In the format provided by the authors and unedited.

## LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in *Bacillus amyloliquefaciens*

Jonathan R. Goodson<sup>1</sup>, Steven Klupt<sup>1</sup>, Chengxi Zhang<sup>2</sup>, Paul Straight<sup>2,3</sup>, Wade C. Winkler<sup>1,3</sup>

<sup>1</sup>The University of Maryland, Department of Cell Biology and Molecular Genetics, College Park, MD;

<sup>2</sup>Texas A&M University, Department of Biochemistry and Biophysics, College Station, TX

<sup>3</sup>Corresponding authors



**Supplementary Figure 1.** Coverage of the *dfnA* leader region in RNA-seq data. RNA-seq coverage across the *dfnA* leader region normalized with DESeq2 normalization factors. Traces represent coverage data smoothed with Gaussian smoothing with a bandwidth of 5 nucleotides. Shading represents standard deviation from libraries from three independent cultures for each condition.



**Supplementary Figure 2.** Conserved RNA elements are present in the *dfn* and *mln* leader regions. (A) A highly similar RNA hairpin with a UUCG tetraloop is present in the upstream portion of both leader regions. (B) Predicted intrinsic terminator sequences are present near the center of both leader regions. (C) Both leader regions contain a similar UNCG tetraloop and intrinsic terminator in long 5' leader regions that do not contain any significant open reading frame. (D) Quantified fluorescence microscopy data of *B. subtilis* strains carrying a xylose-inducible *loaP* and a *dfnA* leader-*yfp* reporter cassette. The

yfp reporter contained a constitutive promoter upstream of the dfnA leader region, which was transcriptionally fused to a downstream yfp gene. Reporter constructs were created with or without a mutation converted the UUCG tetraloop sequence to UUCA, a mutation that is predicted to abolish proper RNA hairpin formation. These data demonstrate that *B. amyloliquefaciens* LoaP antitermination can be recapitulated in the heterologous *B. subtilis* host. Data and error bars represent the mean fluorescence and 95% confidence interval (CI over means of each replicate) for all cells in three fields of view for each of three biological replicate cultures of each strain with and without induction. (E) Representative images of induced reporter strains quantified in (D) showing lack of reporter expression when the tetraloop sequence is mutated.



**Supplementary Figure 3.** RNA transcript levels for genes in the *mln* gene cluster are increased upon LoaP induction. Normalized transcript abundance at the beginning, middle, and end of the *mln* operon (*mlnA*, *mlnD*, and *mlnI*) as measured by qRT-PCR. Filled points represent samples with *loaP* expression and empty points represent samples with no or minimal *loaP* expression. Error bars represent Bayesian 95% highest posterior density estimates of mean expression. Data resulted from four biological replicate cultures.



**Supplementary Figure 4.** Large-scale phylogenetic analysis recapitulates targeted analysis of representative sequences. (A) Phylogenetic tree constructed from multiple sequence alignment of 37 representative protein sequences. (B) Subtree consisting of the same 37 protein sequences extracted from large-scale 1,207 sequence phylogenetic analysis. Both trees show very similar topology.

				FeNusG	1	5	25			
				EcRfaH	i	ō	20			
				BalloaP	9		19			
					Ī		Ĩ			
E. coli NusG	NSEA			PKKR	WYVVOAF	SGFEGR	ATSLR		HIKL-	-HNMEDLEG-
Bsubtilis_NusG	NE			KN	WYVVHTY	SGYENK	KANLE	(	RVES-	-MGMODKIF-
Mtuberculosis_NusG	NTTFDGDTSAGEA	VOLTEANAFQDA	AAPAEEVDPAAAL	KAELRSKPGD	WYVVHS	AGYENK	KANLE	r	RVON-	-LDVGDYIF-
Synechocystis_spNusG	NSFTDDQSPVAEQ			NKKTPSEGH	WFAVOVA	SGCEKR	KLNLE	)	RIHT-	-LDVADRIL-
Tthermophilus_NusG	MS			IE	WYAVHTL	VGQEEK/	KANLE	(	RIKA-	-FGLQDKIF-
ECOl1_RfaH	M			qs	WYLLYCK	REQLER	QEHLEI	·	QAV	
Tpestis_kram	M			Ka	WILL YOU	REUILK/	AKEHLEI	<u></u>	017	
Stypn1_kran				00	MYLL YCK	DCOL OD			OAV.	
S. marcescens Rfall	<u>.</u>			ES	WYLL YC	RGOLLR	OFHLE		00V	
V. cholera Rfall				KR	WYLL YC	RGEOOR	KMHLE	·	osv	
									401	
T. denticola LoaP				ND	YYYYOVS	TGKEKNF	TEDAE	KNKFDE	LSY	
B. anvioliguefaciens LoaP				NK	WYALFVE	SGKEET	OKFLR	OFDE	QAL	
P. polymyxa_LoaP				VS	WYVEEVE	TGREEOV	KOLIN	WLDS	EVY	
Bbrevis_LoaP				LK	WYAIFVE	SGKEEV	OKYLRI	YFNE	OSL	
Ccellulolyticum_LoaP				NY	WYYLFV	TGREEN	KKLLS	(RLDK	DLF	
Twiegelii_LoaP	M			KK	WYVIFT	SGYENK	KDIIE	ICFKQ	EEV	
Csubterraneus_LoaP	M			KK	WYYLFT	SGCEEK	GKIIK	(IWE	NEI	
Ametalliredigens_LoaP					MHAR	SNEEMK/	KKLVE	(EI	EDI	
Bfragilis_UpxY		704		TOCCULDOKS	WLAAYY	LYHEKKI	RORLT		MGI	
Bxylanisolvenens_upxt	MIIKKNUURULLS	104		-1055VAKSKK	WLYALYN	LUMERK	SERLI		MUL	
Pdistasonis_upxf	M			ANN ANN	MTALKY	THREAT	INAILOI	¢	061	
W vanthus Tal	NPCP			DC AEND		VNUEVV	AAGE CI	·	HCV-	
D tigdigi Tak	NTROVI I SRVPEE			PPI DEDLOS	WWWIIT	PNCEKKT	ASVEL		DNT-	
C tenidum Tak	NTNAL			KKDGC	WYAVYV	SRVEKK	HOVILL		KGI	
ci_coprodin_ren									NOL	
Citreicella sp. NgsP	NSVLGSNGKWTSG	EAVRPYLG	LRVGDAVPVEGN	VAIFDRGEIA	WYALLCR	POOERHA	ESWLA		RGV	
H. denitrificans NgsP	M			T	WYAIRTH	POREFLL	AGRYD		NGEWR	PGVLEKKGYD
Roseibium_spNgsP	<b>NSKMDVRVQNRAL</b>	ANDYDF		RALVRVSAGG	WLVACCH	PTKEQH/	LRQLT		RGL	
Rhizobium_spNgsP				MGH	WYVVRTR	AGQQQK/	TREFE	)	NGV	
Pfuriosus_Spt5	MA			GK	IFAVRVI	HGQEETT	AKLIY	S	KVR	
Mjannaschii_Spt5					IFAVRT	VGOEKNI	AGLNA	S	RAE	
Sacidocaldar1_Spt5	MEDF			KYRN	YYYLRV	GGOEIN	ALILE		RIK	
A,_ambivalens_Spt5	MES			KIRN	TYAVKVI	GGQEVS	GLMLE		RAK	
cons										

EcNusG Sec. Structure EcRfaH Sec. Structure

-----

44	59	82	103	117
30	45	68	88	102
33	48	71	99 	113
EVMVPTEEVVEI-RGGQRR-	-KSERKFFPGYVLVQMVMNDA-	SWHLVRSVPRVMGFIG	GTSDRF APISDKEVDAIMN-	RLQQV
RVVVPEEEETDI-KNGKKK-	-VVKKKVFPGYVLVEIVMTDD-	SWYVVRNTPGVTGFVG	SAGSGSKPTPLLPGEAETILK-	RMGMD
OVEVPTEEVTEI-KNGORK-	-OVNRKVLPGYILVRMDLTDD-	SWAAVRNTPGVTGFVG	ATSRPSALALDDVVKFLL-	PRGSTRK
QVEIPKTPIVKI-RK-DGARV-	-OGEEKIFPGYVLIRMINDDD-	AWOVVKNTPHVINFVG	SEQKRHYGRGRGHVLFMPLSHGEVERIFR-	HVDEQ
QVLIPTEEVVEL-RE-GGKKE-	-VVRKKLFPGYLFIQMDLGDEE	EPNEAWEVVRGTPGITGFVG	AGNRF VPLSPDEVRHILE-	VSGLL
NCLAPMITLEKI-VRGKRT-	-AVSEPLFPNYLFVEFDPEVIH-	TTTINATRGVSHFVR	FGASPAIVPSAVIHQLSV-	YKPK
NCWTPIVAIEKI-VRGKRI-	-EVIEALFPNYLFAEFDPENIH-	TTTVSATRGGSHFVR	FGTOFAVIPATVIADMOA-	HAVD
SCLTPMITLEKM-VRGKRT-	-FVSEPLFPNYLFVEFDPEVIH-	TTTINATRGVSHFVR	FGAHFAIVPSSVIHOLSI-	YKPE
NCLAPMITLEKI-VRGKRT-	-AVSEPLFPNYLFVEFDPEVIH-	TTTISATRGVSHFVR	FGASFAIVPSAVIHOLSV-	YKPK
NULSPIITLEKI-VRGKRI-	-AVSEPLEPNTLEVEFUPERIN	IIIISAIRGVSHFVR	PUP DEL OCDI TVOL KO	HASE
ECFTPEVCVERI-LKGRRQ-	-MYVEPLEPSTMFYRFUFENGE	SFILVKSIKGVYDFVK	LGPHPRELQGULITQLKQ-	LDCEQ
				-
SIVFPORILKIR-KAGKVI-	-EKQLPVFAGYLFIGTDEISKD	YQHLRKCKGFYRFLP	NNQEPKFLEGROFEILNQ-	FISFG
TSIIPKKKVIEK-KAGIKT-	-EALKKMPPGYVLFKIKMIERI	HKIKELPISCRIVN	NGATTSKERKITFIIIIDEEILPIIK-	LIGEG
KPFIPLUERLFK-YAGIYK-	-KEWAPLFPSTVFIESNLPULU	VISINSMICISSUIIRLLK	TOKFEASMKUSEKUMLES-	LUNUS
ASTALKER PER-KSULAT-	-NYLKNFFPGTYLLUTENINEM	WENTHEL HESLODITEL VE	NGSTTSQUEGATTSSIEEKEI IPILO-	LAGNG
KII TEKEKIFK-K-AGIYK-	-EKTKI LEPCYVEVNAENSDDI	V KATELAKSLUUTIKLYK-	FC	TKNS
EVI TERRETTER-TKGEER-	-EKTKLLEPGYVEVKTENTEAK	HFTTSVI KOGVELK	EDKNPASVKEEENPVTLN_	TODS
KVTVPORTTPEK-ROGETR-	-HYKKTLETGYLELNVELDVDT	VKI KRTPSTHRELG	FKPEATPLEEMORVLR-	I CN0G
ESEL PVOEETHO_WSDRRK_	-KTERVVTPMMTEVHVDPAERA	EVITI SSVSRV-NVI R	COSTRAVTRDEOMERERE-	NI DYSE-
ENEL PTOOF VHO-WSORRK-	-VVDRVLL PMMTEVHVDPOEOK	VI TI SATSPY-WVI P	GESTPAVVPDOCML REKE-	MI DYSD-
ESYTPAKTRSYP-OP-DGEVV-	-TRRVPL VANL NEL RCDDDVTS	SI NSTI DEKANVYHRP	GTOTPASTPDEEMDMETM-	I TSSNED
				2100-20
FFEL PTYTPPKS-S	-KAKI PLEPGYLECRYOPLNP-	-VPTVPAPCVTPLLC	CDACPEAVPAOELEATRP-	VADSC
SAAT DI ARBETE-ACCI RETE-	-EVEVPLESCYTCEAL DREKH-	-VIL VDTOKEVRTTK	VE	ATASG
SSEL PL TETL ROWSDRKK-	-RVEEPL TRGYVEVNTNYHKEH	-VHVL ETDGVVKETG	TGKTPSVTSERLTOWLKR-	I AHEP
VAEHPVTSPRTR-VR-CKI P-	-EVERRYL POVVEAPEDGTPTP	HR VI TSPEL TOAL C	SDCONGVI GPKPI AAI HEMI	
OVECPTETKERKTTK-KCRRVS	TPVI VPNECOVTEVCOPESM	EL MAENYTTAAVGEPD-	EE	MSGGI T-
LAYLPMRPGRRR-OPRCKKMI-	-DNSQPLIRGYLFYCTDFSTGS	VDEILSCGCVSGLLSFR	ADKYPHRVPSARVVDIID-	HCEOLO-
TVYCPMLRRETR-HFOSKKWL-	-MKECPLFTGYVFAYLRISD-F-	-GTLREMRHVLSVLA-D	AGGTPIPVAGNIVEDIRD-	AQERGD-
TYNLPIYA-	-ILAPSRVKGYIFVEAPNKGVV-	DEAIRGIRHARGVLP	GEVPFKEIEHFL	
KEOLDVYS-	-ILASESLKGYVLVEAETKGDV-	EELIKGMPRVRGIVP	GTIAIEEIEPLL	
TNNINEIFS-	-VVVPPNIKGYVILEATGPHVV-	KLISSGIRHVKGVAH	GLIQKEDVTKFV	
TNNIPEIYS-	-IIVPPGLKGYVIVEASGPHVV-	KLLIAGIRHVRGIAQ	GLVPKDHIVKWV	
	. :		: .	

	132	150	172	
	114	132	152	
	127	145	167	
GOKPRPKTLFEP	GEMVRVNDGPFADENGVV	EEVDYEKSRLKVSVSIFGRATPV	-ELDES-OVEKA	
ERKTDIOFEL	KETVKVIDGPFANFTGSI	LEEIDYDKSKVKVFVNMFGRETPV	-ELEFT-OIDKL	
AAKGAASTAAAAEAGGLERPVVEVDYEV	GESVTVMDGPFATLPATI	<b>ISEVNAEQOKLKVLVSIFGRETPV</b>	-ELTFG-OVSKI	
EPVVKIDNEI	GDHINVLSGPFKDFEGDV	<b>IEVSPERSKLKALLSIFGRETPV</b>	-ELEFT-OVEKON	
GKKEAPKAQVAFRE	GDQVRVVSGPFADFTGTV	TEINPERGKVKVMVTIFGRETPV	-ELDFS-QVVKA	
DIVDPATPYP	GDKVIITEGAFEGFQAIF	TEPDGEA-RSWLLLNLINKEIKH	-SVKNT-EFRKL	
KIIAPEVPKP	GDIVKIIDGVFAGLOAIY	TEPDGEA-RSMLLLNMLNSQIKH	-SLDNR-OFEKO	
GVVDPETPYP	GDSVIITEGAFEGLKAIF	TEPDGET-RSWLLLNLLNKEVKO	-SVKNT-GFRKI	
DIVDPAPPYP	GDKVIITEGAFEGFOAIF	TEPDGEA-RSWLLLNLINKEIKH	-SVKNT-EFRKL	
TYVDPETPQP	GDTVLIVDGVFEGLQAIY	TEPDGEA-RSWLLLNLINKQVSQ	-SIDNR-QFQKM	
LKHATKQLPEK	GQTVRVARGQFAGIEAIY	LEPEGET-RSIMLVKMISQQVPM	-SIENT-DWEVT	
GLAKISQVVFDE	NDRIKVIEGPLSGLEGYI	EVRYNKRKGRATVCLDMCQTAFSI	-DLGFE-ILNKEEKR	
DTVDYS-KVYIE	NSKVTVASGPLKGMEGIJ	<b>EKKIEKRKRRAKICLSFMGLDKMV</b>	-NVGIE-VLSKP	
YCIESS-TGIIE	GDNICILDGPLKGRGSIV	KKICRHKROAVIGLEFNGDIRLV	-RVALE-IISKV	
EVVDYS-KVYLE	NSKVIVESGPLKGLEGII	EKKVEKHKNRAKIFLTFNGVEKTI	-DVGIE-ILSKPL	
NCIESS-SGIIK	GDKIYIIDGPLRGRESIV	<b>WRHINRHKREAKIEIEFMGNIRLV</b>	-SVSLE-IVOKL	
DI TOUS ROUME	GERVKILEGPLKGTEGLI	IKKILKRKKKKKKYMLSIAGELKSY	-DLAIE-VIENVSEUUKSLVTAL-	
ELTOLOKWINKE	GURYKIIEGPLLGTEGPI		DI CTE TTENEE	
ELIGESKYLYKE	ONKAGAA20LEFONEOUT	LINYUNKKGRANYCLOYLGUAKIY	-DEGIE-IIKVSE	
			PHONE FUEDUCINE LUD	
ETTENCION AND	GEOVRVIKGPLAGLEGEL	VILLER-SKVAVRLDMLGCAHV-	-UMPVG-FVERVGRMEAVK	
	GERIRVIRGPLAGLEGEL	PRTKEN KRIVTREECH TIVAT	WITPEC VIOETV	
GFEQIPTUONLYKN	AEKTRYIGGETAGYYGTI	TRRING ARLYIPPEGLLILYAT	TIPKC-TLUELY	
NO CHECON DUIL				
VS-SNPCDYLRV	GORVRITEGPLIGLEGSL	VISKS-ULKFIVSVGLLUKSVSV	-EVSAE-ULEPITU	
DATOTTVICTOR	GKKVRILSGPLEGIEGVI	KRUKTER-UFALSVUMFNUSVIV		
DALGETVISIPV	GKKYRYLAGPFKDMEGYY	KKEGREE-KLLVTFUSLMQGVEL	- TISPE-LLAPIEKGASGQAVEGS	K IGUHEVESALKHLAHS
UKRKAAAARKYRA	GERAMFRAGPFEGMSCEV	VELUAAG-GVKVRFDLFGRENSV	-TITSE-ALVGLAVGSST	
TCCKPOKATCEDT				
	COTVOVEDORI TOUVAAN	THTUCOD ATVAEVETECCI DEV	ETCLE CTRRVA	
CUTERFF TRALKA	ODITERDOPLIONTAR	THITCH ALKAPTELPOSERET	-EIGLE-SIMNYA	
EEVDAVCCI ED		WETDECKDEVNVOETDATNOTOV	TTYCD VUDI TCHI OVER	
TEXATOLE	GOLVETTACPERCEDARY	TRUCKUKEEVTI ELEMAANDTOT	TI DVE CVETVERUED	
	CDI VEVISCRERCHOACY	VOVECTIVEVUL NTLESCVDVOV		
SKEVTOPTTE	CDI VEVVSCPERCHOAD	VENTECKCEVVI NTI ESAEPI OV	TTPVD_OVKPVKKT	
ORATIOF LINE	ODETETTOOPPROMUNET	TRATEGROET TENELESAFFLQT	TALIN SIN INT	

**Supplementary Figure 5.** Multiple sequence alignment of NusG family protein sequences reveals conserved differences between subtypes of specialized paralogs. Alignment contains 33 protein sequences containing representatives of each subtype of NusG paralog (excluding ActX) aligned by T-COFFEE using the "accurate" alignment method combining PSI-COFFEE and EXPRESSO alignment methods utilizing available crystal structure data. Residues are colored according to T-COFFEE consistency score representing pairwise reliability in alignment with all other sequences. Also shown are secondary structure diagrams representing the secondary structure from PDB structures of E. coli NusG

and RfaH (Orange represents beta-strands and green represents alpha-helices, black represents relatively unstructured inter-domain linker).



**Supplementary Figure 6.** LoaP represents a distinct group of NusG specialized paralogs and is commonly associated with large biosynthetic gene clusters. Subset of large-scale phylogenetic analysis showing LoaP homolog sequences. Background shading represents association of gene sequences with large gene clusters. Red shading denotes LoaP near PKS (polyketide synthase) or NRPS (nonribosomal peptide synthase) gene clusters. Purple shading denotes LoaP near polysaccharide gene clusters. Olive shading denotes LoaP near other types of antiSMASH gene clusters. Unlabeled sequences were not found nearby an antiSMASH predicted gene cluster, although some appear to be next to long stretches of coding sequences in one direction.



**Supplementary Figure 7.** Transcription of the *dfnA* leader sequence in vitro. The *B. amyloliquefaciens dfnA* leader sequence, including the endogenous promoter, was PCR amplified using a reverse oligonucleotide primer that would be predicted to result in either a 314-nucleotide transcript (for premature transcription termination at the putative intrinsic termination site) or a 420-nucleotide transcript (for run-off transcription). Reactions included 2 pmol PCR-generated DNA template, 20 mM Tris-HCl pH 8.0, 15 mM NaCl, 4 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 5 mM DTT, 0.01% Triton X100, 1 mM NTPs, 1.5 pmol (5  $\mu$ Ci)  $\alpha$ -<sup>32</sup>P-UTP, and 160 nM  $\sigma$ <sup>A</sup>-saturated *B. subtilis* RNAP. LoaP was added to final concentrations of 300 nM, 150 nM, 75 nM, and 3.7 nM. Purified hexahistidine-tagged *B. subtilis* S10 was added to 150 nM and purified hexahistidine-tagged *B. subtilis* NusA was added to 50 nM, as indicated. The reactions were incubated at 37 °C for 1 hour and products were resolved alongside *dfnA* 

size markers by 6% urea-denaturing polyacrylamide gel electrophoresis. Gel image is representative of multiple transcription experiments.