A genetic and bioinformatic analysis of *Streptomyces coelicolor* genes containing TTA codons, possible targets for regulation by a developmentally significant tRNA

Wencheng Li[1], Jing Wu[1], Weixin Tao[1], Chunhua Zhao[1], Yemin Wang[1], Xinyi He[1], Govind Chandra[3], Xiufen Zhou[1,2], Zixin Deng[1,2], Keith F. Chater[3] & Meifeng Tao[1]

[1]State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China; [2]Laboratory of Microbial Metabolism, Shanghai Jiaotong University, Shanghai, China; and [3]John Innes Centre, Norwich Research Park, Colney, Norwich, UK

Correspondence: Meifeng Tao, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China. Tel.:+86 027 87283702; fax:+86 027 87280670; e-mail: tao_meifeng@yahoo.com

Received 21 May 2006; revised 23 August 2006; accepted 23 August 2006.
First published online 14 November 2006.

DOI:10.1111/j.1574-6968.2006.00494.x

Editor: Jose Gil

Keywords
bldA; TTA codon; bioinformatics; codon adaptation index.

Abstract

The rarest codon in the high G+C genome of *Streptomyces coelicolor* is TTA, corresponding in mRNA to the UUA codon that is recognized by a developmentally important tRNA encoded by the *bldA* gene. There are 145 TTA-containing genes in the chromosome of *S. coelicolor*. Only 42 of these are represented in the genome of *Streptomyces avermitilis*, among which only 12 have a TTA codon in both species. The TTA codon is less represented in housekeeping genes and orthologous genes, and is more represented in functional-unknown, extrachromosomal or weakly expressed genes. Twenty one TTA-containing chromosomal genes in *S. coelicolor* were disrupted, including 12 of the 42 genes that are common to both *S. avermitillis* and *S. coelicolor*. None of the mutant strains showed any obvious phenotypic differences from the wild-type strain under tested conditions. Possible reasons for this, and the role and evolution of the observed distribution of TTA codons among *Streptomyces* genes were discussed.

Introduction

Streptomycetes are Gram-positive mycelial bacteria that have an extensive secondary metabolism and undergo complex morphological differentiation to form a sporulating aerial mycelium. In the model organism *Streptomyces coelicolor* A3(2), and in other species tested, development and many secondary metabolism are pleiotropically defective in mutants of *bldA*, which encodes the only tRNA that can efficiently translate the rare leucine codon UUA (Leskiw *et al.*, 1991b; Chater, 2006; Chater & Chandra, 2006). Thus, although *bldA* mutants show apparently normal vegetative growth, they are defective in the production of at least four known antibiotics and in the formation of aerial mycelium on most media (Merrick, 1976; Champness, 1988). The pleiotropic effects of *bldA* mutations in *S. coelicolor* are at least partially attributable to the presence of UUA codons in the mRNA of critical regulatory genes (Chater, 2006). For example, *actII-4*, which encodes the pathway-specific regulator of actinorhodin production, contains a TTA codon (Fernandez-Moreno *et al.*, 1991), as does *redZ*, a regulatory gene required for undecylprodigiosin production (White & Bibb, 1997; Guthrie *et al.*, 1998). Another TTA-containing transcriptional regulatory gene, *adpA* (also termed *bldH*), is the main route by which *bldA* affects morphological differentiation (Nguyen *et al.*, 2003; Takano *et al.*, 2003). Recent proteomic analyses showed that a *bldA*-deleted mutant had impaired production of several extracellular proteins, including a potentially developmentally significant trypsin-like protease inhibitor SCO0762 (Kim *et al.*, 2005b); and two hypothetical proteins, SCO4244 and SCO4252, were absent (Kim *et al.*, 2005a). *SCO0762* does not contain a TTA codon, and its disruption mutant differentiates normally, but transcription of *SCO0762* depends on the TTA-containing gene *adpA* (Kim *et al.*, 2005b). *SCO4244* and *SCO4252* are in two operons that are located close to each other and the transcription of the operons were inactivated by disruption of the nearby TTA-containing regulatory gene, *SCO4263*, although disruption of *SCO4263* had no obvious phenotype with respect to antibiotic production or morphological differentiation (Kim *et al.*, 2005a; Hesketh *et al.*, in preparation). Expression analysis indicated that the abundance of the *bldA*-encoded tRNA is at its highest in stationary phase, in contrast to what is expected for most tRNA species (Trepanier *et al.*, 1997). In agreement

with this, expression of TTA-containing genes was found to be delayed during early differentiation (Kataoka *et al.*, 1999).

All *Streptomyces* ssp. have a very high G+C content (typically more than 70%), making the TTA codon rare. Many known TTA-containing genes have been found to be associated with morphological and physiological differentiation, and expression of these genes may be limited even in wild-type streptomycetes (Leskiw *et al.*, 1991b; Chater, 2006). The genome sequences of two *Streptomyces* species – *S. coelicolor* (Bentley *et al.*, 2002) and *Streptomyces avermitilis* (Ikeda *et al.*, 2003) – are now available. Analysis of these genomes has shown that the position in mRNAs of UUA codons is biased towards the start of coding sequences, implying that translational selection of codon usage occurs in streptomycetes (Fuglsang, 2005; Chater & Chandra, 2006). In *S. coelicolor*, knowledge of the roles of TTA-containing genes has mostly resulted from investigations of mutants with obvious phenotypic defects. Here, we have attempted to approach this problem using different approaches, a bioinformatics analysis coupled with targeted mutagenesis of 21 of the 145 TTA-containing chromosomal genes.

Materials and methods

Sequences

Genome sequences of *S. coelicolor* (NC_003888 for the chromosome, NC_003903 for plasmid SCP1 and NC_003904 for plasmid SCP2) and *S. avermitilis* (NC_003155) were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Protein functional classification was taken from the *S. coelicolor* genome project at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml). Proteins identified by proteomic approaches, which were taken broadly to represent highly expressed genes, were downloaded from the *S. coelicolor* 2D Gel Protein Database (http://dbk.ch.umist.ac.uk/s_coeli/referencegel/) (Hesketh *et al.*, 2002).

Finding orthologues of TTA-containing genes of *S. coelicolor* in *S. avermitilis*

Stand-alone BLAST (Altschul *et al.*, 1990) (NCBI BLAST package, ftp://ftp.ncbi.nih.gov) was used to search for orthologues of TTA-containing genes. Each TTA-containing gene of *S. coelicolor* or *S. avermitilis* was used to search against all genes of *S. avermitilis* or *S. coelicolor*, respectively, at translated protein levels. For a TTA-containing gene of *S. coelicolor* (gc), an orthologue in *S. avermitilis* (ga) was defined as (i) ga is the best hit of gc in the BLAST; (ii) the E-value is below 1e-10; (iii) the alignable region of the two sequences is at least 50% of the longer sequence; (iv) there is at least 50% amino-acid identity.

Expression prediction

The codon adaptation index (CAI) is a measure of codon usage in a gene relative to that in a reference set of genes (Sharp & Li, 1987). CAI has been used to predict gene expression levels in *S. coelicolor* and *S. avermitilis* (Wu *et al.*, 2005). Here, we used CodonW 1.4.2 (written by John Peden and available from http://codonw.sourceforge.net/) to calculate the CAI value for each TTA-containing gene, using ribosomal protein genes as reference gene set. The CAI values of TTA-containing genes were compared to those of ribosomal protein genes and all genes using a *t* test.

Disruption of some TTA-containing genes in *S. coelicolor*

Genes were disrupted in *Escherichia coli* hosts by the insertion of antibiotic resistance determinants, either by the use of restriction fragments (Kieser *et al.*, 2000) or by PCR targeting (Gust *et al.*, 2003), and then passaged through the non-methylating *E. coli* ET12567 before their reintroduction into *S. coelicolor* M145 by conjugation or protoplast transformation (Kieser *et al.*, 2000). Exconjugants or transformants were screened for double cross-over replacements by loss of vector-encoded antibiotic resistance, and confirmed by PCR or Southern blotting. The phenotypes were observed by culturing the disruption mutant strains on three solid mediums (MM, MS and R2YE, see Kieser *et al.*, 2000) at 30 for up to 10 days. Further details are given in the appropriate section of the Results.

Results

TTA codons in *Streptomyces* genes are rarer than expected

The expected random frequency of TTA codons in the genome of *S. coelicolor* was estimated as 0.095% by multiplying together the overall frequencies of the relevant nucleotides at the three positions of codons (T1, 10.8%; T2, 25.8%; A3, 3.4%). Strikingly, the observed frequency of TTA codons, at 0.006%, is only 6% of the expected frequency. Only 145 of the 7825 chromosomal genes contain TTA codons (see Table 1, three of these are duplicates because they are in the terminal inverted repeats of the chromosome). Other 17/356 are in the large linear plasmid SCP1 and 1/34 in the circular plasmid SCP2. Most of these genes contain only one TTA codon, except for ten in the chromosome that contain two. As shown in Table 1, 31 of the genes fall within groups of genes considered to have been laterally acquired in the relatively recent evolutionary past (Bentley *et al.*, 2002).

TTA codon usage in other organisms was also less than expected (15–52% of expected frequency) in other high

Table 1. TTA-containing genes in the *Streptomyces coelicolor* genome

Genome segment (by gene number)	TTA-containing genes*
SCO0001 to 1000	(0010, 0014, **0020**), 0075, 0101, 0124, 0145, 0182, **0239**, 0308, 0383, 0399, 0588, **0797**, 0856, 0992
SCO1001 to 2000	1004, **1093**, **1187**, 1227, **1242**, 1273, 1331, **1420**, **1434**, **1592**, 1604, **1980**, 1983
SCO2001 to 3000	2320, 2426, 2524, 2603, 2604, **2706**, **2792**
SCO3001 to 4000	**3257**, 3262, 3265, 3268, 3294, **3423**, 3468, 3469, 3487, 3490, 3496, 3498, 3570, 3682, 3693, **3770**, 3776, 3897, 3929, 3930, 3934, **3955**, 3982, 3983
SCO4001 to 5000	**4015**, 4060, 4063, **4114**, **4144**, 4213, 4262, 4263, 4301, **4312**, 4346, 4349, **4395**, 4431, 4464, 4481, **4493**, 4615, **4636**, 4642, 4671, **4794**, 4823
SCO5001 to 6000	5007, 5017, **5040**, 5083, 5085, **5203**, **5222**, 5276, 5345, 5350, 5411, **5460**, **5495**, 5606, 5633, 5786, 5799, 5881, 5913, **5968**, **5970**, 5995
SCO6001 to 7000	6034, 6075, **6209**, **6255**, 6315, 6324, **6384**, 6386, 6387, 6401, **6476**, 6595, **6623**, 6638, **6717**, **6741**, 6925, 6930, 6936
SCO7001 to 7845	7070, 7080, 7091, 7092, 7137, 7212, **7233**, **7251**, **7273**, **7351**, 7465, 7614, 7798, 7801, *7802*, 7807, 7812, *7814*, (7827, 7833, 7837)

***bold-face type** indicates that there is an orthologue in *S. avermitilis*; *italics* indicates that the annotated gene has an inappropriate start or stop codon; underlining indicates that the genes fall within putative laterally acquired gene islands; and (brackets) indicates that the genes are part of the repeated ends of the chromosome.

G+C-content organisms (Table 2), but *S. coelicolor* and *S. avermitilis*, which were the highest G+C-content organisms analysed here, had the lowest ratio of observed/expected frequency (6% and 9%, respectively). There are, altogether, 260 TTA-containing genes in *S. avermitilis*. TTA codons were slightly over-represented in *E. coli* and *Bacillus subtilis*, which have chromosomes of medium or low G+C content, respectively. We found that the G+C content of genome was negatively correlated to the ratio of observed/expected frequency of TTA codon (Pearson's correlation, $r = -0.93$). A phylogenetic tree was drawn based on the 16s rRNA sequences of these genomes (Fig. 1). We suspected that (a) some differences of the obs/exp values in high G+C content organisms are caused by their taxonomies. For example, *Deinococcus radiodurans* and *Halobacterium* sp., two high G+C content organisms grouped together, have relatively high obs/exp values compared with other high G+C content organisms. (b) *Mycobacterium tuberculosis*, with a high G+C content and grouped together with *Streptomyces*, has a relatively high obs/exp value, which might be caused by its very slow growth rate compared with *S. coelicolor* and *S. avermitilis*.

The functional classification of TTA-containing genes in *S. coelicolor*

The distribution of putative function among these TTA-containing genes is skewed in comparison with the whole genome (Table 3). Few of them are likely to function in cell processes, while function-unknown genes, and genes usually associated with mobile genetic elements are over-represented. Ten (7%) of the TTA-containing chromosomal genes and three of those on SCP1 are involved in secondary metabolism, including polyketide synthesis or non-ribosomal peptide synthesis. However, a smaller portion (3.5%) of all chromosomal genes is involved in secondary metabolism. There are nine TTA-containing chromosomal genes in the 22 gene clusters for secondary metabolism proposed by Bentley *et al.* (2002). They are *actII-2* and *actII-4* in the gene cluster of actinorhodin; *redZ* in the gene cluster of prodiginines; *SCO0124* in the gene cluster of eicosapentaenoic acid production; *SCO0383*, *SCO0399* in the gene cluster of deoxysugar synthases/glycosyl transferases; *SCO1273* in the gene cluster of type II fatty acid synthase; *SCO5222* in the gene cluster of sesquiterpene cyclase and *SCO5799* in the gene cluster of siderophore synthetase. The fraction of putative regulatory genes is nearly the same in TTA-containing genes and in all other genes.

S. avermitilis orthologues of TTA-containing genes of *S. coelicolor*

To find out how many of the 145 TTA-containing genes of the chromosome of *S. coelicolor* were present in *S. avermitilis*,

Table 2. The frequency of TTA codons in some bacterial genomes

Organism*	Accession number	G+C content (%)	Observed frequency (%)	Expected frequency (%)	Ratio of observed/ expected
Bacillus subtilis	NC_000964	43.5	1.9	1.5	1.33
Escherichia coli	NC_000913	50.8	1.4	0.84	1.66
Mycobacterium tuberculosis	NC_002755	65.6	0.16	0.32	0.51
Pseudomonas aeruginosa	NC_002516	66.6	0.029	0.20	0.14
Deinococcus radiodurans	NC_001263, NC _001264	67	0.070	0.19	0.36
Ralstonia solanacearum	NC_003295	67	0.025	0.17	0.15
Caulobacter crescentus	NC_002696	67.2	0.035	0.14	0.25
Bordetella pertussis	NC_002929	67.7	0.022	0.17	0.13
Halobacterium sp.	NC_002607	67.9	0.068	0.18	0.39
S. avermitilis	NC_003155	70.7	0.011	0.12	0.09
S. coelicolor	NC_003888	72.1	0.006	0.095	0.06

*Some genomes with G+C content > 65% are included. *Bacillus subtilis* and *Escherichia coli* genomes are also chosen to represent well-studied bacteria with low or medium G+C content, respectively.

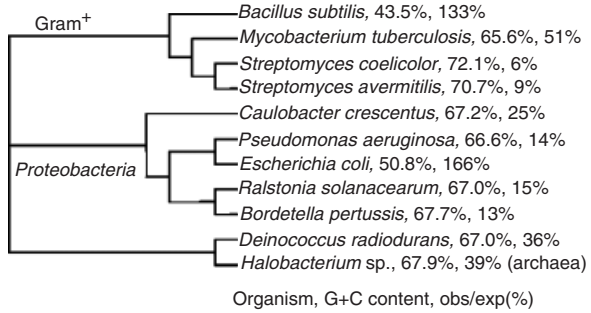

Fig. 1. Phylogenetic tree of microorganisms in Table 3. Their G+C content and the ratio of observed/expected frequency of TTA codon (obs/exp) are also shown.

Table 3. Function classification of all chromosomal genes and of TTA-containing chromosomal genes in *Streptomyces coelicolor*

Function classification	*n* (% in all genes)	*n* (% in TTA-containing genes)
Unknown function	2371 (30.3)	59 (40.7)
Cell processes	802 (10.2)	3 (2.1)
Macromolecule metabolism	496 (6.3)	6 (4.1)
Metabolism of small molecules	1104 (14.1)	15 (10.3)
Cell envelope	1383 (17.7)	21 (14.5)
Extrachromosomal*	139 (1.8)	14 (9.7)
Regulation	965 (12.3)	17 (11.7)
Not classified	565 (7.2)	10 (6.9)
Total	7825 (100.0)	145 (100.0)

*'Extrachromosomal' includes laterally acquired elements, phage-related genes, plasmid-related genes, transposon/insertion element-related genes.

we made a gene-by-gene search, based on reciprocal BLAST hits, at the translated protein level. In all, 30% (42) of the TTA-containing genes had an orthologue in *S. avermitilis* (Table 4), compared with 55% of TTA-free genes.

As indicated by their numbering, many of the TTA-containing genes are in the same order in the chromosomes of the two species, and closer inspection showed that this is invariably associated with substantial local similarity of gene arrangement (synteny). Of the 42 TTA-containing genes also represented in *S. avermitilis*, only one (*SCO4636*) has an apparent orthologue in another sequenced actinobacterial genome (*Thermobifida fusca*, which is fairly closely related to streptomycetes: Chater & Chandra, 2006). Only 12 genes have a TTA codon in both *S. coelicolor* and *S. avermitilis* (Table 4).

Predicted expression levels of TTA-containing genes

To dissect the relation of predicted gene expression level and codon usage, the CAI value of genes was plotted against the G+C content at the third position (degenerate site) of the codon (GC3s) in *S. coelicolor* (Fig. 2). Ribosomal protein genes, having relatively high CAI values (from 0.60 to 0.88, mean value = 0.76), were clustered at the upper end of the plot, as were genes previously identified by proteomic approaches, which had CAI values from 0.40 to 0.89 (mean value = 0.69). In contrast, TTA-containing genes were spread over the lower end of the plot with generally low CAI values (from 0.28 to 0.74, mean value = 0.53). Only three of the 646 proteins listed in the *S. coelicolor* 2D Gel Protein Database are encoded by TTA-containing genes (*SCO4636, SCO6401* and *SCO6638*, with CAI values of 0.67, 0.40 and 0.58 respectively). The differences in CAI values between TTA-containing genes and ribosomal protein genes or all genes are highly significant (*t*-test; $P = 6E\text{-}27$ and 5E-47, respectively). The CAI values of TTA-containing genes for *S. avermitilis* are also significantly lower than those of all genes or ribosomal genes (data not shown).

Table 4. Possible products of 42 TTA-containing chromosomal genes in *S. coelicolor* with an orthologue in *Streptomyces avermitilis*

Protein in *S. coelicolor*	Orthologue in *S. avermitilis*	Annotation in *S. coelicolor*
SCO0020	**SAV7545**	**putative transposase**
SCO0239	SAV818	hypothetical protein
SCO0797	**SAV7430**	**putative integral membrane protein**
SCO1093	SAV1495	putative hydroxylase
SCO1187	SAV555 (CelA1)	putative secreted cellulase B precursor
SCO1242*	**SAV7096†**	**putative DNA-binding protein**
SCO1420	**SAV6926**	**putative integral membrane protein.**
SCO1434*	**SAV6911†**	**putative CbxX/CfqX family protein**
SCO1592	**SAV6746**	**hypothetical protein**
SCO1980	**SAV6252†**	**hypothetical protein**
SCO2706	**SAV5359**	**putative transferase**
SCO2792	**SAV5261†**	**AraC-family transcriptional regulator (AdpA)**
SCO3257	SAV3734† (traSA1)	plasmid transfer protein
SCO3423*	**SAV4648†**	**putative regulator**
SCO3770	SAV1987 (Cyp8)	putative cytochrome P450 oxidoreductase
SCO3955	**SAV4251**	**conserved hypothetical protein SCD78.22c**
SCO4015	**SAV4201**	**hypothetical protein 2SC10A7.19**
SCO4114*	**SAV4113† (Sap)**	**sporulation associated protein**
SCO4144	**SAV4070**	**conserved hypothetical protein SCD84.12c**
SCO4312*	**SAV3919**	**conserved hypothetical protein**
SCO4395*	**SAV3854†**	**putative hydrolase**
SCO4493*	SAV4812	putative AsnC-family transcriptional regulator
SCO4636	SAV4901	hypothetical protein SCD82.07
SCO4794	**SAV3466**	**putative integral membrane protein**
SCO5040*	**SAV3223†**	**conserved hypothetical protein**
SCO5203	**SAV3056**	**hypothetical protein 2SC3B6.27c**
SCO5222	**SAV3032 (Tpc2)**	**putative lyase**
SCO5460	**SAV2785**	**putative AbaA-like regulatory protein**
SCO5495*	**SAV2747†**	**putative phosphodiesterase**
SCO5968	**SAV2328**	**putative bldA-regulated nucleotide binding protein**
SCO5970	**SAV2326**	**hypothetical protein**
SCO6209	**SAV2020**	**hypothetical protein SC2G5.30**
SCO6255	**SAV1985**	**putative dehydrogenase**
SCO6384	SAV6029	putative integral membrane lysyl-tRNA synthetase
SCO6476	**SAV1908**	**hypothetical protein SC9C7.12**
SCO6623*	**SAV1814†**	**putative ATP/GTP binding protein**
SCO6717*	**SAV1691**	**putative acyl-[acyl-carrier protein] desaturase**
SCO6741	**SAV1671**	**putative oxidoreductase**
SCO7233	SAV2604	putative secreted protein
SCO7251*	**SAV1237†**	**conserved hypothetical protein**
SCO7273	**SAV1160**	**hypothetical protein**
SCO7351	**SAV653**	**putative AraC-family transcriptional regulator.**

*The gene has been disrupted in this study.

†The *S. avermitilis* gene also contains at least one TTA codon.

Genes in **bold-face type** show both overall and local synteny between the two chromosomes.

Preference of C3s or T3s in highly expressed genes of *Streptomyces* have been characterized (Wright & Bibb, 1992). In agreement with this, we found that TTA-containing genes have low C3s, T3s and high A3s compared with ribosomal protein genes (*t*-test; $P = 2E\text{-}15$, 0.0007 and 1E-37, respectively). Note that HEG (highly expressed genes identified by proteomic approaches) have high C3s like ribosomal protein genes (not significant in *t*-test; $P = 0.08$), but have low T3s compared with ribosomal protein genes (*t*-test; $P = 8E\text{-}13$), indicating that HEG might have a slightly different codon usage pattern compared with ribosomal protein genes, although they also have high CAI values.

Disruption of 21 TTA-containing genes in *S. coelicolor*

Since *bldA* mutants of *S. coelicolor* grow well, it was not expected that any TTA-containing genes should be essential

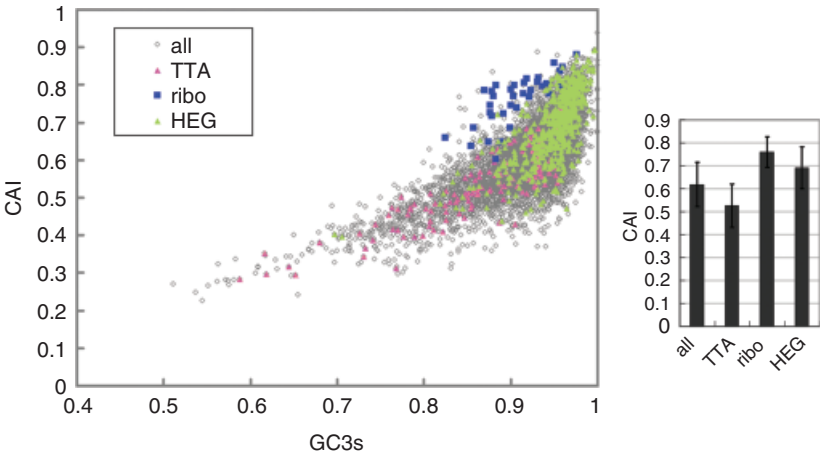

Fig. 2. The codon adaptation index (CAI) values of genes in *S. coelicolor*. (a) CAI plotted against GC3s (G+C content at the 3rd position of codon) for each gene in *S. coelicolor* with a length of longer than 300 bases. Ribosomal genes (ribo), highly expressed genes identified by proteomic approaches (HEG) and TTA-containing genes (TTA) are represented by the blue squares, green triangles and pink triangles, respectively. All other genes are represented by the black circles. TTA-containing genes are clustered at the lower end of the plot, having relatively low CAI values. (b) CAI values of ribosomal genes (ribo), HEG identified by proteomic approaches, TTA-containing genes (TTA) and all genes (all). Error bar is the standard deviation of the mean.

Table 5. Disrupted TTA-containing genes in *S. coelicolor* M145

Gene	Annotation	Gene length	Replaced region or inserted site	Disruption cassette	Methods
SCO0399	possible membrane protein	1224	196	*aac (3)IV* (Tao *et al*., 2002)	traditional
SCO1242	**probable DNA-binding protein**	861	40–861	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO1434	**possible CbxX/CfqX family protein**	1857	72–122	*aac (3)IV* (Tao *et al*., 2002)	traditional
SCO3423	**possible regulator**	465	1–465	*vph* (Blondelet-Rouault *et al*., 1997)	traditional
SCO3496	possible lyase precursor	1482	470	*aadA* (Kieser & Melton, 1988)	traditional
SCO3682	possible delta fatty acid desaturase	1038	6–548	*aac (3)IV* (Tao *et al*., 2002)	traditional
SCO3930	hypothetical protein	567	1–167	*aac (3)IV* (Blondelet-Rouault *et al*., 1997)	traditional
SCO3934	FtsK/SpoIIIE family protein	1302	986	*eryE* (Bibb *et al*., 1985)	traditional
SCO4114	**sporulation associated protein**	1422	1–1422	*aadA* (Gust *et al*., 2003)	PCR-targeting
SCO4301	possible DNA-binding protein	840	2–823	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO4312	**hypothetical protein**	789	214–789	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO4395	**possible hydrolase**	1059	133–1026	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO4493	**probable AsnC-family transcriptional regulator**	504	40–465	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO5040	**conserved hypothetical protein**	2214	472–2205	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO5495	**possible membrane associated phosphodiesterase**	2241	1095	*hyg* (Blondelet-Rouault *et al*., 1997)	traditional
SCO5633	probable fusion protein partially within putative integrated plasmid	2307	808	*aadA* (Blondelet-Rouault *et al*., 1997)	traditional
SCO5913	probable secreted protease	1236	1–599	*hyg* (Blondelet-Rouault *et al*., 1997)	traditional
SCO6034	unknown	1329	20–1246	*aac (3)IV+oriT* (Gust *et al*., 2003)	PCR-targeting
SCO6623	**probable ATP/GTP binding protein**	2196	1–472	*aac (3)IV* (Blondelet-Rouault *et al*., 1997)	traditional
SCO6717	**probable acyl-[acyl-carrier protein] desaturase**	987	416–706	*aadA* (Kieser & Melton, 1988)	traditional
SCO7251	**hypothetical protein**	1038	334	*hyg* (Blondelet-Rouault *et al*., 1997)	traditional

Genes with orthologues in *S. avermitilis* are given in bold-face type (see also Tables 1 and 4).

for vegetative growth; but at least some of these genes must be involved in morphological differentiation or secondary metabolism to account for the defects of *bldA* mutants in development and secondary metabolism (such as *adpA* and some TTA-containing pathway-specific regulatory genes). The TTA-free version of *adpA* gene could only partially restore aerial mycelium formation to a *bldA* mutant (Nguyen *et al*., 2003; Takano *et al*., 2003), indicating that other unknown TTA-containing genes might have a role in morphological differentiation. To investigate the roles of other TTA-containing genes, a further 21 were disrupted in *S. coelicolor* M145. We chose these genes mainly by their annotations, which were considered by us to be possibly related to differentiation (we chose some regulatory genes, enzymes; and avoided laterally acquired genes). The genes targeted include 12 of the 42 that are also represented in *S. avermitilis* and within the 12, 9 contain TTA in both organisms (Table 5). The disruption mutant strains were

cultured on minimal defined medium (MM) and on two rich, undefined media (MS, R2YE). None of the mutant strains showed any obvious phenotypic differences from the wild-type strain. Thus, if these genes are functional, their roles are cryptic under these experimental conditions.

Discussion

Biological roles of TTA-containing genes in *S. coelicolor* and other streptomycetes

Some TTA-containing genes have been shown in previous studies to mediate the *bldA*-dependence of aerial growth and production of certain secondary metabolites. If any other TTA-containing genes are important in the life of *S. coelicolor*, the most likely candidates should be those also present in other species. Just 42 such genes have orthologues in *S. avermitilis*, a species believed to have shared its last common ancestor with *S. coelicolor* some 250 million years ago, fairly early in the evolutionary history of the genus (A. M. Ward, personal communication cited in Chater & Chandra, 2006). Most (33) of these 42 genes occupy essentially the same positions on the chromosome in both organisms, making it very likely that they were part of the chromosome of the last common ancestor. This synteny is particularly true of the 12 orthologues having a TTA in both organisms. One of the 12 orthologues is *adpA*, which is a major target for *bldA* regulation of morphological differentiation in *S. coelicolor* (where it is also known as *bldH*) (Nguyen *et al.*, 2003; Takano *et al.*, 2003). Whether *adpA* provides the same function in *S. avermitilis* remains to be elucidated, but the *S. griseus adpA* orthologue, which also has a TTA, is well characterized as a regulator of both morphological differentiation and secondary metabolism (Chater & Horinouchi, 2003). The confinement of most of the 42 genes to streptomycetes, coupled with their apparent evolutionary conservation within the genus, implies that they should have genus-specific adaptive significance. Our choice of genes for disruption was therefore strongly biased towards this gene set: of the 21 genes disrupted, nine had a TTA-containing orthologue in *S. avermitilis*, and three had a TTA-free orthologue. However, the mutations had no obvious phenotypic effects. Two of the previously studied TTA-containing genes of *S. coelicolor* to which significant roles could be ascribed are absent from *S. avermitilis*, along with the gene sets that they control (i.e. *actII-4* and *redZ*, both antibiotic pathway-specific regulatory genes). We disrupted nine more TTA-containing genes that were absent from *S. avermitilis*. None of the mutants constructed had an obvious phenotype.

A simple explanation of these might be that these 21 genes are all unimportant to growth and development of *Streptomyces*. Other possibilities are as follows:

(1) Perhaps these genes are important for processes that are not seen under normal laboratory conditions, such as responses to a biofilm environment or interactions with the phytosphere. Probably, many unique TTA-containing genes were laterally acquired in the comparatively recent evolutionary past, and have adaptive significance only in specialized ecological or stressed circumstances that are subject only to intermittent selection over evolutionary time, and which are difficult or impossible to detect under normal laboratory conditions.

(2) Disruption of some TTA-containing genes may have a molecular phenotype which was not detected by us. It is noteworthy that proteomic and transcriptomic analyses have shown that a phenotypically 'silent' mutation in another unique TTA-containing gene, *SCO4263* (a regulatory gene), does have a molecular phenotype: genes in a nearby 'function-unknown' operon are inactive in the mutant (Kim *et al.*, 2005a; Hesketh *et al.*, in preparation).

(3) We have found paralogues of the disrupted TTA-containing genes. We used the BLASTP program to search paralogues with overlap $\geq 50\%$ and identity $\geq 30\%$ and found that 12 (SCO1242, SCO3423, SCO3682, SCO3934, SCO4114, SCO4301, SCO4493, SCO5040, SCO5495, SCO5633, SCO5913 and SCO6623) of the 21 disrupted TTA-containing gene products have protein paralogues. However, only two of them (SCO5633 and SCO3682) have paralogues with identity $> 50\%$. It's not a surprise to find these paralogues, as many paralogous proteins were found in *S. coelicolor* genome (Bentley *et al.*, 2002). Although orthologues typically occupy the same functional niche in different species, whereas paralogues tend to evolve toward functional diversification (Tatusov *et al.*, 2003), it is still possible that paralogues may have very similar functions, so the disruption mutant of a single gene in a paralogue may have no phenotype.

How might the present distribution of TTA-containing genes have arisen?

The analysis of codon frequency, function classification and orthologous pairs presented here allows us to extend a simple hypothesis of the evolutionary pathway originally expounded by Leskiw *et al.* (1991a) for the evolution of TTA-containing genes in *Streptomyces*.

(1) TTA codons occurred less and less during evolution as a result of mutation bias towards increased G+C content (Wright & Bibb, 1992), and the abundance of the *bldA*-encoded tRNA became correspondingly reduced.

(2) TTA codons were selectively excluded from housekeeping and highly expressed genes by the force of translational selection, while they were retained by some genes that were subject only to intermittent selection over evolutionary time and/or were lowly expressed, or were functionally

unimportant [selection pressure acting to improve translation efficiency is stronger for highly expressed genes than for weakly expressed genes (Duret, 2002), so an absence of such strong selection in weakly expressed genes may have allowed them preferentially to retain the TTA codon].
(3) Only a limited number of TTA-containing genes acquired a role in morphological and physiological differentiation, and the expression of *bldA* became adapted to be maximized when these developmental genes were expressed, i.e. in severely growth-rate-limited or stressed cells.
(4) Some other genes ('fellow travellers') might be 'useless' for the growth and development of *Streptomyces*, or be 'useful' only in certain physiological conditions.
(5) Because some 'fellow travelling' genes are likely to have adaptive benefits only intermittently over evolutionary time, they are frequently represented in gene sets subject to lateral transfers, as represented by plasmids and chromosomal islands with atypical base composition.

Chater & Chandra (2006) discussed the possibility that the interactions of streptomycetes with bacteriophages might have provided some of the selective pressure for the evolution of the specialized role of TTA codons in streptomycetes.

Not unexpectedly, the fraction of putative regulatory genes is about the same (about 12%) in TTA-containing genes as in TTA-free genes, as the average number of genes regulated per regulatory gene is likely to be independent of the physiological circumstances to which the regulatory gene responds.

There is a relatively high frequency of TTA-containing genes in plasmids. These genes may either have undergone selection for developmentally associated expression, or have been acquired relatively recently from bacteria other than streptomycetes, in which the TTA codon does not have the same significance.

Author contributions

W.L. and J.W. contributed equally to this study.

Acknowledgements

This work was initiated during a Joint Project award to Z.D. and K.F.C. by the National Natural Science Foundation of China and the Royal Society, and was supported by grant NSFC, No. 30200005 from the National Natural Science Foundation of China.

References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bentley SD, Chater KF, Cerdeno-Tarraga AM *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141–147.

Bibb MJ, Janssen GR & Ward JM (1985) Cloning and analysis of the promoter region of the erythromycin resistance gene (*ermE*) of *Streptomyces erythraeus*. *Gene* **41**: E357–E368.

Blondelet-Rouault MH, Weiser J, Lebrihi A, Branny P & Pernodet JL (1997) Antibiotic resistance gene cassettes derived from the omega interposon for use in *E. coli* and *Streptomyces*. *Gene* **190**: 315–317.

Champness WC (1988) New loci required for *Streptomyces coelicolor* morphological and physiological differentiation. *J Bacteriol* **170**: 1168–1174.

Chater KF (2006) *Streptomyces* inside-out: a new perspective on the bacteria that provide us with antibiotics. *Philos Trans R Soc Lond B Biol Sci* **361**: 761–768.

Chater KF & Chandra G (2006) The evolution of development in *Streptomyces* analysed by genome comparisons. *FEMS Microbiol Rev* in press.

Chater KF & Horinouchi S (2003) Signalling early developmental events in two highly diverged *Streptomyces* species. *Mol Microbiol* **48**: 9–15.

Duret L (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**: 640–649.

Fernandez-Moreno MA, Caballero JL, Hopwood DA & Malpartida F (1991) The *act* cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bldA* tRNA gene of *Streptomyces*. *Cell* **66**: 769–780.

Fuglsang A (2005) Intragenic position of UUA codons in streptomycetes. *Microbiology* **151**: 3150–3152.

Gust B, Challis GL, Fowler K, Kieser T & Chater KF (2003) PCR-targeted *Streptomyces* gene disruption identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc Natl Acad Sci USA* **100**: 1541–1546.

Guthrie EP, Flaxman CS, White J, Hodgson DA, Bibb MJ & Chater KF (1998) A response-regulator-like activator of antibiotic synthesis from *Streptomyces coelicolor* A3(2) with an amino-terminal domain that lacks a phosphorylation pocket. *Microbiology* **144**: 727–738.

Hesketh AR, Chandra G, Shaw AD, Rowland JJ, Kell DB, Bibb MJ & Chater KF (2002) Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*. *Mol Microbiol* **46**: 917–932.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M & Omura S (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* **21**: 526–531.

Kataoka M, Kosono S & Tsujimoto G (1999) Spatial and temporal regulation of protein expression by *bldA* within a *Streptomyces lividans* colony. *FEBS Lett* **462**: 425–429.

Kieser T, Bibb MJ, Buttner MJ, Chater K & Hopwood DA (2000) *Practical Streptomyces Genetics*, John Innes Foundation, Norwich, UK.

Kieser T & Melton RE (1988) Plasmid pIJ699, a multi-copy positive selection vector for *Streptomyces*. *Gene* **65**: 83–91.

Kim DW, Chater KF, Lee KJ & Hesketh A (2005a) Effects of growth phase and the developmentally significant *bldA*-specified tRNA on the membrane-associated proteome of *Streptomyces coelicolor*. *Microbiology* **151**: 2707–2720.

Kim DW, Chater KF, Lee KJ & Hesketh A (2005b) Changes in the extracellular proteome caused by the absence of the *bldA* gene product, a developmentally significant tRNA, reveal a new target for the pleiotropic regulator AdpA in *Streptomyces coelicolor*. *J Bacteriol* **187**: 2957–2966.

Leskiw BK, Bibb MJ & Chater KF (1991a) The use of a rare codon specifically during development? *Mol Microbiol* **5**: 2861–2867.

Leskiw BK, Lawlor EJ, Fernandez-Abalos JM & Chater KF (1991b) TTA codons in some genes prevent their expression in a class of developmental, antibiotic-negative, *Streptomyces* mutants. *Proc Natl Acad Sci USA* **88**: 2461–2465.

Merrick MJ (1976) A morphological and genetic mapping study of bald colony mutants of *Streptomyces coelicolor*. *J Gen Microbiol* **96**: 299–315.

Nguyen KT, Tenor J, Stettler H, Nguyen LT, Nguyen LD & Thompson CJ (2003) Colonial differentiation in *Streptomyces coelicolor* depends on translation of a specific codon within the *adpA* gene. *J Bacteriol* **185**: 7291–7296.

Piret JM & Chater KF (1985) Phage-mediated cloning of *bldA*, a region involved in *Streptomyces coelicolor* morphological development, and its analysis by genetic complementation. *J Bacteriol* **163**: 965–972.

Sharp PM & Li WH (1987) The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281–1295.

Takano E, Tao M, Long F, Bibb MJ, Wang L, Li W, Buttner MJ, Bibb MJ, Deng ZX & Chater KF (2003) A rare leucine codon in *adpA* is implicated in the morphological defect of *bldA* mutants of *Streptomyces coelicolor*. *Mol Microbiol* **50**: 475–486.

Tao MF, Zhou XF, Kieser T & Deng ZX (2002) Construction of a temperature inducible shuttle expression vector and its application in *Streptomyces*. *Sheng Wu Gong Cheng Xue Bao* **18**: 420–423.

Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Trepanier NK, Nguyen GD, Leedell PJ & Leskiw BK (1997) Use of polymerase chain reaction to identify a leucyl tRNA in *Streptomyces coelicolor*. *Gene* **193**: 59–63.

White J & Bibb M (1997) *bldA* dependence of undecylprodigiosin production in *Streptomyces coelicolor* A3(2) involves a pathway-specific regulatory cascade. *J Bacteriol* **179**: 627–633.

Wright F & Bibb MJ (1992) Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **113**: 55–65.

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology* **151**: 2175–2187.