Chapter 6

A novel nucleoid-associated protein specific to Actinobacteria that binds to GATC sequences

Chao Du, Joost Willemse, Victor J. Carrion, and Gilles P. van Wezel

Manuscript in preparation

Abstract

Nucleoid-associated proteins (NAPs) are known to affect the development of *Streptomyces*, which is a direct regulating factor in antibiotic production. SCO1839 is a small DNA binding protein which was previously suggested as a novel NAP in *Streptomyces coelicolor*, but no function or binding site was attributed to the protein. Here, we show that SCO1839 is an NAP that plays a role in the control of morphogenesis and antibiotic production in *S. coelicolor*. SCO1839 did not show significant matches against the Pfam database and thus likely represents a novel protein family. Chromatin immunoprecipitation sequencing (ChIP-Seq) showed that SCO1839 has a very large number of binding sites, with one or more sites in more than 2500 distinct genomic locations. ChIP-Seq analysis showed that SCO1839 binds to a short palindromic sequence (GATC), with preferable two A/T bases on one side. Interestingly, by far the largest difference in enrichment between 25 h and 48 h was seen for SCO1311, which encodes a tRNA editing enzyme, and SCOt32, which specifies the tRNA that recognises the rare leucine codon CUA. In both cases, enrichment was much higher in 25 h than in 48 h samples. Disruption of SCO1839 resulted in changes of development speed, colony morphology, and spore shape. The effect of SCO1839 on antibiotic production may in part be caused by the interaction between SCO1839 and the global transcriptional regulator AtrA, which activates antibiotic production. In conclusion, our results highlight the role of SCO1839 as DNA binding protein, we also proved that this protein is involved in the morphological development and antibiotic production of *S. coelicolor*.

Introduction

Streptomycetes are filamentous soil bacteria with a complex life cycle, which are well known for their ability to produce various kinds of antibiotics. Streptomycetes are a major source of clinical drugs (Barka *et al.*, 2016, Hopwood, 2007, Bérdy, 2005). The life cycle of *Streptomyces* starts with the germination of a spore that grows out to form vegetative hyphae. Via tip extension and branching, this results in the formation of a dense mycelial network (Barka *et al.*, 2016, Chater & Losick, 1997). When the environmental situation requires sporulation, for example due to nutrient starvation, streptomycetes start their reproductive growth phase by developing areal hyphae, which eventually differentiate into chains of unigenomic spores (Claessen *et al.*, 2014, Flärdh & Buttner, 2009). The production of antibiotics temporally correlates to the developmental growth phase (Bibb, 2005, van der Heul *et al.*, 2018). The complexity of the underlying regulatory networks is underlined by the fact that the *Streptomyces coelicolor* genome encodes some 900 regulatory proteins, of which only a fraction has been functionally characterized (Bentley *et al.*, 2002). Many of these are involved in the control of development and antibiotic production, such as the *bld* and *whi* genes that are responsible for the control of aerial hyphae formation and sporulation, respectively, and global regulatory genes such as *afsR, dasR* and *atrA* that pleiotropically control antibiotic production (van der Heul *et al.*, 2018).

The control of chromosome structure is also an important regulating factor of gene expression. In bacteria, the organization of chromosome structure is mediated by a diverse group of proteins referred to collectively as nucleoid-associated proteins (NAPs) (Dillon & Dorman, 2010). These are generally small DNA binding proteins involved in processes such as controlling gene expression, nucleoid structure or repair. Well-known NAPs in *Streptomyces* include Lsr2, HupA, HupS, and sIHF. Lsr2 binds non-specifically to AT-rich sequences and can globally repress gene expression (Gehrke *et al.*, 2019). HupA and HupS are homologs of HU (for histone-like factor U) proteins, which are differentially regulated depending on the developmental growth phase (Salerno *et al.*, 2009). sIHF is one of the basic architectural element conserved in many actinobacteria and is able to influence the regulation of secondary metabolism and cell development (Yang *et al.*, 2012). IHF binds a well conserved nucleotide sequence, while HU binds to random DNA sequences (Swinger & Rice, 2004). A detailed proteomic survey identified 24 proteins with NAP-like properties, namely the known Lsr2, HupA, HupS and sIHF and 20 yet unidentified proteins (Bradshaw *et al.*, 2013). Although the functions of many NAPs are still not clear, but there are certainly some have pervasive influences on the transcriptome, such as BldC (Bush *et al.*, 2019, Dorman *et al.*, 2020). One of the putative NAPs was SCO1839, a small protein that was previously identified via pull-down assays as associated with the promoter region of *ssgR* (Kim *et al.*, 2015). SsgR is the transcriptional activator of the cell division

activator gene *ssgA* (Traag *et al.*, 2004). SsgA and its paralogue SsgB are both required for sporulation (Keijser *et al.*, 2003, van Wezel *et al.*, 2000a), and together coordinate the onset of sporulation-specific cell division in *Streptomyces*, whereby SsgB directly recruits the cell division scaffold protein FtsZ to the future sites of septation (Willemse *et al.*, 2011). The association of SCO1839 with the *ssgR* promoter thus implies a possible involvement in the control of sporulation. In this study, we show that SCO1839 represents a novel family of small DNA binding proteins, which plays a role in the regulation of *Streptomyces* development and antibiotic production. SCO1839 is specific to the Actinobacteria, with an HTH DNA binding motif containing three helixes. Deletion of SCO1839 resulted in a significant change in colony phenotype. Chromatin immunoprecipitation coupled with massively parallel DNA sequencing (ChIP-Seq) showed the genome binding sites and a specific binding motif centred around the palindromic sequence GATC.

Results and discussion

SCO1839 is a small NAP specific for Actinobacteria

SCO1839 is a small protein of 73 amino acids (7.6 kDa) and a predicted isoelectric point (pI) of 10.53, indicative of an alkalic protein. A Pfam sequence search (El-Gebali *et al.*, 2018) did not yield any significant matches to known protein families, strongly suggesting that SCO1839 is the first member of a new protein family. Previously, an affinity capture assay using the promoter region of the cell division regulatory gene *ssgR* as a bait pulled down SCO1839 (Kim *et al.*, 2015). SCO1839 was also identified as one of the most abundant possible NAP (nucleoid-associated protein) by Bradshaw and colleagues (Bradshaw *et al.*, 2013).

To obtain more insights into the distribution and phylogeny of SCO1839, a conserved Hidden Markov Model (HMM) domain was constructed using all SCO1839-like proteins from *Streptomyces* species. Consequently, an HMM search against all available bacteria full genomes in the database was performed. No hits were found outside the order Actinomycetales, strongly suggesting that SCO1839 is an Actinobacteria-specific protein (Figure 1A and B). Eight main clusters of similar groups of homologs were found. The largest cluster mainly consists SCO1839 homologs from *Streptomyces*, *Amycolatopsis*, *Pseudonocardia*, *Frankia*, and *Actinomadura*. Other major clusters including a cluster of *Nocardia* and *Rhodococcus*, *Micromonospora* and *Salinispora*, *Geodermatophilus* and *Blastococcus*. *Rhodococcus* itself forms two separate clusters. Interestingly, 24 *Actinomadura* species and 38 *Streptomyces* species contain two paralogues of SCO1839, which divided into two additional clusters. For most genera, more than 90% of the species contains at least one copy of SCO1839 like proteins. *Rhodococcus* stands out as only 28.7% (92 out of 321) of the genomes containing a SCO1839-family protein (Figure 1A), and these proteins divide into three distinct clusters. This could be due to the fact that it is the only nonsporulating genus among all listed genera. The fact that the nonsporulating genus *Rhodococcus* stands out clearly from the sporulating genera, in combination with the low frequency of occurrence, suggests that perhaps SCO1839 has diversified in the non-sporulating *Rhodococcus* and also that SCO1839 may primarily be sporulation-specific.

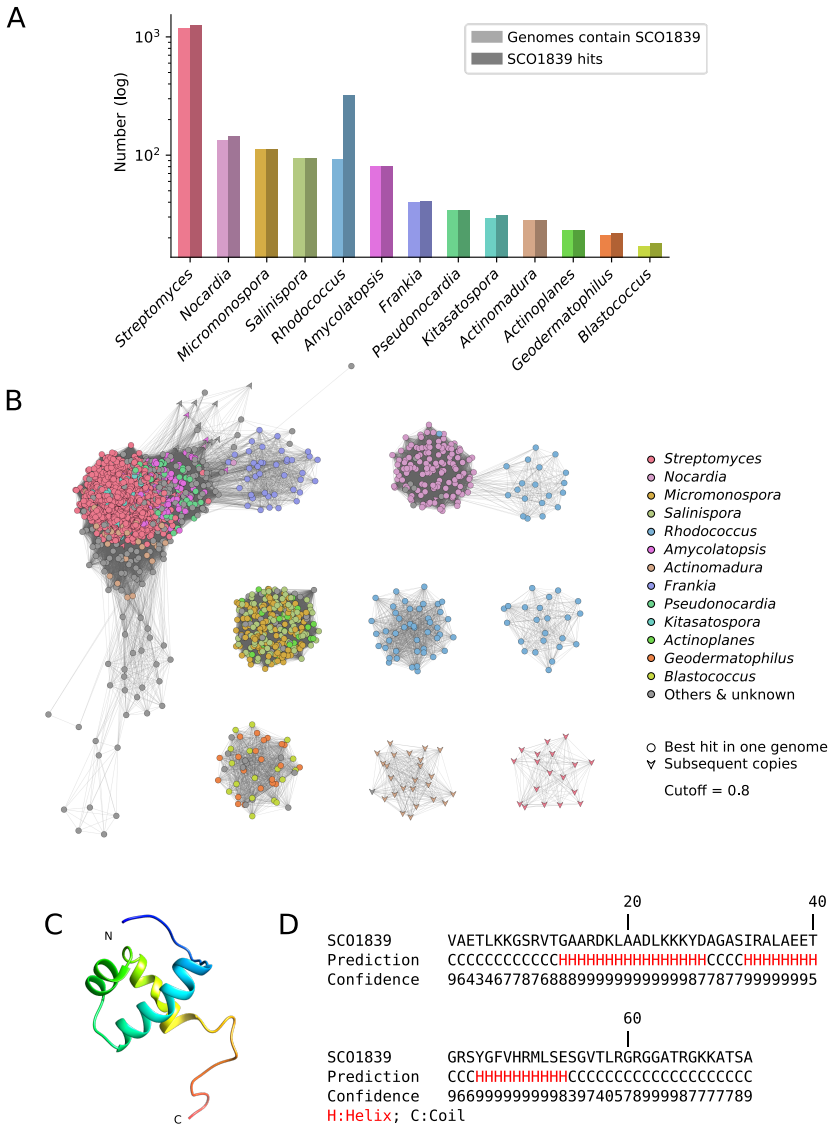

Figure 1. Distribution, phylogenetic network and structural analysis of SCO1839 protein. A) Bar plot representing the distribution of SCO1839-like proteins in all genomes used in this study. Light colours indicate the number of genomes containing SCO1839-like proteins in specific genera, dark colours represent the total number of genomes in each genus. B) Sequence similarity network built with a threshold of 0.8 using all SCO1839-like proteins detected in this study. Nodes represent SCO1839-like proteins. Edges connecting the nodes represent phylogenetic distances. Colours indicate the taxonomic affiliation of each orthologue. Shapes denote whether SCO1839 is the primary copy in one genome. C) *In silico* structural model of the SCO1839-like protein generated with I-TASSER program. D) Predicted structural string for SCO1839, obtained from the I-TASSER output.

In silico structural modelling of SCO1839 was performed using the I-TASSER server (Yang *et al.*, 2015), revealing a putative single DNA binding helix-turn-helix (HTH) motif, in the form of a tri-helical structure (Figure 1C and D). No homology was seen to any other known transcriptional regulator family from bacteria. The regions flanking the residues belonging to the HTH motif are very short. Nine other protein structures found in PDB (Protein data bank, rcsb.org, Berman *et al.*, 2000) share structural analogy and contained DNA binding motifs that structurally resemble SCO1839. These proteins come from a wide range of organisms and functions, including DNA helicase from archaeon bacteria *Pyrococcus furosus* (PDB structure ID: 2ZJ8), *E. coli* (2VA8, 2P6R), human (5AGA); ribosome protein from yeast (5MRC), cell division cycle protein from human (2DIN), terminator binding protein from yeast (5EYB), regulator from *Staphylococcus aureus* (2R0Q), and a tRNA synthetase from *Archaeoglobus fulgidus* (2ZTG). However, none of the analogous proteins were as small as SCO1839 and neither contained only one DNA binding motif. To the best of our knowledge, no other bacterial protein with similar structure has been reported before. Therefore, we propose that the SCO1839-like proteins form a new family of bacterial DNA binding proteins.

Deletion of SCO1839 accelerates sporulation of *S. coelicolor*

To obtain insights into the possible role of SCO1839 in the life cycle of *S. coelicolor*, a knock-out mutant was generated using a strategy published previously (Świątek *et al.*, 2012), whereby the +1 to +207 region of the gene was replaced by the apramycin resistance cassette flanked by *loxP* sites, followed by deletion of the cassette using Cre recombinase so as to avoid polar effects. To genetically complement the mutant and see if the wild-type phenotype would be restored, the -565/+228 region of SCO1839 was amplified from the *S. coelicolor* chromosome and cloned into pHJL401, a shuttle vector that is useful for genetic complementation due to its low copy number in streptomycetes (van Wezel *et al.*, 2000b). To also analyse the effect of enhanced expression of SCO1839, a second strain was constructed using CRISPR-Cas9 wherein the native promoter of SCO1839 (-157 to +4, start codon switched to ATG) was replaced by the strong constitutive *ermE* promoter (Bibb *et al.*, 1985). For details see the Materials and Methods section.

When grown on MM agar plates with mannitol as the carbon source, colonies of the SCO1839 mutant appeared darker in colour and more irregularly shaped as compared with the parental strain *S. coelicolor* M145 (Figure 2A, 6). However, on soya flour mannitol (SFM) agar plates, the mutant colonies showed a smoother surface, devoid of the ridges that are typical of wild-type colonies (Figure 2A, 8). Interestingly, the SCO1839 mutant failed to produce blue-pigmented antibiotic actinorhodin on SFM agar when colonies were grown at higher densities (Figure 2B, right). The SCO1839 mutant phenotype was restored in the complemented strain on minimum medium and SFM medium when growing as single colonies

(Figure 2A, 9-12). However, at higher density, actinorhodin production on SFM was not restored (Figure 2B, right). The strain expressing SCO1839 from *ermE* promoter produced large amounts of undecylprodigiosin on MM agar (Figure 2A, 13, 14) and showed a reduced growth on SFM agar (Figure 2A, 15, 16).

In order to more precisely monitor the differences in the timing of development between SCO1839 mutants and their parent, time-lapse imaging was performed on confluent mycelial lawns. This method allows monitoring multiple morphological characteristics and in particular quantify differences in timing of the individual developmental stages, based on pigmentation of the mycelia. When aerial hyphae are formed, intensity increases due to the white pigmentation, while brightness decreases again when grey-pigmented spores are produced (see Material and Methods section for details). Deletion of SCO1839 resulted in acceleration of development by 2-5 h (Figure 3A). 54 h after inoculation, the light intensity of the SCO1839 mutant again increased (Figure 3A). This may be due to premature germination and renewed growth. The complemented mutant had a far less pronounced acceleration of its life cycle, confirming that this phenomenon was primarily due to the deletion of SCO1839. Conversely, the enhanced and constitutive expression of SCO1839 delayed sporulation by approximately 18 h (Figure 3A and B). Taken together, this strongly suggests that the expression of SCO1839 correlates to the timing of sporulation.

Table 1. Statistics of spore length to width ratio

Strain	n	Median	M145 median	Mann–Whitney *U* test of difference		Variance	M145 variance	Levene's test of equal variance	
				U	*p*-value			*W*	*p*-value
ΔSCO1839	332	1.69		1.2×10^{5}	9.3×10^{-16}	0.145		49.51	3.93×10^{-12}
ΔSCO1839 + SCO1839	556	1.50		1.6×10^{5}	0.40	0.099		13.49	2.52×10^{-4}
pErmE-SCO1839	441	1.60	1.49 (n = 560)	1.5×10^{5}	1.4×10^{-7}	0.125	0.069 (n = 560)	34.89	4.78×10^{-9}
ΔSCO1839 ΔAtrA	565	1.52		1.6×10^{5}	0.20	0.114		13.01	3.23×10^{-4}
ΔAtrA	561	1.48		1.5×10^{5}	0.09	0.076		2.01	0.157

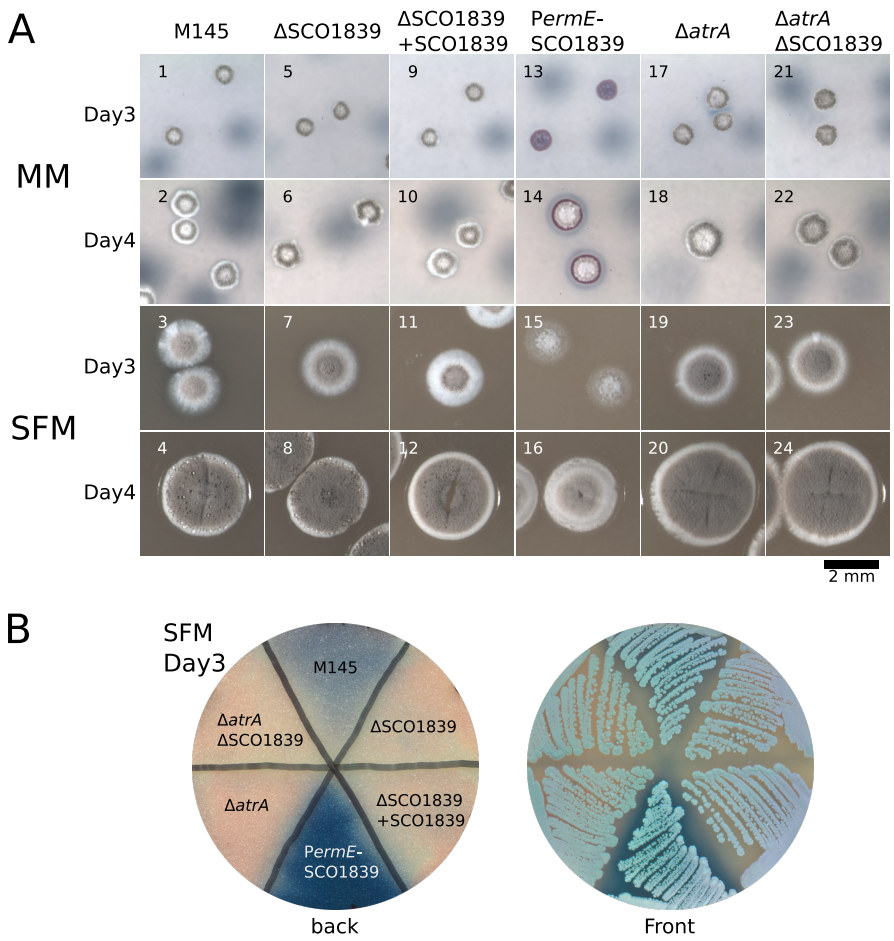

Figure 2. Colony morphology and pigment production of *S. coelicolor* M145 and mutants. A) *S. coelicolor* M145 and its mutants ΔSCO1839, Δ*atrA* and Δ*atrA*ΔSCO1839, as well as the M145 + P*ermE*-SCO1839 and ΔSCO1839 + P*ermE*-SCO1839. Strains were grown as single colonies on MM agar supplemented with mannitol or on SFM agar at 30°C during 3 and 4 days. All pictures were taken at the same scale. B) *S. coelicolor* M145 and derivatives compared on SFM medium. Note the enhanced Act production by M145 + P*ermE*-SCO1839.

Closer examination of the spores by scanning electron microscopy (SEM) revealed that spores were significantly enlarged in both the deletion mutant and the strain with enhanced expression of SCO1839 (Mann-Whitney U test, p-value < 0.01, Table 1, Figure 3 C). The complemented strain produced spores with a normal size. All engineered strains (ΔSCO1839, ΔSCO1839+SCO1839, PermE-SCO1839), including the complementation mutant, had a variance in the distribution of spore length to width ratio that differed significantly from that of the parent strain (Levene's test, p-value < 0.01), with the SCO1839 deletion mutant showing the

largest variance. Complementation of the mutant reduced the large variance to a level closest to the wild-type. Thus, deletion of SCO1839 alters the timing of sporulation and also leads to significant alteration of the spore morphology.

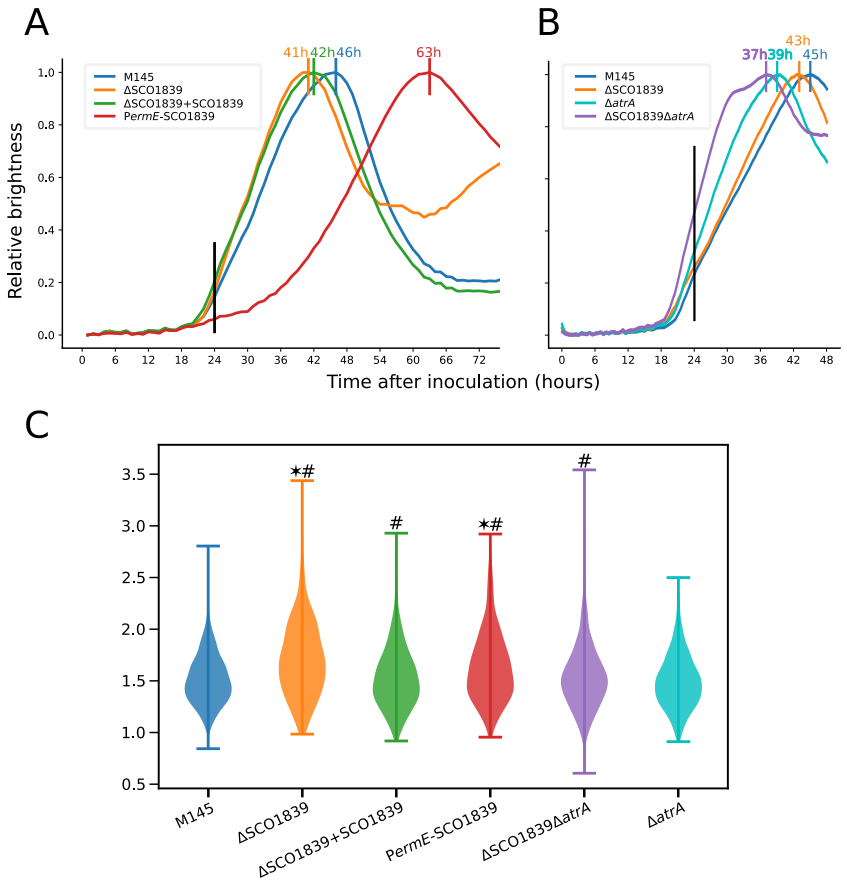

Figure 3. Growth and spore shape analysis of SCO1839 mutants. A) and B) scanner measurements of confluent plate brightness of *S. coelicolor* M145 and mutants (ΔSCO1839, ΔSCO1839+SCO1839, P*ermE*-SCO1839, Δ*atrA* and Δ*atr*AΔSCO1839) grown as confluent lawns. The X-axis represents the time after inoculation and the Y-axis the normalized brightness measured from time lapse scan pictures. Mutants shown as different colours, curve peaks are marked on top. Black vertical lines illustrate 24 h time point. C) Violin plot showing the distributions of the ratio of spore length and width as measured using scanning electron microscope images. Each strain was grown on MM agar for 5 days before imaging. ✶ with significantly ($p < 0.01$) different length to width ratio compared to *S. coelicolor* M145 (Mann–Whitney U test), **#** with significantly ($p < 0.01$) different variance compared to *S. coelicolor* M145 (Levene's test).

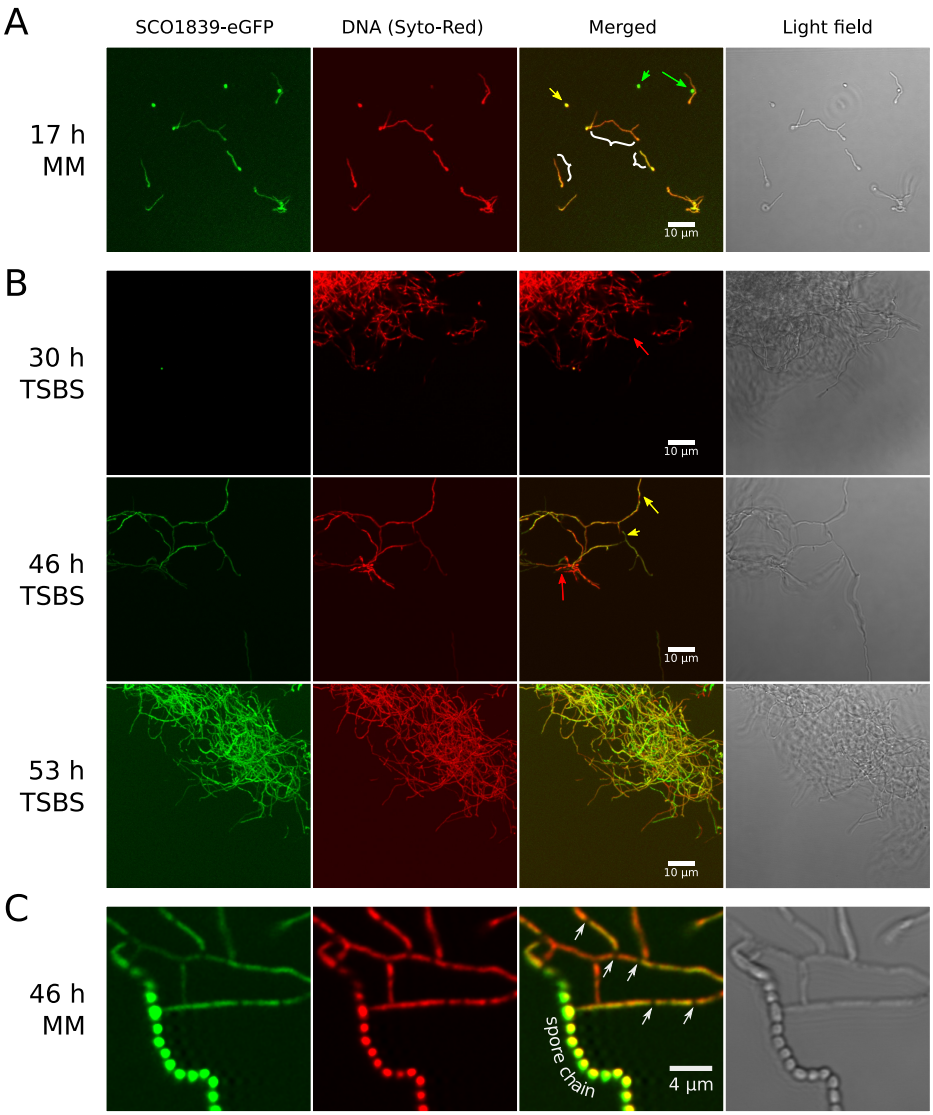

Figure 4. Confocal micrographs showing the expression of SCO19839-eGFP in different developmental stages and its co-localization with genomic DNA. A) Germinating spores on MM agar supplemented with mannitol. Dormant spores are highlighted using green arrows, swelled spores by yellow arrows, and germ tubes with white curly brackets. B) Mycelia grown in liquid TSBS media. SCO1839 expressing mycelium is indicated by yellow arrows, not expressing mycelium is indicated by red arrows. C) Mycelium grown on MM agar. SCO1839-eGFP and DNA were both absent at some location indicated by arrows.

SCO1839 expression after early development stage and is highly abundant in spores

To determine the location and timing of SCO1839 expression, a mutant strain that expresses SCO1839-eGFP fusion protein was constructed. For this, the gene for eGFP was fused to the C-terminal of SCO1839 using CRISPR-Cas9. Confocal images with DNA staining with Syto-Red showed that SCO1839-eGFP protein was strongly expressed in spores (Figure 4A). The fluorescence disappeared once the vegetative hyphae extended after germination, indicating that SCO1839 is not expressed during early stages of growth. In liquid TSBS, SCO1839-eGFP was only expressed approximately 45 h after inoculation and became stronger thereafter (Figure 4B). Transcriptional analysis using quantitative PCR (qPCR) confirmed a two-fold increase in SCO1839 expression at 48 h as compared to 24 h (Figure 7C). The qPCR analysis supported the results of the time series transcriptomics data available in the NCBI GEO database (Barrett *et al.*, 2012). Specifically, we compared our results with available data that include early developmental stages and that had been performed in different media (Gehrke *et al.*, 2019, Castro-Melchor *et al.*, 2010, Nieselt *et al.*, 2010). These authors showed that transcription of SCO1839 is low at early developmental stages, and increases as development progresses, in a medium-independent manner (Gehrke *et al.*, 2019, Castro-Melchor *et al.*, 2010, Nieselt *et al.*, 2010). This indicates that the timing of SCO1839 depends on the growth conditions and the developmental stage.

SCO1839 has thousands of DNA binding sites and specifically recognizes a small motif with GATC as core sequence

To obtain insights into the genome-wide DNA binding capacity of SCO1839, ChIP-Seq analysis was performed after 25 h (vegetative growth) and 48 h (sporulation). This should reveal all binding sites of SCO1839 on the *S. coelicolor* chromosome. For this propose, the original copy of SCO1839 on the genome was fused with a sequence encoding a triple FLAG tag in its C-terminal using CRISPR-Cas9 (see Materials and Methods section). The strain had a phenotype that was highly similar to that of the parent (Figure S1). The ChIP-Seq results showed a wide distribution of SCO1839 binding sites (Figure 5). Half of the binding sites that showed a high fold enrichment (50% of the sites at 25 and 38 h) colocalized with low G+C content regions. In total, 2825 and 2919 binding regions were identified for the vegetative and sporulation growing stage samples, respectively. Interestingly, there was a very high overlap (> 90%, 2402) between the SCO1839 binding regions in both stages. This not only shows that the binding sites of SCO1839 is largely growth phase-independent, but also that the experiments were highly reproducible. This is highlighted by the very similar binding patterns in Figure 5A.

To obtain the consensus sequence required by SCO1839 binding, the binding regions from ChIP-Seq results were extracted and modelled using MACS2 and

MEME-ChIP (Zhang *et al.*, 2008, Machanick & Bailey, 2011). As a result, GATC(A/T)T was found to be the most significant binding motif (Figure 5B). The most conserved binding core was GATC, which is a palindrome known as recognition site for DNA methylation (Barras & Marinus, 1989). The predicted motifs also showed G/C preference on the flanking region separated by two base pairs gaps (Figure 5B). It is important to note that virtually all binding regions that were identified as significant (> 99.8%) contained a GATC motif, and most (88.2% for 25 h, 84.0% for 48 h) contained the consensus sequence GATC(A/T)T. The *S. coelicolor* genome contains in total 6,500 GATC(A/T)T sequences, 64% and 68% were detected as binding regions in the 25 h and 48 h samples, respectively. Note that many binding sites have more than one copy of the motif. The *S. coelicolor* genome contains 43,535 GATC sequences, of which only 24% (25 h) and 30% (48 h) are in the identified SCO1839 binding regions. This indicates that the minimal binding motif is not the only factor that determines *in vivo* binding of SCO1839. Other determining factors might include yet undiscovered elements in the flanking regions, secondary genome structure, and/or the binding of other proteins to the same region.

The affinity of SCO1839 to the DNA binding motif was further tested *in vitro* using electrophoretic mobility shift assays (EMSA). The results showed that SCO1839 could bind to the DNA with one GATC motif with random flanking sequence, but the additional nucleotide (AT)T on one side increased the affinity of SCO1839 by around two-fold (Figure 6B). Furthermore, when four GATC motifs were present in the short 50 bp DNA fragment, simultaneous binding of three SCO1839 proteins was observed. Perhaps the two GATC sequences in close proximity were bound by a single SCO1839 protein due to steric hindrance. Taken together, SCO1839 showed good binding to a motif centred around GATC, and this is probably the only requirement for SCO1839 to bind *in vitro*.

The GATC motif is the target sequence for deoxyadenosine (DAM) methylase, which is essential for DNA mismatch repair in *E. coli* (Barras & Marinus, 1989). Another typical DNA methylation is on the second deoxycytosine of the sequence CCTGG (DCM) (May & Hattman, 1975). *S. coelicolor* lacks both methylation system and degrades methylated exogenous DNA (Flett *et al.*, 1997, Kieser & Hopwood, 1991). Which proteins are involved in the recognition and restriction system in *Streptomyces* remains unknown (Liu *et al.*, 2010, González-Cerón *et al.*, 2009). The fact that SCO1839 binds to GATC sequences could be correlated with the restriction system in *S. coelicolor*. Therefore, the transformation efficiency was tested for methylated DNA and nonmethylated DNA to ΔSCO1839 strain. However, the result indicates the barrier still exists (data not shown). Next, we tested the affinity difference of SCO1839 to GATC and GA^{m}TC in electrophoretic mobility shift assay (EMSA). The results showed an approximately two-fold decreased affinity of SCO1839 for methylated DNA (Figure 6D). As DNA is not

methylated in *S. coelicolor* (Flett *et al.*, 1997), the biological function of this difference if affinity cannot yet be explained. However, it is logical to assume that methylation causes some steric hindrance, which negatively affects the binding of SCO1839.

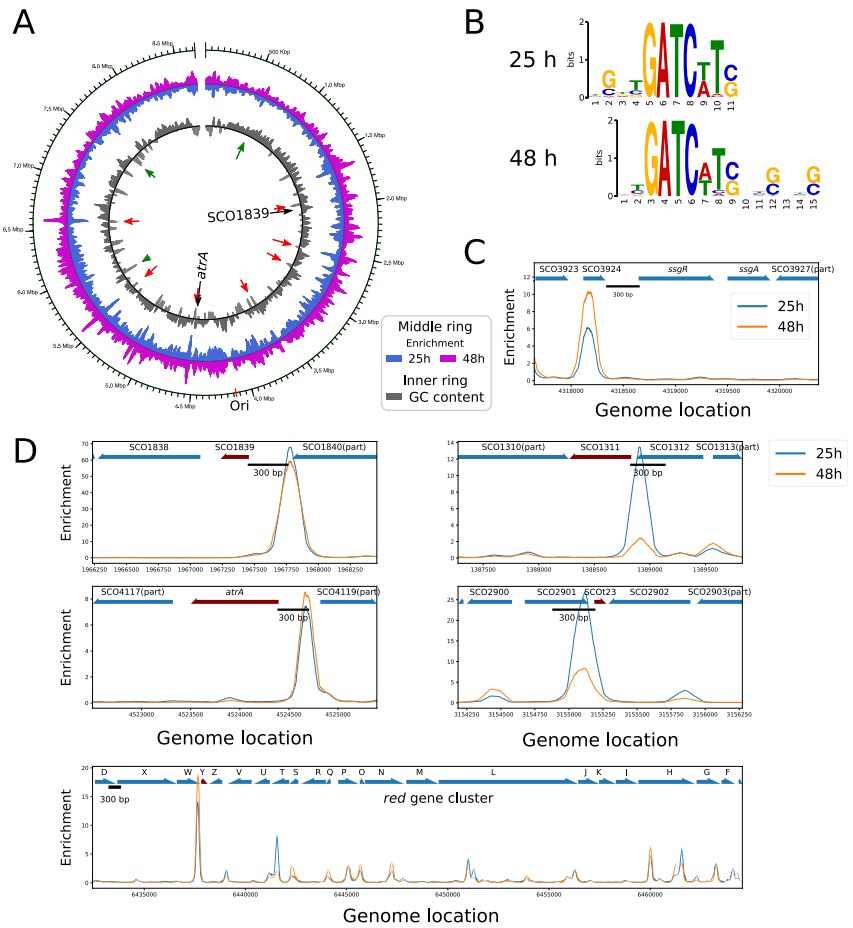

Figure 5. SCO1839 protein binding sites analysis. A) Genome-wide distribution of SCO1839 protein binding sites along the *S. coelicolor* genome. The outer ring shows the genome location; the middle ring shows the local average fold enrichment from ChIP-Seq analysis, 48 h sample oriented outwards, 25 h sample oriented inwards; the inner ring shows the local average G+C content. The G+C percentage above median is plotted outwards, below median inwards. Bin size for local averaging was 20,000 bp. SCO1839 and *atrA* have been indicated by black arrows. B) SCO1839 DNA binding motifs predicted by MEME-ChIP program. C) Enrichment level (compared to the corresponding total DNA control sample) around *ssgR*. D) Enrichment level of genome regions around genes highlighted in Table 2, with target genes coloured red. The *red* gene cluster is shown as a whole, with significant binding only upstream of *redY*. Note that the y-axes of the plots are not at the same scale.

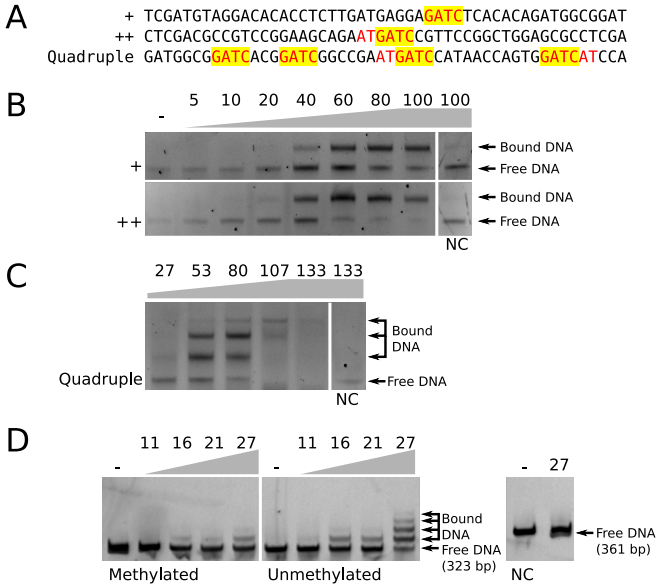

Figure 6. Binding specificity and affinity of SCO1839 protein. A) Sequence of short DNA fragments used in the Electrophoretic mobility shift assays (EMSA). The motifs included in the EMSA sequences were designed with different degrees of binding strength, one weak GATC motif is indicated by "+", one strong GATCWT motif by "++", and four GATCs motifs containing two strong, two weak variances indicated by "Quadruple". GATC motifs are highlighted with the trailing AT sequence coloured red. B) and C) Electrophoretic mobility shift assays (EMSA) using 6×His-tagged SCO1839 protein on synthetic 50 bp dsDNAs containing the different binding strengths GATC motifs described in (A). D) EMSA using 6×His-tagged SCO1839 protein performed on DNA extracted from DNA methylation-deficient *E. coli* ET12567 and methylation-positive *E. coli* JM109. The SCO1839 promoter region (-561 to -239) contains eight GATC motifs and was used as test sequence. Negative Control (NC) sequence derived from a random location with no detectable SCO1839 protein binding as shown by ChIP-Seq analysis, specifically the position 5,052,200 to 5,052,548 (349 bp). In the panels B), C), and D) the protein to DNA molar ratios are shown on top of each gel picture.

SCO1839 affects the expression of about 10% of all genes

To find the genes affected by the binding of SCO1839, the common binding regions were mapped to the translational start site of all genes (-350 to +50 of transcription start site). After removing ambiguous low binding regions, 736 of the SCO1839 binding regions were found in promoter regions, corresponding to 862 genes (Table S1). Of these genes, 47% (406 genes) had at least one binding region with more than 10-fold enrichment in the corresponding promoter region. These genes included the cell division and cell wall (*dcw*) cluster (SCO2077-SCO2088), a 50S ribosomal protein gene cluster (SCO4702-SCO4727), tRNAs (SCOt02-SCOt50), *atrA*, *redY*, SCO1311 and SCO1839 itself (Table 2, Figure 5D). One binding region was found upstream of *ssgR*, with a peak at around nt position -483

relative to the translation start site (Figure 5C). This is in accordance with the previous observation that SCO1839 binds to the *ssgR* promoter region (Kim *et al.*, 2015).

Table 2. Genes strongly bound by SCO1839 at their promoter regions

Gene ID	Strand	Product	Binding width		Binding summit position**		Fold enrichment	
			25h	48h	25h	48h	25h	48h
SCO1311*	-	hypothetical protein with tRNA edit domain	271	ND†	-76	-72†	13.41	2.42†
SCO1839*	-	transcriptional regulator	547	530	-326	-324	67.75	58.18
SCO2077	-	DivIVA	251	300	-90	-91	17.19	13.48
SCO2081	-	YlmD	271	378	-348	-348	10.44	17.18
SCO2084	-	MurG (UDPdiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase)	280	313	-280	-273	13.22	13.41
SCO2086	-	MurD (UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase)	482	471	-90	-100	9.63	10.62
SCO2088	-	MurF (UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate- D-alanyl-alanyl ligase)	227	285	-5	-10	7.37	14.21
SCO3213	-	TrpG (anthranilate synthase component II)	247	299	-240	-236	7.85	10.69
SCO3224	-	ABC transporter ATP-binding protein	349	367	+33	+48	14.23	15.90
SCO3225	+	AbsA1 (two component sensor kinase)	*ab*	*ab*	-168	-183	*ab*	*ab*
SCO4118*	-	AtrA (TetR family transcriptional regulator)	228	255	-284	-289	7.42	8.23
SCO4702	+	RplC (50S ribosomal protein L3)	459	511	-34	-37	12.19	22.16
SCO4707	+	RplV (50S ribosomal protein L22)	290	388	-198	-196	15.80	25.17
SCO4713	+	RplX (50S ribosomal protein L24)	372	406	-7	-5	58.66	44.69
SCO4714	+	RplE (50S ribosomal protein L5)	*ab*	*ab*	-330	-328	*ab*	*ab*
SCO4717	+	RplF (50S ribosomal protein L6)	372	418	-324	-326	25.78	23.66
SCO4718	+	RplR (50S ribosomal protein L18)	252	641	+46	+45	10.27	19.35
SCO4721	+	RplO (50S ribosomal protein L15)	380	432	+22	+21	34.82	29.13
SCO4726	+	RpmJ (50S ribosomal protein L36)	310	761	-3	+21	6.69	15.32
SCO4727	+	RpsM (30S ribosomal protein S13)	*ab*	*ab*	-305	-282	*ab*	*ab*
SCO5880*	+	RedY (RedY protein)	286	347	-168	-161	14.59	18.85
SCOt02	-	tRNA Val (anticodon CAC)	509	559	-136	-138	56.47	55.65
SCOt17	+	tRNA Gly (anticodon UCC)	295	348	-325	-320	20.40	28.92
SCOt23*	+	tRNA Leu (anticodon UAG)	338	280	-78	-90	25.89	8.10
SCOt49	+	tRNA Thr (anticodon GGU)	261	294	-170	-168	15.21	13.95
SCOt50	+	tRNA Met (anticodon CAU)	*ab*	*ab*	-289	-287	*ab*	*ab*

* Shown in Figure 5C

** Relative to gene start site (+1)

† Peak not detected, local summit position and corresponding fold enrichment are shown

ab as above, shared binding region as the gene above

Interestingly, SCO1839 did not only bind to the promoter regions of five tRNA genes, but also to the promoter region of SCO1311, which has a tRNA-editing domain responsible for hydrolysing mis-acylated tRNA (coverage 68%, Pfam: PF04073). Importantly, the promoter region of SCO1311 and SCOt32 showed the largest difference in SCO1839 binding between 25 h and 48 h samples, with enrichment at 25 h more than three times higher than at 48 h (Figure 5D, Table 2). SCOt23 specifies the leucyl-tRNA with anticodon UAG, which is required for the translation of the very rare leucine codon CUA. The rarest codon in *S. coelicolor* is another leucine codon, namely UUA. The tRNA recognizing the UUA codon is specified by *bldA*, and the corresponding TTA codon occurs specifically in many genes involved in development and antibiotic production, making those genes *bldA* dependent (Leskiw *et al.*, 1991, Li *et al.*, 2007, Lawlor *et al.*, 1987, Chater & Chandra, 2008). The CTA leucine codon is also very rare, representing only 0.31% of all leucine codons in *S. coelicolor*. It will be interesting to see if SCOt23 also plays a role in the control of developmental gene expression, and what the role is of SCO1839 in the control of SCOt23 transcription.

Considering that over-expression of SCO1839 enhances undecylprodigiosin (Red) production on MM agar (Figure 2A) it is worth noticing that SCO1839 binds to the promoter region of *redY*, which is part of the BGC for Red (Figure 5D). However, since *redY* is not considered as a key factor in Red production (Cho *et al.*, 2008), we do not want to speculate too much on this topic. The connection (if any) between SCO1839 and Red biosynthesis needs to be investigated further.

Interaction between the global regulators SCO1839 and AtrA

An interesting hit in the list of SCO1839 binding sites was *atrA* (Figure 7C). We previously experimentally validated that both SCO1839 and AtrA bind to the upstream region of the cell division activator gene *ssgR*, suggesting competition or cooperativity (Kim *et al.*, 2015). Furthermore, AtrA is required as transcriptional activator of the actinorhodin BGC (Uguru *et al.*, 2005), and control of *atrA* may therefore explain at least in part the altered antibiotic production in the SCO1839 mutant (Figure 2). To analyse the possible relationship between the two regulators further, we created *atrA* mutants for *S. coelicolor* M145 and its SCO1839 null mutant. For this, the -41 to +46 region relative to the *atrA* translational start site was removed using pCRISPR-Cas9. We then followed growth of the strains using time lapse imaging (Figure 3B). The *atrA* mutant and atrA/SCO1839 double mutant showed accelerated sporulation by 6 and 8 h, respectively, while the SCO1839 mutant developed only slightly earlier (around 2 h). The accelerated growth was also reflected by the larger colony sizes (Figure 2A, 20, 24).

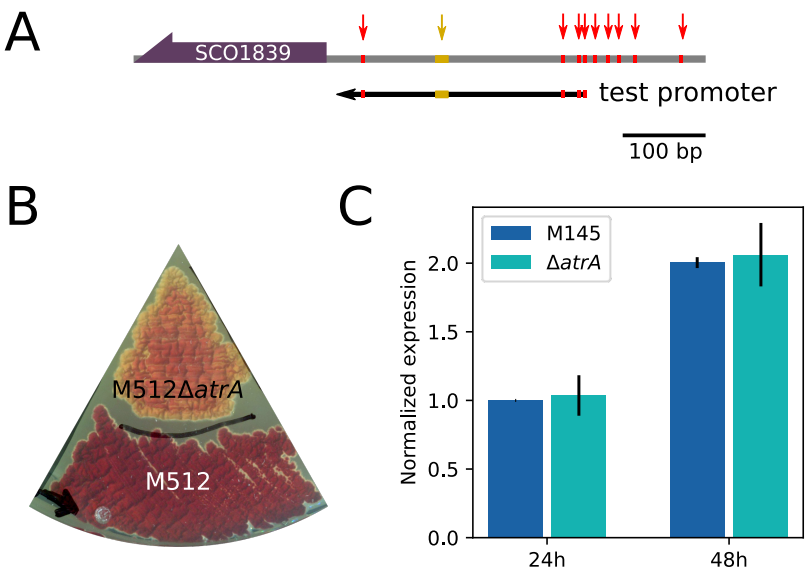

Figure 7. SCO1839 promoter activity assay. A) Graphic representation of the promoter region of SCO1839. Red arrows point at the GATC positions; yellow arrow points at the predicted AtrA binding site. B) Promoter probing experiment showing *S. coelicolor* M512Δ*atrA* (top) and *S. coelicolor* M512 (bottom) harbouring a plasmid wherein *redD* is expressed from the SCO1839 promoter region (redness of the mycelium is a measure for the promoter activity). Patches streaked on R5 agar were grown for 4 days at 30°C. C) Normalized qPCR result showing SCO1839 expression in *S. coelicolor* M145 and its Δ*atrA* mutant on samples taken 24 h and 48 h after inoculation MM agar plates covered with a cellophane. Error bars represent the standard error of the mean derived from two biological replicates. All expression values of SCO1839 were normalized against the expression in M145 at 24 h.

Interestingly, in the SCO1839 promoter region we found an AtrA binding site using PREDetector (Tocquin *et al.*, 2016) with a high score of 13.3. This score is the same as for the confirmed AtrA binding site upstream of *actII*-ORF4, and this therefore suggests direct control of SCO1839 by AtrA. To test the possible transcriptional control of SCO1839 by AtrA, promoter probing experiments were performed based on the *redD* reporter system (van Wezel *et al.*, 2000b). For this, we used the -297 to -10 region relative to the SCO1839 translation start site (Figure 7A). The promoter-probe system is based on the *redD* gene, which activates the red-pigmented antibiotic undecylprodigiosin. When a promoter sequence is inserted in front of the promoterless *redD*, the gene will be transcribed, and the level of red pigment that is produced can be used as a measure of promoter activity. As hosts we used *S. coelicolor* M512 (which is *S. coelicolor* M145 Δ*actII*-ORF4 Δ*redD* and hence non-pigmented) and its *atrA* mutant. The results clearly show that the

SCO1839 promoter is less actively transcribed in the *atrA* mutant (Figure 7B). We therefore conclude that AtrA *trans*-activates SCO1839 transcription. To further validate this, we used quantitative reverse transcription PCR (q-RT-PCR) to analyse the transcription of SCO1839 in *S. coelicolor* M145 and its *atrA* mutant. Surprisingly, the expression of SCO1839 was slightly higher in the mutant at both 24 h and 48 h time points, which conflicts with the promoter probing results (Figure 7C). Thus, expression of SCO1839 likely depends on more than the tested region and involves multiple factors.

Conclusions

In this study, we have shown that SCO1839 is a highly pleiotropic DNA binding protein that plays a role in the (timing of) development and antibiotic production of *S. coelicolor*. Strains in which SCO1839 had been deleted or over-expressed showed accelerated and delayed sporulation, respectively, suggesting that SCO1839 plays a role in the timing (and in particular delay) of development. Localization experiments of SCO1839 showed high concentrations of protein especially inside spores. In both *in-vivo* and *in-vitro* experiments revealed that SCO1839 binds specifically to GATC DNA motif with a preference of (A/T)T on one side. In addition, we have shown that the methylation of adenine in the GATC sequence reduced affinity of SCO1839 for its binding sites. SCO1839 binds to the promoter region of *atrA*, which correlates to the activation of actinorhodin production. In return, AtrA activates SCO1839 expression, which results in an interesting two-way feedback loop between these two important proteins.

Materials and methods

Bioinformatics

The SCO1839 from *S. coelicolor* was used as query in HMMER web server (Potter *et al.*, 2018) to obtain all SCO1839-like proteins from the database, resulting in 727 hits. Sequences with an E value < 0.01 (684 sequences) were selected to generate a Hidden Markov model (HMM) profile using HMMER suit (v. 3.1b2, Eddy, 2011). This profile was used to search against a custom database containing 146,856 genomes with all available bacteria genomes (access date Feb. 9, 2019). Hits with E value $\leqq 5.5 \times 10^{-9}$ (2,317 sequences) were aligned to the generated HMM profile using the hmmalign tool from HMMER suit. Using the alignment, a network was built calculating the pairwise distance between all the detected SCO1839 proteins and the threshold for clustering was settled at 0.8. Network visualizations were constructed using Cytoscape (v. 3.7.1, Shannon *et al.*, 2003).

Four sequences of known AtrA binding motif (Table S2) was used as input for PREDetector (v. 3.1, Tocquin *et al.*, 2016) to find possible AtrA binding motif in *S. coelicolor* genome.

Strains and growth conditions

All strains used in this study are listed in Table S3. *Escherichia coli* strain JM109 was used for routine cloning, *E. coli* ET12567 (MacNeil *et al.*, 1992) for preparing non-methylated DNA, ET12567 containing driver plasmid pUZ8002 (Paget *et al.*, 1999) was used in conjugation experiments for introducing DNA to *Streptomyces*. *E. coli* strains were grown in Luria broth at 37°C supplemented with the appropriate antibiotics (ampicillin, apramycin, kanamycin and/or chloramphenicol at 100, 50, 25 and 25 $\mu g \cdot mL^{-1}$, respectively) depending on the vector used. *S. coelicolor* A3(2) M145 was the parent for all mutants. For promoter probing experiments, *S. coelicolor* M512 (M145 Δ*redD* Δ*actII*-ORF4 (Floriano & Bibb, 1996)) was used. *Streptomyces* strains were grown on soya flour medium (SFM) for conjugation, SFM or MM agar medium supplemented with 0.5% mannitol for phenotype characterization, and R5 agar plates for protoplast regeneration and promoter probing. Solid cultures were grown in a 30°C incubator unless described specifically. For liquid cultures, approximately 10^6 spores were inoculated in 100 mL Erlenmeyer flask equipped with steel spring, containing 15 mL TSBS (tryptone soya broth sucrose) medium (Kieser *et al.*, 2000), and grown at 30°C with constant shaking at 180 rpm. Antibiotics used for screening *Streptomyces* transformants were apramycin and thiostrepton (20 and 10 $\mu g \cdot mL^{-1}$, respectively).

Plasmids and Constructs

Primers used for PCR and short double strand DNA fragment are listed in Table S4. PCR was preformed using Pfu DNA polymerase using standard protocol as described previously (Colson *et al.*, 2007). All plasmids and constructs described

in this study are summarized in Table S5. The constructs generated in this study were verified by sequencing performed in BaseClear (Leiden, The Netherlands).

The SCO1839 knock-out strategy was based on the unstable multi-copy vector pWHM3 (Vara *et al.*, 1989) as described previously (Zhang *et al.*, 2018). Roughly 1.5 Kb of upstream and downstream region of SCO1839 were amplified by PCR from M145 genome. The upstream region (relative to SCO1839 direction) was thereby cloned as a *Bam*HI-*Xba*I fragment, and the downstream region was cloned as a *Xba*I-*Hin*dIII fragment. The apramycin resistance cassette *aac*(3)IV with flanking *loxP* sites was digested from construct pGWS728 made in previous study using *Xba*I (Zhang *et al.*, 2018). And these three fragments were ligated into *Hin*dIII-*Bam*HI-digested pWHM3. The presence of the *loxP* recognition sites allows the efficient removal of the apramycin resistance cassette following the introduction of a plasmid pUWL-Cre expressing the Cre recombinase (Fedoryshyn *et al.*, 2008, Khodakaramian *et al.*, 2006). Knock-out construct pGWS1255 was created for the deletion of nucleotide positions -565 to +228 of SCO1839, where +1 refers to the translation start site. Introducing this construct to *S. coelicolor* A3(2) M145 followed by losing this construct resulted in replacing the respective region with apramycin resistance cassette, which was subsequently removed using Cre expressing construct pUWL-Cre. For complementation of the SCO1839 deletion mutant, the nucleotide positions -565 to +228 relative SCO1839 translation start site, containing entire coding region of SCO1839 (with stop codon) and its promoting region was cloned as a *Hin*dIII-*Bam*HI fragment and inserted into the multiple cloning site of low copy number plasmid pHJL401. This resulted in the complementation construct pGWS1260. This construct was then transformed to SCO1839 null mutant, resulting in strain GAD014.

For FLAG-tagging and enhanced green fluorescent protein (eGFP) fusion expression of SCO1839, the 3×FLAG or eGFP sequences were fused to the end of original copy of SCO1839 on genome using codon optimised CRISPR-Cas9 system (Cobb *et al.*, 2015). The spacer sequence located at the end of SCO1839 was the same for both 3×FLAG tag and eGFP knock-in. The sgRNA with this spacer guides the Cas9 enzyme to make a double strand break after the stop codon. This spacer sequence was inserted into the pCRISPomyces-2 plasmid as described by (Cobb *et al.*, 2015). Templates for homology-directed repair (HDR) were made and inserted to spacer containing pCRISPomyces-2 briefly as follows. For the 3×FLAG tag knock-in, about 1 kb region of SCO1839 and its upstream region were amplified by PCR from the *S. coelicolor* genome, including an additional connecting sequence replacing the stop codon of SCO1839. Additionally, around 1 kb of downstream region was PCR-amplified from the genome, whereby a 3×FLAG sequence was added upstream to ensure in-frame expression of the 3×FLAG tag. The two fragments were connected by overlap extension PCR (Figure S5). *Xba*I sites were added in both ends, and the fragment was then inserted into an *Xba*I site

in the spacer of pCRISPomyces-2, resulting in a SCO1839-3×FLAG knock-in construct pGWS1298. For the eGFP knock-in, the DNA sequence of eGFP was cloned as a *Bam*HI-*Hin*dIII fragment from pGWS526 (Zhang *et al.*, 2016). SCO1839 and its upstream region were cloned as an *Xba*I-*Bam*HI fragment, while the downstream region was cloned as *Hin*dIII-*Xba*I fragment (Figure S2). These three fragments were directly ligated into *Xba*I-digested pCRISPomyces-2, resulting in eGFP knock-in construct pGWS1299.

Mutagenesis was done according to (Cobb *et al.*, 2015). After conjugation of pGWS1298 and pGWS1299 to *S. coelicolor* A3(2) M145, ex-conjugants were patched onto SFM agar plates containing apramycin (50 $\mu g \cdot mL^{-1}$), then positive ex-conjugants were plated onto fresh SFM agar plates and grown at 37°C. Spores were then collected and checked for loss of the construct. Apramycin sensitive strains were incubated for spores and the genomes were checked by PCR for the desired recombination events. A successful 3×FLAG tag knock-in strain was identified and designated GAD043, eGFP knock-in strain was designated GAD099.

For over-expression of SCO1839, the *ermE* promoter was cloned to replace part of the original promoter region of SCO1839 using the same CRISPR-Cas9 system. As spacer sequence we used the promoter region of SCO1839 and this was inserted into pCRISPomyces-2 (Cobb *et al.*, 2015). This guides Cas9 to make double-strand breaks at around nt position -89 relative to the translation start site of SCO1839. To assemble the template for HDR, approximately 1 kb upstream and downstream region of SCO1839 promoter were cloned as *Xba*I-*Sec*I and *Nde*I-*Xba*I fragments, respectively. The fragments were designed removing -157 to +4 region of SCO1839. P*ermE* was digested from pHM10a (Motamedi *et al.*, 1995) as a *Sec*I-*Nde*I fragment. These three fragments were then inserted into pCRISPomyces-2, producing pGWS1295. Following the same procedure as above, strain GAD039 was obtained with expresses SCO1839 from the *ermE* promoter.

To produce protein for EMSA experiments, SCO1839-His_6 was expressed in *E. coli* strain BL21 CodonPlus (DE3)-RIPL (Invitrogen, Massachusetts, U.S.). For this, SCO1839 and its downstream region were cloned as an *Nde*I-*Eco*RI fragment and ligated into pET28a for the expression of 6×His tag at N-terminal of SCO1839 in *E. coli*. To generate methylated and non-methylated DNA for electrophoretic mobility shift assay (EMSA), part of promoter region of SCO1839 and a random region of SCO1839 non-binding region were cloned as 324 bp and 361 bp *Eco*RI-*Hin*dII fragments, respectively, and inserted into the multiple cloning site of pUC19 (Sambrook *et al.*, 1989). Resulting in EMSA testing DNA baring constructs pGWS1300 for SCO1839 binding region and pGWS1451 for non-binding region.

Knock-out of *atrA* was achieved using another CRISPR-Cas9 system that was described by (Tong *et al.*, 2015a). The spacer sequence was designed at the

beginning of *atrA* to guide Cas9 create a double-strand break at around nt position +4 of *atrA*. The template was assembled by PCR amplification of the upstream and downstream regions of *atrA*, so as to remove -41 to +46 position relative to the translation start site of *atrA*, with overhangs for Gibson assembly (Gibson *et al.*, 2009). The two fragments together with *Stu*I digested spacer containing pCRISPR-Cas9 construct was assembled using Gibson assembly kit (NEB, Massachusetts, U.S.) and the resulting construct was designated pGWS1452, which was further used for the knock-out of *atrA* in *S. coelicolor* M145, SCO1839 null mutant GAD003, and *S. coelicolor* M512. The knock-out procedure was the same as using pCRISPmyces-2. The obtained M145 Δ*atrA* strain was designated GAD075, ΔSCO1839ΔatrA strain was designated GAD078, M512 Δ*atrA* strain was designated GAD093.

For testing the promoter efficiency in *S. coelicolor* strain M512 and its *atrA* null mutant, the promoter region of SCO1839 (-297 to -10) was cloned as a *Bam*HI-*Sac*I fragment and ligated into *Bam*HI-*Sac*I digested pIJ2587 (van Wezel *et al.*, 2000b). The resulting construct pGWS1454 was then transformed into *S. coelicolor* M512 and *S. coelicolor* M512 Δ*atrA*.

Time-lapse confluent plate morphology monitoring

Approximately 10^7 spores were plated on MM agar supplemented with mannitol. The plates were then placed upside down in Perfection V370 scanner (Epson, Nagano, Japan) located inside 30°C incubator. A scanning picture was taken every hour, and images were processed using custom python script to get the brightness value of the plate. Specifically, the pictures were first converted to grey scale. To avoid the interference of mycelium grown in the edge which might show different developmental properties, 70% the diameter of the plate from the centre was selected as the region of interest (ROI). The average grey value of all the pixels within ROI was used as the brightness of the mycelium lawn. The measured values from one plate were then normalized to range 0 to 1.

Scanning electron microscopy (SEM)

Mycelia were grown on MM agar supplemented with mannitol and grown for 5 days. Sample preparation and imaging was done as described before (Piette *et al.*, 2005), using JSM-7600F scanning electron microscope (JEOL, Tokyo, Japan). For each strain, 5 images with 7,500 × magnification were taken in randomly selected spore-rich areas. The length and width of spores in each picture were measured using ImageJ version 1.52p strictly according to a randomized file list, in order to minimize human error. Only spores which are approximately parallel to the focal plane were measured.

Confocal imaging

Sterile coverslips were inserted at an angle of 45° into minimum medium agar plates, spores of SCO1839-eGFP fusion protein expressing strain was carefully inoculated at the intersection angle, and incubate at 30°C. Before imaging, a drop of water containing 5 μM cell-permeant Syto-Red nucleic acid dye (Invitrogen, Massachusetts, U.S.) and stained for 5 min at room temperature. For imaging liquid grown cultures, 5 μL of TSBS culture were dropped on a glass slide containing a 5 μL water containing 10 μM Syto-Red and stained for 5 min at room temperature. Imaging was performed with a Zeiss LSM system as described before (Willemse & van Wezel, 2009). eGFP fluorescent were excited using 488 nm laser and monitored with a 505-530 nm band-pass filter, the DNA Syto-Red red fluorescent was excited using 633 nm laser and monitored using 635 nm secondary dichroic beam splitter. All images were processed in ImageJ version 1.52p.

DNA-protein cross-linking and chromatin immunoprecipitation

10^8 spores of strain GAD043 were plated on MM medium covered with cellophane. After 25 h or 48 h growth, cellophane disks were soaked up-side-down in PBS solution containing 1% formaldehyde for 20 min for DNA-protein crosslinking. Ten plates were collected for 25 h samples; four plates were collected for 48 h samples. Then the disks were moved to PBS solution containing 0.5 M glycine for 5 min to stop crosslinking reaction. The mycelium was then collected, washed in PBS and resuspended in 0.5 mL lysis buffer (10 mM Tris-HCl pH 8.0, 50 mM NaCl, 15 $mg \cdot mL^{-1}$ lysozyme, 1× protease inhibitor (Roche, Bavaria, Germany) and incubated at 37°C for 20 min. After incubation, 0.5 mL IP buffer (100 mM Tris-HCl pH 8.0, 250 mM NaCl, 0.8% v/v Triton-X-100) was added to the sample, and chromosomal DNA sheared to 100 to 500 bp using Bioruptor Plus water bath sonication system (Diagenode, Liège, Belgium). After centrifugation to remove cell debris, lysates were incubated with 40 μL Anti-FLAG M2 affinity gel (cat A2220, Sigma-Aldrich, St. Louis, U.S.) suspension according to the manufacturer's instructions and incubated at 4°C overnight. After centrifugation and washing, the pellets and 50 μL of untreated total extracts (controls) were incubated in 100 μL IP elution buffer (50 mM Tris-HCl pH7.5, 10 mM EDTA, 1% m/v SDS) at 65°C overnight to reverse cross-links. Beads were then removed by centrifugation before DNA extraction with phenol-chloroform. The DNA sample was then extracted with chloroform and the water layer was further purified using DNA Clean & Concentrator kit (Zymo Research, California, U.S.). The samples were then sent for next generation sequencing using BGI-Seq platform (BGI, Hong Kong, China)

ChIP-Seq data analysis

Clean reads received from sequencing contractor were aligned to the *S. coelicolor* M145 genome with RefSeq accession number NC_003888.3 using bowtie2 (version 2.3.4, Langmead & Salzberg, 2012). Resulted SAM files are sorted using

SAMtools (version 1.9, Li *et al.*, 2009) producing BAM files. MACS2 (version 2.1.2, Zhang *et al.*, 2008) was then used for binding peak prediction and peak modelling by comparing the chromatin immunoprecipitated DNA sample with the corresponding whole genome sample. The models for both samples are shown in Figure S4. The enrichment data used in Figure 5 was calculated for each nucleotide using MACS2 'bdgcmp' command with '-m FE' switch. The peak summit positions including sub peak positions of each predicted binding region were then extracted, and the ± 150 bp region of each summit was extracted from genome sequence using python script dependent on the Biopython module (version 1.70, Cock *et al.*, 2009). Extracted sequences were subjected to MEME-ChIP (version 5.0.2, Machanick & Bailey, 2011), which is suitable for large sequence sets, for binding motif prediction.

The enrichment data of two samples was averaged separately in a moving bin of 20,000 bp and plotted at opposite directions as the middle ring of the circular genome diagram. The G+C content was calculated using the same moving bin and centred at the middle of maximum and minimum value before plotting as the inner ring on the plot (Figure 5A). For determining the overlap of low G+C content regions and high enrichment regions, the genome was divided into 1,000 bp long sections, the G+C content and average enrichment levels were calculated. The sections with G+C content below the first quartile was considered low in G+C, and those with average enrichment level above the third quartile were considered high in enrichment. To find genes possibly regulated by SCO1839, the locations of promoter regions (-350 to +50) of all genes were extract from the GenBank file containing annotations and checked for overlap with ± 150 bp location of the summit of SCO1839 binding peaks. This was done using python script dependent on module Biopython and pybedtools (version 0.8, Dale *et al.*, 2011) and external BEDTools (version 2.27, Quinlan, 2014).

Electrophoretic mobility shift assay (EMSA)

Construct pGWS1286 expressing SCO1839-His_6 was introduced into *E. coli* BL21 CodonPlus (DE3)-RIPL and expressed and purified as described before (Mahr *et al.*, 2000). Purified protein was dialyzed over-night at 4°C against EMSA buffer (10 mM Tris-HCl, pH 7.9, 0.1 mM EDTA, 50 mM KCl). 50 bp double strand DNA was generated by gradual cooling of reverse complemented single strand oligonucleotides in T4 DNA ligase buffer (NEB, Massachusetts, U.S.) from 95°C to 12°C in 45 min. For testing the methylation effect on SCO1839 binding, long DNA fragments of the SCO1839 promoter region were cloned into pUC19, forming pGWS1300. The non-binding control fragment was randomly selected from the genome that showed no enrichment in ChIP-Seq experiment, this fragment was then cloned into pUC19, to give pGWS1451. pGWS1300 was then transformed to both DAM methylation effective *E. coli* strain JM109 and DAM deficient *E. coli* strain ET12567, while pGWS1451 for negative control was transformed to strain

ET12567 only. The target fragments were them digested from extracted constructs and blunt-ends created using DNA polymerase I Klenow fragment (NEB, Massachusetts, U.S.) and dNTPs. The *in-vitro* DNA-protein interaction tests were done in EMSA buffer in a total reaction volume of 10 µL and incubated at 30°C for 15 min. The reactions were then loaded on 5% polyacrylamide gels and separated by electrophoresis. The gel was briefly staining with ethidium bromide and imaged in a Gel Doc imaging system (BioRad, California, U.S.).

RNA extraction and quantitative reverse transcription PCR

S. coelicolor strains were inoculated on minimal medium agar plates covered with cellophane disks with a final inoculum of 10^7 CFU per plate. Biomass from two (for 24 h) or one (48 h) agar plate was collected in 15 mL tube. The RNA was extracted using modified Kirby mix (Kieser *et al.*, 2000). The total nucleic acid was precipitated using 1 volume isopropanol and 0.1 volume of 3 M sodium acetate and then dissolved followed by purification using RNeasy columns (Qiagen, Venlo, The Netherlands) according to vendor's instructions, and the remainder of the DNA was removed by DNase I (NEB, Massachusetts, U.S.) treatment followed by phenol chloroform extraction. The complete removal of DNA was confirmed by PCR. cDNA was synthesized using iScript cDNA synthesis kit (BioRad, California, U.S.) according to vendor's instructions. Quantitative PCR was done using iTaq Universal SYBR green qPCR kit (BioRad, California, U.S.). The PCR program was set as following: 95°C 30 s; 40 cycles of 95°C 10 s, 60°C 30 s, plate read; melt curve from 65°C to 95°C with 5 s per 0.5°C increment. House-keeping genes *gyrA* (SCO3873) and *rpmE3* (SCO5359) that are known to show constitutive expression were used as inner controls. The data was analysed using CFX Manager software (version 3.1, BioRad, California, U.S.) using $\Delta\Delta C_q$ method, which is an implementation of the method described by (Vandesompele *et al.*, 2002).

Acknowledgments

We are grateful to Prof.dr. K.J. McDowall and to Prof.dr. R.T. Dame for stimulating discussions.