAGTT TGTTCGGTCGATCA L G L E 252 CAACCCGCCCCAAC GTGGGCCGGCTTG G A W S	P V TCGGTGC AGCCACC T F TGTCGGC ACAGCCT H R 1 TCGGCGC AGCCGCT E A 1 CTGCGGC SACGCCC SACGCCC SACGCCC C R C	E S CCCGTG CCCCAC R T * AACCGTA TGCCAT Y T * * * AACCACC CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCACT CCCCAC CCCACC CCCAC CCCACCCACC CCCACCCACCCACCCACC CCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCACCCC	A TATA I Y GAGGG CTCCCC S P CCGAAA GCCTT G F * * AAAATC: TTTAG L D * * * TCGAA. ACCTT E F *	TTCG A CATA GTAT M CCCCG GCCC A CCAT. T CCA. P I A CCATA CCATA CCATA CCATA CCATA CCATA CCCCG A CCATA A CCATA A CATA A CATA A CATA A CATA A CATA A CATA A CATA A CATA A CATA A CATA CATA CATA A CCCCG C CCCG C CCCCG C CCCCG C CCCCG C CCCCG C CCCCG C CCCC C CCCCG C CCCCG C CCCCG C CCCC C CCCCG C CCCC C CCCCG C CCCC C CCCCG C CCCC C CCCC C CCCC C CCCC C CCCC C CCCC	CGCGG R F TCCGG AGGCC D P AACGC F P * * TCACC AGGCG AGGCG AACCCC K A *	GCGA R K GCAG CGTC L GCAC CGTG V X X CATTG V X X X X X X X X X X X X X X X X X X	AGTA M * CCCA GCCA A CCCCA GCCCA GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC GCCCC A CCCCCA CCCCCA CCCCCA CCCCCA CCCCCA CCCCCA CCCCCC	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E * ATGAA TACTT I F * * GCGCC CGCGC A G *	C C C C C C C C C C C C C C C C C C C	GGCA/ G N GTCG/ CAGCT T S * AAAGC TTTCC L A * CGCAC GCGTC R L * * GGCACC CCTGC P G	AAAGCA R AGCGT TCCGCA R * CCCGGA GGCCT P * CCCGGA CGTTG L CATAG GTATC Y	AGCT. E I ACAC TGTG V R * * GCCA. CGGT A L CCCA GGGT. G M * * *	AGTT/ L GAACCO 2TTGO V AAGCC TTCGO A TGTG/ A CTCGO GAGCC R	CCCCCG S G CCCCCG GGGGCC/ G T * CCCGGA C GGACCT S A CCCTGC S A CCCTGC S A CCCTGC S A CCCTGC S A CCCCCG S A CCCCCG S C S C S C S C S C S C S C S C S C S C	GACGA GACGA ACTGC V TCGAG GACTCC S TATGG CTATGG TATGGA TATCA TATCA TATCA	CCCGG PW ATAT TATAJ IN TGAGG ACTCC HP CCTT GGAAJ GCTT GGAAJ GCTT GGAAJ GCTT CTTC SL V 244	TGAG E * TGGA ACCT S GGAT CCTA I * TTGA AACT S GTTC CAAG E 6	S GATA CTAT I G GCCA A TCGCA CGGT G * CGCA CGCA CGCA A	GTAAG M I GGC/ G I * AAGCC TTCG L C GTAAG CATTC Y / * *	CCAC P F TCTC AGAC D F CCCT GGGA G E GCAA CGTT A I TCCC D S	CGGCA GGGCA GGGCA CCCGT CCAG GGCCGT L CCCAG A GGCCG A CCAGGC A CCAGGC A CCAGGC A CCCAGG C A CCCAGG C CCAGG C CCAGGCA C CCCGT C CCAGG C C CCCGT C CCAGGCA C CCGCA C C CCGCA C C CCGCA C C CCGT C C CCGCA C C CCGT C C C C C C C C C C C C C C C	GAATAC I AGAAT F GATCC CTAGC D * CCCGGC GGGCCC R * STTCC E	P F * AGAGA R E CGGCCC CCCGC P F * * CGCTA R E * * *	TTAT TTAT TCCA AGGT E V CGCA AGGC E P TCCG AGGC E P TCCG AGGC E P TCCG AGGC A AGGT A A A A A A A A A A A A A	GGTCCT GGACCAA (H N GCACCAA (H N GCCCCGT GCGGCA C Y R A GCCCCCGAC R A C Y C Y C Y C Y C Y C Y C Y C Y	1080 1200 1320 1440 1560
GTCGTCTATTTGCGACAGATACGATCCGCCTCCGCTCGCT	A E D F CGAGCACGTCTI GCTCGTGCAGAA L V D E TCCGCATAGCAT AGGCGTATCGTA R M A F ACAAGCCAAGTI TGTTCGGTCCAAG L G L E 252 CACCCGCCCCAAC G G L S CACCCGCCCCAAC GGGGCGGGTTG G A W S AGGTTATTTTCCC TCCAATAAAAGG T I K W	P V TCGGCGCACC T F F T F F TCGCGCGC TCCGCCGCC TCCGCCGC TCCGCCGC TCCGCCGCC TCCGCCGCC TCCGCCGC TCCGCCGC TCCGCCGC TCCGCCGC TCCGCCGC TCCGCCGCC TCCGCCGC TCCGCCGC TCCGCCGCC TCCGCC TCCGCCGCC TCCGCC TCCGCCGCC TCCGCCC TCCGCC TCCGCC TCCGCC TCCGCC TCCGCCC TCCGCCC TCCGCC TCCGCC TCCGCCC TCCGCCCC TCCGCCCCC TCCGCCCCC TCCGCCCCC TCCGCCCCCC TCCGCCCCCC TCCGCCCCCCC TCCGCCCCCCCCCC	E S COCCATO R T ACCGGTA R T CCCAT R T CCCATO R T CCCATO CCCATO CCCCA R T CCCCA CCCA CCCA CCCA CCCA CCCA CCCA C	ATATA I Y GAGGG CTCCCC S P CCGAA GGCTT G F AAATCC TTTAG L D * * TCGAA. AGCTT E F * * * *	TTCG A CATA GTAT. M CGCG GCGC A GCGC A CCAT. T T CATG STACO M	R F TCCGG AGGCC D P AACGC TTGCC F P * * TCACC AGTGC D G TTGGC K A * *	GCGA R K GCAG CGTC L GCAC CGTG V CATTO STAAC N SCTCC CGAGO	AGTA M + CCCA GCGC7/ A A CCCCC/ GCGC GCGG A I TTTT/ AAAA K * GAATC GAATC GTTA/	TCGAC L E TCTGG AGACC D P AGCTC TCGAC L E ATGAA TACTTI I F * * SCCGCC CCGCGC A G * *	CCTGCC GAAGG CTTCC L CCTGCC GGACG V VTCGA T CCTCCA CTCAG CCTCC GGGAG	GGCA/ G N GTCG/ CAGCT T S AAAGA AAGCA/ R L A GGCACG CCTGC P G AGCCA/ TCGGT	AAGCAA R AGCGTT CCCGCA R * CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGGCA CCCGCA CCCGCCCCC CCCGCA CCCCCCGCA CCCCGCA CCCCCCA CCCCCCCA CCCCCCA CCC	AGCT. E I ACAC TGTG V R * * GCCA. CGGT A L CCCA G M * SGTTG CAAC T A GCCCC, CGGG	AGTT/ L GAACC 2TTGC V AAGCC TTCCC A ACACC H * 2TCCC GAGCC R AAAGC TTCCC	CCCCCG S G CCCCCG G T + CCCCGA CG CGCCT/ R I + CCCCCC S A CCCTCG C S AAAAAA FTTTTT F F	STCTT: S IGACG. VCTGC V TCGAG GAGCTC. S TTTCAI I I I I I I I I I I I I I I I I I I	CCCGG PW * TATAT TATAJ IN TGAGI ACTO GAAA CCTTT GGAA SACAA CCTTT GGAA SACAA CCTTT GGAA SACAA CCTTT GGAA	TGAG E * TGGA AACCT S GGAT CCTA I * TTGA AACT S GTTC CCAAG E 6 TCAG AGTC	IGATA S IGATA CTA1 I I GCCA CGGT G * ATCC G * CGCA A CGCA A CGCA GCGT	M I I ACCCG ¹ (GGC/) GGC/ GGC/ GGC/ GGC/ GGC/ GGC/ CATTO CATTO CATTO CATTO CATCO GGC/ GGC/ GGC/ CATCO GGC/ GGC/ CATTO CATCO CATTO CATTO CATCO CATTO CATTO CATCO CATCO CATTO CATCO CATCO CATTO CATTO CATCO CATTO CATTO CATCO CATTO CATTO CATTO CATCO CATTO CATO CATTO CAT	CCAC P F TCTC AAGAC D F CCCT GGGA G E GCAA CGTT A I TCCCG AGGC D S SCTG CCAC	CGGCA CGGCA CCCCGT CCCAG GGCTC L CCCAG GGCTC L L TTCCC AACG A * * *	AGAAT CTTA F SATCC CTAGO D * CCCGG GGGCC R * STTCG E SAAGC E	P F F TCTCT AGAGA R E CGGCCC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGC CGCCCGCC	TTAT TTAT TTAT TCCA AGGT AGGT AGGCA AGGCA TCAT AGTA AGTA AGTA AGCA	GGTCCT / L \ \ CCGTGGT GGCACCA / H \ \ * CCGCTCG CCGCACCA / H \ \ * CCGCTCG CCGCACCA / H \ \ * CCGCTCG CCGCTT	1200 1220 1320 1440 1560
	A E D F CGAGCACGTCTI GCTCGTGCAGAV L V D E TCCGCATAGCAT AGGCGTATCGTA R M A H ACAAGCCAAGTI TGTTCGGTTCAA L G L E 252 CACCCGCCCAAG G A W S AGGTTATTTTCC TCCAATAAAAC T I K W GTCGTCTATTTG CAGCAGATAAAC	P V TCGGTGG AGCCACC T F F TGTCGGG CACACCC TGTCGGGG TGTCGGGG TCGGCGGT TCGGCGGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGT TCGGCGCGCGCGT TCGGCGCGCGT TCGGCGCGCGCGCGT TCGGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGC	E S CCCCGTG CCCCCCA T ACCGTAA TGCCAT TGCCAT CCCACC TGGTCG G V L H H H H H CCCACCCA T CCCCCAC CCCCCA T CCCCCCA CCCCA CCCCCA CCCCCA CCCCCA CCCCCA CCCCCA CCCCCA CCCCA CCCCCA CCCCA CCCCCA CCCCA CCCCCA CCCCCA CCCCCA CCCCA CCCCA CCCCA CCCCA CCCCA CCCCA CCCCA CCCCCC	ATATA I Y GAGGG CTCCC S P CCGAA GGTT G F * * AAATC: TTTAG L D * TCGAA. ACTT F * SCAGAC CGTCTC A S NTCCGC	A CATA GTAT M CGCG GCGC A GGCA CCAT. T CATGO GGTA CCAT M CCTCO GGAGO	R F TCCGG AGGCC D P AACGC TTGCC F P TCACC AGGGG C AACCC C AACCC C AACCC C C C	GCGA R K GCAG CGTC L GCAC CGTG V * CATTA STAAC N SCTCC CGAG CGCT/ SCGA	AGTA' M * CGCAT/ A I CGCCG GCGCG A I TTTTI AAAA' K * * CAATI GAATI GAATI GTTAA AGTTI TCAAA	TCGAC L E TCTGC AGACC D P AGCTC TCGAC L E * * * * * * * * * * * * * * * * * * *	C C C C C C C C C C C C C C C C C C C	GGCA/ G N GTCG/ CAGCT T S * AAAAGCTTTTCC L A * CGCAAGCCA CCCGCT GCCTTC GCCTTT CCGCAA	AAAGCA R AGCGGT TCCGCA R * CCCGGA R * CCCGGA CCTTG CGTTG CGTAG CCTAT	AGCT. ACAC TGTG V R * + CCGGT A L CCCA GGGT. G M * GTTG CCAC GCCC/ CCGGG AGTTC CCGGG	AGTT/ L GAACCO 2TTGO V AAAGCO A ACACT H * CTCGG GAGCO R AAAAGC CCGGGCCO	CCCCCG S G GGGGCC/ G T * CCCCGGA CCCCCG R I CCCCCG CCCCC CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCG CCCCCG CCCCCG CCCCCG CCCCCG CCCCCG C CCCCGG CCCCG CCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCCG CCCG CCCCG CCG CCC	STCTT S IGACG ACTGC V ICGAG AGCTC V ICGAG AGCTC S I I I I I I I I I I I I I I I I I I	CCCGG PW * ATAT TATAJ IN TGAGI ACTCC HP CCTT GGAA SK * * SACAA CCTTT GGAA SK * *	TIGAG E * TIGGA AACCT S GGAT CCTA GGAT CCTA GTTC CAAG AACTC AACTC AACTC	IGATA S IGATA CTAT I I I I I I I I I I I I I I I I I I	ACCG ACCG ACCC	CCAC P F TCTC AGAC D F CCCT GGGA G E GGCAA CGTT A I TCCC GGCAA CGTT A I TCCC CGAC	CGGCA CGGCA CCCGT CCCAG GGTC L CCCAG GGTC CCAG GGTC CCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCAG CCCGT CCCAG CCC	AGAAAT I AGAAAT F SATCC D SATCC D SATCC D SATCC D SATCC D SATCC D SATCC D SATCC D SATCC D SATAC D SATAC D SATAC D SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATAC SATCC D SATCC D SATCC SATCC D SATCC	COCTT P F F * CTCTC1 AGAG4 R E CGGCCC CCCGC P F * CCCGT1 CGCAA R E * CCCGT1 CGCAA L E CTAC CTAC	TTATT TTAT TCCA AGGT TCCA AGGT AGGCA TCCCG AGGCA TCCCG AGGCA	Getter	1200 1320 1440 1560 1680
	A E D F CGAGCACGTCTI GCTCGTGCAGAV L V D E TCCGCATAGCAI AGGCGTATCGTA R M A A ACAAGCCAAGTTI TGTCCGGTCAAG CACCGGCCCAAC G A W S AGGCTATTTGC GCACTATTTGC CACCGGCCTATTTGC CACCGCCTATTTGC CACCGCCTATTTGC CACCGCCTATTTGC CACCGCCTATTTGC CACCGCCTATTTGC CACCGCCTATTTGC CACCGCCTATACAC	P V TCGGTGG AGCCACC T T F TGTCGGG CAGCCCC T A N TCGGCGG A N TCGGCGG A N TCGGCG A N TCGGCGCG A N TCGGCGCGC A N TCGGCGCGCGC A N TCGGCGCGCGC A N TCGGCGCGC A N TCGGCGCGC A N TCGGCGCGCGC A N TCGGCGCGCGC A N TCGGCGCGCGCGC A N TCGGCGCGCGCGC A N TCGGCGCGCGCGCGCGC A N TCGGCGCGCGCGCGC A N TCGGCGCGCGCGCGCGCGCGCGCGC A N TCGGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGC	E S CCCCTCC R T ACCGTA TGCCAT CCCCCAT CCCCCCA CCCCCCA CCCCCCA CCCCCCA CCCCCCA CCCCCC	ATATA I Y GAGGG CTCCC S P CCGAA GGCTT: G F AAATC: TTTAG L D * * * * * * * * * * * * * * * * * * *	TTCG A CATA GTAT M CCCCG GCCC A GCCAT. T CATG GGTA CCAT. * CATG GGTACI M CCTCC GGAG * *	R F TCCGG AGGCC D P AACGC TTGCC F P * * TCACC AGTGC D G TTGGC AACCG K A * CAAAAC GTTTC CGCTCC GCGAC *	GCGA R K GCAG CGTC L GCAC CGTG V * CATTC STTAAC N SCTCC CGAG CGAC CGAC CGAC CGAC CGAC CGAC	AGTA M CGCA CGCA CGCG A L CGCCC GCGG GCGG A L CGCCC GCGG GCGG A L CGCCC GCGG CGCC CGCCC GCGG CGCCC CGCCC CGCC CGCCC CGCCC CGCCC CGCCC CGCC CGCC CGCC CGCCC CGCC CGCC	TCGAC L E TCTGC AGACCO D P AGCTO TCGAC L E ATGAC TACTI I F * * SCCCCC CCCCCC CCCCCC CCCCCC CCCCCC CCCCCC	C GAAGG C TTCCC C CTCC C C CTCC C C C C	GGCA/ G N GTCG/ CAGC ¹ T S * AAAAGC TTTCC R L * * CGCAAG CCTCG CCTCG P G AGCC/ TCGGT CCCTT CCGCA * *	AAAGCAA R AGCGTT FCGCAA R BCCGGAA GGCCT P BCCAACC GGTTG L CATAG Y CGATA CCTAT * V	AGCT. E I ACAC TGTG V R * * CCCA GGGT. G M * GTTG CCCGGG CCCC, CCGGG AGTTG CCGGG	AGTT/ L GAACC CTTGC V AAGCC TTCGC A ACACT H + + CTCGC GAGCC R AAAGC TTTCC CCGGC GGCCC + C	CCTAG S G CCCCCG GGGGC/ GG T * CCGCGA CGCCA CGCCA S A CCTTGC GAACCA CGCAACCA SAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GACGA IGACGA IGACGA ICCGAG ICCCGAG ICCCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCCGAG ICCCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCGAG ICCCCGAG ICCCGAG ICCCCGAG ICCCCGAG ICCCGAG ICCCCGAG ICCCCGAG ICCCCGAG ICCCCGAG ICCCGAG ICCC	CCGG PW * ATATT TATAJ IN TGAG ACTCC HP * TGGT SL QAAAA CTGTC SL QAAAA CTGTC SL QAAAA CTGTC SL QAAAAA CTGTC SL QAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	TGAG E * TGGA ACCT S GGAT CCTA GGAT CCAAG E 6 6 TCAG ACTC TGAG *	GGATA GGATA CTAT CTAT CGGGT G * CGGCA CGGGT A CGGCA CGGAA GCGT	AGATTO	CCAC P F TCTC AGAC D F CCCT GGGA G E GCAA CGTT A I TCCC GAC D S CCTC CCAC	CGGCAGC CGGCAGC CCCGT CCCAG AGGTC CCCAG AGGTC CCCAG AGGTC CCCAG AGGTC CCCAG AGGTC CCCAG CCCAG CCCAGC CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCGT CCCAG CCCAG CCCGT CCCAG CCCAG CCCGT CCCAG CCCCAG CCCCCAG CCCC	AGAATA GAAATA F SATCCTTA F SATCCTAGE D CCCGG GGGCCC R * * STTCG E SAAGC E SAAGC E	CCCTT PFF * CCCTTC CCCCCCCCCCCCCCCCCCCCCCCCCCCC	TTATT TTATT TCCA AGGT TCCG AGGT TCCG AGGC TCCG AGGC TCCG AGGC TCCG AGGC TCCG AGGC TCCG AGGC TCCG AGGT	Getter	1200 1200 1320 1440 1560 1680

• • • • • • • • • • • • • • • • • • • •	
.S Q P A M S A A I G R L R A Y F N D E L F L M Q Q R R L V P T P R A E A L A P A	
CAGTCAGCCAGCCATGAGCGCAGCCATCGGCAGGTTGCGCGCCTACTTCAACGACGAACTGTTTCTGATGCAGCAGCGCAGACTTGTCCCGACACCGCGAGCGGAGGCGCCCCCCCC	2040
GTCAGTCGGTCGGTACTCGCGTCGGTAGCCGTCCGCCGCGGATGAAGTTGCTGCTTGACAAGACTACGTCGTCGCGTCGCGTCGGCGCTCGCCCCCCCC	
245/274	
AGTICGCGGGGGCTCTICTGCACATICAGCTCTCTGTGATIGGCGGGGATCCATIAGTCCCGGCGGAATCTGGCCGGCGCTTCAGAATCGTACTGTCCGATTCATGACGCCGCGGATCCATA	2160
TCAAGCGCTCCGAGAAGACGTGTAAGTCGAGAGACAGTAACGCACCCTAGGTAATCAGGGCCGCCTTAGACTGGCCGCGAAGTCTTAGCATGACAGGCTAAAGTACTGCCAGGACATAAGAA	
EKVIKRVAREAPGVSFELLHVNDDPDERLRSGDLDFLILP	
IGAAAAGGIGA IAAAGCGCGAGCGCAGGGACCI GCCGICACCI TIGAA TIGCICCACCITATGA TCCCGGACGACGICICCCGCAGGGACTI TGATTITTGA TCC TICC	2280
D Q F M S A T H P S A K L F E D K L V C V G C P S N Q Q L R G K L S L K R F M S	
GGATCAGTTCATGTCAGCCACTCATCCCAGCGCGAAGCTGTTTGAGGACAAACTCGTATGCGTCGGCTGCCCTAGCAATCAGCAGCTGCGCGGGAAACTTTCCCTCAAGCGATTCATGTC	2400
CCTAGTCANGTACAGTCGGTGAGTAGGGTCGCGCTTCGACAAACTCCTGTTTGAGCATACGCAGCGACGGGATCGTTAGTCGTCGACGCGCCCTTTGAAAGGGAGTTCGCTAAGTACAG	
	2520
CTACCCCGTGCAGCGCGCGCACAACCTCGATGAACTCGCAGCAACTTGCTACCGACAATGACCTCGTGCCAAAATTCGCCCGACACCTTTAGCACCAGGGCCCAAAGTTGACCTA	2720
▲ ₈₅₁ ▲ _{K9}	
• • • • • • • • • • • • • • • • • • • •	
PMLLQGTNRIATLPLLLVRHFEPTIPLQIVDHPLPPLSFT	
CCCAA IG LACT LOAGGGGGGACGA TO GA TAGCACCC TECC TTO TO	2640
E A L Q W P L L H N S D P G N I W M R N I I L E E A S R I E T S S F R C S Q F P	
AGAAGCACTACAGTGGCCGTTGCTTCACAATTCTGACCCCCGGGAACATCTGGATGCGCAACATAATACTGGAGGAGCGCTTCGCGCATCGAGACATCCTCCGGAAAGATGCTCTCAAGAGCC	2760
TCT TCGTGAT GTCACCGGCAACGAAGTGT TAAGAC TGGGGCCC TT GTA GACC TAC CCGTTGTAT TAT GACC TCC GAACCCCGTACC TCTGTAGAGCCTT TC TACGAGAGTT TCTCGG	
Characterization of the industry of the construction of the industry of the construction of the constructi	2880
GAAAGGTGCCATCCCTCGCATTCCGCGTTTCCACCCATCCGAACTTGGATAAAATAGTGTGTTTTTGGCCGGCAGCCACGCTCGATCAGTTGTCAAAGTCGCGAACCCGA	3000
CTTTCCAC6GTA6G6A6CGTAA66C6CAAA6GT6G6TA66CTT6AACCTATTTTTATCACACAAAAACC66CCGTC6GT6CGA6CCGA6C	
ACCTC ICCCCCICCC IGCATA TCCCCCTTCATA ICCCCCCCACATETACAAGGTACAAGGTACAAGGTACCACGGGGCTCATTCCTTCATCCATACTCCGGATECTTTCGATCCAATC	3120
TO A DISCUSSION OF A DISCUSSI	
ANTEL ACCANTOC TIC GOOD TO CONTRACT OF THE ACCANT AND A CONTRACT AND A	3240
TTAAAA TGGTTAGGAAGCCGTACGAGGTATCGTGTTTTGGTCTGCA TCCGGCCGGGGAGGGCGTCCTCGTAGCTTGCACGAGTTCGGGACGAACGGGCCTACCACGGCCCACAGTCCTGA	5240
M2 MADQLTVEIIAAIKNLAQSENGGRIPAAIGDI	
AND TGGAN TGATGGAAGGAACATGATGGCCGA TCACE TCACAGTAGAAT TCATTCCT CCALA TCACAGTACAATACC TCGCCCAATCGGCGGAGGATTCCTGCAGCAATCGGAGGATTCCTGCGAGGATTCCTGCAGCAATCGGAGGATTCCTGCAGCAATCGGAGGATTCCTGCAGCAATCGGAGGATTCCTGCGAGGATTCCTGCAGCAATCGGAGGATTCCTGCAGCAATCGGAGGATTCCTGCGAGGATTCCTGCAGCAATCGGAGGATTCCTGCGAGGATTCCTGCGAGGATTCCTGCGAGGATTCCTGGAGGATTGGAGGATTCCTGGAGGATTGCTGGAGGATTCCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGCTGGAGGATTGGAGGAGGATTGCTGGAGGATTGGAGGAGGATTGCTGGAGGAGGATTCCTGGAGGAGGATTGGAGGAGGATTGGAGGAGGAGGATTGGGAGGA	3360
THE CENTRE THE THE THE CALCER FROM THE CALCENT ACTION AND A CALCENT ACTION AND A CALCENT ACCOUNT ACCOUNT ACTION AND A CALCENT ACCOUNT	
TADROLTSLGIDSLALADVI W DIFOAYGIRIFMNTADAWS	
CACCOCTGATAGECAATTGACCTCCCTTGECATCGATTCACTGECTTTGECCCATGTCCCCTGEGACCTGEGAGCAGECCTACGGTATCAGEATCGAGATGAACACGECCGATGCTTGETC	3480
GTGGCGAC TA TCCGTTAACTGGAGGGAACCGTAGCTAAGTGACCGAAACCGGCTACACGAGACCCTGGACCCTGGACCTGCGCGATGCCATAGTCCTACCTA	
MDRRVVITGIGGLCGLGTN	
	74.00
ATTAMACTITITATAGCCGTACAGAGCTTCGGCAAGGCCGAACTAGCGCTTCCTCCGCACATACCGCGCACCATTAGCCGCTACAGAGACTAGCGCATAGCGCACAGAGACTAGCGCACAGAGAGAG	0000
A A S I W K E M R E G P S A I S P I I T T D L	
GCCGCATCTATTTGGAAAGAATGGCGAAGGCCCGTCCGCAATCAGCCCGATCATCACGACAGATCT	3668
CONCERNMENTAACCET FICE FIRE DUC GOOGLAGOCGET AGEC COGOCIA GEAGECCETECE TAGA	

Figure 1. Nucleotide sequence of the 3668-bp <u>EcoRI- BgIII</u> nod gene fragment of <u>R</u>. trifolii strain ANU843. The predicted amino acid sequences of the <u>nod</u> gene products are indicated (see Fig. 5 for summary). Asterisks indicate the invariant amino acids in the homologous gene products of <u>R</u>. trifolii, <u>R</u>. leguminosarum, <u>R</u>. meliloti and Parasponia Rhizobium sp. Arrowheads indicate the Tn5 insertion points which were determined by Tn5 primer-directed, chain-termination sequencing. The 75/76-bp conserved sequences, flanking the <u>nodD</u> gene, are indicated by solid lines drawn between the two strands.

23.5 kD; <u>nodD</u>, 36.0 kD; <u>nodF</u>, 9.8 kD and partial sequence of <u>nodC</u> and <u>nodE</u>) were compared to those of <u>R</u>. <u>leguminosarum</u> (16, 20), <u>R</u>. <u>meliloti</u> (17, 4, 21) and <u>Parasponia</u> <u>Rhizobium</u> (22). Invariant amino acids are indicated by asterisks in Figure 1. Comparisons of <u>Rhizobium</u> <u>nod</u> gene products reveal 70-80% amino acid homology between pairs of species. With 50-60% of the amino acids being invariant, between any pair of species, there remains sufficient homology between at least the <u>nodA</u>, <u>B</u>, <u>C</u> and <u>D</u> gene products to allow functional interspecies complementation (23, 3, 24, 25).

For the <u>nodA</u> gene, our data are consistent with that obtained from <u>R</u>. <u>meliloti</u> (17, 21). In both instances, the predicted <u>nodA</u> gene product is 196 amino acids long, whereas the proposed <u>R</u>. <u>leguminosarum nodA</u> gene product differs at both the amino- and carboxy-terminal ends, but these differences could be accounted for by single nucleotide insertions or deletions. In both <u>R</u>. <u>leguminosarum</u> and <u>R</u>. <u>meliloti</u>, several potential initiation codons are present at the beginning of the <u>nodC</u> coding sequences. Comparison with the <u>R</u>. <u>trifolii</u> <u>nodC</u> sequence, which has only one initiation codon, would suggest that the most likely translation start site for <u>nodC</u> in the other two species is the more upstream initiation codon.

Location of Tn5 insertions in nod genes

Fifteen independent Tn5 insertion points were determined (Fig. 2) and some insertional specificity was noted (see Fig. 1), as found in E. coli (26). The locations of these Tn 5 insertions, and their corresponding mutant phenotypes (5), correlate with the predicted nod gene coding regions. Insertions in nodA (246, 252, H9 and M7), nodB (249, M16, C21 and F3) and nodC (C8, L11, C10, L21, A29, L9, 277, K13, O7 and J7) result in a root hair curling deficient (Hac) and Nod phenotype. The four nodB mutants, however, induce some root hair distortions (5), which suggest that, not only are these mutants located in a second nodulation gene, but that nodC may have an independent promoter. The proximity of nodB and C (22-bp intergenic space) and the lack of any promoter-like sequences, do not support this conclusion. Mutants of nodD (245, 274, A22, 276, C12, 851 and K9) are all Nod and Hac. Two Tn5 insertions (246 and K7), located in the 232-bp intergenic region between nodD and nodA, are Nod⁺. Both of these Nod⁺ mutants, however, are located very close to the point of translational initiation (246 is 51 bp upstream of nodA; K7 is 18 bp upstream of nodD). Whilst the Tn 5 insertions have disrupted either leader or promoter sequences, the observed nodulation may have resulted from weak expression



Figure 2. Tn 5 primer-directed sequencing strategy. DNA fragments, extending from the HindIII (H) sites within Tn 5 to EcoRI(E) (BamHI or BgIII) sites in flanking R. trifolii DNA, were cloned into M13mp18. The synthetic 17-mer Tn 5 primer (see Materials and Methods) was then used to direct chain-termination sequencing reactions using single-stranded recombinant M13 DNA as templates.

of the <u>nod</u> genes from a promoter within the arms of Tn5. These mutants show slightly reduced and delayed nodulation (5) further suggesting that the Tn 5 insertions may have affected the transcription of these genes. In <u>E</u>. <u>coli</u>, low-level constitutive expression of the <u>lac</u> (27) and <u>dnaG</u> (28) genes has been observed downstream of Tn5 insertions. Weak transcription from a promoter within the arms of Tn5, is only detectable when the Tn5 insertion is located close to a ribosome binding site, thus preventing <u>rho</u>-dependent termination and allowing the expression of downstream genes (29).

Host range mutants (M2, M23, C7, 297, H7, K11 and 258) nodulate subterranean clover, peas and beans and (with the exception of M2) are unable to nodulate white clover (5). Mutants M2 (see below) and M23 appear to map within the <u>nodF</u> gene whilst the others (C7, 297, H7, K11 and 258) are thought to be located within the <u>nodE</u> gene. The amino-terminal portions of the <u>nodE</u> gene of <u>R</u>. <u>trifolii</u> and <u>R</u>. <u>leguminosarum</u> are homologous (20) and the <u>nodE</u> gene of <u>R</u>. leguminosarum encodes a 48 kD protein (18).

Mutant M2 has a unique phenotype, being able to nodulate white and subterranean clover, as well as peas and beans (5). In this mutant, Tn5 is located between the proposed ribosome binding site (30) and the initiation codon of the nodF gene (Fig. 3). This insertion results in a duplication of



Nature of the $Tn \underline{5}$ -induced mutation in host-range mutant M2. Figure 3. (a) DNA sequence analysis (see Fig. 1) predicts an open reading frame (nodF) preceded by a ribosome-binding site (boxed) in the wild-type sequence. (b) Insertion of Tn5 results in the direct duplication of 9bp of host DNA (underscored arrows) (15). (c) The presence of a potential ribosome-binding site (boxed) in the arm of Tn5 may allow translation of the putative Tn5-promoted nodF transcript. The unique phenotype of mutant M2 (see text) may be a consequence of differential levels of transcription or intervals between the differential translation due to differing ribosome-binding site and initiation codon in the wild-type (5bp) and mutant M2 (8bp) sequences.

9 bp of host DNA, which regenerates the <u>nodF</u> gene initiation codon. A potential ribosome binding site (AGAG) is located in the terminal nucleotides of Tn5 (Fig. 3), and therefore weak transcription from within the arms of Tn5 (see above) could allow translation of the <u>nodF</u> gene product. The ability of mutant M2 to nodulate white clover, unlike the other host range mutants, may be a consequence of the unique location of Tn5 in this derivative. For example, the extended host range of this mutant may result from plant-independent, Tn5 - promoted transcription of the <u>nodFE</u> genes.

<u>R</u>. <u>trifolii</u> host range genes have previously been identified by their ability to confer altered host range on other <u>Rhizobium</u> species (5,7). However, two putative host-range genes, <u>nodF</u> and <u>nodE</u> appear to be conserved between <u>R</u>. <u>trifolii</u> and <u>R</u>. <u>leguminosarum</u> (20), suggesting that

host range may be determined, not by widely-differing gene products, but by differences in the active sites of the <u>nod</u> gene products. The <u>nodFE</u> gene products may target or modify a bacterial signal for the appropriate host plant nodulation response. An incorrectly processed signal may elicit nodule initiation on legumes other than the normal host. The unique host range of mutant M2 may result, not from an incorrectly-processed signal, but from abnormal (i.e. Tn<u>5</u>-promoted) transcription of the <u>nodFE</u> genes. <u>Operon structure of nod genes</u>

Previously, the <u>nodA</u> and <u>nodB</u> genes have been shown to overlap in <u>R</u>. <u>meliloti</u> by the sequence \overline{ATGA} (17). We observe an identical overlap in the <u>nodA</u> and <u>nodB</u> genes of <u>R</u>. <u>trifolii</u>, at position 1001-1004. Additionally, we observe an overlap between the <u>nodF</u> translation stop and the <u>nodE</u> translation initiation codons <u>TGATG</u> at position 3542-3546. The presence of these overlapping sequences may indicate that translational coupling (31) occurs.

The presence of overlapping nodA-nodB and nodF-nodE genes, as well as the proximity of nodB and nodC, suggest that these <u>R</u>. trifolii nod genes are present in two separate operons, namely: <u>nodABC</u> and <u>nodFE</u>. The <u>nodD</u> gene coding sequence is read divergently from <u>nodABC</u> indicating that the <u>nodD</u> gene constitutes a separate operon. The presence of a terminator structure (32) (two inverted repeats located between positions 2815 and 2833, forming an 8 out of 9-bp stem and a 2-bp loop, followed by a T-rich region) at the 3' end of the <u>nodD</u> gene suggests that the <u>nodFE</u> operon is transcribed separately from the <u>nodD</u> gene. A second terminator - like structure is located within the coding region of <u>nodD</u> forming an 8-bp stem between positions 2244 and 2260 followed by a T-rich sequence (8 of 11 bases). The significance of this sequence is not clear, but it may be that <u>nodD</u> expression is controlled by this sequence via a mechanism such as <u>rho</u>-dependent termination (33).

Examination of the sequences located 5' of each of the <u>nod</u> gene operons does not reveal homology with either the <u>E</u>. <u>coli</u> (32) or <u>nif</u> (34) consensus promoter sequences.

The <u>R</u>. <u>trifolii nod</u> gene region contains two reiterated sequences, the first located between nucleotide 1756 - 1830 and the second between nucleotides 3076 - 3151. These two sequences are inverted with respect to each other and are 78% homologous. The sequences are located 235 bp upstream of the <u>nodABC</u> genes and 114 bp upstream of the <u>nodFE</u> genes. In <u>R</u>. <u>meliloti</u> there is an analogous sequence, located 192 bp upstream of the

Consensus

Rhizobium trifolii	nodABC	COCATTC-TCGATCCACOCTGTAGATGATTOCGATCCAAACAATCAATTTTACCAATCTTTCGGAGTGCTTATTAG
Rhizobium trifolii	nodFE	CTCATTCCTTCATCCATACTCCGGATCCTTTCGATCCAATCAAT
Rhizobium meliloti	nodABC	9CATGT9C99CATCCATATC9CAGATGATCGTTATCCAAACAATCAATTTTACCAATCTT9CAGACT-CCTATTAG

-----T-----A TCCA-----G--GATG-T----A TCCAA-CAA TCAA TTTTACCAA TCTT-C-G--T-C----TAG

Figure 4. Comparison of the conserved 75/76-bp nucleotide sequences preceding the <u>nodABC</u> and <u>nodFE</u> genes in <u>R</u>. <u>trifolii</u> and <u>R</u>. <u>meliloti</u>. The consensus sequence indicates the nucleotides conserved in the three copies, including a highly-conserved (96%) 27-bp core sequence (underlined).

<u>nodABC</u> genes (21). A comparison of the <u>R</u>. <u>meliloti</u> sequence with those of <u>R</u>. <u>trifolii</u> reveals a total of 59% homology between all copies (Fig. 4). In a core sequence of 27 bp, the homology is 96%. Additionally, this conserved sequence is found preceding the <u>nodKABC</u> genes of a slow-growing <u>Rhizobium</u> sp. isolated from the non-legume <u>Parasponia</u> (22).

To determine if additional copies of this sequence occur in the 14 kb <u>nod</u> region, the 557-bp <u>PstI</u> fragment (positions 2785-3342) was hybridized to a Southern blot of various restriction digests of pRt587 DNA. Two additional fragments hybridized with this probe. The first carries the reiterated sequence preceding the <u>nodABC</u> genes whilst the second carries a region that includes other <u>nod</u> genes (6) (Region IV in Figure 5).

Regulation of nod gene expression

To examine the possibility of this reiterated sequence being involved in the regulation of nod gene expression, the locations of the operons defined by DNA sequence analysis (Fig. 5) were correlated with transcriptional and regulatory information derived from R. trifolii (6). Using Mud lac fusions, Innes et al. (6) identified four R. trifolii transcriptional units. One of these, corresponding to the nodD gene, is constitutively expressed. The other three are activated in the presence of clover root exudate. These plant-activatable transcriptional fusions correlate with the locations of the reiterated nod-gene consensus sequences on the 5' sides of the plant-inducible nodABC and nodFE genes, as well as in the vicinity of the Region IV gene(s).

In <u>R</u>. meliloti and <u>R</u>. leguminosarum, the nodD gene has been shown to be a positively-acting autoregulatory gene whose product, in the presence of plant root exudate, activates expression of the <u>nodABC</u> and <u>nodFE</u> operons (35,36). The promoters of many positively-regulated genes are characterized by the presence of conserved nucleotide sequences located upstream of the normal consensus promoter sequence (37). These observations suggest that the <u>nod</u> gene-specific consensus sequences may be involved in the



Figure 5. The 14kb <u>HindIII nod</u> gene region of <u>R</u>. <u>trifolii</u> strain ANU843 cloned in pRt587(9). Regions I-V carry genes involved in clover host-specific nodulation (5,7). The lower part of the figure indicates the locations of Tn <u>5</u> insertions (arrows) and the extent of the predicted <u>nodA</u>, <u>B</u>, <u>D</u> and <u>F</u> coding regions (solid arrows) as well as the expected lengths of the <u>nodC</u> and <u>E</u> genes (broken arrows). Open boxes indicate the locations of the 75/76-bp conserved sequences. Restriction sites are indicated as : $B(\underline{BamHI})$; $Bg(\underline{BgII})$; $C(\underline{ClaI})$; $E(\underline{EcoRI})$; $H(\underline{HindIII})$; $P(\underline{PstI})$; $S(\underline{SphI})$ and $X(\underline{XmaI})$.

plant-activated, coordinate expression of <u>Rhizobium</u> nod genes.

The identification of a <u>nod</u> gene-specific, reiterated sequence preceding the putative <u>nodABC</u> and <u>nodFE</u> operons, and its correlation with plant root exudate-activated transcriptional units, suggests that <u>Rhizobium</u> nodulation genes are under complex regulatory controls that involve not only a bacterially-encoded activator, but also a signal molecule from the plant partner in the symbiosis. The data presented in this paper should facilitate definition of the promoters of each of the <u>nod</u> gene transcriptional units and, by comparison of various <u>Rhizobium</u>-legume symbioses, enable elucidation of the roles played by individual bacterial <u>nod</u> genes in the complex process of host-specific nodulation.

ACKNOWLEDGEMENTS

We thank Christine Hayes for assistance with DNA sequencing, Judy Tellam for oligonucleotide synthesis, Sharon Long and colleagues , Kieran Scott, Allan Downie, Lone Rossen and Claire Shearman for sharing <u>R</u>. <u>meliloti</u>, <u>Parasponia</u> <u>Rhizobium</u> and <u>R</u>. <u>leguminosarum</u> <u>nod</u> gene sequences prior to publication, Michael Djordjevic and Barry Rolfe for some Tn5 mutant DNAs, Lina Onorato for preparation of the manuscript and John Shine for his continued support of this work. We also thank Siiri lismaa and Allan Downie for their constructive comments during the preparation of this manuscript. This work has been supported in part by an Agrigenetics Sponsored Research Program and by the Wool Research Trust Fund on the recommendation of the Australian Wool Corporation.

*To whom correspondence should be addressed

¹Present address: Department of Molecular Biology, Genentech, Inc., 460 Point San Bruno Boulevard, South San Francisco, CA 94080, USA

REFERENCES

- Schofield, P.R., Ridge, R.W., Rolfe, B.G., Shine, J. and Watson, J. (1984) Plant Mol. Biol. 3, 3-11.
- Downie, J.A., Hombrecher, G., Ma, Q-S., Knight, C.D., Wells, B. and Johnson, A.W.B. (1983) Mol. Gen. Genet. 190, 359-365.
- Kondorosi, E., Banfalvi, Z. and Kondorosi, A. (1984) Mol. Gen. Genet. 193, 445-452.
- 4. Jacobs, T.W., Egelhoff, T.T. and Long, S.R. (1985) J. Bacteriol. 162, 469-476.
- 5. Djordjevic, M.A., Schofield, P.R. and Rolfe, B.G. (1985) Mol. Gen. Genet. 200, 463-471.
- Innes, R.W., Kuempel, P.L., Plazinski, J., Canter-Cremers, H., Rolfe, B.G. and Djordjevic, M.A. (submitted for publication).
- 7. Schofield, P.R., Ph.D. Thesis, (1984) Australian National University, Canberra.
- 8. Bossi, L. and Ciampi, M.S. (1981) Mol. Gen. Genet. 183, 406-408.
- Schofield, P.R., Djordjevic, M.A., Rolfe, B.G., Shine, J. and Watson, J. (1983) Mol. Gen. Genet. 192, 459-465.
- 10. Yanish-Peron, C., Vieira, J., Messing, J. (1985) Gene 33, 103-119.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- 12. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5476.
- 13. Staden, R. (1984) Nucleic Acids Res. 12, 551-567.
- 14. Beaucage, S.L. and Caruthers, M.H. (1981) Tetrahedron Lett. 22, 1859-1862.
- 15. Auerswald, E.A., Ludwig, G. and Schaller, H. (1981) Cold Spring Harbor Symp. Quant. Biol. 45, 107-113.
- 16. Rossen, L., Johnston, A.W.B. and Downie, J.A. (1984) Nucleic Acids Res. 12, 9497-9508.
- 17. Torok, I., Kondorosi, E., Stepkowski, T., Posfai, J. and Kondorosi, A. (1984) Nucleic Acids Res. 12, 9509-9524.
- Downie, J.A., Knight, C.D., Johnston, A.W.B. and Rossen, L. (1985) Mol. Gen. Genet. 198, 255-262.
- 19. Fickett, J.W. (1982) Nucleic Acids Res. 10, 5303-5318.
- 20. Downie, J.A. and Shearman, C. pers. comm.

- 21. Egelhoff, T.T., Fisher, R.F., Jacobs, T.W., Mulligan, J.T. and Long, S.R. (1985) DNA 4, 241-248.
- 22. Scott, K.F., this volume.
- 23. Banfalvi, Z., Sakanyan, V., Koncz, C., Kiss, A., Dusha, I. and Kondorosi, A. (1981) Mol. Gen. Genet. 184, 318-325.
- Djordjevic, M.A., Schofield, P.R., Ridge, R.W., Morrison, N.A., Bassam, B.J., Plazinski, J., Watson, J.M. and Rolfe, B.G. (1985) 24. Plant Mol. Biol. 4, 147-160.
- Fisher, R.F., Tu, J.K. and Long, S.R. (1985) Appl. Env. Microbiol. 25. 49, 1432-1435.
- 26. Berg, D.E., Schmandt, M.A. and Lowe, J.B. (1983) Genetics 105, 813-828.
- 27. Berg, D.E., Weiss, A. and Crossland, L. (1980) J. Bacterial. 142, 439-446.
- 28. Lupski, J.R., Smiley, B.L. Blattner, F.R. and Godson, G.N. (1982) Mol. Gen. Genet. 185, 120-128.
- Schmid, M.B. and Roth, J.R. (1983) J. Bacteriol. 153, 1114-1119. 29.
- 30.
- Shine, J. and Dalgarno, L. (1975) Nature 254, 34-38. Normark, S., Bergstrom, T., Jaurin, B., Lindberg, F.P. and Olsson, O. (1983) Ann. Rev. Genet. 17, 499-525. 31.
- 32. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
- 33. Holmes, W.M., Platt, T. and Rosenberg, M. (1983) Cell 32, 1029-1032.
- 34. Ausubel, F.M. (1984) Cell 37, 5-6.
- 35. Mulligan, J.T. and Long, S.R. Proc. Natl. Acad. Sci. USA 82, 6609-6613.
- 36. Rossen, L., Shearman, C.A., Johnston, A.W.B. and Downie, J.A. EMBO J. (in press).
- 37. Raibaud, O. and Schwartz, M. (1984) Ann. Rev. Genet. 18, 173-206.